In Defense of Color Names for Small-Scale Person Re-Identification

Yang Yang[1]*, Zhen Lei[1], Jinqiao Wang[1], Stan Z. Li[1]
[1]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{yang.yang, zlei, jqwang, szli}@nlpr.ia.ac.cn

Abstract

In this paper, we propose an efficient image representation strategy for addressing the task of small-scale person re-identification. Taking advantages of being compact and intuitively understandable, we adopt color names descriptor (CND) as our color feature. To solve the inaccuracy by comparing color names with image pixels in Euclidean space, we propose a new approach – soft Gaussian mapping (SGM), which uses a Gaussian model to bridge their semantic gap. We further present a cross-view coupling learning method to build a common subspace where the learned features can contain the transition information among different cameras. Experiments on the challenging small-scale benchmark public datasets demonstrate the effectiveness of our proposed method.

1. Introduction

Person re-identification is to match the persons across multiple cameras with non-overlapping views [31, 19, 30]. It is challenging because the appearance of a person's surveillance images in different cameras may exhibit dramatic changes caused by illumination variation, as well as different camera views and body poses. Recently, convolutional neural network (CNN) is the main focus because of its higher performance on large-scale data and end-to-end manner [24, 35, 14, 4, 34]. But it still faces two problems: (1) their performance requires sufficient labeled training data, which is an extremely difficult task because of there being significant changes in surveillance environments, and (2) it is time-consuming for training. In view of this, we shun ourselves away from deep learning methods on larger-scale datasets and mainly consider the following task in this paper: given a small amount of labeled training data, how to efficiently match unseen persons?

Image representation is arguably the most fundamental task because it determines the upper limit of the overall performance. The appearance based low-level features can be roughly divided into (1) color (color histogram [13, 12, 6, 29, 19] and color names based representation [31, 33]), (2) texture, e.g., SILTP [15, 3], LBP [12] and Gabor filters [9], (3) shape [20], and (4) gradient[19]. However, due to the fact that it is extremely complicated in unconstrained surveillance conditions, no single feature can be qualified completely for the task of person re-identification. A common strategy is to combine the features with complementary information to build richer signatures. In existing image representation methods, local maximal occurrence (LOMO) [15] and Gaussian of Gaussian (GOG) [19] show impressive performance and are widely used. Both of them contain color and texture information. However, the dimensions of them are over 20,000, which take up a lot of storage space and increase the training time. In [31], Yang et al. propose salient color name based color descriptor (SCNCD) which applies 16 pre-defined color names to represent images. This kind of CND is not only compact (dimension: 1000+) but also has shown good robustness to illumination. Hence, we firstly concentrate on designing a novel low-level color feature – CND, which has complementary information with LOMO and GOG.

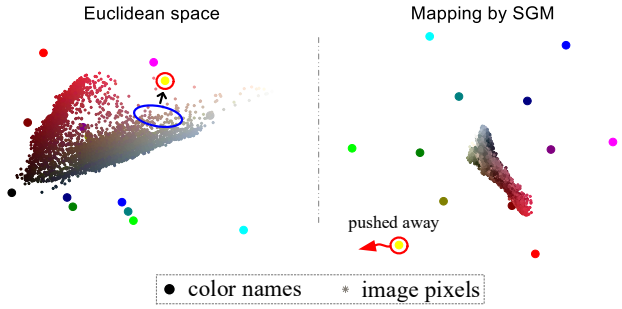

Figure 1. Examples of color names and image pixels in original space (left) and transformed space by SGM (right). Image pixels come from an image in VIPeR dataset.

We note that color names are pre-defined and image pixels are from surveillance cameras. Therefore, the underlying distribution of them are different, which leads to an unreliable comparison in the original Euclidean space. This is because Euclidean distance treats three color channels as an isotropic one, and thus being unable to exactly reflect

*Corresponding author.
978-1-7281-3640-0/19/$31.00 ©2019 IEEE

the underlying relationship between color names and image pixels. In Fig. 1, when the Euclidean distance is regarded as a dissimilarity measure between color names and image pixels, we can see that there will be a set of image pixels (circled by a blue ellipse) being assigned to the color name *yellow*. In fact, the set of image pixels visually appear totally different from color name *yellow*. This observation reflects that although the image pixel stays 'close' to one color name, it does not definitely imply that it has the same semantic information with the color name. The inaccurate semantic measure between image pixels and color names will further limit the performance of CND.

