SDALF+C: Augmenting the SDALF Descriptor by Relation-based Information for Multi-shot Re-identification

Sylvie Jasmine Poletti[†], Vittorio Murino[⊕†], and Marco Cristani[†⊕]

[†] Department of Computer Science, University of Verona (IT)
[⊕] Pattern Analysis and Computer Vision Dept., Istituto Italiano di Tecnologia (IT)

Abstract. We present a novel multi-shot re-identification method, that merges together two different pattern recognition paradigms for describing objects: feature-based and relation-based. The former aims at encoding visual properties that characterize the object per se. The latter gives a relational description of the object considering how the visual properties are interdependent. The method considers SDALF as feature-based description: SDALF segregates salient body parts, exploiting symmetry and asymmetry principles. Afterwards, the parts are described by color, texture and region-based features. As relation-based description we consider the covariance of features, recently employed for re-identification: in practice, the parts found by SDALF are additionally encoded as covariance matrices, capturing structural properties otherwise missed. The resulting descriptor, dubbed SDALF+C, is superior to SDALF by about 2% and to the covariance-based description by a 53%, both in terms of average rank1 probability, considering 5 different multi-shot benchmark datasets (i-LIDS, ETHZ1,2,3 and CAVIAR4REID).

Keywords: re-identification, SDALF, covariance of features

1 Introduction

People re-identification (re-id) has definitely become a primary module for the multi-camera video surveillance systems, allowing to recognize individuals across different locations and times. The re-id literature can be partitioned in different ways: *direct* vs. *learning-based*, and *single-shot* vs. *multi-shot* methods. Direct approaches [2, 1, 3] are on-line feature extractors, while learning-based techniques [11, 7, 12, 10, 4] require a training phase prior to work. Single-shot [2, 11, 7, 10, 4] and multi-shot [2, 12, 1, 3] approaches differ for the number of images exploited to describe each probe or gallery subject: multi-shot strategies employ several shots (images) for building an individual signature.

In this paper, we present an approach for direct, multi-shot re-identification, that aims at joining two different ways to represent objects, employing *feature-based* and *relation-based* descriptions. Features serve to encode the tangible aspects of an entity, while relation-based descriptions explain how these aspects are inter-related. Both approaches have their pros and cons. Features are intuitive to understand and easy to extract, but cannot usually describe structural

information. Relation-based representations are mostly suitable to encode structural information, but their effectiveness is usually limited to this purpose. In re-id, most of the descriptors are feature-based, while the sole relation-based representation is the covariance of features [1].

Our approach aims at joining both paradigms, exploiting SDALF [2] as feature-based descriptor. SDALF is a symmetry-based description of the human body, and it is inspired by the well-known principle that natural objects manifest symmetry in some form. Using symmetry and asymmetry principles, SDALF isolates three human body regions, usually corresponding to the head, the torso and the legs. After that, torso and legs regions are described by heterogeneous features, and matched by minimizing a proper distance. Our approach complements this scheme, by adding relation-based descriptions: essentially, the body regions found by SDALF are encoded as Mean Riemannian Covariances (MRCs) [1], which are semidefinite positive descriptors built by fusing multiple covariances of features, these latter encoding each shot available of an individual. MRCs are then added to the final descriptions (one for each body region). This produces a novel method, dubbed here SDALF+C.

In the experiments, we show that SDALF+C is an effective solution for direct multi-shot re-identification, allowing to get better results than their single components, on five different multi-shot benchmark datasets (i-LIDS, ETHZ1,2,3 and CAVIAR4REID).

The rest of the paper is organized as follows. In Sec. 2, SDALF and the Mean Riemannian Covariance Grid (MRCG [1], from which the MRC descriptor is extrapolated) are briefly summarized. Sec. 3 details our approach, and Sec. 4 presents the experimental results. Finally, in Sec. 5, conclusions are drawn and future perspectives are envisaged.

2 Fundamentals

2.1 Symmetry Driven Accumulation of Local Features (SDALF)

Let us suppose to have M images portraying an individual: the SDALF descriptor starts by isolating the foreground (the human body) employing the STEL generative model [9]. After that, SDALF individuates three main body parts (head, torso, legs) by exploiting horizontal asymmetry principles: the rationale is that the head and the torso are horizontally asymmetric (with respect to area and color), and the same applies for the torso and the legs. On the other hand, vertical symmetry criteria allow to weight more those features which are located near the vertical axis of symmetry of the human body, thus pruning out distracting background clutter that lies on the peripheral portions (see Fig. 1 for some examples).

