Exploiting feature representations through similarity learning and ranking aggregation for person re-identification

Julio C. S. Jacques Junior[1,2], Xavier Baró[2,3] and Sergio Escalera[1,2]
[1] Department of Mathematics and Informatics - University of Barcelona, Spain
[2] Computer Vision Center - Universitat Autònoma de Barcelona, Spain
[3] Faculty of Computer Science, Multimedia and Telecommunication - Universitat Oberta de Catalunya, Spain

***Abstract*— Person re-identification has received special attention by the human analysis community in the last few years. To address the challenges in this field, many researchers have proposed different strategies, which basically exploit either cross-view invariant features or cross-view robust metrics. In this work we propose to combine different feature representations through ranking aggregation. Spatial information, which potentially benefits the person matching, is represented using a 2D body model, from which color and texture information are extracted and combined. We also consider contextual information (background and foreground data), automatically extracted via Deep Decompositional Network, and the usage of Convolutional Neural Network (CNN) features. To describe the matching between images we use the polynomial feature map, also taking into account local and global information. Finally, the Stuart ranking aggregation method is employed to combine complementary ranking lists obtained from different feature representations. Experimental results demonstrated that we improve the state-of-the-art on VIPeR and PRID450s datasets, achieving 58.77% and 71.56% on top-1 rank recognition rate, respectively, as well as obtaining competitive results on CUHK01 dataset.**

I. INTRODUCTION

Person re-identification is the task of assigning the same identifier to all instances of a particular individual captured in a series of images or videos, even after the occurrence of significant gaps over time or space. It has a wide range of applications, but most of them are focused on surveillance and forensic systems. Even though the proposed models and reported results in this field have considerably advanced in recent years [1], [2], [3], this task still presents main open challenges, mainly due to the influence of numerous real-world factors such as illumination problems, occlusions, camera settings, as well as many factors associated with the dynamics of the human being, like the great variety of appearance features, pose variations and strong visual similarity between different people. These difficulties are often compounded by low resolution images or poor quality video feeds with large amounts of unrelated information, making re-identification even harder.

As related in [4], given a query person image, in order to find the correct matches among a large set of candidate images captured by different cameras, two crucial problems have to be addressed. First, good image features are required to represent both the query and the gallery images. Second, suitable distance metrics are indispensable to determine whether a gallery image contains the same individual as the query image. An ideal measurement is a matching rule that yields higher matching score for the image pairs belonging to the same person than the pairs belonging to different persons. Chen et al. [5] also highlighted that similarity measurements which are learned (e.g., [6], [7]) from training samples generally enjoy better accuracy performance than learning free methods [8].

In order to address the re-identification problem, existing methods exploit either feature representation [9], [10], [11] or metric learning [12], [7]. In feature representation, robust and discriminative features are constructed such that they can be used to describe the appearance of the same individual across different camera views under various conditions [13], whereas distance metric learning methods attempt to learn a metric in the space defined by image features that keep features coming from same class closer, while, the features from different classes are farther apart [2]. Recently, Convolutional Neural Networks (CNN) have been adopted in person re-identification [14], [9], providing a powerful and adaptive tool to handle computer vision problems without excessive usage of handcrafted image features. However, as mentioned in the work of Wu et al. [9], hand-crafted concatenation of different appearance features sometimes would be more distinctive and reliable, due to significant changes in view angle, lighting, background clutter and occlusion.

In this work we exploit the best of different state-of-the-art models to advance the field of person re-identification. The proposed model is inspired by the work of Chen et al. [5], which enforces similarity learning with spatial constraints, and achieved (up to now) the best score (i.e., top rank recognition rate) on VIPeR [15] dataset (which is one of the most challenging datasets employed in person re-identification). In this paper, by combining new and complementary features within [5], followed by a ranking aggregation strategy [16], we advance the state-of-the-art in person re-identification on two public datasets, VIPeR and PRID450s [17] (by 8.89% and 6.9%, respectively) as well as achieve competitive results on CUHK01 [18] dataset.

The new and complementary adopted features can be briefly enumerated as follows: (i) Salient Color Names based Color Descriptor (SCNCD) [6] (to encode color information) combined with Histogram of Oriented Gradients (HOG) [19] and Scale Invariant Local Ternary Patterns (SILTP) [20] (to encode texture information); (ii) SCNCD combined with

978-1-5090-4023-0/17/$31.00

contextual information (background and foreground data), automatically extracted via Deep Decompositional Network (DDN) [21]; (iii) Convolutional Neural Network (CNN) features constrained by hand-crafted color histograms [9] and combined with Local Maximal Occurrence (LOMO) [22] features. A quantitative analysis regarding the effectiveness of each complementary feature is presented on Sec. IV-D. In particular, experimental results obtained when only the proposed SCNCD based descriptor was employed, demonstrated that the inclusion of context information within SCNCD improves the top-1 rank recognition performance by 11.68%, 10.49% and 12.71%, on VIPeR, PRID450s and CUHK01 dataset, respectively. Compared to the baseline features [5], the usage of Deep features (iii) combined with SCNCD (i) and context information (ii), improved the top-1 rank recognition performance by 6.48%, 13.0% and 19.07%, on the same datasets. The proposed new features demonstrated to complement each other, being very powerful when combined with a ranking aggregation strategy.