Different with Euclidean distance, Mahalanobis distance takes into account the correlation among different dimensions based on a Gaussian model, the covariance matrix of which is estimated by aggregating them together. However, in real applications, the number of image pixels is far larger than that of color names. As thus, the estimated covariance matrix may approach to reflecting the distribution of image pixels while neglecting, to some extent, the color names' distribution. In addition, it assumes that image pixels and color names obey the same Gaussian model, which is not the truth. To overcome it, we establish a connection between an image pixel and each color name by taking their discrepancy as a new sample. By doing so, the imbalance between image pixels and color names can be alleviated. Since there are a large number of new samples, we can reasonably model them using a Gaussian (according to the central limit theorem). The inverse of Gaussian's covariance matrix only reflects the difference between image pixels and color names and can then be employed to bridge their semantic gap. Fig. 1 shows that based on the Gaussian model of discrepancy, the color name *yellow* is pushed away while in the transformed space, the distance between an image pixel and a color name is consistent with their semantic relationship.

The contribution of this paper is as follows:

- We propose a novel and efficient method named soft Gaussian mapping (SGM) to learn the description of an image pixel over color names. It has complementary information with other color/texture features.
- We introduce cross-view coupling learning (CCL) to build a common subspace with cross-view information. Based on it, the learned image representation is low-dimensional and discriminative.

2. Color Names based Image Representation

2.1. SGM for CND of a Pixel

Let $\boldsymbol{Z}$ be a set of three-dimensional image pixels, i.e., $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_n] \in \mathcal{R}^{3\times n}$. Meanwhile, we assume $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_{16}] \in \mathcal{R}^{3\times 16}$ denotes a set of 16 three-dimensional color names defined in [31]. To compute the CND for a pixel, a probability distribution over color names is often used. For example, SCNCD uses a saliency coding with an $\exp(.)$ function and an index table is established by taking all 256×256×256 colors as image pixels. However, SCNCD compares the image pixels and color names in Euclidean space, which is not unreliable.

To take the distribution into consideration, Mahalanobis distance is well defined by using the inverse of covariance matrix. It first creates a variable $\boldsymbol{g}_M$ with the mean subtracted and then assumes a zero-mean Gaussian model for $\boldsymbol{g}_M$. But it is inappropriate to assume that image pixels and color names follow the same distribution. To address it, we create a novel variable $\boldsymbol{g}$: the set of discrepancies between image pixels $\boldsymbol{z}_i, i = 1, 2, ..., n$ and color names $\boldsymbol{c}_j, j = 1, 2, ..., 16$, i.e., $\boldsymbol{g}_{ij} = \boldsymbol{z}_i - \boldsymbol{c}_j$. It is obvious that the variable $\boldsymbol{g}$ is zero-mean. We then model the variable $\boldsymbol{g}$ by using a Gaussian:

$$P((\boldsymbol{z}_i, \boldsymbol{c}_j)|\Theta) = \sigma \exp(-\frac{1}{2}\boldsymbol{g}_{ij}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{g}_{ij}), \tag{1}$$

where σ is a constant, i.e., $(2\pi)^{-3/2}|\boldsymbol{\Sigma}|^{-1/2}$ with $|\boldsymbol{\Sigma}|$ being the determinant of matrix $\boldsymbol{\Sigma}$ and $\Theta = (\boldsymbol{0}, \boldsymbol{\Sigma})$ is the Gaussian model parameter. Note that the set of discrepancies is symmetric with zero mean. $\boldsymbol{\Sigma}$ is estimated by

$$\boldsymbol{\Sigma} = \frac{1}{16n} \sum_{i=1}^{n} \sum_{j=1}^{16} \boldsymbol{g}_{ij} \boldsymbol{g}_{ij}^T, \tag{2}$$

where $\boldsymbol{\Sigma}$ is a 3×3 symmetric matrix. $\boldsymbol{\Sigma}^{-1/2}$ can be treated as the mapping matrix.

