Efficient Person Re-identification by Hybrid Spatiogram and Covariance Descriptor

Mingyong Zeng, Zemin Wu, Chang Tian, Lei Zhang, and Lei Hu
College of Communications Engineering, PLA University
of Science and Technology, Nanjing 210007, China
zengmingyong1987@gmail.com

Abstract

Feature and metric researchings are two vital aspects in person re-identification. Metric learning seems to have gained extra advantages over feature in recent evaluations. In this paper, we explore the neglected potentials of feature designing for re-identification. We propose a novel and efficient person descriptor, which is motivated by traditional spatiogram and covariance descriptors. The spatiogram feature accumulates multiple spatial histograms of different image regions from several color channels and then extracts three descriptive sub-features. The covariance feature exploits several colorspaces and intensity gradients as pixel features and then extracts multiple statistical feature vectors from a pyramid of covariance matrices. Moreover, we also propose an effective and efficient multi-shot re-id metric without learning, which fuses the residual and coding coefficients after collaboratively coding samples on all person classes. The proposed descriptor and metric are evaluated with current methods on benchmark datasets. Our methods not only achieve state-of-the-art results but also are straightforward and computationally efficient, facilitating real-time surveillance applications such as pedestrian tracking and robotic perception in various dynamic scenes.

1. Introduction

Person re-identification (re-id) aims to identify an individual in different time, locations or different views, considering a large set of candidate targets recognized before. It has attracted much attention recently for its vital importance, especially in video surveillance for tasks such as person retrieval, long-term pedestrian tracking [8], and robotic perception. For example, an interesting android APP called *Insight* has been developed on the famous *Google Glass* to recognize and re-identify certain persons in market or airport [22]. Person re-id is a non-trivial problem due to many uncontrollable variations in illumination, viewpoint or occlusion, which usually make intra-personal variations even larger than inter-personal variations.

Existing re-id solutions focus either on designing proper features [6, 7, 11, 15, 16] or on learning distance metrics from training samples [17, 19, 20, 36]. On the whole, leaning based methods (*i.e.* metric learning or feature learning, supervised or unsupervised learning) have gained increasing advantages over pure feature based methods in recent re-id evaluations [19, 20, 32]. However, their results may be more biased by parameter selection, thus making them less generic or flexible to different scenarios. In real-time online applications, the extra big-data labeling, the expensive training complexity of time or storage, and the unavoidable over-fitting or retraining are often not favored. On the contrary, the feature based methods suffer no such troubling issues. Though more practical and generic, feature based methods seem to be underestimated and attract much less attention in very recent years. The reason behind is that their results haven't got significantly improved since the earlier famous SDALF feature in 2010 [7] and the results can't outperform or even compete with recent learning based methods.

Learning models undoubtedly deserve more research for their greater potentials but the feature designing should not be ignored in the meantime. It's not just an important re-id approach, but also an indispensable step or precondition for later distance computing. With more excellent features, metric learning methods may be expected to achieve better than just adopting ordinary features. However, many feature based methods choose to combine with SDALF for better results, but SDALF's feature extracting and matching actually cost massive computing time. The adopted heavy preprocessing measures (*e.g.* foreground extraction) make SDALF and related methods inefficient and unpractical.

In this paper[1], we attempt to explore the neglected potentials for feature designing strategy and try to

[1]Code is released at https://github.com/Myles-ZMY/HSCD

win one round back ambitiously even against learning methods. Specifically, we present a novel efficient person re-id descriptor named HSCD , which is short for *Hybrid Spatiogram and Covariance Descriptor*. As the name suggests, HSCD is fused with a spatiogram feature and a covariance feature. Motivated by the excellent representation ability of spatiogram in tracking [5], a spatial histogram based feature named MCSH (*Multi-Channel Spatio-Histogram*) is constructed. The spatiograms are firstly calculated and accumulated on several different image layers/regions within multiple color channels. Then the histogram and two distinctive spatial statistical vectors are creatively decomposed from these spatiograms and are further combined. Then motivated by the Sigma Points [10] from covariance matrix, a covariance based feature named MSPC (*Multi-Statistics on Pyramid of Covariance*) is proposed. Multiple color and gradient components are used as pixel features to compute the pyramid of region covariance matrices. Then four vectorized statistic features are extracted from each covariance for people description, which depict all pixel features' mean, deviation, correlation and Sigma vector.

Besides, to deal with the multi-shot case where multiple images for each person are available, we further investigate the collaborative representation scheme which utilizes coding samples of all person classes. Still, without using learning strategy, both the coding residual and coefficients are proposed to combine for an efficient multi-shot metric.

The contributions of our paper are two-folds. Firstly, the proposed HSCD descriptor is straightforward and efficient. It possesses good intra-class invariance and does not require complicated preprocessing, outperforming all other feature based methods and most of learning based methods on the VIPeR [7] and iLIDS [6] datasets. Secondly, the multi-shot method with HSCD descriptor and multi-shot metric also achieves significant result on the Caviar4reid dataset [6] with rather small computational cost. Above all, the performance of our methods can get significantly improved without learning models, which demonstrates the great potentials of feature designing and other strategies.

The rest of the paper is organized as follows. Sec. 2 gives a taxonomy of the present re-id literature. Sec. 3 details our HSCD descriptor. Then multi-shot metric is introduced in Sec. 4. Experiments on three public datasets are reported in Sec. 5. Finally, conclusions are drawn in Sec. 6.

2. Related Work

Feature extracting and distance computing are two successive steps of person re-identification. According to different emphasis, existing re-id work can be roughly separated into feature based and distance based approaches. Feature based methods focus on designing or selecting excellent features and may adopt some regular distances. Distance based methods focus on learning discriminative metrics from training samples with some ordinary features.

