A structural approach to Person Re-identification problem

Amal Mahboubi, Luc Brun, Donatelo Conte

A structural approach to Person Re-identification problem

Amal Mahboubi
Normandie Univ,
UNICAEN, ENSICAEN,
CNRS, GREYC,
F-14050 Caen, France
Email: amal.mahboubi@unicaen.fr

Luc Brun
Normandie Univ,
UNICAEN, ENSICAEN,
CNRS, GREYC,
F-14050 Caen, France
Email: luc.brun@ensicaen.fr

Donatelo Conte
Université de Tours
LI EA 6300
64, Avenue Jean Portalis
F-37200, Tours, France
Email: donatello.conte@univ-tours.fr

***Abstract*—Although it has been studied extensively during past decades, object tracking is still a difficult problem due to many challenges. Several improvements have been done, but more and more complex scenes (dense crowd, complex interactions) need more sophisticated approaches. Particularly long-term tracking is an interesting problem that allow to track objects even after it may become longtime occluded or it leave/re-enter the field-of-view. In this case the major challenges are significantly changes in appearance, scale and so on.**

At the heart of the solution of long-term tracking is the re-identification technique, that allows to identify an object coming back visible after an occlusion or re-entering on the scene. This paper proposes an approach for pedestrian re-identification based on structural representation of people. The experimental evaluation is carried out on two public data sets (ETHZ and CAVIAR4REID datasets) and they show promising results compared to others state-of-the-art approaches.

I. Introduction

The purpose of re-identification (re-id) is to identify people coming back into the field of view of a camera or to recognize an individual through different cameras in a distributed network. The re-id task can be organized in two families of approaches: biometric approaches and appearance-based approaches. Biometric approaches based on some fine characteristics of each person are inadequate in video surveillance scenarios. Indeed, the low camera resolution and pose variation in video surveillance scenario does not allow to perform a reliable identification based on biometric techniques. Conversely, appearance-based methods detailed in Section II are based on global features such as the distributions of colors or textures which allow to capture the global appearance of a person. Conversely to biometric techniques appearance based method suppose that the re-id occurs in a short time delay so that the global appearance of a person remains unchanged.

In order to address the person re-id problem, three aspects are involved: (1) The detection of blobs enclosing persons in a video stream, (2) the extraction of significant features from each blob in order to get informative signatures, (3) the re identification task from signatures given probe and gallery sets. A gallery set is an image dataset of people with known identities. It encodes the set of persons which may be re identified. A probe set contains the query images of a person whose identity has to be identified or who appears for the first time in the video stream.

In this paper, we propose two structural signatures : one based on a RGB string centered on a person and the other one based on a segmentation of the blob enclosing a person and the encoding of this segmentation through a Region Adjacency Graph (RAG). For each structural description we propose an appropriate kernel which is used for the re identification task.

The remainder of this paper is organized as follows. Section II reviews related works and describes the re-id problem. Sections III and IV present the details of the proposed approaches, one based on RGB string kernel and the other based on region with edition kernel. We also outline the available datasets for performance evaluation in Section V. Finally, we conclude the paper in Section VI.

II. Related Works

Re-identification consists in identifying a person who came back in a scene where it has been previously detected. Re-identification is indispensable either for on-line tracking of an individual over a network of cameras, or for off-line retrieval of all scenes containing a given person. On this basis, we can define more formally this problem.

Given a set I of images of H known individuals a re-id problem is defined by a triplet (L, G, P), where:

$\surd$ L is a set of H identities $L = \{1, 2, \ldots, H\}$

$\surd$ G is a set of gallery images whose label is known $G = \{g_1, g_2, \ldots, g_N\}$; $N \geq H$

$\surd$ P is a set of probe images (whose labels have to be identified). $P = \{p_1, p_2, \ldots, p_M\}$

$\surd$ $I = G \cup P$ (the gallery and probe sets are disjoint)

Furthermore, we define a map identity function:

$$l : I \rightarrow L$$
$$\forall x \in I, l(x) \in L$$

Thus, for a person x we can define the three following sets:

$$I_x = \{i \in I / l(i) = x\}$$
$$G_x = \{i \in G / l(i) = x\}$$
$$P_x = \{i \in P / l(i) = x\}$$

Note that we have $I_x = G_x \cup P_x$.