The rest of the paper is organized as follows: Section II presents the state-of-the-art concerning person re-identification. The proposed model is described in Section III, and experimental results are provided in Section IV. Finally, conclusions are given in Section V.

II. RELATED WORK

Existing research on person re-identification has concentrated either on the development on sophisticated and robust features to describe the visual appearance of a person under significant visual variabilities or on the development of new learning distance metrics. In this section we present the state-of-the-art on person re-identification, briefly describing the works that achieved the best recognition rates on three broadly employed public datasets, VIPeR, PRID450s and CUHK01, without focusing on the standard taxonomy (i.e., feature representation or metric learning).

As related in the work of Paisitkriangkrai et al. [13], one simple approach to exploit multiple visual features is to build an ensemble of distance functions, in which each distance function is learned using a single feature and the final distance is calculated from a weighted sum of these distance functions. However, the usage of predetermined weights is undesirable as highly discriminative features in one environment might become irrelevant in another one. In their work, a model to learn weights of these distance functions by optimizing the relative distance or by maximizing the average rank-k recognition rate is proposed.

Prates and Schwartz [16] presented a Color-based Ranking Aggregation (CBRA) method, which explores different feature representations to obtain complementary ranking lists, and combine them in order to improve person re-identification. In their work, the KISSME [12] metric learning was adopted and different strategies for ranking aggregation, based on the Stuart rank aggregation method [23], were proposed.

In order to consider spatial information, a common usage in person re-identification is to divide the person image in few regions/stripes and concatenate dense local features, extracted for each region, to implicitly encode the spatial layout of the person. Chen et al. [5] proposed a model for person re-identification that combines spatial constraints and the recently proposed polynomial feature map [7] into a unified framework. They consider that breaking down the variability of global appearance regarding the spatial distribution potentially benefits person matching (i.e., the region containing the head of a person should be compared with the region containing the head rather than the region containing the feet). Authors mention that enforcing the matching within corresponding regions can effectively reduce the risk of mismatching and become more robust to partial occlusions. In addition, their framework can benefit from the complementarity of global and local similarities.

In relation to domain adaptation in machine learning, Chen et al. [10] proposed a schema called Mirror Representation to address the view-specific feature distortion problem in person re-identification. It embeds the view-specific feature transformation and enables alignment of the feature distributions across disjoint views for the same person. Zhang and collaborators [24] argue that most existing approaches focus on learning a fixed distance metric for all instance pairs, while ignoring the individuality of each person. They formulate person re-identification as an imbalanced classification problem and learn a classifier specifically for each pedestrian such that the matching model is highly tuned to the individual appearance. To investigate the intrinsic relationship between the feature space and classifier space, authors proposed the Least Square Semi-Coupled Dictionary Learning (LSSCDL) algorithm in order to learn a dictionary pair and mapping function simultaneously.

Considering the recently proposed CNN based methods for person re-identification, in [9] a deep Feature Fusion Network (FFN) is proposed in order to use hand-crafted features to regularize CNN process so as to make the convolutional neural network extract features complementary to hand-crafted ones. As mentioned by the authors, different to other deep methods for person re-identification (e.g., [14], [25]) which are based on pairwise input, they can directly extract deep features on single images, being able to be learnt by any conventional classifier. Xiao et al. [11] presented a pipeline for learning deep feature representations from multiple domains with CNN. Authors argue that when training a CNN with data from all domains, some neurons learn representations shared across several domains, while some others are effective only for a specific one. Based on this observation they proposed a Domain Guided Dropout algorithm (a method of muting non-related neurons for each domain). Cheng et al. [4] presented a multi-channel parts-based CNN model under the triplet framework to jointly learn both the global full-body and local body-parts features of the input persons.

Although a large number of existing algorithms have exploited state-of-the-art visual features, advanced metric learning algorithms, domain adaptation based models or even CNN based ones, state-of-the-art results on commonly

evaluated person re-identification benchmarks is still far from the accuracy performance needed for most real-world surveillance applications [13].

III. PROPOSED MODEL

In this section, we propose to exploit different feature representations[1], through ranking aggregation, to advance the state-of-the-art in person re-identification. In the proposed model, each image is represented in different ways, which include hand-crafted descriptors (based on color and texture cues) and deep features (extracted via CNN). To describe the matching between a probe image and a gallery set, a similarity learning metric built on the polynomial feature map [7] is adopted, also taking into account spatial (local and global) information. As each image has different descriptors, different similarities are computed, according to each representation. This way, for each probe image and gallery set, different rank lists are generated, each one assigned to each feature representation. The final rank list is obtained through ranking aggregation, which combines such complementary ranking lists. An overview of the proposed model is illustrated in Fig. 1.

Next, we briefly revisit the polynomial feature map and the spatially constrained techniques [5][2], as they are the basis of the proposed model. In a second stage, we describe the proposed complementary features. Finally, we describe the adopted ranking aggregation strategy, which exploits such complementary information.

