IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON MISSION CRITICAL SENSORS AND SENSOR NETWORKS (MC-SSN)

Received September 20, 2019, accepted October 3, 2019, date of publication October 7, 2019, date of current version October 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2946021

Person Re-Identification via Contextual Region-Based Metric Learning in Camera Sensor Networks

ZHONG ZHANG, (Senior Member, IEEE), TONGZHEN SI, MEIYAN HUANG, AND SHUANG LIU, (Senior Member, IEEE)

Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China

Corresponding author: Shuang Liu (shuangliu.tjnu@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61501327 and Grant 61711530240, in part by the Natural Science Foundation of Tianjin under Grant 19JCZDJC31500 and Grant 17JCZDJC30600, in part by the Fund of Tianjin Normal University under Grant 135202RC1703, in part by the Open Projects Program of National Laboratory of Pattern Recognition under Grant 201800002, and in part by the Tianjin Higher Education Creative Team Funds Program.

ABSTRACT Person re-identification in camera sensor networks is a challenging issue due to significant appearance variations of pedestrian images captured by different camera sensors. The contextual information of pedestrian images is a vital cue to overcome appearance variations. However, many existing approaches learn the distance metric in a global way or restrict to corresponding sub-regions, which discards the contextual information of pedestrians or learns the contextual information inadequately. In this paper, we propose an effective method to tackle the problem for person re-identification in camera sensor networks. Firstly, we propose the Contextual Region-based Metric Learning (CRML) to fully learn the contextual information in a local manner, which simultaneously utilizes three kinds of sub-region pairs to learn a discriminative transformation matrix. Secondly, we employ the greedy axis rotation algorithm to optimize the transformation matrix in the framework of mutual information. Thirdly, in the process of local similarity integration, we further propose the Context-Constrained Match (CCM) to overcome the misalignment problem by seeking the optimal match in the neighboring sub-regions. Fourthly, we further present the nCRML to avoid the dimensionality curse and fuse similarity scores in different low-dimensional subspaces. The experimental results on three challenging datasets (VIPeR, QMUL GRID and CUHK03) demonstrate the effectiveness of our method.

INDEX TERMS Person re-identification, contextual region-based metric learning, context-constrained match, camera sensor networks.

I. INTRODUCTION

Person re-identification in camera sensor networks [1]–[3] is an essential issue in the filed of intelligent surveillance and its target is to spot the same person under different camera views as shown in Fig. 1. It has attracted much attention due to its wide range of applications, such as person retrieval, group behavior analysis, long-term person tracking, and so on [4], [5]. However, person re-identification in camera sensor networks is a very challenging task because the same pedestrian observed in different camera views often undergoes significant variations in illumination, poses, viewpoints and occlusions, which usually results in extra-personal difference even larger than intra-personal difference.

To deal with the above-mentioned challenges, many different methods have been proposed, and the most common research directions are feature representation and metric learning. Some researchers focus on designing features that are discriminative to distinguish extra-personal difference and robust against intra-personal difference. Texture and color features are commonly used in feature representation. There are many useful features, like Local Maximal Occurrence (LOMO) [6], Gaussian of Gaussian (GOG) [7], etc, which achieves promising performance for person re-identification. Furthermore, recently deep learning methods [8], [9] obtain promising results in feature learning. As for person re-identification [10]–[13], many researchers utilize

The associate editor coordinating the review of this manuscript and approving it for publication was Qilian Liang.

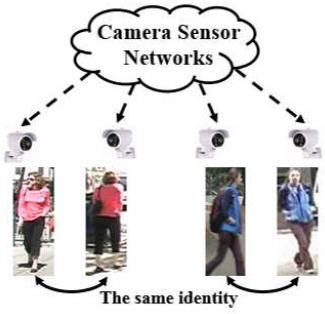

FIGURE 1. Person re-identification in camera sensor networks.