Three different re-id modalities are usually distinguished [8]:

- **Single-versus-All** re-id (SvsAll): in this scenario, the probe set contains a single occurrence of a person. The gallery set may contain several examples of a same person but in this case each occurrence is considered as a singleton. Hence an input example in mapped onto a given person is the re identification is performed with at least one of its occurrences in the gallery set.
- **Multiple-versus-Single** re-id (MvsS): in which multiple images of each individual are given in the gallery set, while a single image of each person is used in the probe set. The gallery set requires the extraction of several images of the same person to build the representation of the person, while the probe set uses a single image. Thus for each person x we have $|G_x| = k > 1$ and $|P_x| = 1$. The MvsS modality is little used in the literature.
- **Multiple-versus-Multiple** re-id (MvsM): in which both of the gallery and the probe sets of each individual have k images ($k > 1$). Thus for each person x we have $|G_x| = k$ and $|P_x| = k$

The task of re-id consists in defining for each probe individual a ranked list of identities using the gallery. Then, we refer to SvsAll as single-shot and to MvsM as multiple-shot.

Appearance-based re-identification methods can be divided into two groups: the learning methods and the feature oriented methods. The former group usually requires an initial training phase and needs to be frequently re-trained while the latter group works on each person independently.

Common distance metrics (Bhattacharyya, Euclidean) or nearest-neighbor algorithm are traditionally adopted in learning-based approaches. In [13], the individual signature (color, texture and edges) is learned using a partial least squares model. The classification is performed using the Euclidean distance. Bak [1] uses multiple images of each person to build a person's signature. The similarity measure between signatures is based on a mean Riemannian covariance. Another group of work falling in the learning approaches is distance metric learning [9, 12]. The basic idea of this technique is to learn an optimal metric under which instances belonging to a same person are closer than instances belonging to different persons.

In feature-based approaches, a set of non learned features are extracted, the whole set of features defining the signature of each person. Then a matching between probe signatures and gallery signatures is realized. The work [4], proposes an ID signature based on pictorial structures in one-shot case and custom pictorial structures in multiple-shot case. The signatures of the probe and the gallery images are matched using a Bhattacharyya distance. In [2], a symmetry-driven appearance-based descriptor is proposed. The matching is carried out by a log-likelihood estimation.

Classically, the person re-id framework has to cope with three aspects; First, a person detector module is needed to determine which parts relate to a person. Second, features extraction and reliable signatures computation. Third, matching the probe images with the gallery images.

A. Person segmentation

Different segmentation strategies can be considered for re-id as background subtraction or foreground detection with shadow removals or appearance-based pedestrian detectors In the re-id literature, it is generally assumed that the bounding boxes for all the individuals in the dataset are available. Consequently, in this study we assume that the mask of the objects have already been extracted.

B. Features and Representations

In real video surveillance scenario the gallery set only provides few example of each persons. This last point is problematic for many machine learning methods which implicitly assume large training sets. In this context, feature based approaches appear to us more effective. The problem of feature approaches is to find a good signature that captures most of the relevant information. However, one drawback of such an approach is the definition of a signature by a bag of features which do not allows to properly capture the 2D spatial coherence of these features. The idea behind our proposal consists in adding some structure to these bags using two different approaches. Furthermore, as the expressiveness of the signature lies mainly on the structural aspect, it allows us to use simple features (like color, area, etc.).

Features descriptors such as color (histogram, Gaussian models), texture (covariance matrices, SURF descriptors), shape (width, height, contours) or a combination or those features are usually used in appearance based method.

Re-id methods make the assumption that the clothing of the person does not change. The main features used in the literature are color histograms , Gaussian mixtures , histogram of oriented gradients (HOG) , scale-invariant future transform (SIFT) , covariance matrix, etc. In our work, we propose two representations: RGB string kernel and Region-based with edition kernel. The people description using RGB string and Region-based with edition kernel will be detailed in section III and section IV respectively.

C. Matching

Once features have been extracted from both probe and gallery sets the matching step consists in identifying each image of the probe set from the gallery set. This last step may be performed by using an SVM classifier or by using a closest neighbor method both using the gallery set as train set. Thus, we propose the following method: from the features we build a distance that returns the identity of the individual based on one of the protocols defined above in section II. So as soon as the visual features are extracted for the probe images set (the query) the matching can be started.