A. Polynomial Feature Map

In order to measure the similarity between image descriptors $\mathbf{x}_a, \mathbf{x}_b \in \mathbb{R}^{d\times 1}$, we learn the similarity function as:

$$f(\mathbf{x}_a, \mathbf{x}_b) = \langle \phi(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W} \rangle_F, \tag{1}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. To take advantage of both Mahalanobis distance and bilinear similarity metric, we decompose $f(\mathbf{x}_a, \mathbf{x}_b)$ as follows:

$$f(\mathbf{x}_a, \mathbf{x}_b) = \langle \phi_M(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}_M \rangle_F + \langle \phi_B(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}_B \rangle_F. \tag{2}$$

The part $\langle \phi_M(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}_M \rangle_F = (\mathbf{x}_a - \mathbf{x}_b)^\top \mathbf{W}_M (\mathbf{x}_a - \mathbf{x}_b)$ is connected to the Mahalanobis distance. The part $\langle \phi_B(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}_B \rangle_F = \mathbf{x}_a^\top \mathbf{W}_B \mathbf{x}_b + \mathbf{x}_b^\top \mathbf{W}_B \mathbf{x}_a$ corresponds to bilinear similarity. Both parts ensure the effectiveness of $f(\mathbf{x}_a, \mathbf{x}_b)$. The dimensionality of the feature map is reduced by means PCA for $\mathbf{x}_a$ and $\mathbf{x}_b$ before its generation[3].

[1]An evaluation about different color spaces and their combinations for person re-identification can be found in [26].

[2]Implementation provided by the authors, available at `http://dapengchen.com/files/SCSP/SCSP_page.html`

[3]A detailed explanation about how $\mathbf{W}_M$ and $\mathbf{W}_B$ are learned, using the ADMM optimization algorithm, can be found in [5].

B. Spatially Constrained Similarity Function

1) Regional feature map: the input image is partitioned into R non-overlap horizontal stripe regions. Each region is divided into a collection of overlapped patches, from which we extract color and texture histograms. The extracted histograms belonging to a same stripe region are concatenated together. After that, PCA is applied to reduce the dimensionality and to obtain the region descriptor $\mathbf{x}^r$ for the r-th stripe, where $r \in \{1, ..., R\}$. A stripe region r can be described by C visual cues $\{\mathbf{x}^{r,1}, ..., \mathbf{x}^{r,c}, ..., \mathbf{x}^{r,C}\}$, thus $\mathbf{x}_a$ and $\mathbf{x}_b$ accordingly form C polynomial feature maps for the r-th region, i.e., $\{\phi^{r,1}(\mathbf{x}_a, \mathbf{x}_b), ..., \phi^{r,c}(\mathbf{x}_a, \mathbf{x}_b), ..., \phi^{r,C}(\mathbf{x}_a, \mathbf{x}_b)\}$, where $\phi^{r,c}(\mathbf{x}_a, \mathbf{x}_b) = \phi(\mathbf{x}_a^{r,c}, \mathbf{x}_b^{r,c})$.

2) Local similarity integration: in order to exploit the complementary strengths of multiple visual cues within a local region, a linear similarity function is employed to combine them together for the r-th region:

$$s^r(\mathbf{x}_a, \mathbf{x}_b) = \sum_{c=1}^{C} \langle \phi^{r,c}(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}^{r,c} \rangle_F, \tag{3}$$

where $\mathbf{W}^{r,c} = [\mathbf{W}_M^{r,c}, \mathbf{W}_B^{r,c}]$ and $\mathbf{W}_M^{r,c}$, $\mathbf{W}_B^{r,c}$ correspond to $\phi_M^{r,c}(\mathbf{x}_a, \mathbf{x}_b)$ and $\phi_B^{r,c}(\mathbf{x}_a, \mathbf{x}_b)$, respectively. The local similarities scores are integrated as:

$$s^{local}(\mathbf{x}_a, \mathbf{x}_b) = \sum_{r=1}^{R} s^r(\mathbf{x}_a, \mathbf{x}_b). \tag{4}$$

3) Global-local collaboration: in order to describe the matching of large patterns across the stripes, the polynomial feature map is also used for the whole image, yielding global similarity:

$$s^{global}(\mathbf{x}_a, \mathbf{x}_b) = \sum_{c=1}^{C} \langle \phi^{G,c}(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}^{G,c} \rangle_F, \tag{5}$$

where $\mathbf{W}^{G,c} = [\mathbf{W}_M^{G,c}, \mathbf{W}_B^{G,c}]$ and $\mathbf{W}_M^{G,c}$, $\mathbf{W}_B^{G,c}$ correspond to $\phi_M^{G,c}(\mathbf{x}_a, \mathbf{x}_b)$ and $\phi_B^{G,c}(\mathbf{x}_a, \mathbf{x}_b)$, respectively. Here, $\phi^{G,c}(\mathbf{x}_a, \mathbf{x}_b) = \phi(\mathbf{x}_a^{G,c}, \mathbf{x}_b^{G,c})$ and $\mathbf{x}_a^{G,c}, \mathbf{x}_b^{G,c}$ are the c-th type global visual descriptors for image a and b. Finally, the global similarity and local similarity are linearly combined, and the overall similarity score is given by:

$$s(\mathbf{x}_a, \mathbf{x}_b) = s^{local}(\mathbf{x}_a, \mathbf{x}_b) + \gamma s^{global}(\mathbf{x}_a, \mathbf{x}_b), \tag{6}$$

where γ is the hyper-parameter that mediates the local and global similarities (experimentally set to $\gamma = 1.1$).