In feature based methods, several dividing criteria exist, such as 1) global or patch-based features; 2) low-level, mid-level or semantic features; and 3) feature designing (or handcrafting), feature selection, or feature learning. The commonly exploited features include color, texture and keypoints. Different proper features can be fused and preprocessing steps may also be employed. Farenzena *et al.* [7] proposed the famous SDALF method which combined weighted HSV histogram with stable color region and salient texture. The later CPS [6] and SCEAF [11] methods both tried to improve SDALF with detected human parts or structural constraints. Ma *et al.* proposed novel LDFV [15] and eBiCov [16] descriptors based on fisher vectors and bio-inspired features. Bak *et al.* [2] extracted spatial covariance matrices of body parts for features. Xu *et al.* [28] represented the body as an articulated assembly of compositional parts and matched the assembly by cluster sampling. Gray *et al.* [9] used AdaBoost to select a subset of features for matching persons. Bak *et al.* [1] selected the most descriptive features by correlations in a covariance metric space. Zhao *et al.* [32] learned an unsupervised saliency model for patch matching and then combined with SDALF to form the better eSDC method. Recently Zheng *et al.* [35] proposed to evaluate feature effectiveness in a query-adaptive manner and observe consistent improvement by fusing multiple effective features.

In distance based methods, metric learning is formulated to learn the optimal similarity between a pair of person images. Zheng *et al.* [36] introduced a Relative Distance Comparison (RDC) model to maximize the probability of a pair of true matches having a smaller distance than a wrong match pair. Mignon and Jurie [17] proposed a new PCCA approach from sparse pairwise constraints. Kostinger *et al.* [12] introduced a KISSME metric from equivalence constraints based on statistical inference. Tao *et al.* [20] further presented regularized smoothing KISS (RS-KISS) by integrating smoothing and regularization techniques. Pedagadi *et al.* [19] applied PCA to reduce dimension and used Local Fisher discriminative analysis (LF) to match the visual features in a low dimension. The high dimensional features can thus be represented in an efficient way and achieve high performance. Some works consider both feature and metric learning. Zhao *et al.* [31] estimated patch salience and matched it with a unified RankSVM model. Li *et al.* [13] proposed a deep filter pairing neural network to jointly learn feature, metric and photometric transformation directly from data.

Besides, multi-shot re-id is attracting researchers' interest because multiple images for each person are often available rather than just one image in the single-shot case. How

to exploit the sample set of each person for better result appeals a new re-id approach. Bazzani *et al.* [4] introduced Asymmetry Histogram Plus Epitome (AHPE) feature to incorporate global and local appearances of multiple images. Bak *et al.* [3] combined information from multiple images to obtain a signature named Mean Riemannian Covariance Grid (MRCG). The recent LCSA and CRNP methods proposed by Wu *et al.* [25, 26, 27] model the sample set of each person as a affine hull. The geometric distance between two hulls was computed by sparse or collaborative representation on all gallery samples.

3. The Proposed HSCD Descriptor

The proposed feature representation is constructed based on the simple idea that only useful feature components and descriptors should be exploited and properly fused for enhancement. Several color spaces, gradients, and patch extraction based on human structure are considered to be useful. Since spatiogram and covariance descriptors have been acknowledged and widely used for their excellent representation abilities, they are adopted and adapted here according to the challenges in the re-id domain.

3.1. Multi-Channel Spatio-Histogram

Accumulation of Spatiogram. Spatiogram is an excellent image descriptor which has been well applied in object tracking [5]. It is a generalization of the histogram but captures richer target appearance information. It includes high order spatial moments, which can be expressed as below for an image region R with B quantization bins

$$\begin{aligned} &\mathbf{S}_R(b) = < \hat{n}_b, \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b >, b = 1, 2, \cdots, B \\ &n_b = \sum\nolimits_{k=1}^{N} \delta_{kb} \ , \ \hat{n}_b = n_b \Big/ \sum\nolimits_{i=1}^{B} n_i \\ &\boldsymbol{\mu}_b = \left(\sum\nolimits_{k=1}^{N} \mathbf{c}_k \delta_{kb} \right) \Big/ n_b \\ &\boldsymbol{\Sigma}_b = \left[\sum\nolimits_{k=1}^{N} (\mathbf{c}_k - \boldsymbol{\mu}_b)(\mathbf{c}_k - \boldsymbol{\mu}_b)^{\mathrm{T}} \delta_{kb} \right] \Big/ n_b \end{aligned} \tag{1}$$

where N is the total pixel number of region R, $\mathbf{c}_k = (x_k, y_k)^T$ is the k-th pixel's coordinate in region R. $\delta_{kb} = 1$ if pixel k is within the b-th bin, or else $\delta_{kb} = 0$. n_b is the count of pixels whose value belong to b-th bin in the quantified color (or feature) space (*i.e.* b-th bin histogram). For each bin b, $\mathbf{S}_R(b)$ includes a normalised histogram $\hat{n}_b$, a spatial mean vector $\boldsymbol{\mu}_b$ and a spatial covariance matrix $\boldsymbol{\Sigma}_b$.

The spatiogram similarity metric ρ is usually based on the Bhattacharyya distance of histogram and weighted by the pixel's spatial position information as in [5]

$$\rho(\mathbf{S}, \mathbf{S}') = \sum\nolimits_{b=1}^{B} \varphi_b \sqrt{\hat{n}_b \hat{n}'_b} \tag{2}$$

Figure 1. The proposed layers and regions. From left to right: original image, the chosen image centre and its four layers.

$$\varphi_b = \eta \exp\{-\tfrac{1}{2}(\boldsymbol{\mu}_b - \boldsymbol{\mu'}_b)^T \hat{\boldsymbol{\Sigma}}_b^{-1} (\boldsymbol{\mu}_b - \boldsymbol{\mu'}_b)\} \tag{3}$$

where the weight φ_b is the b-th bin's spatial similarity between two spatiograms $\mathbf{S}$ and $\mathbf{S}'$, (2) will become the Bhattacharyya distance if all $\varphi_b = 1$, η is a Gaussian constant, and $\hat{\boldsymbol{\Sigma}}_b^{-1} = \boldsymbol{\Sigma}_b^{-1} + (\boldsymbol{\Sigma'}_b)^{-1}$.