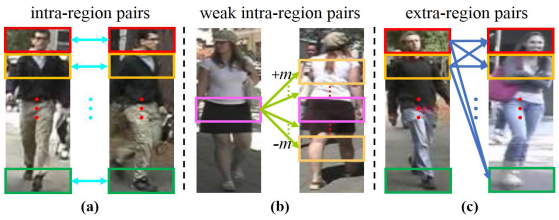

FIGURE 2. Three kinds of sub-region pairs. (a) shows the corresponding sub-region pairs belonging to the same pedestrian under different camera sensors called intra-region pairs; (b) shows the *m* neighbouring non-corresponding sub-region pairs belonging to the same pedestrian under different camera sensors called weak intra-region pairs; (c) shows the corresponding and non-corresponding sub-region pairs belonging to different pedestrians under different camera sensors called extra-region pairs.

the output of fully connected layer of deep model to represent pedestrian images, and then calculate the similarity between pedestrian features. The similarity calculation is vital to the performance of person re-identification. So, in this paper, we mainly focus on designing an effective metric learning method to calculate the similarity based on the extracted pedestrian features.

Metric learning could learn a distance metric to adapt to different data distributions. It has been applied to many research fields, such as face recognition, ground-based cloud classification, scalable image retrieval and so on [14], [15]. Inspired by its extensive applications, many researchers propose metric learning methods for person re-identification in order to ensure high similarity between images of the same pedestrian and low similarity between images of different pedestrians. The metric learning methods can be divided into global metric learning and local metric learning. The representative global metric learning methods include Large Margin Nearest Neighbor (LMNN) [16], Probabilistic Relative Distance Comparison (PRDC) [17], Cross-view Quadratic Discriminant Analysis (XQDA) [6], Kernelized Random KISS (KRKISS) [18] and so on. The global metric learning methods learn the similarity between image pairs in a holistic way, which discards the contextual information of pedestrians. The contextual information reflects the spatial structure of body parts and therefore it is essential to overcome viewpoint and pose variations. The local metric learning methods [19]–[21] focus on learning the relationship of body parts and the contextual information of pedestrian. However, most local metric learning methods are restricted to the limited contextual information learned from corresponding sub-region pairs.

In this paper, we propose a novel metric learning method named Contextual Region-based Metric Learning (CRML) to learn the distance metric in a local manner for person re-identification in camera sensor networks. Specifically, to overcome the variations in viewpoints and poses, we define three kinds of sub-region pairs (see in Fig. 2), i.e., intra-region pairs, weak intra-region pairs and extra-region pairs, which could model the completed relationship among sub-regions including misalignment and correspondence. Afterwards, we simultaneously utilize these three kinds of sub-region pairs to fully learn the contextual information for pedestrian images within the framework of mutual information. We maximize the mutual information to obtain a discriminative transformation matrix and correspondingly employ the greedy axis rotation algorithm to optimize the transformation matrix. We further present the nCRML to avoid the dimensionality curse and fuse similarity scores in different low-dimensional subspaces using several times of Random Projection (RP) on sub-region features. In the process of local similarity integration, we apply the Context-Constrained Match (CCM) to further overcome the misalignment problem caused by viewpoint and pose variations, which locates sub-region pairs using the maximum similarity in the neighborhood.

The rest of the paper is organized as follows. We first review related work in Section II. In Section III, we show the proposed CRML and nCRML in detail. In Section IV, we utilize the proposed CCM to integrate the similarity scores of sub-regions. Then, we compare our method with the state-of-the-art methods on three public datasets in Section V. Finally, we make a conclusion in Section VI.

II. RELATED WORK

Metric learning acts as a key step for person re-identification in camera sensor networks. An ideal metric learning method could yield higher similarity score for the image pairs belonging to the same class than that of the image pairs belonging to different classes. Global metric learning methods formulate a holistic similarity score between pedestrian images, while local metric learning methods learn the local similarity score between image regions. Next, we review global and local metric learning methods, respectively.

