ABORT: Accumulation of Bag Of featuRe descripTors for person re-identification

Loris Bazzani, *Ph.D. student*

***Abstract*—Person re-identification problem consists in recognizing an individual in diverse locations over different non-overlapping camera views, considering a large set of candidates. The appearance model plays a fundamental role for the success or the failing of the recognition task. In this work, we investigate a particular type of appearance representation for person modeling based on the bag of feature paradigm. In particular, we investigate a variant that builds a affine-covariant feature-based image descriptor, incorporating spatial relations between features, called spatially-sensitive bag of feature [1]. As appearance description, we employ a novel, color-based affine-covariant feature, *i.e.*, the maximally stable color region descriptor, that capture the interesting regions of the images in terms of stable color blobs. A symmetry-driven approach [2] is employed to extract feature from interesting part of the image. Finally, we prove that accumulating more than one descriptor for each pedestrian increases the robustness of the re-identification method. We named our re-identification protocol ABORT, Accumulation of Bag Of featuRe descripTors. A comparison between our method and the state of the art results is carried out on several compelling publicly available datasets for person re-identification: iLIDS [3], and ETHZ [4].**

***Index Terms*—Bag of Feature, Spatially-Sensitive Bag of Feature, Person Re-identification, Object Modeling**

I. Introduction

In recent years, the interest of many researcher has been captured by the deployment of automated systems for video surveillance. Automatic surveillance task can be broken down into a series of subproblems [5]: 1) *object detection and categorization* which detects and classifies the *interesting* objects in the field of view of the camera(s); 2) *Multi-Target Tracking* (MTT) where the objective is to estimate the trajectories of targets, keeping the identification of each target; 3) *MTT across cameras* tracks the targets while observing them through multiple overlapping or non-overlapping cameras. The common denominator of all these problems is the kind of representation of the objects that have to be detected, categorized, tracked, re-identified.

In this work, we focus our attention on for MTT across cameras problem that involve a task called person re-identification, that consists in recognizing an individual in diverse locations over different non-overlapping camera views (Fig. 1), considering a large set of candidates. The object representation plays a fundamental role and has to deal with several problems: pose, viewpoint, and illumination changes, occlusions and very low-resolution information.

We investigate a particular type of appearance representation based on the Bag of Feature (BoF) paradigm and an its spatial-sensitive variant. BoF is a quite popular approach in Computer Vision and it has been successfully applied to different applications such as object recognition, stereo matching, and so forth.

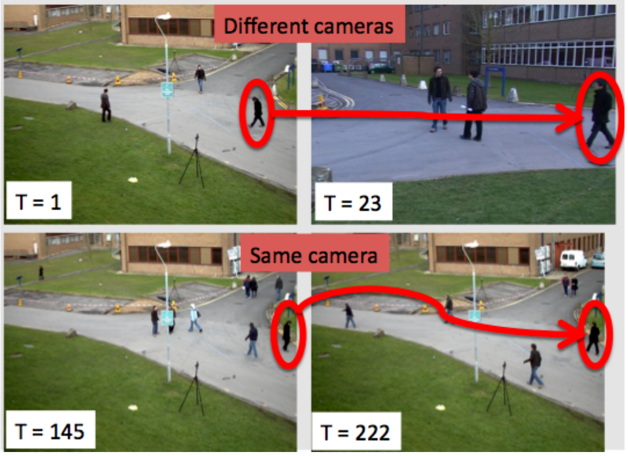

Fig. 1. Person re-identification problem for multi-target tracking across different cameras (first row) or the same camera (second row).

However, the main disadvantage of this approach is the loss of spatial relations between features, which often carry important information about the image. The main focus of this work is to investigate a method that builds a affine-invariant feature-based image descriptor that incorporate spatial relations between features, called Spatially-Sensitive Bag of Feature (SSBoF) [1].