4) Visual Cues and Parameter settings: in the original model of [5], four visual cues are used (i.e., $C = 4$). First, images are resized to 48×128. Each region r (from R, experimentally set to $R = 4$)[4] is divided into a set of local patches (with 8×16 of size and stride of 4×8). For each patch, six types of features are extracted: HSV_1, LAB_1 (are 8×8×8 joint histograms), HSV_2, LAB_2 (are 48

[4]A default R value was adopted from [5] in order to focus on feature representation, as they performed an in-depth evaluation of the term.

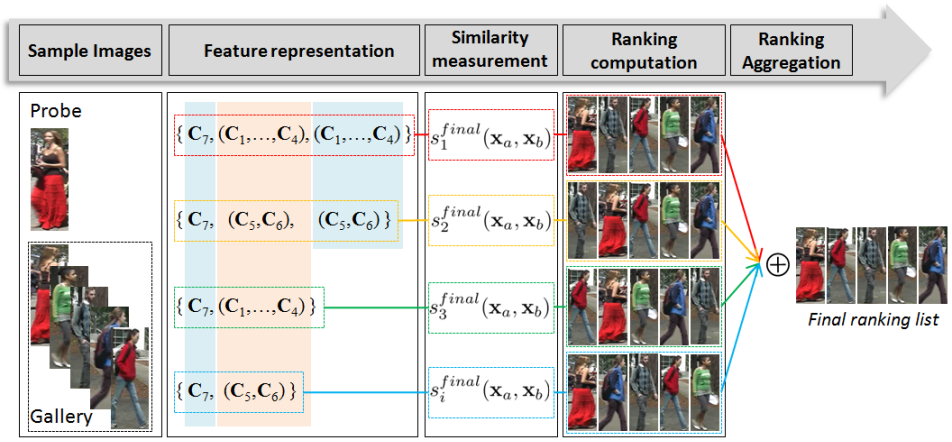

Fig. 1. Overview of the proposed model. For each sample image, different visual cues are defined ($\mathbf{C}_1, ..., \mathbf{C}_7$). Features are represented in different ways, also taking into account global (blue regions) and local (salmon region) information. For each probe image and gallery set, different similarity measures are computed, using different feature representations. Each representation produces a different ranking list, based on the adopted similarity function. The final ranking list is obtained through ranking aggregation, which combines complementary ranking lists obtained from different feature representations.

bin concatenated histograms with each channel having 16 bins), HOG [19] and SILTP [20] (texture descriptors). The four visual cues $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ and $\mathbf{C}_4$ concatenate both color and texture features, which are organized as HSV_1/HOG, HSV_2/SILTP, LAB_1/SILTP and LAB_2/HOG, respectively.

Regarding each visual cue, descriptors generated for each patch, within a specific region r, are concatenated to compose the descriptor of such region. Similarly, the global descriptor is generated through the concatenation of the descriptors computed for all patches. Finally, for each descriptor, PCA is applied to reduce the dimensionality, as well as a whitening process to limit the impact of co-occurrence. The resulting descriptors are then normalized to have unit L_2 norms. As mentioned in [5], the PCA reduced dimension d depends on the size of training data. In our experiments we adopted d to be 120 for all evaluated datasets.

C. Complementary features

In order to improve the state-of-the-art recognition performance in person re-identification, we propose to include new and complementary features within the similarity function presented in [5], as described next.

1) SCNCD [6]: for each color to be named, salient color names indicate that a color only has a certain probability of being assigned to several nearest color names, and that the closer the color name is to the color, the higher probability the color has of being assigned to this color name. Through this way, we can assign multiple similar colors to the same index with the same color descriptor.

Color distributions over color names in different color spaces are then obtained and fused to generate a feature representation. In this work, SCNCD are extracted using the original RGB, normalized *rgb*, $l_1l_2l_3$ and HSV color models (the number of bins for each channel is set to 32). To be specific, SCNCD are extracted similarly as in [6], except that in our model the image is divided in 4 regions ($R = 4$). Such procedure is performed locally, regarding each region r, as well as globally, regarding the whole image.

Two new visual cues are then proposed, $\mathbf{C}_5$ and $\mathbf{C}_6$. Both concatenate color and texture features, which are organized as SCNCD/HOG and SCNCD/SILTP, respectively. In this case, HOG and SILTP are extracted in the same way as in [5]. As before, PCA is applied to reduce the dimensionality of both descriptors, which are then normalized.

2) Context information: due to the fact that the background in person re-identification is not constant and may even include disturbing factors, background feature representation combined directly with the foreground feature representation may reduce classification accuracy. To address this problem, [6] proposed an image-foreground feature representation, which can be seen as that the foreground information is employed as the main information while the background information is treated as a secondary one. Differently from [6], we propose to extract the foreground mask with a more powerful segmentation model, based on Deep Decompositional Network (DDN) [21].

The DDN was developed to tackle the problem of pedestrian parsing, and designed to segment pedestrian images into semantic regions, such as hair, head, body, arms, and legs. It directly maps low-level visual features (HOG) to the label maps of body parts, being able to accurately estimate complex pose variations with good robustness to occlusions and background clutters. In a nutshell, DDN jointly estimates occluded regions and segments body parts by stacking three types of hidden layers: occlusion estimation layers, completion layers, and decomposition layers. The occlusion estimation layers estimate a binary mask, indicating which part of a pedestrian is invisible. The completion layers synthesize low-level features of the invisible part from the original features and the occlusion mask. The decomposition layers directly transform the synthesized visual features to label maps. Fig. 2 illustrates some binary masks automatically obtained using [21][5].