A. GLOBAL METRIC LEARNING

Many global metric learning approaches [16]–[18], [22]–[25] have been proposed for person re-identification. Weinberger *et al.* [16] propose the Large Margin Nearest Neighbor (LMNN) based on the KNN rule. The goal of LMNN is to learn a Mahalanobis distance metric, which pulls the pedestrian image with the k-nearest neighbors belonging to the same identity and meanwhile pushes pedestrian images

from different identities. Davis *et al.* [22] minimize the differential relative entropy between two Gaussian distributions to learn the Mahalanobis distance function and then utilize Bregman's method to optimize the convex model. Guillaumin *et al.* [23] learn a metric using the logistic discriminant, which enables the distances between positive pairs smaller than that of negative pairs. Similarly, Zheng *et al.* [17] propose the Probabilistic Relative Distance Comparison (PRDC) model which maximizes the probability of correct image pairs with a smaller distance than incorrect pairs. In [24], Koestinger et al. learn a distance metric with equivalence constraints through a simple and effective optimization strategy. Liao *et al.* [6] propose the Cross-view Quadratic Discriminant Analysis (XQDA) to learn a discriminative subspace and similarity measurement simultaneously. Wang *et al.* [25] propose the EquiDistance constrained Metric Learning (EquiDML), in which the intra-person distance is forced to be zero and meanwhile the inter-person distance is fixed to a constant positive value. Zhao *et al.* [18] present the Kernelized Random KISS (KRKISS) which could enhance the Gaussian distribution of samples by converting original features into kernelized features. However, the global metric learning methods learn the metric using the whole pedestrian images, and therefore they neglect the contextual information provided by sub-regions.

B. LOCAL METRIC LEARNING

In order to impose the contextual information on similarity measurement, the local metric learning methods [19]–[21], [26] are proposed to learn the distance metric in a local manner. Zhao *et al.* [19] apply adjacency constraints to build dense correspondence between image pairs in a patch level. Chen *et al.* [20] learn multiple sub-similarity measurements for regions based on the polynomial feature maps in order to take the spatial constraint into consideration. Li *et al.* [21] learn a discriminative region-to-point metric with positive regions generated from positive neighbors. Yang *et al.* [26] exploit the privileged information to learn a distance metric by building a locally adaptive decision rule. However, these local metric learning methods only learn the contextual information restricted to the corresponding sub-region pairs, but always ignore the contextual information provided by the non-corresponding sub-region pairs. Our work takes advantage of three kinds of sub-region pairs to fully learn the contextual information in a local manner.

III. CONTEXTUAL REGION-BASED METRIC LEARNING

In this section, we first define the distance between sub-region pairs of pedestrian images. Then, we utilize three kinds of sub-region pairs to learn the transformation matrix in the framework of mutual information. Finally, we introduce the corresponding optimization algorithm to obtain the transformation matrix and further present the nCRML to avoid the dimensionality curse.

A. DISTANCE BETWEEN SUB-REGION PAIRS

It could increase the model robustness against viewpoint and pose variations when incorporating the contextual information into the distance metric. In this paper, we present CRML to fully learn the contextual information in a local manner and learn the transformation matrix using three kinds of sub-region pairs. We firstly partition each pedestrian image into K sub-regions, where each sub-region is a horizontal strip. We define the distance between a sub-region pair as

$$d_k(x_k, z_k) = (x_k - z_k)^{\mathrm{T}} A (x_k - z_k), \tag{1}$$

where $x_k \in \mathbb{R}^{d\times 1}$ and $z_k \in \mathbb{R}^{d\times 1}$ are the feature vectors of the k-th sub-regions from an image pair. $A = MM^{\mathrm{T}}$ is a positive semi-definite matrix and $M \in \mathbb{R}^{d\times r}(r < d)$ is the transformation matrix satisfying $M^{\mathrm{T}}M = I$. $M^{\mathrm{T}}M = I$ could avoid the trivial and the rank one solution.

From Eq. (1), we can see that the key issue is how to obtain an effective transformation matrix. We introduce how to learn the transformation matrix M by maximizing mutual information in Section III-B, and correspondingly propose an optimization algorithm in Section III-C.