The spatiogram is often calculated on the whole image region within a single colorspace, which lacks robustness. Since re-id often assumes that people wear the same clothes, several colorspaces are jointly considered and the adopted spatiogram is enriched with more regions and color channels. The exploited channels are selected as Y, C_b, C_r, H, S, nR, nG, nB, which are from the YCbCr, HSV, and normalized RGB colorspaces (*i.e.* nR=R/(R+G+B), *etc.*). The employed statistical regions of four layers are illustrated in Figure 1, where the centre area is chosen as the top layer to reduce some background affection and the four layers have 1, 2, 4, and 8 horizontal regions. The statistical bins on each region from the top to the bottom layer are set as 32, 24, 16 and 8 for a hierarchical quantization. Though above parameters may not be optimal but are proved quite good approximate solutions in our supporting tests. By accumulating spatiograms on all these regions among 8 channels, the proposed spatiogram feature with total 1664 bins is formed (*i.e.* 1664 = 8×(32×1+24×2+16×4+8×8)).

Decomposition of Spatiogram. Through preliminary experiments, the spatiogram feature proposed above wasn't so efficient. Firstly, the traditional metric in (2) uses spatial Gaussian model to enhance the histogram metric, which involves complicated computation. Besides, the spatial information of the horizontal x axis and the vertical y axis are actually not equally useful, which can be illustrated in Figure 2. From two images of a same person, two patches of the same coordinates are extracted, the brightest pixels (*i.e.* the white stripe) are plotted and modeled as two ellipses.

As revealed in Figure 2, the spatial information of y axis exhibits much better intra-class invariance than x axis due to viewpoint or pose variations. In fact, we extract only horizontal regions on each layer in Figure 1 with the same consideration. And our preliminary tests supported that using vertical regions usually achieved decreased results. Therefore, we ignore the trivial x axis and decompose the spatiogram into three vectorized parts, which include the histogram, the first and the second order spatial information (*i.e.* the mean and the standard deviation vector) of y axis.

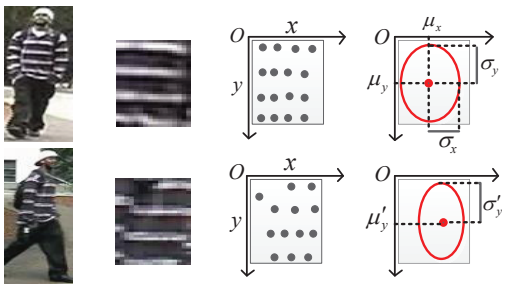


Figure 2. Spatial distributions of two patches' brightest pixels.

The traditional spatiogram in (1) can thus be simplified as

$$\mathbf{S}(b) =< \hat{n}_b, \mu_{by}, \sigma_{by} >, b = 1, 2, \cdots, B$$
$$\mathbf{S} = \{\mathbf{h} = \{\hat{n}_b\}; \boldsymbol{\mu}_y = \{\mu_{by}\}; \boldsymbol{\sigma}_y = \{\sigma_{by}\}, b = 1 \sim B\}$$
$$\mu_{by} = \frac{1}{n_b}\sum_{k=1}^{N} y_k \delta_{kb}, \sigma_{by} = \sqrt{\frac{1}{n_b}\sum_{k=1}^{N}(y_k - \mu_{by})^2 \delta_{kb}} \tag{4}$$

Above three vectors $\boldsymbol{h}$, $\boldsymbol{\mu}_y$, and $\boldsymbol{\sigma}_y$ in (4) stand for three different kinds of features which are rather complementary. They constitute a simplified but more distinctive spatiogram descriptor, which is denoted as *Multi-Channel Spatial Histogram* (MCSH) in this paper.

3.2. Multi-Statistics on Pyramid of Covariance

Pyramid of Region Covariance. Region covariance matrix is a matrix of covariance of several image statistics calculated in an image region. By extracting d dimensional feature vector z_i from each pixel at (x, y) with a mapping function $\mathbf{U}$, region R can be represented by a covariance matrix $\mathbf{C_R}$, where $\boldsymbol{\mu} = \sum_{i=1}^{n} \mathbf{z}_i / n$ and n is the total pixel number.

$$\mathbf{C_R} = \frac{1}{n-1}\sum_{i=0}^{n}(\mathbf{z}_i - \boldsymbol{\mu})(\mathbf{z}_i - \boldsymbol{\mu})^{\mathrm{T}} \tag{5}$$

The performance of covariance depends on the employed pixel features in the mapping function, the calculated regions and the covariance distance. Similarly as MCSH, several colorspaces (YCbCr, HSV, Lab, nRGB) are jointly used to construct $\mathbf{U}$ together with spatial and gradient features.

$$\mathbf{U}(x, y) = [y, Y, C_b, C_r, H, S, a, b, nR, nG, Y_x, Y_y, Y_{xx}, Y_{yy}]^{\mathrm{T}} \tag{6}$$

where y axis is used for pixel spatial position regardless of the useless x axis (as revealed in Figure 2). Y stands for the intensity channel, so other intensity channels V and L are not used again for a compact mapping function. Similarly, one redundant channel nB is omitted in the covariance as nRGB is a normalised colorspace. The last four channels represent the first and the second order intensity gradients.