[5]Implementation provided by the authors, available at `http://mmlab.ie.cuhk.edu.hk/projects/luoWTiccv2013DDN/`

Fig. 2. Input images and respective binary masks obtained using [21].

3) Deep feature [9]: the recently proposed Feature Fusion Net (FFN) is used to allow deep feature representation in the adopted framework, as it demonstrated to be very effective in person re-identification tasks. Their FFN consists of two parts. The first part deals with traditional convolution, pooling and activation neurons for input images. It is composed of 5 convolutional layers. Every convolutional layer is followed by a pooling layer and a local response normalization layer, except the 3rd layer. Finally, the output of the 5th pooling layer is a 4096D vector. The second part of the network processes additional hand-crafted feature representations of the same image. Both, CNN features and the hand-crafted features are followed by a fully connected layer (Buffer layer) and then linked together in order to produce a full-fledge image description from the last convolutional layer.

Regarding the hand-crafted features, authors first modified the Ensemble of Local Features (ELF) [27] by improving the color space and stripe division (denoted as ELF16). Input images are equally partitioned into 16 horizontal stripes, and the features are composed of color features including RGB, HSV, LAB, XYZ, YCbCr and NTSC, and texture features including Gabor, Schimid and LBP. A 16D histogram is extracted for each channel and then normalized by L_1 norm. All histograms are concatenated together to form a single vector. The FFN was then trained on the recently proposed Market-1501 [28] dataset, which is the largest public person re-identification dataset, composed of 38195 images from 1501 identities.

The authors of [9] also mention that even though the proposed CNN-based feature performs better when compared to LOMO [22] features, the combination of both kind of features demonstrates to have higher discriminative power. Thus, the concatenation of both (CNN-based feat.+LOMO) is defined in their work as the final representation (31056D vector - denoted in our work, from now, by just *Deep feature*[6]). We also apply PCA ($d = 120$, as previously mentioned) to reduce the dimensionality of the resulting Deep feature, which is then normalized by L_2 norm. This final representation is used as another complementary cue ($\mathbf{C}_7$). Note that $\mathbf{C}_7$ composes a representation for the whole image, so it will be only used as a global descriptor.

4) Integrating the complementary features: to integrate the new and complementary features we compute four similarity measures using different descriptors (for each pair of images being compared), which are then exploited next by the ranking aggregation strategy. To be specific, we compute $s_i^{final}(\mathbf{x}_a, \mathbf{x}_b)$ using Eq. 6, where $i \in \{1, 2, 3, 4\}$, as described next.

- When $i = 1$, we employ $\mathbf{C}_1$ to $\mathbf{C}_4$ (locally and globally, as in [5]) and include $\mathbf{C}_7$ only in the global part of the equation. In this case, s^{local} is computed using four visual cues and s^{global} is computed using five cues.
- When $i = 2$, we employ $\mathbf{C}_5$ and $\mathbf{C}_6$ (locally and globally, as in [5]) and include $\mathbf{C}_7$ only in the global part of the equation. In this case, s^{local} is computed using two visual cues and s^{global} is computed using three visual cues.
- In a more simplified way, when $i = 3$, we just employ $\mathbf{C}_7$ as a global descriptor and use $\mathbf{C}_1$ to $\mathbf{C}_4$ as local descriptors. In this case, s^{local} is computed using four visual cues and s^{global} is computed using only one cue.
- Similarly, when $i = 4$, we just employ $\mathbf{C}_7$ as a global descriptor and use $\mathbf{C}_5$ and $\mathbf{C}_6$ as local descriptors. Here, s^{local} is computed using two visual cues and s^{global} is computed using only one cue.

[6]Available at `http://isee.sysu.edu.cn/resource`

D. Ranking Aggregation Strategy

We propose to explore different feature representations to obtain complementary ranking lists and combine them using the Stuart ranking aggregation method [23]. The Stuart ranking aggregation method, which was originally designed to define a gene-coexpression network over DNA microarrays from humans, flies, worms, and yeast, is a probabilistic method based on order statistics to evaluate the probability of observing a particular configuration of ranks across the different organisms, even when there are irrelevant and noise inputs. The significance of the interactions in the network is verified by means of a variety of statistical tests. An optimized solution of [23] is presented in [29].

Let first denote by $\oplus$ the aggregation operator, for instance if $L_n = L_1 \oplus L_2 \oplus ... \oplus L_{n-1}$, then L_n is a ranking list computed by the aggregation of ranking lists from L_1 to L_{n-1}. As we use different descriptors to represent each image, and have adopted a strategy in which we can measure the similarity $s_i^{final}(\mathbf{x}_a, \mathbf{x}_b)$ of image pairs using different ways ($i \in \{1, 2, 3, 4\}$), we are also able to compute different ranking lists for each probe image and gallery set, as illustrated in Fig. 1.

IV. EXPERIMENTAL RESULTS

In order to demonstrate the effectiveness of the proposed model, this section presents experimental results on three broadly employed public datasets for person re-identification, i.e., VIPeR [15], PRID450s [17] and CUHK01 [18]. Three case studies were performed. First, the proposed model was compared against the state-of-the-art person re-identification models using a well known evaluation protocol. Then, we decomposed the proposed complementary features and performed the following two experiments: (i) the influence of the context information within SCNCD and (ii) the accuracy performance obtained by each complementary feature.