With Eqs. 1 and 2, we can easily estimate the likelihood of the image pixel $\boldsymbol{z}_i$ belonging to the color name $\boldsymbol{c}_i$. This estimated likelihood can be served as the color names descriptor of an image pixel. It describes the membership of an image pixel to color names from a probabilistic perspective. In [31, 25, 11, 16], an 'early cut-off' is often used to remove the adverse impact of dissimilar factor and can handle the underlying manifold structure when local descriptors are learned. To make it more generic, we defined our soft gaussian mapping in a more flexible manner:

$$\boldsymbol{s}_j = \begin{cases} P((\boldsymbol{z}_i, \boldsymbol{c}_j)|\Theta) & , \quad if\ \boldsymbol{c}_j \in \mathcal{N}_k(\boldsymbol{z}_i) \\ 0 & , \quad else, \end{cases} \tag{3}$$

where $\mathcal{N}_k(\boldsymbol{z}_i)$ denotes k most similar color names of $\boldsymbol{z}_i$ defined by their similarities $P((\boldsymbol{z}_i, \boldsymbol{c}_j)|\Theta)$. We further employ sum normalization [25]

$$\boldsymbol{s}^T 1 = 1 \tag{4}$$

to make the descriptor stable. In consequence, an image pixel's CND obtained by SGM can be taken as its probability distribution over color names $\boldsymbol{s}_j, j = 1, 2, ..., 16$. Here, 'soft' means that given an image pixel, several color names are considered.

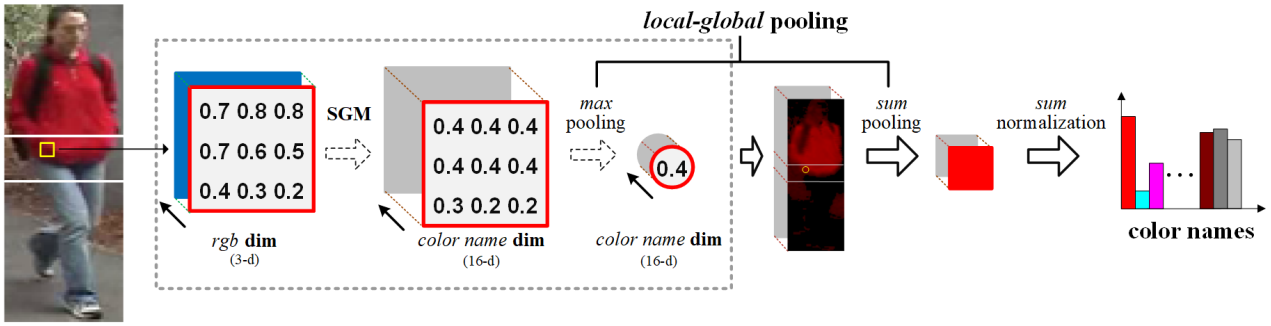

Figure 2. Flowchart of extracting a robust color names descriptor in one strip.

2.2. Image Representation: from Pixel-Level to Image-Level

Based on SGM, an image is converted to 16 soft Gaussian maps, within each of which the global spatial layout is still preserved. We then present a local-global pooling scheme to characterize the signature of each stripe. It includes maxpooling in a 3×3 patch with a stride of 3, followed by sumpooling in one stripe. By taking the maximum in a local patch, small-scale deviation and mapping noises can be reduced, thus enjoying, to some extent, invariance properties [16]. Meanwhile, sumpooling is further utilized to describe the statistical information in one stripe. To make it owning the concept of probability distribution over color names, sum normalization in Eq. 4 is also employed to form the CND of a stripe (see Fig. 2). Then, the image-level feature is obtained via concatenating CNDs of all stripes.