Since a global image region is not discriminative enough to describe a person image, a four-layer pyramid of region covariance matrices (only the y axis) within the image centre area is proposed. The pyramid is almost the same as in Figure 1 except in the third layer where 4×2 patches are extracted, forming total 19 image regions. We adopt similar regions because both features can share part of computing and avoid calculating repeated color channels or regions.

Combining Covariance Statistics. Unlike vector feature, covariance matrix does not lie in the Euclidean space. The distance proposed in [21] is often used to measure the dissimilarity of two covariance matrices.

$$\rho(\mathbf{C}_1, \mathbf{C}_2) = \sqrt{\sum_{i=1}^{d} \ln^2 \lambda_i(\mathbf{C}_1, \mathbf{C}_2)} \tag{7}$$

where $\lambda_i(\mathbf{C}_1, \mathbf{C}_2)$ are the generalized eigenvalues of two matrices. Bak et al. [2] proposed to sum up all the covariance distances between respective regions in two images and then discarded 20% of the larger distances. However, this measure for covariance pyramid is not discriminative enough because it is almost equivalent to the average of several covariance distances. It's also too expensive as it involves complicated eigenvalue calculating.

Motivated by the Sigma Set or Sigma Points [10] that are converted from covariance matrix and lie in Euclidean space, we attempt to extract useful parts from covariance and avoid the direct use of above strategy. Four vectorized statistical features are proposed to form a simpler but more descriptive representation, which are detailed below.

The first statistical feature for region descriptor is called Sigma vector (*i.e.* vectorized Sigma points). It's a $d(d+1)/2$ dimensional vector constructed by vectorizing the lower triangular matrix $\boldsymbol{L}$ after the Cholesky decomposition of covariance matrix as in [10], $\mathbf{C} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$. And $d = 14$ since we adopt the mapping function in (6). The second one is the d dimensional mean vector $\boldsymbol{\mu}$ in (5), which captures the first-order information. The third one is the d dimensional standard deviation vector. It can be extracted from the diagonal entries of the covariance, reflecting the varying extents of adopted pixel features. The last one is the $d(d-1)/2$ dimensional vector constructed by vectorizing the lower triangular matrix (excluding the diagonal entries) of the normalized covariance matrix, *i.e.* the correlation matrix. It can represent the correlation coefficients between employed pixel features in (6).

By cascading on all the covariances of the pyramid regions, four vectorized sub-features are thus formed to jointly describe a person image. These four sub-features are denoted as $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3$, $\mathbf{f}_4$ and are gathered to form the MSPC descriptor $\mathbf{F} = \{\mathbf{f}_i, i = 1, 2, 3, 4\}$ (or *Multi-Statistics on Pyramid of Covariance*) with 4256 dimensions (*i.e.* $4256 = [d(d+1)/2 + d + d + d(d-1)/2] \times 19,\ d = 14$).

3.3. The Fused HSCD Descriptor

Although MCSH and MSPC both exploit color and structure information, they differ largely in the description

scheme. In MCSH, multiple color channels are used for cascading and in every channel three spatiogram components are extracted. While in MSPC, multiple regions are used for cascading and in each region four statistical vectors are extracted from region covariance. Notice that the pixel color channels are fused with covariance in a natural manner in MSPC, which is different from the cascading way in MCSH. The description difference provides potentials for complement and improvement by feature fusing, which finally forms the proposed HSCD descriptor.

A person descriptor can't become a re-id method without proper metrics. The two feature components in HSCD are fused by weighted combination as below

$$d(\mathbf{I}, \mathbf{I}') = w_1 d(\mathbf{S}, \mathbf{S}') + w_2 d(\mathbf{F}, \mathbf{F}') \tag{8}$$

where $\mathbf{I} = \{\mathbf{S}, \mathbf{F}\}$ and $\mathbf{I}' = \{\mathbf{S}', \mathbf{F}'\}$ represent HSCD descriptors of two images. Further, $\mathbf{S} = \{\mathbf{h}, \boldsymbol{\mu}_y, \boldsymbol{\sigma}_y\}$ and $\mathbf{S}' = \{\mathbf{h}', \boldsymbol{\mu'}_y, \boldsymbol{\sigma'}_y\}$ are respective MCSH features, $\mathbf{F} = \{\mathbf{f}_i, i = 1, 2, 3, 4\}$ and $\mathbf{F}' = \{\mathbf{f}_i', i = 1, 2, 3, 4\}$ are respective MSPC features, their metrics can be written as

$$d(\mathbf{S}, \mathbf{S}') = w_{s1} d(\mathbf{h}, \mathbf{h}') + w_{s2} d(\boldsymbol{\mu}_y, \boldsymbol{\mu'}_y) + w_{s3} d(\boldsymbol{\sigma}_y, \boldsymbol{\sigma'}_y) \tag{9}$$

$$d(\mathbf{F}, \mathbf{F}') = \sum\nolimits_{i=1}^{4} w_i d(\mathbf{f}_i, \mathbf{f'}_i) \tag{10}$$

The related weights can be set according to the components' relative importance with part of samples in experiments as in SDALF method [7]. The estimated weights are in fact rather stable in most datasets or scenarios for their relative importance tends to be fixed or similar.

As a feature based method, we can use some simple distances as basic metrics to testify the superiority of our descriptor. The l_1 vector distance is adopted here for its simpleness. As for the single-shot re-id case, distance between two images equals to the distance between two persons. The single-shot method with HSCD descriptor and l_1 based vector distances in (9) and (10) is also called HSCD method below. Other better metrics may be available but the feature designing is our main concern. In fact, the simple l_1 metric is proven efficient in later experiments.