The adopted datasets are presented in two disjoint camera views, with significant misalignment, light changes and body part distortion. Table I summarizes the three datasets. Challenging image samples (due to illumination problems,

TABLE I

SUMMARY OF THE ADOPTED DATASETS.

	VIPeR	PRID450s	CUHK01
Images	1264	900	3884
Individuals (ID)	632	450	971
Images per ID (per view)	1	1	2

pose variation, occlusions or even by high similarity between different people, etc) are illustrated in Fig. 3.

(a) VIPeR (b) PRID450s (c) CUHK01

Fig. 3. Sample images of the adopted datasets. Images on the same column represent the same person.

A. Evaluation Protocol

Our experiments follow the evaluation protocol defined in [13] for a single-shot scenario, i.e. we randomly partitioned each dataset into two parts, 50% for training and 50% for testing, without overlap on person identities (as the CUHK01 dataset contains 971 individuals, 485 of them were randomly sampled for training and the rest for testing, as in [24]). Images from camera A are used as probe and those from camera B as gallery. For the CUHK01 dataset, in which each individual has two images per camera view, we randomly selected one image of the individual taken from the camera A as the probe image and one image of the same individual taken from the camera B as the gallery image. For all evaluated datasets, each probe image is matched with every image in gallery and the rank of correct match is obtained. This procedure is repeated 10 times and the average of Cumulative Matching Characteristic (CMC) curves, which is the most widely used evaluation methodology for person re-identification [2], across 10 partitions is reported.

B. Case 1: State-of-the-art comparison

This experiment compares the overall accuracy performance of the proposed model in relation to the state-of-the-art. Different feature representations were integrated, as described in Sec. III-C.4, followed by the ranking aggregation strategy described in Sec. III-D. Table II summarizes the obtained results. As it can be seen in Table II, the proposed model outperforms the state-of-the-art on both VIPeR and PRID450s datasets, and achieved competitive results on CUHK01 dataset. We can also observe that two competitive approaches (i.e., LSSCDL and FT-JSTL+DGD), which obtained promising results on CUHK01 dataset, were outperformed by the proposed model (27.41% and 34.32%, respectively) on VIPeR dataset (while our method still achieved better accuracy performance than LSSCDL method on PRID450s dataset, with an improvement of 15.47%).

TABLE II

STATE-OF-THE-ART COMPARISON. TOP MATCHING RANK (%) ON THE THREE ADOPTED DATASETS.

Rank	1	5	10	20
VIPeR				
Our	**58.77**	**86.39**	**93.48**	**97.82**
SCSP [5]	53.54	82.59	91.49	96.65
Deep+LOMO [9]	51.06	81.01	91.39	96.90
TCP [4]	47.80	74.70	84.80	91.10
CMC [13]	45.90	77.50	88.90	95.80
Mirror [10]	42.97	75.82	87.28	94.84
LSSCDL [24]	42.66	-	84.27	91.93
FT-JSTL+DGD[11]	38.60	-	-	-
CBRA [16][7]	31.20	60.80	74.30	85.90
PRID450s				
Our	**71.56**	**90.58**	**94.40**	**96.98**
Deep+LOMO [9]	66.62	86.84	92.84	96.89
LSSCDL [24]	60.49	-	88.58	93.60
Mirror [10]	55.42	79.29	87.82	93.87
CBRA [16][7]	26.40	57.10	71.00	83.20
CUHK01				
FT-JSTL+DGD[11]	**66.60**	-	-	-
LSSCDL [24]	65.97	$\approx$ **88.0**	$\approx$ **92.0**	$\approx$ 96.0
Our	59.63	83.66	89.71	94.39
Deep+LOMO [9]	55.51	78.40	83.68	92.59
3TCP [4][8]	53.70	84.30	91.00	**96.30**
CMC [13]	53.40	76.40	84.40	90.50
Mirror [10]	40.40	64.63	75.34	84.08

The work proposed in [11] was designed to learn feature representations from multiple domains, and a very large training set was adopted. As the authors mentioned it would be insufficient to learn such CNN when a quite small dataset is employed. To this end, part of the CUHK03 [14] dataset (which is composed by 13164 images) was also included in their training set. Notice that CUHK03 dataset was captured in the same environment as in the CUHK01 dataset, which could benefit person re-identification when CUHK01 dataset is adopted as both share similar features. Regarding [24], it learns a classifier specifically for each pedestrian such that the matching model is highly tuned to the individual's appearance. This model's characteristic can benefit when large training sets are employed (i.e., CUHK01).

C. Case 2: SCNCD with/without context information

This experiment analyzed the accuracy performance of the context information within SCNCD (described in Sec. III-C.2). To this end, we set up the adopted framework to load only the following visual cues, $\mathbf{C}_5$ and $\mathbf{C}_6$ (detailed in Sec. III-C.1, i.e., without deep features), both without and with the context information. Fig. 4 shows the CMC curves obtained for this experiment (for the first rank values). As it can be observed, the context information significantly improved the overall accuracy on the three evaluated datasets,

[7]Last accuracy performance provided by the authors is described in `www.ssig.dcc.ufmg.br/reid-results`

[8]For the CUHK01 dataset, authors employed a different configuration with additional convolution layers.

being effective to remove the background noise. Yang et al. [6] obtained same conclusion when evaluating both representations (image-foreground and image-only, i.e., with and without context information) on VIPeR and PRID450s datasets. However, differently from their work, in which the evaluation was performed using only RGB information, combined with the segmentation model proposed in [30] and the KISSME [12] metric learning, we adopted a more powerful segmentation strategy, as well as a different similarity function.