Given the two regions $Reg_{\text{torso}}, Reg_{\text{legs}}$ (the head is discarded as only a few pixels do not contain enough discriminative content), SDALF extracts complementary visual aspects of the human body appearance, highlighting: i) the global chromatic content by the color histogram (in the multiple-shot case, M

histograms for each part are considered); ii) the per-region color displacement employing Maximally Stable Colour Regions (MSCR) [6]; iii) the presence of *Recurrent Highly Structured Patches* (RHSP), estimated by a per-patch similarity analysis. in the multiple-shot case, it is worth noting that 1) the MSCRs are opportunely distilled from the M images by employing a Gaussian clustering procedure [5], which automatically selects the number of components keeping the means, and 2) the RHSP descriptors are extracted considering different frames.

This process applies for all the M individuals of the probe and the gallery sets, obtaining M different signatures. Each signature of the probe set is then compared with the gallery set, looking for a match. To this aim, a proper distance d_{SDALF} is employed. For further details, please refer to [2].

2.2 Mean Riemannian Covariance Grid (MRCG)

Let I be an image and F be a d-dimensional feature image extracted from I,

$$F = \theta(I)$$

where function θ can be any set of d mappings, such as color, intensity, gradients, filter responses, etc.. For a given rectangular region $Reg \subset F$, let $\{f_h\}_{h=1,\dots,n}$ be the d-dimensional feature points inside Reg (n is the number of feature points, e.g. the number of pixels). We represent region Reg by the $d \times d$ covariance matrix of the feature points

$$C_{\mathrm{Reg}} = \frac{1}{n-1}\sum_{h=1}^{n}(f_h - \mu)(f_h - \mu)^{\intercal} \tag{1}$$

where μ is the mean of the feature points.

In the original approach, each of the M images of the subject A is decomposed in K patches, where each patch has a fixed location in the image plane. For each patch instance (intended as the patch content of a single image), d dense features are extracted, so that a $d \times d$ covariance matrix can be built for each patch instance. To distill a single descriptor for each patch, which takes into account all the M images of the same subject (i.e., all his patch instances), the Mean Riemannian Covariance (MRC) is calculated, by computing the Karcher mean [8] on all the local covariances. In practice, for patch k, a "mean" covariance $\mu_{A,k}$ is built, which summarizes all the correspondent patch instances. Then, in order to weight each MRC, a discriminant index is computed, which considers how different is a particular patch (i.e., its related MRC), from all the correspondent patches of all the other probe images that should be taken into account. In practice, for patch k, a discriminant $\sigma_{A,k}$ is created. The same approach is applied on the gallery images. At the end of the process, each patch is described by an MRC, and a discriminant index. To match the probe with a gallery subject B, a distance is calculated, which has the following form

$$d_{MRC}(A,B) = \sum_{k=1,\dots,K} \frac{\sigma_{A,k} + \sigma_{B,k}}{\rho(\mu_{A,k}, \mu_{B,k})} \tag{2}$$

where ρ is a proper distance between covariance matrices. Minimizing such distance gives the best match. For further details, please refer to [1].

3 Our approach

Our approach wants to combine SDALF and MRC, since the two descriptors are highly complementary. While SDALF extracts heterogeneous visual properties from the human appearance, MRC tells how visual properties are related with each other. For this purpose, SDALF is run in its original version, obtaining for person A a given descriptor. After that, on the torso and the legs regions $Reg_{\text{torso}}, Reg_{\text{legs}}$ found by SDALF, for all the M images of the same individual, d-dimensional covariances matrices are built, following Eq. 1. The following $d = 11$ features are taken into account:

$$\left[x, y, R_{xy}; G_{xy}; B_{xy}; \nabla^R_{xy}, \theta^R_{xy}, \nabla^G_{xy}, \theta^G_{xy}, \nabla^B_{xy}, \theta^B_{xy}\right] \tag{3}$$