Most of the work employ the SIFT descriptor [6] as sparse feature for the object representation in the BoF approach. However, this descriptor suffers of two problems: 1) poor description with a very low resolution images, 2) it does not employ the color information. For this reasons, in this work we employ a novel, color-based feature, *i.e.*, the Maximally Stable Color Region (MSCR) [7], that capture the interesting regions of the images in terms of stable color blobs.

Similar to [2], we select salient parts of the body figure by adopting perceptual principles of symmetry and asymmetry extracting the descriptors from each part. 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. Moreover, 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.

We tested the descriptor on several compelling publicly available datasets for person re-identification : iLIDS [3], and ETHZ [4]. We have carried out a comparison among different methods: the BoF method using the MSCR descriptor, the SS-BoF method using the MSCR descriptor, and a “crude” method that employes the MSCR without any BoF representation. And finally we compared each method adding the symmetry-based description.

The technical report is organized as follows: in Sec. II, we detail the protocol for the extraction of the accumulation of bag of feature descriptors. Then, Sec. III reports the proposed matching strategy. The experiments and the results are reported in Sec. IV. Finally, the conclusions are drawn in Sec. V.

II. ABORT: Accumulation of Bag Of featuRe descripTors

The typical steps of most of the BoF methods are detailed in the following. The first step is the *feature detection* that finds the interesting regions or points in the image (Sec. II-A). Then, the extracted features are transformed into a *canonical representation* in order to make them comparable (Sec. II-B). A *feature descriptor* is computed for each feature, in the third step (Sec. II-C). At the end, the features are *quantized* using a previously learned codebook (Sec. II-D and Sec. II-E). After the computation of the descriptors, the re-identification is carried out by a matching algorithm (Sec. III), that compares the extracted descriptors with a dataset of previously-seen pedestrians.

Moreover, we introduce two additional steps for extracting the descriptor from significative parts of the image (Sec. II-F) and for accumulating several descriptors over the time (Sec. III).

A. Maximally Stable Color Region Detection

The MSCR operator[1] [7] 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 m, covariance matrix $\mathbf{C}$ 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 (consecutive) views of an object, MSCRs are likely to occur in the same correspondent location.

In order to be robust to the background (BG) clutter, the MSCR operator has been applied only in the foreground part of the image, that is, the pixels belong to the human body. We obtain the foreground (FG) silhouette for each person by inferring over the STEL generative model [8]. STEL model captures the general structure of an image class as a blending of several *component* 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 [8] for further details). In the case that a video sequence is available, a motion-based BG subtraction strategy is enough to obtain the silhouette.

Let's denote as $\mathbf{F}_\mathbf{I}^j = \{F_1, F_2, \ldots, F_N\}$ the set of N features F_n extracted from the image I of the j-th pedestrian. In our case, F_n corresponds to the n-th stable region (or blob) extracted by the MSCR operator. In particular, a interesting advantage of the MSCR descriptor is that is covariant with the group of the affine transformations, *i.e.*, $\mathbf{F_{TI}} = \mathbf{TF_I}$, where $\mathbf{T}$ is an affine transformation and $\mathbf{TI}$ denote the application of the transformation $\mathbf{T}$ to the image.

B. Maximally Stable Color Region Canonization

Once features have been detected, they are often normalized or canonized by means of a transformation into some common system of coordinates. We apply the standard canonization proposed in [9] (called affine normalization). Given the centroid m and the eigenvalue decomposition of the mask covariance matrix $\mathbf{C} = \mathbf{RDR}^T$ ($\mathbf{R} > 0$) of the n-th stable region F_n, the rectifying transform is defined as:

$$F_n = s\mathbf{A_F}\hat{F}_n + m \quad \text{for } \mathbf{A_F} = 2\mathbf{RD}^{1/2} \qquad (1)$$

where the point $\hat{F}_n$ lies in the normalized space, *i.e.*, $\hat{F}_n \in [-1, 1]^2$, and s is a scaling factor that determines how much wider the MSCR should be compared to the covariance matrix (set 1 in our experiments). We refer to $\mathbf{A}_\mathbf{F}^{-1} F_n / s$ as to a canonical representation of the feature F_n.