4. The Multi-shot Metric

4.1. Coding based Metric Model

In the multi-shot case, both the probe and the gallery sets may have multiple images for each person, so the distance between two image sets is needed. Earlier works [6, 7] often adopt MPD strategy (Minimum Point-wise Distance) which simply calculates the minimum of all the sample distances between two persons. But MPD is sensitive to outliers as it only depends on a single sample from each set. The APD distance (Average Point-wise Distance) [25] performs more robust by calculating the average of all pairwise distances, but it still cannot get satisfactory result.

MPD and APD exploit only samples from two sets to get their direct distance, while lots of samples from other sets are not involved. Recently, sparse coding has been proven promising in face recognition, especially the SRC [23] or CRC [29] models. These models code a signal by samples from all classes and assume it can be better re-expressed by a linear combination of samples from its own class.

In this section, we will use the above coding strategy to design a proper multi-shot metric in the re-id domain. The probe set of a person is denoted as $\mathbf{P}$ and the gallery set of all n persons (or person classes) is denoted as $\mathbf{G} = \{\mathbf{G}_1, \cdots, \mathbf{G}_i, \cdots, \mathbf{G}_n\}$, where $\mathbf{G}_i$ is the gallery set for the i-th person. The distance metric between $\mathbf{P}$ and $\mathbf{G}_i$ depends on the coding of all samples in $\mathbf{P}$ with SRC or CRC, which is detailed here. For a probe image in $\mathbf{P}$ which is represented by a vector feature $\mathbf{z}$, it can be firstly coded on the whole gallery dictionary $\mathbf{G}$ as

$$\min_{\boldsymbol{\rho}} \|\mathbf{z} - \boldsymbol{G\rho}\|_2 + \lambda \|\boldsymbol{\rho}\|_k \tag{11}$$

where $\boldsymbol{\rho}$ is the coding vector to be optimized, $\|\boldsymbol{z} - \mathbf{G}\boldsymbol{\rho}\|_2$ and $\|\boldsymbol{\rho}\|_k$ stand for coding residual and k norm of $\boldsymbol{\rho}$, λ is a trade-off parameter between these two parts.

SRC adopts l_1 norm (*i.e.* $k = 1$) to constrain the sparse coding vector whereas CRC adopts l_2 normalization ($k = 2$). It's because CRC verifies that the collaborative strategy actually plays the key role rather than the sparseness brought by the l_1 norm [24, 29]. With CRC, a closed-form coding solution can be efficiently computed via (12), where $\mathbf{I}$ is the identity matrix and $\mathbf{Q}$ only needs to be computed once for all probe images.

$$\boldsymbol{\rho} = \mathbf{Q}\boldsymbol{z}, \mathbf{Q} = (\mathbf{G}^T\mathbf{G} + \lambda\mathbf{I})^{-1}\mathbf{G} \tag{12}$$

Different from SRC which determines a probe sample's class label only by its coding residual, CRC also considers the respective coding coefficients $\boldsymbol{\rho}_i$ in $\boldsymbol{\rho}$ on the i-th class $\mathbf{G}_i$. CRC fuses these two parts with the quotient form below

$$d(\boldsymbol{z}, \mathbf{G}_i) = \|\boldsymbol{z} - \mathbf{G}_i\boldsymbol{\rho}_i\|_2 / \|\boldsymbol{\rho}_i\|_2 \tag{13}$$

4.2. The Proposed Multi-shot Method

Motivated by the success of SRC and CRC, we propose to investigate a proper metric for multi-shot re-id task. We adopt the CRC coding model as it is effective and efficient. However, the metric (13) in CRC is mainly based on such assumption that if $\boldsymbol{z}$ belongs to class $\mathbf{G}_i$, the related residual and coding coefficients should be more discriminant at the same time, *i.e.* the residual is often smaller and the coefficients are often denser (or bigger) than those of other unrelated classes. It does make sense in face recognition as the intra-class variation of facial images is usually not large. In

[23], either part in (13) can achieve above 90% true matches at the first matching rank, assuring the enhancement of metric (13) with a rather high probability.

However, in the re-id domain, the intra-class variation is so significant that even humans may make wrong judgements. The first rank matching rate is often less than 30% on the Caviar4reid dataset [6]. The premise in CRC does not hold any more as the two metric parts are always not discriminative enough at the same time, which results in the unpredictability of metric (13). We even find in our tests that each single part (residual or coding coefficients) may perform better than (13), which implies the quotient form in (13) degrades the re-id result. Considering that both parts exhibit discriminative ability and may be well complemented, we thus propose to fuse them in the subtraction form in (14). In fact, it resembles an equivalent form of weighted combination of these two recognition metrics.

$$d(\boldsymbol{z}, \mathbf{G}_i) = \|\boldsymbol{z} - \mathbf{G}_i\boldsymbol{\rho}_i\|_2 - \delta\|\boldsymbol{\rho}_i\|_2 \tag{14}$$

where δ is a trade-off parameter which can be estimated by experiment on specific datasets or occasions. The distance between two persons can thus be averaged as below

$$d(\mathbf{P}, \mathbf{G}_i) = \frac{1}{m}\sum_{j=1}^{m} d(\mathbf{P}_j, \mathbf{G}_i) \tag{15}$$

where m is the number of probe images in $\mathbf{P}$ and $d(\mathbf{P}_j, \mathbf{G}_i)$ represents the distance between a probe image and a gallery person $\mathbf{G}_i$ as in (14). The proposed metric is briefly denoted CRC-S (CRC in Subtraction form) in this paper.

The proposed CRC-S metric can be applied on any vector feature. Therefore, we apply it on the proposed HSCD descriptor and derive the proposed multi-shot metric below

$$D(\mathbf{P}, \mathbf{G}_i) = \sum\nolimits_{f=1}^{7} w_f d_f(\mathbf{P}, \mathbf{G}_i) \tag{16}$$

where the vector components are represented by subscript f (7 sub-features), and the weights are determined in the same way as in Sec 3.3. Our final multi-shot re-id method is denoted as HSCD-CR (HSCD with CR metric), which can achieve significant results benefiting from both its excellent feature and multi-shot metric.