For RGB string the matching between the query and all the persons in the gallery images set is achieved by maximizing the similarity between each possible pair of persons contained in both sets. This process will be detailed in sections III-B and III-C. The same matching is used using the Region-based representation, replacing RGB string Kernel by region-based Kernel with edition.

III. RGB String Kernel

Our RGB string kernel is based on a previous contribution [10]. This approach is based on the description of each person by a RGB string descriptor. Let us consider the bounding box $W \times H$ of an object obj_a whose top-left corner's coordinates are denoted (tl_x,tl_y). The obj_a, should be delineated by two main contours (for example using Deriche filter). For each value $h \in \{0, \ldots, H-1\}$, we consider the horizontal line segment defined as the intersection between the bounding box of obj_a and the line $y_h = h + tl_y$. The x coordinate of the central point of obj_a at height y_h is denoted $\overline{x}_h$ and is defined as the x coordinate of the weighted mean of all points along the line segment. More precisely, $\overline{x}_h$ is defined as:

$$\forall h \in \{0, \ldots, H-1\} \quad \bar{x}_h = \frac{\sum_{w=0}^{W} |\nabla I(x_w, y_h)|^2 . x_w}{\sum_{w=0}^{W} |\nabla I(x_w, y_h)|^2}, \quad (1)$$

where $x_w = tl_x + w$, $I(x_w, y_h)$ denotes the pixel's value of (x_w, y_h) and $|\nabla I(x_w, y_h)|$ is the amplitude of its gradient.

Naturally equation 1 is not used as it stands but optimized to mitigate in problems of perturbations of the gradient and curve discontinuities using equation 2.

$$\jmath(x) = \sum_{y=1}^{n} (\bar{x}_y - x_y)^2 + \lambda (x_y - x_{y-1})^2 \quad (2)$$

where λ is a tuning parameter. The average coordinate $\bar{x}_y$ is given by equation 1 and x_y is the corresponding final coordinate. The symbol n denotes the height of the bounding box.

The energy function equation 2 combines two terms: the first one encodes the distance to the average coordinate. The second term is a regularization term which enforces the continuity of the curve. The energy regularization of equation 2 is achieved through a gradient based descent minimization and may be found in [10].

A. People description

RGB string encoding of people's appearance may be altered by the addition of erroneous extremities encoding, for example, a part of the floor or a difference of sampling due to the variations of the distance between a person and the camera. In order to cope with such variations we consider each curve as a string and encode the similarity between two strings using the global alignment kernel defined by [3]:

$$K_{GA}(s_1, s_2) = \sum_{\pi \in A(n,m)} e^{-D_{s_1,s_2}(\pi)}, \quad (3)$$

where n and m, denote the length of the first string s_1 and the second string s_2 respectively.

The symbol D denotes the Dynamic Time Warping distance. It measures the discrepancy between two strings s_1 and s_2 according to an alignment π. Function D is defined [3] as:

$$D_{s_1,s_2}(\pi) = \sum_{i=1}^{|\pi|} \varphi(x_{\pi_1(i)}, y_{\pi_2(i)}), \quad (4)$$

where $s_1 = (x_i)_{i \in \{1,..,n\}}$, $s_2 = (y_i)_{i \in \{1,...m\}}$ and function φ corresponds to a distance function defined in [3] as follows:

$$\varphi(x, y) = \frac{1}{2\sigma^2} \|x - y\|^2 + \log(2 - e^{-\frac{\|x-y\|^2}{2\sigma^2}}), \quad (5)$$

where x and y denote the RGB values of the first object and the second object respectively. Symbol σ denotes a tuning parameter. The log term is added to the squared Euclidean distance $\|x-y\|^2$ in order to ensure the definite positiveness of K_{GA} (equation 3) [3]. Note that, using equation 4, equation 3 may be computed using a slightly modified version of the classical string edit distance algorithm. The computational complexity of equation 3 is thus bounded by $\mathcal{O}(nm)$ where n and m denote respectively the length of s_1 and s_2.