C. Maximally Stable Color Region Descriptor

The descriptor we use in the BoF approach is computed using the canonization transform. After that, we extract the mean color of each MSCR descriptor. An alternative descriptor for each MSCR could be the color histogram of the blob, but in practice the average color is more robust to low-resolution images. Since the MSCRs are usually very small in size and that they are stable regions, the resulting histograms would be very sparse, that is, just few bin will be "activated". Therefore, a more compact and meaningful descriptor is preferred, that is, the average color composed by a 3-dimensional patterns in the HSV color space.

D. Bag of Feature

Similarly to [1], descriptors of its features are aggregated into a single statistic that describes the entire image. For that purpose, descriptors are vector-quantized in a visual vocabulary $\mathbf{V} = \{v_1, \ldots, v_m\}$ containing m representative descriptors, which are usually found using clustering algorithms. We denote by $\mathbf{Q_V}$ a quantization operator associated with the visual vocabulary $\mathbf{V}$ that maps a descriptor into a distribution over $\mathbf{V}$, represented as an m-dimensional vector. The simplest hard quantization is given by

$$(\mathbf{Q_V} v)_i = \begin{cases} 1 & d(v, v_i) \leq d(v, v_j) \; j = 1, \ldots, m \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

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

where $d(v, v')$ is the distance in the visual descriptor space, usually the Euclidean distance $||v - v'||$. Summing the distributions of all features,

$$\mathbf{B}_I = \sum_{F \in \mathbf{F}_I} \mathbf{Q_V} v_F \tag{3}$$

yields an affine-invariant representation of the image called a bag of features, which with proper normalization is a distribution of the image features over the visual vocabulary. Bags of features are often L2-normalized and compared using the standard Euclidean distance or correlation.

E. Spatially-Sensitive Bag of Feature

The extension of the BoF paradigm to the spatially-sensitive case has been introduced in [1]. The spatial information has been introduced because the major disadvantage of the BoF approach is the fact that they discard information about the spatial relations between features in an image. The motivation behind this approach is similar to the extension of the color histogram feature to the spatial color histogram (called spatiograms). The reason to doing this is that spatial information in the form of expressions is useful in disambiguating different uses of a word in text search. For example, in person re-identification the spatial displacement of the MSCRs in the pedestrian image is constrained by the human body geometry. Hence, the absence of a spatial feature may make weak the descriptor, because for example two similar (in appearance) pedestrian could have similar MSCRs (in terms of affine transformation) but with different position. The goal is to capture this information exploiting the Spatially-Sensitive Bag of Feature (SSBoF).

A straightforward generalization of the notion of combinations of words and expressions to images can be obtained by considering pairs of features. For this purpose, we define a visual vocabulary on the space of pairs of visual descriptors, and use the quantization operator $\mathbf{Q}^2_\mathbf{V} = \mathbf{Q_V} \times \mathbf{Q_V}$ assigning to a pair of descriptors a distribution over $\mathbf{V} \times \mathbf{V}$. $(\mathbf{Q}^2_\mathbf{V}(v, v'))_{ij}$ can be interpreted as the joint probability of the pair (v, v') being represented by the expression (v_i, v_j).

Same way as expressions in text are pairs of adjacent words, visual expressions are pairs of spatially-close visual words. The notion of proximity is expressed using the idea of canonical neighborhoods. Fixing a disk M of radius $r > 0$ centered at the origin of the canonical system of coordinates, we define $N_F = \mathbf{A}_F M$ to be a canonical neighborhood of a feature F. Such a neighborhood is affine-covariant,*i.e.*, $N_{\mathbf{T}F} = \mathbf{T} N_F$ for every affine transformation $\mathbf{T}$. The notion of a canonical neighborhood induces a division of pairs of features into near and far. Using this notion, we define a bag of pairs of features simply as the distribution of near pairs of features,

$$\mathbf{B}^2_I = \sum_{F \in \mathbf{F}_I} \sum_{F' \in N_F} \mathbf{Q_V}(v_F, v_{F'}) \tag{4}$$