5. Experimental results

In this section, extensive experiments are conducted on three public datasets to evaluate the proposed methods.

5.1. VIPeR Dataset Evaluation

The HSCD descriptor is evaluated on the single-shot benchmark dataset VIPeR [8], which contains exactly two images of 632 pedestrians. The same setting as in SDALF is used for fair comparison, *i.e.* the same probe/gallery protocol and 10 dataset splits with random 316 testing pedestrians. The CMC curve is used as the validation method which represents the expectation of finding the correct match in the top n matches [8].

Method	Feature Extracting	Metric Learning	Distance Computing
MCSH	84s	0	40s
TSM	121s	0	8h
MSPC	114s	0	25s
TCM	114s	0	5min
HSCD	185s	0	56s
SDALF [7]	1.6h	0	1.4h
RDC [36]	131s	1.5h	50s

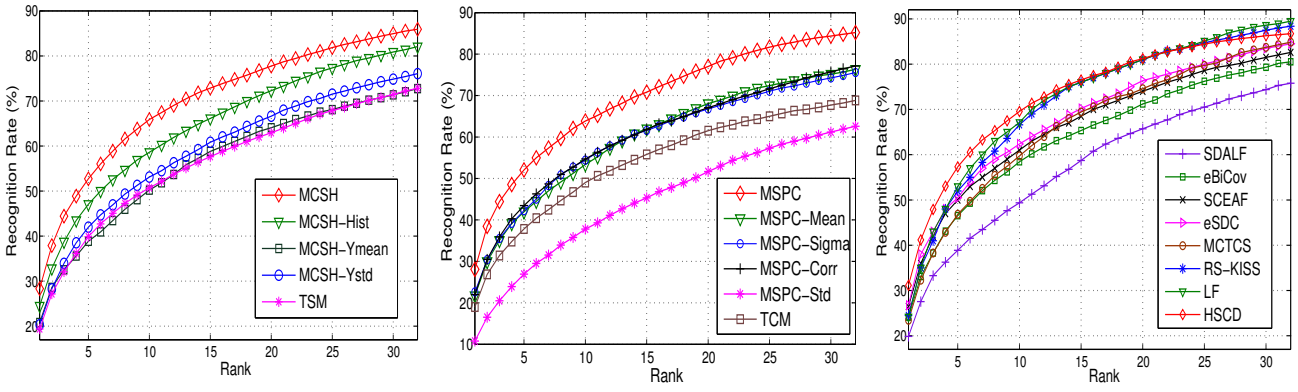
Table 1. Time comparison of different methods on VIPeR dataset.

Component Analysis. A detailed analysis of HSCD components is provided firstly. Figure 3(a) and 3(b) present the CMC results of several sub-features in MCSH and MSPC as well as two traditional methods. All the related weights are estimated using the first 100 image pairs in VIPeR as in SDALF [7]. Specifically, MCSH adopts 0.5, 0.35, and 0.15 for its components $\mathbf{h}$, $\boldsymbol{\mu}_y$ and $\boldsymbol{\sigma}_y$. The method with traditional spatiogram metric (denoted as TSM) as in [5] is included. MSPC adopts 0.35, 0.25, 0.1 and 0.3 for its four vectors $\mathbf{f}_i, i = 1, 2, 3, 4$. The method with traditional covariance metric (denoted as TCM) as in [2] is compared.

Figure 3 indicates that the histogram, Sigma vector and correlation vector perform better than other components. More importantly, both MCSH and MSPC get substantially promoted by fusing their components. Besides, traditional TSM and TCM methods don't get impressive results with expensive Gaussian or covariance metrics. In Figure 3(c), the fused HSCD method (both weights in (8) are set as 0.5) achieves further enhancement, with rank-1 rate reaching 31% from 28%.

Time comparison is then presented in Table 1. All compared methods are tested on the same Matlab platform. The codes of SDALF and RDC [36] are downloaded and reproduced from their papers. The related time is recorded about all testing images in one fold test. TSM costs 8 hours on expensive Gaussian metric, while MCSH just takes 84 seconds with the simple l_1 based metric in (8). MSPC is almost 12 times faster than the TCM method. Moreover, the time of extracting the HSCD features on the whole dataset (1264 images) is as short as 3 minutes while the time reaches about 1.6 hour for SDALF. The reason behind is that HSCD doesn't employ heavy preprocessing steps such as foreground or parts segmentation as in SDALF. Also, unlike our unsupervised HSCD method, RDC spends extra hours and consumes massive memory resource for training its metric.

Comparison with Current Methods. We then evaluate HSCD against other published methods, including SDALF



(a) (b) (c)

Figure 3. CMC curves on VIPeR. (a) MCSH components analysis, (b) MSPC components analysis, (c) Current methods comparison.

[7], SCEAF [11], eBiCov [16], MCTCS [28], eSDC [32], RS-KISS [20] and LF [19]. The CMC curves are shown in Figure 3(c). Besides, as practical applications such as pedestrian tracking just take several most promising targets as potential candidates, we also propose using new ACI (Application based CMC Index) to quantify the practical or useful re-id performance. For all compared methods, three indexes are firstly extracted: matching rates at rank 1 and rank 10% (*i.e.* percentage of the gallery set, for VIPeR, rank 10% of 316 testing people equals rank 32), and nAUC-10% (*i.e.* normalized Area Under CMC between rank 1 and rank 10%). Three indexes are normalized and added to form ACI. Unlike PUR (Proportion of Uncertainty Removed) index [19] in which a whole CMC curve is needed, methods which just publish frontal partial curves (no public CMC data file) are also available for comparison. A clearer comparison with ACI is presented in Table 2.