It can also be noticed from Fig. 4, that the proposed feature representation based on SCNCD slightly improved obtained results (for the VIPeR dataset) reported in [5] (see Table II), demonstrating its effectiveness.

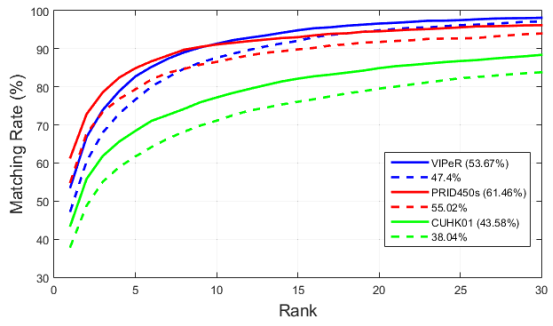

Fig. 4. Accuracy performances based on SCNCD (i.e., using only $\mathbf{C}_5$ and $\mathbf{C}_6$), with and without context information (solid and dashed lines, respectively). Top-1 rank values, for each case, are also provided.

D. Case 3: Complementary feature representations

This experiment evaluated the complementary features individually. Each proposed feature representation was integrated as detailed in Sec. III-C.4, and s_i^{final} is adopted as the similarity function related to each representation i ($i \in \{1, 2, 3, 4\}$). For the sake of simplicity, lets denote the proposed representations as F_1, F_2, F_3 and F_4, and the the baseline [5] feature representation as F_0. Obtained results are shown in Fig. 5 in terms of top-1 rank recognition rate.

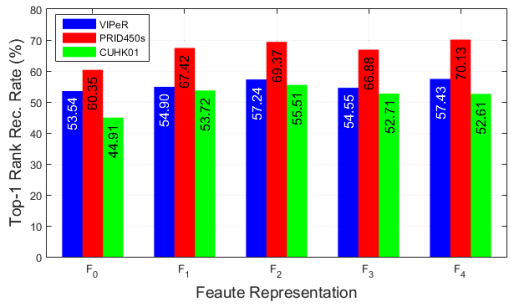

Fig. 5. Accuracy performance obtained for each feature representation F_i (Sec. III-C.4), for the VIPeR, PRID450s and CUHK01 datasets. In this case, F_0 is related to the baseline feature representation [5].

From Fig. 5 we can make the following observations:

- All complementary features outperformed the baseline feature representation (F_0).
- F_1 and F_3 obtained very similar accuracy performances, as well F_2 and F_4 (in this case, in at least two of the three adopted datasets). F_3 and F_4 are simplifications of F_1 and F_2, respectively (i.e., they only use deep features to describe global information). Based on these observations we can conclude the proposed simplifications still have strong discriminative power for person re-identification applications, and require less computation resources when compared to their respective and complete representations.
- The previous observation also emphasizes the benefits of the inclusion of deep features if we only consider F_1 and F_3, for example, and their obtained accuracy performances (compared to F_0). Notice that F_1 and F_3 (despite the inclusion of deep features) are based on the same visual cues as F_0.
- F_2, which exploits SCNCD (with context information) and deep features, obtained the best overall accuracy performance in the three adopted datasets. It should be noticed that, despite feature extraction procedures, F_2 is more compact than F_1 (i.e., it use five cues instead of nine) and employ the same number of visual cues as F_3 (both are composed by five cues).

The previously mentioned observations indicate that the proposed complementary feature representations have strong discriminative power in person re-identification applications, mainly when combined through a ranking aggregation strategy, as shown in Sec. IV-B. In addition, different integration strategies (from those described in Sec. III-C.4) were also evaluated in other experiments (e.g., the integration of all features, $\mathbf{C}_1$ to $\mathbf{C}_7$, using the simplified and complete representations), however, no significant accuracy performance improvements were observed.

E. Computational cost

We adapted the MATLAB implementation provided in [5] to consider the proposed complementary features. The deep features were provided by the authors [9]. However, they reported their FFN approach requires about 1s (one second) per image to extract deep features[9] on VIPeR dataset (also taking into account the concatenation with LOMO features).

Taking the VIPeR dataset as example, our implementation[10] requires (per image) 0.146s to compute the binary mask, 0.131s with SCNCD feature extraction and 0.069s with baseline features extraction. In the feature representation stage, it takes 5.348s to build $\{\mathbf{C}_1, ..., \mathbf{C}_4\}$ for each image, 0.174s to build $\{\mathbf{C}_5, ..., \mathbf{C}_6\}$ and 0.063s to reduce the dimensionality of deep feature and build $\mathbf{C}_7$.

In the learning stage of a single run on VIPeR dataset (see Sec. IV-A for details), it takes 239.738s, 125.857s, 194.289s and 102.863s, when using F_1, F_2, F_3 and F_4, respectively. In the test stage, each probe image requires 0.014s, 0.007s, 0.011s and 0.006s, when using F_1, F_2, F_3 and F_4, respectively. Finally, to compute the ranking aggregation and generate the final rank list of each probe image, it requires 0.1473s.