B. CONTEXTUAL REGION-BASED METRIC LEARNING

In order to learn the contextual information between sub-region pairs, we define three kinds of sub-region pairs, i.e., intra-region pairs, weak intra-region pairs and extra-region pairs, and the differences between the same kind of sub-region pairs. Note that the intra-region pairs are the corresponding sub-regions from the same pedestrian under different visual sensors as shown in Fig. 2 (a), and we define the difference between an intra-region pair as

$$\Delta I = M^{\mathrm{T}}(x_k^I - z_k^I), \tag{2}$$

where $x_k^I \in \mathbb{R}^{d\times 1}$ and $z_k^I \in \mathbb{R}^{d\times 1}$ represent the feature vectors of an intra-region pair. We regard the differences between intra-region pairs as the positive sub-region samples $\Delta I \in \mathbb{R}^{r\times 1}$ and assign the labels $l_I = 1$. The weak intra-region pairs denote the neighbouring non-corresponding sub-region pairs belonging to the same pedestrian under different camera sensors as shown in Fig. 2 (b). We utilize the parameter m to control the number of the weak intra-region pairs so that they could provide the contextual information for pedestrian images. The weak intra-region pairs can provide weakly supervised information and spatial structure information, which is beneficial to person re-identification in camera sensor networks. We express the difference between a weak intra-region pair as

$$\Delta W = M^{\mathrm{T}}(x_k^W - z_k^W), \tag{3}$$

where $x_k^W \in \mathbb{R}^{d\times 1}$ and $z_k^W \in \mathbb{R}^{d\times 1}$ indicate the feature vectors of a weak intra-region pair. The differences between weak intra-region pairs are treated as weak positive sub-region samples $\Delta W \in \mathbb{R}^{r\times 1}$ and their labels l_W are assigned to 2. Finally, we define the extra-region pairs as the corresponding and non-corresponding sub-region pairs belonging

to the different identities under different camera sensors as shown in Fig. 2 (c) and the difference between an extra-region pair is defined as

$$\Delta E = M^{\mathrm{T}}(x_k^E - z_k^E), \tag{4}$$

where $x_k^E \in \mathbb{R}^{d\times 1}$ and $z_k^E \in \mathbb{R}^{d\times 1}$ represent the feature vectors of an extra-region pair. We regard the differences between extra-region pairs as negative sub-region samples $\Delta E \in \mathbb{R}^{r\times 1}$ and assign the labels $l_E = -1$.

From the above formulations, we can see that directly optimizing the transformation matrix M in Eq. (1) is difficult because three kinds of sub-region pairs should be considered simultaneously. Hence, we optimize the transformation matrix M in a roundabout way. We expect to maximize the discriminative ability between different kinds of sub-region pairs by optimizing the transformation matrix M. Under the mutual information framework, the problem is defined as

$$\max_M I(v_k; l_v) + \varepsilon I(e_k; l_e), \tag{5}$$

where $v_k \in \{\Delta I, \Delta E\}$ is a set that contains positive and negative sub-region samples, and $l_v \in \{l_I, l_E\}$, i.e., $l_v \in \{1, -1\}$. $e_k \in \{\Delta W, \Delta E\}$ is also a set including weak positive and negative sub-region samples, and $l_e \in \{l_W, l_E\}$, i.e., $l_e \in \{2, -1\}$. I is the mutual information which measures the degree of dependence between two random variables. The larger the value of the mutual information is, the greater the dependence is between the features of sub-region samples and their labels. Moreover, ε is the coefficient to balance the two kinds of mutual information of Eq. (5).