From Figure 3(c) and Table 2, we can conclude that HSCD outperforms other methods on such a challenging dataset where 1% improvement may be deemed great due to its numerous samples. HSCD achieves as high as 58% rank-1 improvement (*i.e.* 31.2% vs. 19.8%) compared with SDALF, which is the famous benchmark method that combines three different color and texture features.

Among recent feature based methods, eBiCov, SCEAF, eSDC and MCTCS focus on feature combination, feature designing, salience learning and graph matching. But their whole CMC curves perform poorer than MCSH or MSPC. As for RS-KISS and LF, the recent two outstanding metric learning methods, though performing better in the latter CMC ranks, they can't achieve better ACI index than HSCD. Though not compared here, another two recent learned metrics SSCDL [14] and LMF [33] still can't compete with HSCD in the first few ranks (25.6% and 29.11% rank-1 rate respectively), which may indicate metric learning optimizes the overall CMC curve but not necessarily the useful front part. So we address that HSCD outperforms all current feature based methods substantially and can compete with most metric learning methods in frontal CMC.

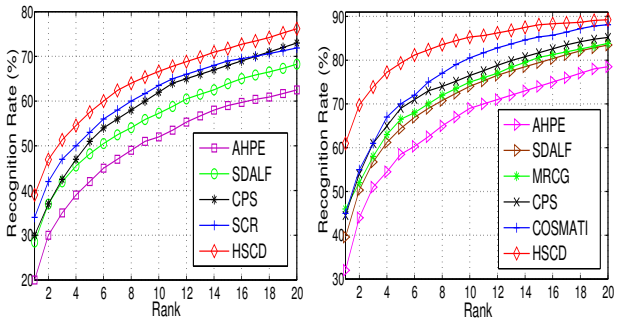
Method	R-1	R-10%	nAUC-10%	ACI
MCSH	28.3	85.9	70.0	96.3
MSPC	28.1	85.2	69.1	95.8
HSCD	**31.2**	86.5	**73.2**	**98.8**
SDALF	19.8	75.8	57.3	87.1
eBiCov	24.3	80.5	63.4	91.9
MCTCS	23.4	84.9	66.3	92.5
SCEAF	26.5	82.6	66.4	94.2
eSDC	26.7	84.5	68.0	95.1
RS-KISS	24.5	88.4	71.5	94.3
LF	24.2	**89.5**	72.2	94.5

Table 2. ACI comparison of different methods on VIPeR dataset.

5.2. iLIDS Dataset Evaluation

The iLIDS dataset [6] contains 479 images of 119 pedestrians in an airport scenario. The images undergo quite large illumination changes and subject to severe occlusions. It's used to evaluate HSCD method in both single shot case and multi-shot case. The same setups as in [6, 7] are adopted. Specifically, in the single shot case, one image of each person is randomly selected to build the gallery set while the others form the probe set. For the multi-shot case, two images are used to form both sets for each person. Simple MPD strategy is used for HSCD (HSCD-CR is not tested here due to the scarce samples per person). The CMC results in Figure 4 are averaged with 100 trials.

Figure 4 reveals that HSCD outperforms other methods on both cases, achieving almost 40% and 60% matching rate at rank 1. In the single shot case, AHPE [4] performs worst. SDALF [7], CPS [6], and SCR [2] can't compete with HSCD. In the multi-shot case, the feature based AHPE,



(a) (b)

Figure 4. CMC curves of different methods on iLIDS dataset. (a) Single shot evaluation, (b) Multi-shot evaluation.

Metric	R-1(%)	R-10(%)	R-25(%)	nAUC	Time(s)
MPD	18.6	66.2	91.1	82.4	**16.8**
APD	23.1	75.9	94.7	87.1	19.2
SRC	30.0	77.4	93.0	87.4	158
CRC	25.2	72.2	92.8	85.8	33.5
CRC-R	30.6	78.8	94.8	88.5	32.4
CRC-C	32.2	80.2	95.4	89.4	33.2
CRC-S	**35.2**	**84.2**	**97.6**	**91.1**	34.1

Table 3. Multi-shot metrics comparison on Caviar4reid.

SDALF, CPS, and MRCG [3] fall much behind HSCD (with MPD strategy). The learning based COSMATI [1] performs poorer than HSCD in most of the CMC ranks.

5.3. Caviar4reid Dataset Evaluation

We continue to evaluate the proposed HSCD-CR on the multi-shot dataset Caviar4reid [6]. In this dataset, 50 people are captured by two different views and 10 samples from each view are selected for each person. As in [6], the probe and gallery sets are taken from different views and 5 samples are randomly chosen for each person in both sets.

To validate the proposed multi-shot metric, different multi-shot metrics with the same MCSH feature are firstly compared in Table 3. The CMC values of rank 1, 5, 10, and 25 are listed, followed by the nAUC index and the metric computing time. Parameters are well tuned for every metric. For CRC-S, distances of residual and coefficients are computed and normalized respectively and then combined with weights 0.45 and 0.55 (to avoid estimating δ in (14)). CRC-R and CRC-C denote two metrics that just use residual or coefficients, which both perform better than traditional CRC metric. CRC-S achieves the best result and costs much less time than the sparse coding based SRC metric.

At last, the proposed HSCD-CR method is evaluated against other current multi-shot methods in Figure 5. For HSCD-CR, the combination weights of HSCD's MCSH and

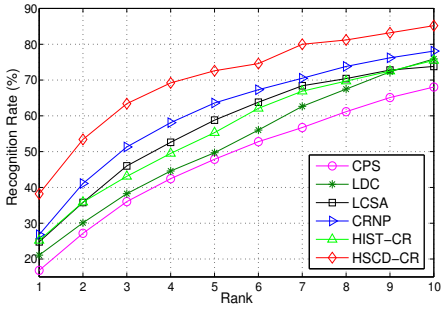


Figure 5. CMC curves of current methods on Caviar4reid dataset.