[9]Using a 2.00GHz Xeon CPU with 16 cores.

[10]Using a 2.30GHz Intel Core i7 CPU and 8Gb of memory, without considering I/O procedures and image resize operations.

V. Conclusion

In this work we exploited different feature representations, combined with a ranking aggregation strategy, to advance the state-of-the-art in person re-identification. Our model was built on a very robust framework, which combines similarity learning metric with spatial constraints. The proposed SCNCD-based feature representation (F_2), which exploits context information, also combined with deep features, demonstrated to have strong discriminative power, even when its simplified version is employed. In particular, when considering the SCNCD-based representation individually, the inclusion of context information on it improved the top-1 rank recognition performance by an average value of 11.63% (± 1.1) in all adopted datasets. Compared to the baseline features [5], the proposed F_2 feature representation improved the top-1 rank recognition by 6.48%, 13.0% and 19.07%, on VIPeR, PRID450s and CUHK01 datasets, respectively.

The ranking aggregation strategy improved the best obtained recognition performance for each new feature representation when treated individually, i.e., F_4 for the VIPeR, F_4 for the PRID450s and F_2 for the CUHK01 datasets, by 2.28%, 0.6% and 6.9%, respectively, demonstrating to be very effective on integrating complementary ranking lists.

The proposed new features demonstrated to complement each other, being very powerful when combined with a ranking aggregation strategy. We show that handcrafted and deep features fusion enhance re-identification performance especially in domains where there is a reduced amount of available data. Quantitative evaluation based on three broadly employed datasets demonstrated the proposed model outperformed the state-of-the-art in at least two of them (VIPeR and PRID450s), as well as obtained competitive results in the third one (CUHK01). In particular, two competitive approaches [24], [11], which obtained promising results on CUHK01 dataset, were outperformed (i.e.,27.41% and 34.32%, respectively) by the proposed model, on the VIPeR dataset, while our method still outperformed [24] by 15.47% on PRID450s dataset.

Acknowledgements

This work has been partially supported by the Spanish projects TIN2015-66951-C2-2-R and TIN2016-74946-P (MINECO/FEDER, UE), by the European Comission Horizon 2020 granted project SEE.4C under call H2020-ICT-2015, and by the CERCA Programme/Generalitat de Catalunya.

References

[1] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Computing Surveys*, vol. 46, no. 2, pp. 29:1–29:37, Dec. 2013.

[2] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270 – 286, 2014.

[3] S. Gong, M. Cristani, S. Yan, and C. L. (Eds.), *Person Re-Identification*. Advances in Computer Vision and Pattern Recognition. Springer, 2014.

[4] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *CVPR*, 2016, pp. 1335–1344.

[5] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *CVPR*, 2016, pp. 1268–1277.

[6] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *ECCV*, 2014, pp. 536–551.

[7] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *CVPR*, 2015, pp. 1565–1573.

[8] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Computer Vision and Image Understanding*, vol. 117, no. 2, pp. 130–144, 2013.

[9] S. Wu, Y. C. Chen, X. Li, A. C. Wu, J. J. You, and W. S. Zheng, "An enhanced deep feature representation for person re-identification," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2016.

[10] Y.-C. Chen, W.-S. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in *International Joint Conf. on Artificial Intelligence*, 2015, pp. 3402–3408.

[11] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, 2016, pp. 1249–1258.

[12] M. Kstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012, pp. 2288–2295.

[13] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *CVPR*, 2015, pp. 1846–1855.

[14] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014, pp. 152–159.

[15] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *IEEE Int. Workshop on Performance Evaluation for Tracking and Surveillance*, 2007.

[16] R. F. de Carvalho Prates and W. R. Schwartz, "CBRA: Color-based ranking aggregation for person re-identification," in *IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 1975–1979.

[17] P. M. Roth, M. Hirzer, M. Koestinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*, S. Gong, M. Cristani, S. Yan, and C. C. Loy, Eds. London, United Kingdom: Springer, 2014, pp. 247–267.

[18] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *ACCV*, 2012, pp. 31–44.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, pp. 886–893.

[20] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikinen, and S. Z. Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *CVPR*, 2010, pp. 1301–1306.

[21] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in *ICCV*, 2013, pp. 2648–2655.

[22] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015, pp. 2197–2206.

[23] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science*, vol. 302, no. 5643, pp. 249 – 255, 2003.

[24] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific svm learning for person re-identification," in *CVPR*, 2016, pp. 1278–1287.

[25] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015, pp. 3908–3916.

[26] Y. Du, H. Ai, and S. Lao, "Evaluation of color spaces for person re-identification," in *ICPR*, 2012, pp. 1371–1374.

[27] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, 2008, pp. 262–275.

[28] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

[29] R. Kolde, S. Laur, P. Adler, and J. Vilo, "Robust rank aggregation for gene list integration and meta-analysis," *Bioinformatics*, vol. 28, no. 4, p. 573, 2012.

[30] N. Jojic, A. Perina, M. Cristani, V. Murino, and B. Frey, "Stel component analysis: Modeling spatial correlations in image class structure," in *CVPR*, 2009, pp. 2044–2051.