2.3. Cross-View Coupling Learning

Due to the changes of viewing conditions in different cameras and the high-dimensional image representation, a lower subspace with cross-view information should be learned from labeled training data. LDA is an efficient way of learning a discriminative subspace $\boldsymbol{w} \in \mathcal{R}^{d\times r}$, where d and r are the dimensions of original feature and subspace, respectively. Its objective is as follows:

$$J(\boldsymbol{w}) = \sigma_E(\boldsymbol{w})/\sigma_I(\boldsymbol{w}), \tag{5}$$

where $\sigma_E(\boldsymbol{w})$ and $\sigma_I(\boldsymbol{w})$ denotes inter-personal and intra-personal variances, respectively. Since both of inter-/intra-personal variables have zero mean, we ignore the mean while solving the objective like [15]. Eq. 5 means that in the learned subspace, the intra-personal variance is suppressed with respect to inter-personal variance. However, it only considers *difference* of an image pair. Motivated by [28] which deems that more discrimination can be expected on account of both *commonness* and *difference*, we propose a new subspace learning method using both *commonness* and *difference*.

Given an image pair $(\boldsymbol{x}, \boldsymbol{y})$ from two different cameras, we have *commonness*, i.e., $\boldsymbol{m} = \boldsymbol{x} + \boldsymbol{y}$ and *difference* i.e., $\boldsymbol{e} = \boldsymbol{x} - \boldsymbol{y}$. The coupled variables $\boldsymbol{m}$ and $\boldsymbol{e}$ (zero-centered) are negatively correlated, i.e., when $\|\boldsymbol{m}\|_2$ is small, $\|\boldsymbol{e}\|_2$ is large and vice versa. Then, we expect to learn a subspace with the following traits: (1) for $\boldsymbol{e}$, the intra-personal variance is suppressed with respect to inter-personal variance and (2) for $\boldsymbol{m}$, the inter-personal variance is suppressed with respect to intra-personal variance. That is to say we expect to learn a subspace where the same persons both stay more compact (for e) and own higher similarities (for m) than different persons. To that end, our objective is defined to maximize $J_e(\boldsymbol{w})$ and $J_m(\boldsymbol{w})$ jointly:

$$\begin{cases} J_e(\boldsymbol{w}) = \sigma_{eE}(\boldsymbol{w})/\sigma_{eI}(\boldsymbol{w}) \\ J_m(\boldsymbol{w}) = \sigma_{mI}(\boldsymbol{w})/\sigma_{mE}(\boldsymbol{w}) \end{cases} \tag{6}$$

where σ_{eE} and σ_{mE} denote the inter-personal variance covariance matrices over $\boldsymbol{e}$ and $\boldsymbol{m}$, respectively while σ_{eI} and σ_{mI} denote the intra-personal variance covariance matrices over $\boldsymbol{e}$ and $\boldsymbol{m}$, respectively. According to [28], σ_{eE} equals to σ_{mE} in the same subspace. Consequently, our objective can be simplified to maximize $J_0(\boldsymbol{w})$

$$J_0(\boldsymbol{w}) = \sigma_{mI}(\boldsymbol{w})/\sigma_{eI}(\boldsymbol{w}), \tag{7}$$

which reflects that when only similar pairs are considered, we wish to learn a subspace where the intra-personal variance for $\boldsymbol{e}$ is suppressed with respect to intra-personal variance for $\boldsymbol{m}$. We rewrite Eq. 7 to

$$J_0(\boldsymbol{w}) = \boldsymbol{w}^T\Sigma_{mI}\boldsymbol{w}/\boldsymbol{w}^T\Sigma_{eI}\boldsymbol{w}. \tag{8}$$

The maximization of $J_0(\boldsymbol{w})$ can be solved by the generalized eigenvalue decomposition problem, i.e., the subspace is composed of the eigenvectors corresponding to r largest eigenvalues of $\Sigma_{eI}^{-1}\Sigma_{mI}$. It has a closed-form solution. In consideration of the *small sample problem* [7], we add a small regularizer to the diagonal elements of Σ_{eI} to avoid Σ_{eI}^{-1} being singular. Finally, we calculate the similarity score for any two images based on LSSL [28].

2.4. Run-time Complexity

Here, we discuss the run-time complexity of our proposed method: (a) In SGM, the computation complexity

Table 1. Comparison with SCNCD on VIPeR. Rank-1 results are shown under different similarity learning methods.