B. Single-shot matching

Given a person x ($I_x \subset I$) whose label should be identified, the probe set P_x is one singleton, while the gallery set is the union of singleton of each individual. We will detail the choice of singletons in section V-B. In order to compare the similarity between two individuals, we normalize the kernel between two strings using the following formula:

$$\tilde{k}(s, s') = \frac{K_{GA}(s, s')}{\sqrt{K_{GA}(s, s) K_{GA}(s', s')}} \quad (6)$$

Then, the matching between the probe and each gallery sub set is measured using the distance defined by the equation:

$$d^2(s_P, s_{G_i}) = 1 - \tilde{k}(s_P, s_{G_i}) \; i = 1..H \quad (7)$$

where s_P is the string of the probe set and s_{G_i} is the string of individual i.

Finally, the identity of the probe $l(P)$ is the smallest element in the ranked list of identities obtained using equation 7. Thus $l(P) = \arg\min_{i=1}^{N} d^2(s_P, s_{G_i})$ where N is the number of sub set in G.

C. Multiple-shot matching

The appearance of a person is established using a set of N frames, where each person is described by one RGB string in each frame. As in single-shot matching, we will detail the choice of images which establish the sets G and P in section V-B.

Let S_A and S_B denote two sets encoding two persons A and B. The similarity between A and B is given by comparing each possible pair of RGB string contained in both sets S_A and S_B in order to keep only the closest pair. Based on equation 3, the similarity between A and B is defined as follows:

$$SIM_{MvsM}(A, B) = \max_{i=1,j=1}^{N} \tilde{k}(s_{A_i}, s_{B_j}) \quad (8)$$

where N is the number of frames for each person.

Re-id consists in using equation 8 to compute for each probe individual P of an unknown person x a ranked list of individuals from the gallery. Thus, given a probe P and the gallery set G the re-id is carried out by:

$$l(P) = \max_{i=1}^{N} (SIM_{MvsM}(P, G_i)) \quad (9)$$

where N is the number of sub set in G.

IV. Region-based with edition Kernel

First of all it is important to define the graph representation for understanding the second proposed kernel. Detected foreground regions are segmented using Statistical Region Merging (SRM) algorithm [11]. Finally, the segmentation of the mask within each rectangle is encoded by a Region Adjacency Graph (RAG). Two nodes of this graph are connected by an edge if the corresponding regions are adjacent.

A. Attributes

Each node of the graph have the following attributes: the RGB average color of the region, the size S (in pixels) of the region and the proportion η of the region area with respect to the overall (object) image area.

Therefore, first we define a color distance between two nodes a and b as:

$$d_c(a,b) = \sqrt{(2+\frac{\overline{r}}{256})\delta_R^2 + 4\delta_G^2 + (2-\frac{255-\overline{r}}{256})\delta_B^2} \quad (10)$$

where $\overline{r} = \frac{R_a+R_b}{2}$ and δ_R, δ_G and δ_B encode respectively the differences of coordinates along the red, green and blue axis. This definition of color distance better fit human perception than a simple Euclidean distance between the three channels.

Such a distance is used to define the Kernel between two nodes:

$$K_v(a,b) = e^{\frac{|\eta_A-\eta_B|}{2\sigma_1^2}} \cdot e^{\frac{-d_c(a,b)}{2\sigma_2^2}} \quad (11)$$

In Eq. 11 σ_1 and σ_2 are tuning variables to weight the similarity between two regions. Whit this kernel two regions are similar if theirs size and color are similar.

Edges represent only adjacency between regions and they do not have attributes. We have decided to exclude attributes for edges in order to have invariance to rotation of the object.

B. The Kernel definition

Let us consider a graph $G = (V, E)$ where V denotes the set of vertices and $E \subset V \times V$ the set of edges. A bag of paths P associated to G is defined as a set of paths of G whose cardinality is denoted by $|P|$. GivenK_{path} a path kernel. Given two graphs G_1 and G_2 and two path $h_1 \in P_1$ and $h_2 \in P_2$ of respectively G_1 and G_2, $K_{path}(h_1, h_2)$ may be interpreted as a measure of similarity between h_1 and h_2. The aim of a Bag of Path Kernel consists in aggregating local measures between pairs of paths into a global similarity measure between the two graphs.

Let us consider a generic Path Kernel $K_{path}(h_1, h_2)$ between two paths from two graphs: $h_1 \in G_1$ and $h_2 \in G_2$. We can design a convolution kernel called *mean kernel* as follows:

$$K_{mean}(P_1, P_2) = \frac{1}{|P_1|}\frac{1}{|P_2|}\sum_{h_1 \in P_1}\sum_{h_2 \in P_2} K_{path}(h_1, h_2). \quad (12)$$

where P_1 and P_2 denote the two bags of paths of, respectively, the graphs G_1 and G_2.