We consider the canonical relation $\mathbf{S_{F,F'}} = \mathbf{A}_F^{-1} \mathbf{A}_F$, in order to encode the spatial relation into the BoF descriptor. Being an invariant quantity, the canonical spatial relation can be used to augment the information contained in visual descriptors in a bag of pairs of features. For that purpose, we construct a vocabulary of spatial relations, $\mathbf{S} = \{S_1, ..., S_n\}$. A quantization operator $\mathbf{Q_S}$ associated with the spatial vocabulary can be constructed by plugging an appropriate metric into 2. The easiest way of defining a distance on the space of transformations is the Frobenius norm on transformations represented in homogeneous coordinates,

$$d^2(\mathbf{S}, \mathbf{S}') = ||\mathbf{S} - \mathbf{S}'||^2_F = \mathrm{tr}((\mathbf{S}' - \mathbf{S}')^T(\mathbf{S} - \mathbf{S}')), \tag{5}$$

which is equivalent to considering the 3×3 transformation matrices as vectors in $\mathcal{R}^9$ using the standard Euclidean distance.

Coupling the spatial vocabulary $\mathbf{S}$ with the visual vocabulary $\mathbf{V} \times \mathbf{V}$of pairs of features, we define the spatially-sensitive bag of features

$$\mathbf{B}^3_I = \sum_{F \in \mathbf{F}_I} \sum_{F' \in N_F} \mathbf{Q_V}(v_F, v_{F'})\, \mathbf{Q_S}(\mathbf{S}_{F,F'}) \tag{6}$$

Spatially-sensitive bags of features are again affine-invariant by construction.

F. Symmetry-based Silhouette Partition

In this section, we discuss the idea of the symmetry-based silhouette partition proposed in [2]. 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 [10], [11]. 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*:

$$C(i, \delta) = \sum_{B_{[i-\delta, i+\delta]}} d^2\left(p_i, \hat{p}_i\right) \tag{7}$$

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:

$$S(i, \delta) = \frac{1}{J\delta} \left| A\left(B_{[i-\delta, i]}\right) - A\left(B_{[i, i+\delta]}\right) \right|, \tag{8}$$

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:

$$i_{TL} = \operatorname*{argmin}_{i} \; (1 - C(i,\delta)) + S(i,\delta), \tag{9}$$

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:

$$i_{HT} = \operatorname*{argmin}_{i} \left(-S(i,\delta)\right). \tag{10}$$

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:

$$j_{LRk} = \operatorname*{argmin}_{j} \; C(j,\delta) + S(j,\delta). \tag{11}$$

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).

We use this part-based model in this work, simply accumulating the BoF descriptors extracted from each of the three vertical body parts. It is worth noting that for the SSBoF, the choice of the canonical neighborhood is automatic, limited by the horizontal axes. Using the symmetry-based approach permits to fix a parameter that is usually crucial for the success of the experiments. Moreover, we use the vertical axes to delete out the blobs that are too far from them, because they usually come out from the background, due to the noisy foreground extraction.

III. Matching Methods

In the person re-identification problem, we have two sets of pedestrian images: a gallery set A and a probe set B. Re-identification and therefore the matching consist in associating each person of B to the corresponding person of A, minimizing a certain distance d. We discuss about how to define this distance in Sec. III-A and in Sec. III-B for the MSCR feature without any BoF procedure and for the BoF approach, respectively.