Method	KISSME	LSSL	XQDA	CCL
SCNCD	35.0%	41.7%	40.0%	44.6%
SGM	**40.9%**	**47.3%**	**44.4%**	**50.0%**

of learning the covariance matrix in Eq. 2 is reduced to $\boldsymbol{O}(n3^2)$, where n is larger than 16; (b) CCL requires the computation of covariance matrix, matrix multiplication and eigenvalue decomposition. Its computation complexity is $\boldsymbol{O}(Nd^2 + d^3 + d^3)$, where N denotes the number of labeled persons.

3. Experiments

In this section, we evaluate our method on three popular small-scale benchmark datasets (VIPeR [8], PRID 450S [21] and GRID [17]. When we compare with the state-of-the-art methods, the best matching rate is shown in red, the second best is in blue and our methods are shown in bold. All experiments are evaluated on a PC with the 3.40 GHz Core I7 CPU with 8 cores.

3.1. Datasets

VIPeR Dataset. VIPeR dataset has 632 persons captured with two disjoint cameras in outdoor environments. There is one image for each person in each camera view. It is challenging due to arbitrary viewpoints, pose changes and illumination variations. Images are mostly captured from 0 degree to 90 degree in Camera A while those from Camera B are mostly from 90 degree to 180 degree. All images are normalized to 128×48.

PRID450S Dataset. PRID450S dataset consists of 450 persons captured from two spatially disjoint camera views. Each person has one image in each view. Due to different viewpoint changes, background interference, partial occlusion and illumination variations, it is also a challenging dataset. All images are normalized to 168×80.

GRID Dataset. GRID dataset contains 1025 persons captured in a busy underground station. It is captured from 8 disjoint camera views. There are 250 person image pairs. For each of them, there is one image in each camera view. Besides, there are additional 775 gallery persons that are different with the former 250 persons. All images are normalized to 160×60 pixels.

3.2. Setup

Training/test. In experiments, we report our results in the form of Cumulated Matching Characteristic (CMC) curve [26]. On all datasets, we randomly choose 50% of all persons for training while the remaining is used for test. We conduct it for 10 random splits and the average results are reported.

Features. As in [31], we use the image-foreground representation and the same 4 color spaces including RGB, rgb, $l_1l_2l_3$ and HSV. When comparing with the state-of-the-art approaches, we employ another two simple and commonly used features: color histogram and SILTP.

Parameters. Unless otherwise specified, we empirically set the parameters as follows: (1) We use 10 non-overlapping stripes. (2) We set k to 5 for SGM. (3) We adopt 16 bins in each channel for color histogram. (4) We project each type of features to 100 subspace by CCL.

3.3. Evaluation on VIPeR

In this subsection, we make an evaluation of our method on the widely used VIPeR dataset.

Comparison with SCNCD. It has been demonstrated that SCNCD shows better performance than color histogram and existing CND. In Table 1, we mainly compare SGM with SCNCD under different similarity learning methods. For a fair comparison, 10 non-overlapping stripes are used for SCNCD, i.e., both of SGM and SCNCD are 1280-dimensional features. Rank-1 results are shown. We can observe that based on the same similarity learning method, SGM outperforms SCNCD at least 4.4% at Rank 1. In addition, we observe that an improvement can be achieved by using CCL instead of PCA used in LSSL and that CCL, combined with LSSL, achieves the best results.

Euclidean v.s. Mahalanobis v.s. SGM distance. Our proposed color names descriptor is simply named as SGM. We compare SGM with SGM(Eu) and SGM(Ma) which employ Euclidean and Mahalanobis distances under our framework, respectively. Specifically, SGM(Eu) set the covariance matrix $\boldsymbol{\Sigma}$ in Eq. 1 to be an identity matrix while SGM(Ma) compute it by aggregating image pixels and color names together. All of them use CCL for subspace learning and LSSL for similarity learning. In Table 2, we list the comparison results. In comparison with SGM(Eu), SGM(Ma) can eliminate the inaccurate representation to a certain degree and improves the results by 3.3% at Rank 1. By eliminating the discrepancy in a better manner, SGM performs the best at all Ranks.

3.4. Comparison with the State-of-the-art Methods

To make it comparable with the state-of-the-art approaches, we fuse **S**GM, **S**ILTP and **c**olor histogram, named by SSC, as image representations. They only cost 0.036s, 0.004s and 0.018s to represent an image of 128 × 48, respectively. Then, we use CCL for subspace learning and LSSL for similarity learning. .