MSPC sub-features are set as 0.6 and 0.4. The compared counterparts include CPS, LDC [30], LCSA [27] and CRNP [25]. CPS and LDC focus on feature designing and metric learning while LCSA and CRNP are coding based methods. Apart from the HSCD feature, we also apply CRC-S on a similar histogram feature which is adopted in LDC and denote it as HIST-CR (Histogram based CRC-S). Our HSCD-CR outperforms other methods including the recent CRNP greatly, achieving 38.2% and 85.2% recognition rate at rank 1 and 10. Though without metric learning, HIST-CR can still achieve comparable result with LDC. However, it is much poorer than HSCD-CR for its feature is not as excellent as HSCD, implying that HSCD-CR benefits from both its well-designed feature and metric. Furthermore, HSCD-CR is especially efficient for not involving time-consuming training or l_1 constrained optimization as LDC and LCSA.

6. Conclusions

In this paper, we take a deep exploration to the potentials of feature designing strategy. Confronted with recent unfair competition with learning based methods, we address non-learning strategies instead and achieve the goal of winning one round back against learning methods somehow. Specifically, we propose a novel efficient HSCD descriptor and a collaborative representation based multi-shot metric for the re-id task. The proposed feature and metric are both well-constructed based on the simple idea that only useful components should be exploited and fused for enhancement.

Our proposed methods not only achieve state-of-the-art results in the experiments, but also are straightforward and computationally efficient. They don't require complicated training or any pre-processing, which will facilitate practical real-time surveillance applications. Further comprehensive comparisons will be considered on recent challenging datasets (*e.g.* [34]), including dimension reduction and metric learning. Besides, fusion with more sensors (*e.g.* infrared or thermal) [18] and robotics deserve further study.

References

[1] S. Bak, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In *ECCV*. 2012. 2, 8

[2] S. Bak, E. Corvee, F. Brémond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *AVSS*, 2010. 2, 4, 6, 7

[3] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Multiple-shot human re-identification by mean riemannian covariance grid. In *AVSS*, 2011. 3, 8

[4] L. Bazzani, M. Cristani, A. Perina, and V. Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognition Letters*, 33(7):898–903, 2012. 3, 7

[5] S. T. Birchfield and S. Rangarajan. Spatiograms versus histograms for region-based tracking. In *CVPR*, 2005. 2, 3, 6

[6] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011. 1, 2, 5, 6, 7, 8

[7] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 1, 2, 5, 6, 7

[8] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, 2007. 1, 6

[9] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*. 2008. 2

[10] X. Hong, H. Chang, S. Shan, X. Chen, and W. Gao. Sigma set: A small second order statistical region descriptor. In *CVPR*, 2009. 2, 4

[11] Y. Hu, S. Liao, Z. Lei, D. Yi, and S. Z. Li. Exploring structural information and fusing multiple features for person re-identification. In *CVPRW*, 2013. 1, 2, 7

[12] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 2

[13] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 2

[14] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *CVPR*, 2014. 7

[15] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV*, 2012. 1, 2

[16] B. Ma, Y. Su, and F. Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*, 32(6):379–390, 2014. 1, 2, 7

[17] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 1, 2

[18] A. Mogelmose, C. Bahnsen, T. B. Moeslund, A. Clapés, and S. Escalera. Tri-modal person re-identification with rgb, depth and thermal features. In *CVPR Workshop on PBVS*, 2013. 8

[19] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013. 1, 2, 7

[20] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li. Person re-identification by regularized smoothing kiss metric learning. *TCSVT*, 23(10):1675–1685, 2013. 1, 2, 7

[21] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *ECCV*. 2006. 4

[22] H. Wang, X. Bao, R. R. Choudhury, and S. Nelakuditi. Insight: recognizing humans without face recognition. In *MobiSys*, 2013. 1

[23] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI*, 31(2):210–227, 2009. 5, 6

[24] Y. Wu, V. Jarich, M. Mukunoki, and M. Minoh. Collaborative representation for classification, sparse or non-sparse? *arXiv preprint arXiv:1403.1353*, 2014. 5

[25] Y. Wu, M. Minoh, and M. Mukunoki. Collaboratively regularized nearest points for set based recognition. In *BMVC*, 2013. 3, 5, 8

[26] Y. Wu, M. Minoh, M. Mukunoki, W. Li, and S. Lao. Collaborative sparse approximation for multiple-shot across-camera person re-identification. In *AVSS*, 2012. 3

[27] Y. Wu, M. Mukunoki, and M. Minoh. Locality-constrained collaborative sparse approximation for multiple-shot person re-identification. In *ACPR*, 2013. 3, 8

[28] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu. Human re-identification by matching compositional template with cluster sampling. In *ICCV*, 2013. 2, 7

[29] D. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? In *ICCV*, 2011. 5

[30] G. Zhang, Y. Wang, J. Kato, T. Marutani, and K. Mase. Local distance comparison for multiple-shot people re-identification. In *ACCV*. 2013. 8

[31] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013. 2

[32] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 1, 2, 7

[33] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 7

[34] L. Zheng, L. Shen, L. Tian, S. Wang, J. Bu, and Q. Tian. Person re-identification meets image search. *arXiv preprint arXiv:1502.02171*, 2015. 8

[35] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian. Query-adaptive late fusion for image search and person re-identification. In *CVPR*, 2015. 2

[36] W. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *PAMI*, 35(3):653–668, 2013. 1, 2, 6