Following Haussler [6], this kernel is positive definite on the bag of paths domain if and only if K_{path} is positive definite on the path domain.

The above definition of the kernel (Eq. (12)) does not consider the relevance of a path in the bags. As a consequence, irrelevant paths may corrupt the overall value of the kernel due to the "meaning" effect. As in the literature we limit this effect by using a relevance measure of paths. Let P_1 and P_2 two bags of path, the *weighted mean kernel* $K_{weighted}(P_1, P_2)$ between these two bags is then defined as:

$$\frac{1}{|P_1|}\frac{1}{|P_2|}\sum_{h_1 \in P_1}\sum_{h_2 \in P_2} w(h_1)w(h_2)K_{path}(h_1, h_2) \quad (13)$$

where h_1 and h_2 denote the two paths in the sets P_1 and P_2 and $w(h_i)$ encode the relevance of the path h_i within the set P_i.

The importance weight of h in the set P is defined as the average of the area of the regions corresponding to the nodes belonging to the path:

$$w(h) = avg(\eta_1, \cdots, \eta_n) \quad (14)$$

where η_i is the area weight of the i-th vertex belonging to the path.

The kernel defined by eq. (13) is still a convolution kernel and so is positive definite if K_{path} is positive definite.

The kernel between two paths $h_1 = (v_1^1, \ldots, v_n^1)$ and $h_2 = (v_1^2, \ldots, v_p^2)$ is defined as 0 if both paths have not the same size and as follows otherwise:

$$K_{classic}(h_1, h_2) = \prod_{i=1}^{|h|} K_v(v_i^1, v_i^2) \quad (15)$$

The terms K_v denotes the kernel for node's attributes. We use the Kernel defined by equation 11 (section IV-A).

In the comparison of two images of a person, the effect of small perturbations on their appearance, due to illumination or pose changes, is not negligible. Consequently, the graphs representing the two images may be very different even if they belong to a same person. Therefore, in the definition of a similarity between two graphs, we have to deal with this kind of problem. The principal change in appearance is the fusion of adjacent regions. In a graph representation this effect results in an edge contraction operation.

The edge contraction operation contracts an edge and merges its two extremity nodes. An edition cost is associated to this operation. This cost encodes the relevance of the merging of the two nodes and it can be derived by the merging criterion of the segmentation step according to the algorithm SRM in [11]. Let define, for a node v, the function (from [11]):

$$b(v) = 256\sqrt{\frac{1}{2Q|v|}\ln\frac{|\mathcal{R}_{|v|}|}{\delta}} \quad (16)$$

where $\delta = 1/(6|I|^2)$ ($|I|$ is the image size), $|v|$ is the node size, Q is a parameter of the segmentation and $\mathcal{R}_{|v|}$ is the

set of nodes with the same size than v. Therefore the edge contraction cost is defined as:

$$w_e(v_1, v_2) = \frac{\max_{k=r,g,b} \left| \overline{R}_k(v_1) - \overline{R}_k(v_2) \right|}{\sqrt{b^2(v_1) + b^2(v_2)}} \quad (17)$$

where v_1 and v_2 are the extremity nodes of the contracting edge and the function $\overline{R}_k(v)$ states for the average of the color channel k within the region represented by the node v. A high value of this cost means that it is unlikely that the two regions merge each others.

Moreover, we suppose that this attribute is additive: the weight of two consecutive edges along a path is the sum of both weights.

Let us denote by κ the function which applies the cheapest edge contraction on a path and D the maximal number of reductions. The successive applications of the function κ associates to each path h a sequence of reduced paths $(h, \kappa(h), \ldots, \kappa^D(h))$. Each $\kappa^k(h)$ is associated to a cost: $cost_k(h)$ defined as the sum of the costs of the κ operations yielding $\kappa^k(h)$ from h. Using $K_{classic}$ for the path comparison, we define the kernel K_{edit} as a sum of kernels between reduced paths.

Given two paths h_1 and h_2, the kernel $K_{edit}(h_1, h_2)$ is defined as:

$$\frac{1}{2D} \sum_{k=0}^{D} \sum_{l=0}^{D} e^{-\frac{cost_k(h_1) + cost_l(h_2)}{2\sigma_{cost}^2}} K_{classic}(\kappa^k(h_1), \kappa^l(h_2)) \quad (18)$$

where σ_{cost} is a tuning variable.