On VIPeR and PRID450S datasets, we use the same mask as [31]. On others, we use the method in [18] to automatically generate the masks (0.13s for an image of 160×60).

Table 2. Evaluation of different distance methods: Euclidean, Mahalanobis and SGM on VIPeR.

Rank	1	5	10	20
SGM(Eu)	43.6%	73.3%	82.6%	90.0%
SGM(Ma)	46.9%	75.1%	84.0%	90.8%
SGM	**50.0%**	**78.5%**	**88.1%**	**94.2%**

Table 3. Comparison with the state-of-the-art methods on VIPeR dataset. *denotes deep learning based methods.

Rank	1	5	10	20
NK3ML	99.8%	-	100%%	100%
SCSP+PCN*	54.2%	82.8%	91.4%	99.1%
Spindle*	53.8%	74.1%	83.2%	92.1%
SSM	53.7%	-	91.5%	96.1%
SCSP	53.5%	82.6%	91.5%	96.7%
STNs*	50.1%	73.1%	84.4%	-
GOG	49.7%	79.7%	88.7%	94.5%
MetricEmsemble	45.9%	77.5%	88.9%	95.8%
LOMO	40.0%	68.1%	80.5%	91.1%
SCNCD+CH	37.8%	68.5%	81.2%	90.4%
SGM	**50.0%**	**78.5%**	**88.1%**	**94.2%**
SSC	**59.2%**	**85.0%**	**92.1%**	**96.4%**

State-of-the-art: VIPeR. VIPeR is a classic benchmark dataset and may be most widely used for person re-identification. The compared approaches include NK3ML [1], SCSP+PCN [5], Spindle [32], SSM [2], SCSP [3], STNs [10], GOG [19], LSSL [28], MetricEmsemble [20], LOMO [15] and SCNCD+CH [31]. Among the previous approaches in Table 3, NK3ML achieves the best result at Rank 1. It is a metric learning method. SCSP+PCN obtains the second best result at Rank 1. It combines the PCN with SCSP which uses 6 types of basic features including two types of HSV and LAB, as well as HOG and SILTP. Compared with SCSP+PCN, our SSC improves the result by 5.0% at Rank 1. Ours is mainly based on feature extraction. We can achieve better results when we use NK3ML instead of LSSL to compute the similarity. We do not report this results because the code of NK3ML has not been released yet and we have not reproduced its result.

State-of-the-art: PRID450S. We compare our method with the state-of-the-art approaches on PRID450S dataset, including NK3ML [1], SSM [2], GOG [19], LOMO [15], MED_VL [27], CSL [22], TSR [23] and SCNCD+CH [31].

Among the previous approaches in Table 4, NK3ML [1] achieves the best results at Rank 1 and SSM achieves the best results at Ranks 10-20. SSM is based on the fuse of GOG and LOMO. Note that LOMO is a combination of joint histogram and SILTP. The second best result at Rank 5 is achieved by GOG which is based on pixel location, gradient information and color information. Our SSC achieves a new state-of-the-art result (74.8%) at Rank 1.

Table 4. Comparison with the state-of-the-art methods on PRID450S dataset.

Rank	1	5	10	20
NK3ML	73.4%	-	96.3%	98.6%
SSM	73.0%	-	96.8%	99.1%
GOG	68.4%	88.8%	94.5%	97.8%
LOMO	62.6%	85.6%	92.0%	96.6%
MED_VL	45.9%	73.0%	82.9%	91.1%
TSR	44.9%	71.7%	77.5%	86.7%
CSL	44.4%	71.6%	82.2%	89.8%
SCNCD+CH	41.6%	68.9%	79.4%	87.8%
SGM	**66.1%**	**86.9%**	**91.4%**	**95.3%**
SSC	**74.8%**	**91.4%**	**94.8%**	**97.2%**

Table 5. Comparison with the state-of-the-art methods on GRID dataset.