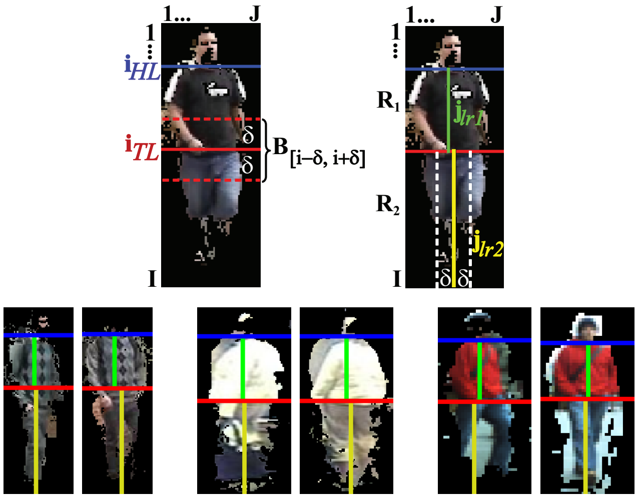

Fig. 2. Symmetry-based Silhouette Partition. On the top row, overview of the method: first the asymmetrical axis Ax_{TL} is extracted, then Ax_{HT}; afterwards, for each R_k region the symmetrical axis j_{LRk} are computed. On the bottom row, examples of symmetry-based partitions on images from the datasets.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

Accumulating more than one image for each pedestrian into the sets seems a good approach to increase the robustness of the matching algorithm [2]. Therefore, we follow two kind of matching procedures: 1) *single-shot vs single-shot* (SvsS), if each image represents a different individual; 2) *multiple-shot vs multiple-shot* (MvsM), if both A and B contain signatures from multiple images for each individual. Groups of images of the same individual can be obtained from tracking information, if available. Moreover, it is possible to define a hybrid approach, that is, *single-shot vs multiple-shot*, but it is not discussed in this work; see [2] for additional information.

A. "Crude" Maximally Stable Color Region Matching

This matching procedure has been proposed for the SDALF descriptor in [2] and it has been proven to carry out the best state of the art performances in several datasets. Here is reported a brief overview of the matching method.

The MSCRs are extracted from the FG region of each pedestrian, without processing any canonization and BoF approach. In the multiple-shot case, we opportunely accumulate the MSCRs coming from the different images by employing a Gaussian clustering procedure [12], which automatically selects the number of components using the Bayesian Information Criterion. 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.

In this case, the matching of two signatures I_A and I_B is carried out by estimating the following distance:

$$d_{\text{MSCR}} = \sum_{b \in I_B} \min_{a \in I_A} \gamma \cdot d_y^{ab} + (1-\gamma) \cdot d_c^{ab} \tag{12}$$

where d_y^{ab} compares the y component of the MSCR centroids, d_c^{ab} compares their mean color, and γ is a weighting parameter that takes values between 0 and 1. In both cases, the comparison is carried out using the Euclidean distance. In our experiments, we fix the value of the parameter as follows: $\gamma = 0.4$. These value has been estimated using the first 100 image pairs of the iLIDS dataset (see Sec. IV-B), and left unchanged for all the experiments.

When we have more than one image per pedestrian, we first calculate d_{MSCR} on each MSCR b of I_B and each cluster representative of I_A, in order to speed up the computation. The representative that gives the lowest distance indicates the cluster, *i.e.*, the set of MSCRs, with which b must be compared with.

B. *Spatially-Sensitive Bag of Feature Matching*

For the BoF approaches is very simple to define a distance measure between signatures, because we have a histogram-based representation, that if normalized it could be treated as a probability mass function. In the state of the art, we can find a lot of measure for this type of representation, such as Kullback-Leibler divergence, Euclidean distance, Bhattacharyya distance, and so forth. In our experiments, we have found that the Bhattacharyya distance d gives the best results for the application reported in this work.

The fact that the descriptor, in this case, naturally enables us to choose a simple way to compare signature is an advantage compared with the "crude" MSCR matching, in which it has been defined a complex, not very intuitive matching procedure. Moreover, another advantage is that the BoF matching is very efficient.