Take the first item of Eq. (5) as an example. According to the chain rules of entropy, $I(v_k; l_v)$ can be written in terms of the differential entropy

$$\begin{aligned} I(v_k; l_v) &= H(v_k) - H(v_k|l_v) \\ &= H(v_k) - P(l_v = 1)H(\Delta I) \\ &\quad -P(l_v = -1)H(\Delta E), \end{aligned} \tag{6}$$

In Eq. (6), we approximate differential entropy $H(v_k)$ using the positive and negative sub-region samples. Assuming that the sub-region samples follow the Gaussian distribution, we reformulate $H(v_k)$ as

$$H(v_k) = \frac{1}{2} ln(2\pi e)^r det\, \Sigma_{l_v}, \tag{7}$$

where *det* represents the determinant of a matrix, and Σ_{l_v} is the covariance matrix of all positive and negative sub-region samples, which can be estimated from all ΔI and ΔE. Hence, we approximate the objective Eq. (6) using

$$I(v_k; l_v) = lndet\, \Sigma_{l_v} - \mu_1 lndet\, \Sigma_{l_I} - \rho lndet\, \Sigma_{l_E}, \tag{8}$$

where Σ_{l_I} and Σ_{l_E} are the covariance matrices of positive and negative sub-region samples, respectively, and μ_1 and ρ are the corresponding prior probabilities for positive and negative sub-region samples. In the same way, the second item of Eq. (5) is written as

$$I(e_k; l_e) = lndet\, \Sigma_{l_e} - \mu_2 lndet\, \Sigma_{l_W} - \rho lndet\, \Sigma_{l_E}, \tag{9}$$

where Σ_{l_e} is the covariance matrix of all weak positive and negative sub-region samples, Σ_{l_W} is the covariance matrix of weak positive sub-region samples, and μ_2 is the prior probability for weak positive sub-region samples. Note that there is a rank deficiency problem in any of covariance matrices in practice. Hence, we first determine the minimum rank η among all covariance matrices, and use the product of the top η large eigenvalues of each matrix to approximate its determinant. Furthermore, the number of negative sub-region samples is much larger than the number of positive and weak positive sub-region samples, which may increase the risk of over-fitting. To overcome the drawback, we simply set $\mu_1 = \mu_2 = \rho = 1/2$.

C. OPTIMIZATION ALGORITHM

In this section, we optimize the objective function Eq. (5) to obtain the transformation matrix M. For simplicity, we denote the objective function as $\Phi(M)$ in the following discussion.

We employ the greedy axis-rotating approach to search the transformation matrix M iteratively that maximizes $\Phi(M)$. Let $M(t-1)$ to be the estimation for M at iteration $t-1$. We seek matrix $Y(t) \in SO(d)$, so that the estimation at step t is $M(t) = Y(t)M(t-1)$, where $SO(d)$ is the d-dimensional special orthogonal group. Due to $SO(d)$ corresponding to a set of rotation operations in $\mathbb{R}^d$, the resulting $M(t)$ will be orthogonal matrix as well satisfying $M^{\mathrm{T}}M = I$. Essentially, we find $Y(t)$ to provide a steep ascent in $\Phi(M)$. According to the Lie algebra, the optimal rotation direction for M is found by

$$Y_\gamma = exp(\gamma\beta \sum_{p,q} \lambda_{p,q}(B_{p,q} - B_{q,p})), \tag{10}$$

where $2 \le p \le d$, $p+1 \le q \le d$, β is step length, and γ is the step number for searching optimal rotation direction. $B_{p,q}$ is a matrix whose (p, q)-th element is one and all others are zero. In addition, $\lambda_{p,q}$ is expressed as:

$$\lambda_{p,q} = \frac{\Delta\Phi_{p,q}}{(\sum_{p,q} \Delta\Phi_{p,q}^2)^{1/2}}, \tag{11}$$

The $\Delta\Phi_{p,q}$ is approximated by:

$$\Delta\Phi_{p,q} = [\Phi(Y_{p,q}M(t-1)) - \Phi(M(t-1))]/\alpha, \tag{12}$$

where

$$Y_{p,q} = exp(\alpha(B_{p,q} - B_{q,p})), \tag{13}$$

where α is a small positive number. The iterative algorithm terminates when $|M(t) - M(t-1)| \le \delta$. Here, δ is a small positive number. The above process is illustrated in Algorithm 1. More details about $SO(d)$ can be found in [27].