where x and y are pixel location, R_{xy}, G_{xy}, B_{xy} are RGB channel values and ∇ and θ correspond to gradient magnitude and orientation in each channel, respectively. We voluntarily exploit the dense features employed in [1], in order to understand the exact added value that the two descriptors bring in the joint framework. Once the covariances are extracted, the related MRCs (one for the torso, another for the legs) and the associated discriminants σ described in Sec. 2.2 are also computed. The two MRCs together with their discriminant indexes compose the relation-based description. After computing the descriptors on all the probe and gallery subjects into play, the matching can be performed considering two subjects A and B. To this end, the two distances reported above for the SDALF and the COV descriptors are joined together in a weighted linear fashion, as follows:

$$d_{SDALF+C}(A, B) = \alpha d_{MRC}(A, B) + (1 - \alpha) d_{SDALF}(A, B) \tag{4}$$

where the α coefficient serves to weight the importance of the single description. Estimating the value of α giving the maximum performance will help to understand the interplay of the two components. It is important to note that the two distances are opportunely normalized to sum up to one.

4 Experiments

Experiments have been performed on different multi-shot datasets (i-LIDS for re-id [11], ETHZ[1] 1, 2, and 3 , and CAVIAR4REID[2]), in order to evaluate our proposal against diverse re-id problems, as explained in the following. As metrics, we adopt the standard Cumulative Matching Characteristic (CMC) curve, which represents the probability of finding the correct match in the top n ranks; in practice, after calculating the distance of a probe individual with all the gallery subjects, a ranking is made, and the position of the correct match is kept. On the CMC curve, the rank 1, rank 10 and rank 20 probabilities are usually reported numerically, as so as the normalized Area Under the Curve (nAUC), which is the area under the entire CMC curve normalized over the total area of the graph. As

[1] `http://www.liv.ic.unicamp.br/~wschwartz/datasets.html`.
[2] `http://www.lorisbazzani.info/code-datasets/caviar4reid/`

comparative approaches, we consider SDALF, the MRC part taken alone, and the MRCG approach [1], when the results are available.

In order to assess how the two components of the approach interact, we perform an explorative analysis by mediating the nAUC scores obtained on all the datasets (for the experimental protocol for each benchmark, see below) with different multi-shot cardinalities, i.e., number of images that compose a signature, i.e., $M = 2, 5$.

Fig. 1. Example of partitions obtained with the SDALF approach (best viewed in colors).

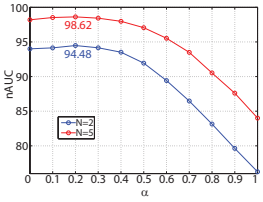

Fig. 2. Analysis of the influence of the α value on the SDALF+C performance: high α means high weight for the MRC part of the descriptor (best viewed in colors).

As visible in Fig. 2, we have for $M = 2$ the best nAUC for $\alpha = 0.2$ and the same happens with $M = 5$: this witnesses that SDALF plays a primary role, but MRC furnishes a complementary information which produces the best performance, independently on the cardinality of the multi-shot signature. Therefore, in all the next experiments, we report the performance of SDALF+C employing this α value as fixed parameter. Using this setting, we overcome in all the datasets the performances of SDALF and MRC. In addition, for each dataset, we report the performance with α_{best}, i.e., the alpha value for which SDALF+C gives its best on that benchmark (that is, the best nAUC): this provides an upper bound of the SDALF+C performances.

In the following, we discuss the results obtained on each dataset.

i-LIDS for re-identification Dataset. The i-LIDS Multiple-Camera Tracking Scenario dataset is a public video dataset captured at a real airport arrival hall in the busy times under a multi-camera CCTV network. In [11], i-LIDS for re-identification dataset has been built from i-LIDS Multiple-Camera Tracking Scenario. The dataset is composed by 479 images of 119 people. The images, normalized to 64×128 pixels, derive from non-overlapping cameras, under quite large illumination changes and subject to occlusions. This dataset a critical multi-shot scenario because the average number of images per person is 4, and thus some individuals have only two images.