IV. Experiments and Results

In this section we show extensive experiments for person re-identification providing a comparison among the three presented approaches: the BoF method for MSCR (BoMSCR), the SSBoF method for MSCR (SSBoMSCR), and the "crude" method (classical MSCR) that will be our baseline. In addition, we compare all the methods, using the part-based model[2] proposed in [2] and discussed in Sec. II-F. We consider two different datasets, that covers different aspects and challenges for the person re-identification problem: iLIDS Dataset [3] and ETHZ Dataset [4]. First to analyze the results, let's detail the evaluation protocol used for the cross-validation of the methods.

A. *Evaluation Protocol*

The results are usually reported in terms of recognition rate, by the Cumulative Matching Characteristic (CMC) curve. The CMC curve represents the expectation of finding the correct match in the top n matches. But, since the large amount of experiments, we prefer to summarize the results using a more compact and useful measure: the normalized Area Under the Curve (nAUC). The nAUC is the percentage of the area under the CMC with respect to the ideal case (equivalent to a step function in the graph). The more the nAUC is high, the more the method perform well. This follows the validation method suggested in [13] for the person re-identification problem. As to single-shot (SvsS) case, we reproduce the same settings of the experiments in [14]. We randomly select one image ($N = 1$) 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 100 times, and the CMC is averaged and so the nAUC is. A similar protocol has been followed for the MvsM case, where both gallery and probe sets are made up of multi-shot signatures. The multiple-shot signatures are built from N images of the same pedestrian randomly selected. We test our algorithm with $N = \{2, 3, 4, 5\}$. Moreover, we tested the BoF approaches with different quantization of the feature space, namely $K = KV = \{8, 32, 64, 128\}$ and $KS = \{2, 4, 8\}$ bins. We report here the best result (i.e., $K = 128$ for BoMSCR, and $KV = 64$ and $KS = 8$ for SSBoMSC), for simplicity.

Testing set	Codebook Dataset
iLIDS	ETHZ3
ETHZ1	iLIDS
ETHZ2	ETHZ1
ETHZ3	ETHZ2

TABLE I
Datasets with the respective training dataset used in our experiments.

For the BoF approaches, that is, BoMSCR and SSBoMSCR, the codebook has been computed using different datasets, in order to not overtrain the BoF feature space and for generalization purposes. Table I reports the training datasets employed for each tested dataset.

B. *iLIDS Dataset [3]*

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. 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 [14]. The images of this dataset, normalized to 128×64 pixels, derive from non-overlapping cameras, under quite large illumination changes and subject to occlusions.

In Table II, it has been reported the nAUC results for this dataset. The first row corresponds to the SvsS case, because we selected only one image for each pedestrian view. In this case, the best result are performed by the part-based MSCR. If we do not use the part-based reasoning, the BoMSCR gives the best performances. It seems that for the standard MSCR the part-based model is very important as is highlighted in [2].

Considering the MvsM cases, we conclude similar observations. The spatial information provided by SSBoMSCR does improve the results only combined to the part-based model. This is because in the part-based model we automatically select the ray for the spatial relations. In fact, for $N = 4$

[2] In the experiments, we insert the prefix PB to the method name to refer to the part-based method.

the best result are given by part-based SSBoMSCR. In general, the part-based SSBoMSCR performances are very close to the state of the art performances. Moreover, it is worth noting that the performances increase with the increasing of the images used for building the descriptor, in all the methods. Hence, adding (accumulating) information improve the performance for the person re-identification task. This is a similar results reached by the authors of [2].

Another interesting conclusion in our experiments[3] is that increasing the size of the codebook, that is, increasing the quantization resolution, seems to give better results. In fact, we have obtained the best results in this dataset using 128 bins for BoMSCR and $KV = 64$ and $KS = 8$ for SSBoMSCR.

C. ETHZ Dataset [4]

This dataset is captured from moving cameras, and it has been used originally for pedestrian detection. Schwartz and Davis in [15] 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: ETHZ 1 contains 83 pedestrians, for a total of 4.857 images; ETHZ 2 contains 35 pedestrians, for a total of 1.936 images; ETHZ 3 contains 28 pedestrians, for a total of 1.762 images.