Rank	1	5	10	20
NK3ML	27.2%	-	61.0%	71.0%
OL-MANS	30.2%	-	49.2%	59.4%
SSM	27.2%	-	61.1%	70.6%
GOG	24.7%	47.0%	58.4%	69.0%
SCSP	24.2%	-	54.1%	65.2%
LOMO	16.6%	-	41.8%	52.4%
SGM	**26.6%**	**49.0%**	**57.9%**	**68.1%**
SSC	**31.9%**	**51.1%**	**59.6%**	**69.2%**

State-of-the-art: GRID. On GRID, the compared methods include NK3ML [1], OL-MANS [36], SSM [2], GOG [19] SCSP [3] and LOMO [15]. All of them are traditional feature / similarity learning methods. Among the previous approaches in Table 5, OL-MANS achieves the best result at Rank 1. Compared with OL-MANS, SSC improves the rank-1 result by 1.7%.

4. Conclusion

In this paper, we propose a new method to learn the color names descriptor for small-scale person re-identification. It addresses the semantic gap between color names and image pixels based on a Gaussian model and uses a local-global pooling strategy to make the descriptors enjoying some invariance properties. Finally, a new subspace learning method based on positive samples is presented by a cross-view analysis on *commonness* and *difference*. We make an evaluation of our method on VIPeR and demonstrate its effectiveness on three small-scale datasets.

Acknowledgements. This work was supported by the Chinese National Natural Science Foundation Projects #61806203, #61872367 and #61572501.

References

[1] T. M. F. Ali and S. Chaudhuri. Maximum margin metric learning over discriminative nullspace for person re-identification. In *ECCV*, 2018. 5

[2] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, 2017. 5

[3] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, pages 1268–1277, 2016. 1, 5

[4] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017. 1

[5] Z. Cheng, X. Li, and C. C. Loy. Pedestrian color naming via convolutional neural network. In *ACCV*, pages 35–51, 2016. 5

[6] 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

[7] K. Fukunaga. Introduction to statistical pattern recognition, 2nd ed. *New York: Academic*, 1990. 3

[8] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007. 4

[9] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 1

[10] Y. Guo and N.-M. Cheung. Efficient and deep person re-identification using multi-level similarity. In *CVPR*, 2018. 5

[11] Y. Huang, K. Huang, Y. Yu, and T. Tan. Salient coding for image classification. In *CVPR*, 2011. 2

[12] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 1

[13] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(7):1622–1634, 2013. 1

[14] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *CVPR*, volume 1, page 2, 2018. 1

[15] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 1, 3, 5

[16] L. Liu, L. Wang, and X. Liu. In defense of soft-assignment coding. In *ICCV*, pages 2486–2493, 2011. 2, 3

[17] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *CVPR*, pages 1988–1995, 2009. 4

[18] P. Luo, X. Wang, and X. Tang. Pedestrian parsing via deep decompositional network. In *ICCV*, 2013. 4

[19] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016. 1, 5

[20] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015. 1, 5

[21] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof. Mahalanobis distance learning for person re-identification. In *Advances in Computer Vision and Pattern Recognition*, 2014. 4

[22] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang. Person re-identification with correspondence structure learning. In *ICCV*, 2015. 5

[23] Z. Shi, T. M. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *CVPR*, 2015. 5

[24] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 1

[25] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 2

[26] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007. 4

[27] Y. Yang, Z. Lei, S. Zhang, H. Shi, and S. Z. Li. Metric embedded discriminative vocabulary learning for high-level person representation. In *AAAI*, 2016. 5

[28] Y. Yang, S. Liao, Z. Lei, and S. Z. Li. Large scale similarity learning using similar pairs for person verification. In *AAAI*, 2016. 3, 5

[29] Y. Yang, S. Liao, Z. Lei, D. Yi, and S. Z. Li. Color models and weighted covariance estimation for person re-identification. In *ICPR*, 2014. 1

[30] Y. Yang, L. Wen, S. Lyu, and S. Z. Li. Unsupervised learning of multi-level descriptors for person re-identification. In *AAAI*, pages 4306–4312, 2017. 1

[31] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *ECCV*, 2014. 1, 2, 4, 5

[32] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017. 5

[33] 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. 1

[34] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[35] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2018. 1

[36] J. Zhou, P. Yu, W. Tang, and Y. Wu. Efficient online local metric adaptation via negative samples for person re-identification. In *ICCV*, pages 2439–2447, 2017. 5