The signatures are built from M images of the same pedestrian, randomly selected. Due to the average number of images per pedestrian, we tested SDALF+C with $M = 2$, running 10 independent trials for each case. It is worth noting that some of the pedestrians have less than 4 images: therefore, in such a case, we simply build a multi-shot signature composed by less instances. The results

i-LIDS M=2	**rank1**	**rank5**	**rank10**	**rank20**	**nAUC**
SDALF	45.04	69.13	78.30	86.55	93.02
MRC only	9.78	29.46	40.27	52.35	74.43
MRCG [1]	46.25	67.50	76.00	83.75	**-**
SDALF + C ($\alpha = 0.20$)	**47.40**	**72.55**	**80.43**	**87.66**	93.36
SDALF + C ($\alpha_{\mathbf{best}} = 0.10$)	47.14	72.24	80.13	87.26	**93.41**

Table 1. Performances on i-LIDS for re-identification

show that SDALF+C gives better performances (in terms of nAUC) of all its separate components, overcoming also the MRCG approach: this happens either with the α value kept fixed at 0.2, and with the best value for this dataset, i.e., $\alpha = 0.1$.[1]

ETHZ Dataset. The data are captured from moving cameras in a crowded street. 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 mirror a genuine re-identification scenario (a single camera is employed), it still carries important challenges not exhibited by other public dataset, as the high number of images per person. The protocol is the same as the one employed for i-LIDS, but here we also include $M = 5$ (as more per-person images are available). As visible in Table 2, in all the cases the nAUC performances of SDALF+C, both choosing the best α, or keeping it fixed at $\alpha = 0.2$, are better that the SDALF and the MRC ones. Please note that here, being the nAUC scores near 100%, it is more difficult to get a strong improvement.

CAVIAR for re-identification Dataset. CAVIAR4REID dataset contains images of pedestrians extracted from the CAVIAR repository, and consists of several images captured in a shopping centre in Lisbon. A total of 72 unique pedestrians have been identified: 50 with both the camera views (20 images per pedestrian) and 22 with one camera view (10 images per pedestrian). The challenging features of this dataset are a broad change in the image resolution, with a minimum and maximum size of 17×39 and 72×144, respectively; pose variations are severe, as so as the illumination changes and the occlusions.

In this case, we took only the 50 individuals for which 20 images are available, 10 per camera: images taken from one camera form the probe set, the other camera individuates the gallery. This way, chromatic dissimilarity between probe

[1] We remember here that as best performance for an approach we consider that one which gives the best nAUC, irrespective of the other figure of merits.

(a) ETHZ1

ETHZ1 M=2	**rank1**	**rank5**	**rank10**	**rank20**	**nAUC**
SDALF	**74.12**	**89.20**	92.24	**95.08**	96.68
MRC only	18.48	33.61	44.41	59.13	75.20
SDALF + C ($\alpha = 0.20$)	73.86	88.39	**92.72**	95.04	**96.71**
SDALF + C ($\alpha_{\mathbf{best}} = 0.20$)	73.86	88.39	**92.72**	95.04	**96.71**
ETHZ1 M=5	**rank1**	**rank5**	**rank10**	**rank20**	**nAUC**
SDALF	86.36	94.07	95.81	96.60	97.80
MRC only	23.47	43.47	54.87	70.12	81.53
SDALF + C ($\alpha = 0.20$)	**86.70**	**94.36**	**95.93**	**96.80**	**97.99**
SDALF + C ($\alpha_{\mathbf{best}} = 0.20$)	**86.70**	**94.36**	**95.93**	**96.80**	**97.99**

(b) ETHZ2

ETHZ2 M=2	**rank1**	**rank5**	**rank10**	**rank20**	**nAUC**
SDALF	83.71	95.77	**98.69**	99.43	98.11
MRC only	20.00	51.14	72.57	91.77	80.54
SDALF + C ($\alpha = 0.20$)	84.51	96.17	98.63	**99.83**	98.36
SDALF + C ($\alpha_{\mathbf{best}} = 0.10$)	**84.91**	**96.46**	98.51	99.71	**98.57**
ETHZ2 M=5	**rank1**	**rank5**	**rank10**	**rank20**	**nAUC**
SDALF	90.97	97.71	**99.26**	99.43	98.94
MRC only	29.89	66.00	84.29	96.11	86.93
SDALF + C ($\alpha = 0.20$)	**92.57**	**98.69**	**99.26**	**99.83**	**99.26**
SDALF + C ($\alpha_{\mathbf{best}} = 0.20$)	**92.57**	**98.69**	**99.26**	**99.83**	**99.26**