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 [12] 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- ($N = 1$) and multiple-shot case ($N > 1$) for ETHZ 1, ETHZ 2, and ETHZ 3 are reported on Table. III, IV, and V, respectively. For those datasets, the resolution of the codebook that gives the best performances is the same highlighted for iLIDS datasets. In Table III, it is shown that the part-based MSCR reaches the best performances. Considering the methods that do not employ the part-based reasoning, we observe that BoMSCR outperforms MSCR and SSBoMSCR, even thought the results of the latter are very similar. MSCR increases the performances introducing the part-based model. It seems that the BoMSCR and the SSBoMSCR do not gain improvements, exploiting the part-based model. From Table IV, we can conclude similar observation of the iLIDS dataset, but this time SSBoMSCR outperforms the other methods, even though the performances are not significantly better. The last experiment session (Table V) highlights that we obtain similar performances to the state of the art with the SSBoMSCR, but not better.

These tables report that for some cases, the SSBoMSCR method outperforms or the results are very close to the state of the art method. This is highlighted for iLIDS, ETHZ 2, and ETHZ 3. The only dataset where the SSBoMSCR method fails is ETHZ 1, giving low performances with respect to the part-based MSCR. The main reason why this happens is that ETHZ 1 dataset contains a lot of partially and fully occluded images. Probably, the SSBoMSCR matching fails in this case compared with the handcrafted matching procedure for the MSCR, because it is not able to cluster out the blobs yielded by the occluded object.

V. Conclusions

In this work, different approaches for person re-identification have been discussed. We proposed a new protocol for dealing with this problem, called ABORT, that consists in accumulate bag of feature descriptor over the time and over spatial meaningful parts of the object of interest extracted using a symmetry-driven approach. This protocol can be naturally extended to spatially-sensitive bag of feature method proposed in [1]. As appearance descriptor of the person we used a powerful affine-covariant feature, namely the Maximally Stable Color Region descriptor, instead of the classic SIFT feature that gives very poor in our application, due to the low resolution of the images.

We have compared the "crude" method proposed in [2] that so far gives the state of the art results with ABORT. In our experiments, we have highlighted that the "crude" method outperforms the bag of feature methods in some cases. However, the bag of feature results are very close to the state of the art performance. This suggests that the bag of feature methods can outperform the crude method introducing further improvements. For example, a problem of the SSBoMSCR approach is due to the ambiguity of rotation transform, given by the canonical transformation. A future work will investigate a spatial canonization and a metric defined in the spatially-sensitive feature space that are robust to blobs rotation. Another interesting problem is concerned to the occlusions. In our experiments, we highlighted that (SS)BoMSCR is strongly affected by this problem. A solution could be to accumulate the descriptors over the time in a smart way, for example considering directly on the volume of MSCRs and adding a descriptor of temporal depth on the BoF codebook.

Even though the results for ABORT do not outperform the state of the art results [2], our method remains more general and simple. It can be applied to different problem, as it has been proved in [1]. Instead, the method in [2] consists on a handcrafted, complex solution for the specific problem. This suggests to try to test the proposed protocol on other problems: such as image retrieval, object recognition, and so forth.

References

[1] A. M. Bronstein and M. M. Bronstein, "Spatially-sensitive affine-invariant image descriptors," in *Proc. European Conf. Computer Vision (ECCV)*, 2010.

[3] We did not report the results here for keeping clear the main results.