The kernel $K_{classic}$ is a tensor product kernel based on positive-definite kernels, so it is positive-definite. The kernel over edition costs is constructed from a scalar product and is thus positive-definite. These two last kernels form a tensor product kernel. Finally K_{edit} is proportional (by a factor $2D$) to a R-convolution kernel [6], thus is positive-definite.

V. Experiments

A. Datasets

We evaluated the performance of our kernels on several publicly available re-id datasets. We assume that the mask of the objects have already been extracted by [5].

$\surd$ ETHZ [13] for this dataset, images are extracted from a single camera. The most challenging aspects of ETHZ are the illumination changes and the occlusions.

$\surd$ CAVIAR4REID dataset [4] contains 1220 images of 72 people distributed as follow: 50 peoples with 20 images (captured by two surveillance cameras) and 22 peoples with 10 images (captured by one surveillance camera). The challenge of this dataset are: the images are extracted from two cameras, the illumination changes and the occlusions, an important change in the image resolution, with a minimum and maximum size of 17×39 and 72×144, respectively.

The ETHZ and CAVIAR4REID datasets can be used in single-shot and multiple scenario. Besides the CAVIAR4REID dataset is more challenging due to the characteristics summarized above.

B. Experimental setup

In reviewing the literature, the authors [5] and [2] adopt the following random experimental protocol for both single-shot and multiple-shot scenarios.

$\surd$ **Multiple-shot:** We randomly select a subset of N images for each individual to build the gallery set and probe set ($|P_x| = N$, $|G_x| = N$). For each N, we repeat all experiments k times ($k = 100$) to obtain reliable statistics. Regarding ETHZ dataset, $N = \{2, 5, 10\}$. Considering CAVIAR4REID dataset, $N = 5$. Fixing N is not always an easy task. Indeed, usually, adding more information enrich the model. However, there is a limit above which adding more information does not enhance the results. For [2] this limit is $N = 5$ for ETHZ and CAVIAR4REID.

$\surd$ **Single-shot:** We randomly select one image for each person to build the gallery set, while the other images form the probe set $|P_x| = 1$, $|G_x| = 1$. For each image in the probe set the position of the correct match is obtained. The whole procedure is repeated 10 times.

C. Random Protocol results

Results are shown according to re-identification rates extracted from CMC curves. The re-identification rate at rank r represents the probability to find the right match among the first r matches. Table I shows the performances for the three sequences of ETHZ and CAVIAR4REID using the single-shot protocol. As can be seen in these table, the one-shot protocol results using edit kernel are worse than those obtained by using RGB string. Throughout this point, we will simplify the analysis by considering only the RGB string kernel. We refer to the results obtained using RGB string kernel as our result.

We compare our method with several representative methods in the literature: SDALF [2] and SURF [7]. The rank 1,5,10, 15 matches rates are reported in Table II. We first observe that in single-shot protocol we are not performing. This finding seems to be imputable to the fact that, in the case of strong occlusions, the string RGB provides a corrupt representation of the person. This drawback explains the bad results of seq1 in the single-shot protocol. Nevertheless, this drawback is reduced by the use of the sliding window HTW, Indeed the performances show an improvement in the MvsM protocol. Regarding MvsM random protocol, It can be seen that our method achives moderate performance on seq1 and seq2. However, our results for seq3 and CAVIAR4REID datasets are better than [2] and [7]. This finding may be explained by the fact that, in absolute terms there is no perfect method. Generally speaking, the more robust a method is, the less precise it is. Our model conveys a good spatial description of the person. The matching deployed gives good results in the presence of moderate occlusions (seq3 and CAVIAR4REID); unfortunately it is not robust to the strong occlusions and large changes of illumination existing in seq1 and seq2 (table III). Further investigations on occlusions and the addition of a color normalization step are planned in our future developments.