(c) ETHZ3

ETHZ3 M=2	**rank1**	**rank5**	**rank10**	**rank20**	**nAUC**
SDALF	88.79	97.86	99.64	**100.00**	98.86
MRC only	33.29	74.54	86.21	96.07	86.33
SDALF + C ($\alpha = 0.20$)	**92.79**	**99.50**	99.71	**100.00**	99.40
SDALF + C ($\alpha_{\mathbf{best}} = 0.30$)	92.14	99.43	**100.00**	**100.00**	**99.46**
ETHZ3 M=5	**rank1**	**rank5**	**rank10**	**rank20**	**nAUC**
SDALF	95.14	99.21	**100.00**	**100.00**	99.30
MRC only	42.50	82.43	90.79	98.07	90.28
SDALF + C ($\alpha = 0.20$)	96.43	**100.00**	**100.00**	**100.00**	99.76
SDALF + C ($\alpha_{\mathbf{best}} = 0.30$)	**97.50**	**100.00**	**100.00**	**100.00**	**99.87**

Table 2. Performances on ETHZ1 (a), ETHZ2 (b), ETHZ3 (c)

and gallery images is maximized. All the images are resampled at 64×32 pixels, and ten independent trials have been run. Results are reported in Table 3. In this case, the best performances of SDALF+C are obtained exploiting the standard $\alpha = 0.2$ (so α and α_{best} do coincide), overcoming SDALF and the MRC ones.

5 Conclusions

In this paper, we provide a novel hybrid descriptor for re-id, SDALF+C, which joins together a feature-based and a relation-based description of the human appearance. The former focuses on characterizing visual properties of the human body, the latter captures how visual properties are interrelated. The experimental results show that, in terms of nAUC, SDALF+C overcomes the single parts

CAVIAR4REID M=2	**rank1**	**rank5**	**rank10**	**rank20**	**nAUC**
SDALF	32.16	57.20	70.64	84.12	83.34
MRC only	8.32	24.88	38.60	58.72	64.75
SDALF + C ($\alpha = 0.20$)	**34.96**	**60.80**	**72.68**	**85.24**	**84.55**
SDALF + C ($\alpha_{\mathbf{best}} = 0.20$)	**34.96**	**60.80**	**72.68**	**85.24**	**84.55**
CAVIAR4REID M=5	**rank1**	**rank5**	**rank10**	**rank20**	**nAUC**
SDALF	72.04	89.20	94.28	**98.08**	96.52
MRC only	15.72	42.04	58.60	78.08	77.30
SDALF + C ($\alpha = \alpha_{\mathbf{best}} = 0.20$)	**74.40**	**91.32**	**96.16**	96.68	**97.42**

Table 3. Performances on CAVIAR4REID

(visual-based and relation-based) of which it is composed, in a systematic way. Therefore, our proposal paves the way for further studies, aimed at providing hybrid solutions for the single-shot re-identification case. In addition, we plan to embed SDALF+C in a learning framework, in order to automatically infer the best value for alpha for a given scenario.

References

1. S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Multiple-shot human re-identification by mean riemannian covariance grid. In *AVSS*, 2011.
2. L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *CVIU*, 117(2):130–144, 2013.
3. L. Bazzani, M. Cristani, A. Perina, and V. Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *PRL*, 33(7):898–903, 2012.
4. D. Figueira, L. Bazzani, M.H. Quang, M. Cristani, A. Bernardino, and V. Murino. Semi-supervised multi-feature learning for person re-identification. In *AVSS*, 2013.
5. M. Figueiredo and A.K. Jain. Unsupervised learning of finite mixture models. *TPAMI*, 24(3):381–396, 2002.
6. P.E. Forssen. Maximally stable colour regions for recognition and matching. In *CVPR*, 2007.
7. M. Hirzer, P. M. Roth, M. Kostinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012.
8. X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *IJCV*, 66(1):41–66, 2006.
9. A. Perina, N. Jojic, M. Cristani, and V. Murino. Stel component analysis: Joint segmentation, modeling and recognition of objects classes. *IJCV*, 100(3):241–260, 2012.
10. R. Satta, G. Fumera, F. Roli, M. Cristani, and V. Murino. A multiple component matching framework for person re-identification. In *ICIAP*, pages 140–149, Berlin, Heidelberg, 2011. Springer-Verlag.
11. W. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009.
12. W. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *TPAMI*, (99), 2012.