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

	MSCR [2]	BoMSCR	SSBoMSCR	PB MSCR [2]	PB BoMSCR	PB SSBoMSCR
		K=128	KV=64 KS = 8		K=128	KV=64 KS = 8
N=1	74.8402	75.9280	74.2017	**76.4466**	74.7031	75.6211
N=2	79.1683	82.0326	80.3891	**82.8571**	81.4293	82.5881
N=3	80.9754	82.7319	81.5755	**82.8289**	81.8276	82.5034
N=4	81.6583	82.8006	81.6475	82.9532	81.4314	**83.0266**

TABLE II
nAUC FOR iLIDS DATASET. NOTE: PB = PART-BASED METHOD, THAT IS, WE HAVE A DESCRIPTOR FOR EACH BODY PART. THE MATCHING IS DONE INDEPENDENTLY FOR EACH PART; HEAD DESCRIPTOR IS NOT CONSIDERED.

	MSCR [2]	BoMSCR	SSBoMSCR	PB MSCR [2]	PB BoMSCR	PB SSBoMSCR
		K=128	KV=64 KS = 8		K=128	KV=64 KS = 8
N=1	81.7243	86.6962	84.9310	**90.8710**	85.7918	82.9583
N=2	87.5191	93.9714	91.3776	**95.4043**	91.8363	90.6968
N=3	89.6150	95.2159	93.3749	**96.8036**	93.3227	92.2442
N=4	91.0266	95.5465	93.8003	**97.2957**	93.5811	92.7609
N=5	91.9003	95.6053	94.1733	**97.8139**	93.6580	92.7740

TABLE III
nAUC FOR ETHZ1 DATASET.

	MSCR [2]	BoMSCR	SSBoMSCR	PB MSCR [2]	PB BoMSCR	PB SSBoMSCR
		K=128	KV=64 KS = 8		K=128	KV=64 KS = 8
N=1	87.4604	88.9698	**92.0082**	89.8122	88.8490	90.5143
N=2	90.4678	94.7290	95.0694	95.1020	95.5918	**97.1429**
N=3	91.1184	96.4131	97.5265	95.8531	96.4980	**98.0000**
N=4	91.6269	96.5543	97.8449	96.2204	96.9551	**98.7755**
N=5	92.7371	96.7608	**98.1796**	97.1429	96.8327	98.6531

TABLE IV
nAUC FOR ETHZ2 DATASET.

	MSCR [2]	BoMSCR	SSBoMSCR	PB MSCR [2]	PB BoMSCR	PB SSBoMSCR
		K=128	KV=64 KS = 8		K=128	KV=64 KS = 8
N=1	88.9643	86.2054	91.1480	**93.3163**	88.0357	88.0230
N=2	94.5446	94.1352	95.2423	95.5867	93.3036	**96.8750**
N=3	95.3954	95.3763	96.1480	**97.3597**	93.8265	96.5434
N=4	96.5077	95.5753	95.7143	**97.7296**	94.0051	97.0281
N=5	96.5319	95.3253	95.9566	**98.1760**	94.2857	96.8367

TABLE V
nAUC FOR ETHZ3 DATASET.

[2] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2010)*. San Francisco, CA, USA: IEEE Computer Society, 2010.

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

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

[5] O. Javed and M. Shah, *Automated Multi-Camera Surveillance: Algorithms and Practice*, 2008.

[6] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the International Conference on Computer Vision (ICCV)*. Washington, DC, USA: IEEE Computer Society, 1999, p. 1150.

[7] P.-E. Forssén, "Maximally stable colour regions for recognition and matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society. Minneapolis, USA: IEEE, June 2007.

[8] 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.

[9] P.-E. Forssén and D. Lowe, "Shape descriptors for maximally stable extremal regions," in *IEEE International Conference on Computer Vision*, vol. CFP07198-CDR. Rio de Janeiro, Brazil: IEEE Computer Society, October 2007.

[10] K. L. M. Cho, "Bilateral symmetry detection and segmentation via symmetry-growing," in *BMVC 2009: Proceedings of the 20th British Machine Vision Conference*, 2009.

[11] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context-free attentional operators: The generalized symmetry transform," *International Journal of Computer Vision*, vol. 14, no. 2, pp. 119–130, 1995.

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

[13] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recongnition, reacquisition and tracking." in *Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2007.

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

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