TABLE I
SINGLE-SHOT RESULTS

method	r=1	r=5	r=10	r=15
CAVIAR4REID RGB string	16.66	40.27	51.38	56
CAVIAR4REID edit kernel	31.37	42.15	51.96	53.92
ETHZ-seq1 RGB string	36.14	56.62	69.87	78.31
ETHZ-seq1 edit kernel	37.72	43.71	52.09	56.88
ETHZ-seq2 RGB string	77.14	90	94	97
ETHZ-seq2 edit kernel	42.46	61.64	79.45	86.30
ETHZ-seq3 RGB string	39.28	78.57	89	92
ETHZ-seq3 edit kernel	52.83	64.15	75.47	77.35

TABLE II
COMPARISON WITH STATE OF THE ART, WHERE '-' DENOTES NO RESULT REPORTED FOR THE METHOD

ETHZ dataset

	seq1		seq2		seq3	
method	r=1	r=5	r=1	r=5	r=1	r=5
our single-shot	36.14	56.62	77.14	90	39.28	78.57
[2] single-shot	65	81	64	85	76	90
our multiple-shot	62.65	85.54	77.14	98	85.71	100
[2] multiple-shot	90	94	90.5	98	94	97

CAVIAR4REID dataset

method	r=1	r=5	r=10	r=15
our single-shot	16.66	40.27	51.38	56
[2] single-shot	10	25.8	45	60
our multiple-shot	23.16	56.94	77.77	87.5
[2] Multiple-shot	18	50	68	80
[7] Multiple-shot	20	-	-	-

TABLE III
THE RESULT OF THE QUERY-MISCLASSIFIED EXAMPLES

seq1		seq2	
probe	match r=1	probe	match r=1
			
			

VI. CONCLUSIONS

In this paper, we focus on the problem of people re-identification using appearance based re-identification methods. Furthermore, we propose two structural descriptions to provide a discriminating signature for each person. For each signature we propose an appropriate kernel which is used for the re-identification task. The results are promising and experiments show that the proposed approach is comparable, often better, that some good existing approaches. In the future we plan to use some more complex structural representation in order to keep more information on the appearance model and so having improved results.

REFERENCES

[1] Slawomir Bak et al. "Boosted human re-identification using Riemannian manifolds". In: *Image and Vision Computing* 30.6-7 (2012), pp. 443–452.

[2] Loris Bazzani, Marco Cristani, and Vittorio Murino. "Symmetry-driven accumulation of local features for human characterization and re-identification". In: *Comput. Vis. Image Underst.* 117.2 (Feb. 2013), pp. 130–144.

[3] Marco Cuturi. "Fast Global Alignment Kernels". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. Ed. by Lise Getoor and Tobias Scheffer. ICML '11. Bellevue, Washington, USA: ACM, June 2011, pp. 929–936.

[4] Cheng Dong Seon et al. "Custom Pictorial Structures for Re-identification". In: *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, 2011, 68.1–68.11.

[5] M. Farenzena et al. "Person re-identification by symmetry-driven accumulation of local features". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2010, pp. 2360 –2367.

[6] David Haussler. *Convolution Kernels on Discrete Structures*. Technical Report UCS-CRL-99-10. Department of Computer Science, University of California at Santa Cruz, 1999.

[7] Mohamed Ibn Khedher, Mounim El Yacoubi, and Bernadette Dorizzi. "Fusion of appearance and motion-based sparse representations for multi-shot person re-identification". In: *Neurocomputing* 248 (July 2017), pp. 94 –104.

[8] Svebor Karaman et al. "Leveraging local neighborhood topology for large scale person re-identification". In: *Pattern Recognition* 47.12 (2014), pp. 3767 –3778.

[9] Martin Köstinger et al. "Large scale metric learning from equivalence constraints". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. 2012, pp. 2288–2295.

[10] Amal Mahboubi et al. "Tracking System with Re-identification Using a RGB String Kernel". In: *Structural, Syntactic, and Statistical Pattern Recognition - Joint IAPR International Workshop, S+SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings*. 2014, pp. 333–342.

[11] Richard Nock and Frank Nielsen. "Statistical Region Merging". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 26.11 (Nov. 2004), pp. 1452–1458.

[12] Bryan Prosser et al. "Person Re-Identification by Support Vector Ranking". In: *Proceedings of the British Machine Vision Conference*. BMVA Press, 2010, pp. 21.1–21.11.

[13] W.R. Schwartz and L.S. Davis. "Learning Discriminative Appearance-Based Models Using Partial Least Squares." In: *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing*. IEEE Computer Society, 2009, pp. 322–329.