Based on the contextual information and sub-region pair learning strategy, there are two advantages of the proposed CRML. Firstly, to fully mine the contextual information, we consider three kinds of sub-region pairs, simultaneously. Secondly, we obtain the optimal transformation matrix M

Algorithm 1 Greedy Axis Rotation

Input: $M(0), \alpha > 0, \beta > 0, \Gamma > 0, \delta \geq 0$
Output: $M(t)$
Initialize $M(0)$ is initialized by the PCA method for $(x_k^I - z_k^I)$, $(x_k^W - z_k^W)$ and $(x_k^E - z_k^E)$;
While 1 **do**
If $|M(t) - M(t-1)| \leq \delta$, break;
1. For all the p, q calculate:
   1) $Y_{p,q}$ according to Eq. (13)
   2) $\Delta\Phi_{p,q}$ according to Eq. (12)
   3) $\lambda_{p,q}$ according to Eq. (11)
2. The optimal rotation direction Y_γ can be computed by Eq. (10), where $\gamma^* = \arg \max_{0 \leq \gamma \leq \Gamma} \Phi(Y_\gamma M(t-1))$
3. $Y(t) = Y_{\gamma^*}$
4. $M(t) = Y(t)M(t-1)$
end

by maximizing mutual information, which could increase the discrimination of sub-region representations.

Since the high dimensionality of sub-region feature may result in the curse of dimensionality, we first perform the dimensionality reduction using the RP [28] which has been proved to avoid information loss and meanwhile preserve the relative distances between samples in a low dimensionality space. In addition, to fuse similarity scores in different low dimensionality subspaces, we conduct n times of RP to map the sub-region features into randomly chosen low-dimensional subspaces. For each RP, the reduced features are utilized to learn one transformation matrix using Algorithm 1, and then we calculate a global similarity score between two pedestrian images using CCM. Finally, we add n similarity scores obtained by all RPs as the final similarity score. We term this extension as nCRML.

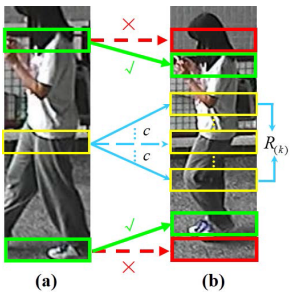

FIGURE 3. Context-Constrained Match between sub-regions.

IV. CONTEXT-CONSTRAINED MATCH

After obtaining the transformation matrix M and the local similarity scores between sub-region pairs, we should integrate them into a global similarity score for the subsequent matching. The traditional method obtains the global similarity score between two pedestrian images by directly adding the local similarity scores of corresponding sub-region pairs. However, in the process of integration, viewpoint and pose variations cause uncontrolled misalignment between images. For example in Fig. 3, the sub-region pairs marked by red are the corresponding sub-region pairs, but they are wrongly matched. The correct matching should be the sub-region pairs marked by green. Thus, corresponding sub-region pairs cannot be directly compared. In our method, a novel integration method named CCM is applied to tackle the misalignment problem. The proposed CCM searches the most matched sub-region for each sub-region in a certain range. Formally, we obtain the maximal similarity score of a sub-region by computing the minimum distance between sub-regions in the range of CCM, and the global similarity score between an image pair is given by

$$d(x, z) = \sum_k \min_{R_{(k)}} d_k(x_k, z_{R_{(k)}}). \tag{14}$$

$$R_{(k)} = \begin{cases} \{1, ..., k, ..., k+c\}, & 1 \leq k \leq c \\ \{k-c, ..., k, ..., k+c\} & c+1 \leq k \leq K-c \\ \{k-c, ..., k, ..., K\}, & K-c+1 \leq k \leq K \end{cases} \tag{15}$$

where k is the index of the k-th sub-region, K is the total number of sub-regions in one pedestrian image, c is the number of searching sub-regions, and $R_{(k)}$ is a set of sub-regions for matching as shown in Fig. 3. Here, $z_{R_{(k)}}$ represents one of feature vectors of sub-regions in $R_{(k)}$, and d_k indicates the Euclidean distance between x_k and $z_{R_{(k)}}$.

V. EXPERIMENTS

In this section, we evaluate our algorithm on three public datasets, the VIPeR dataset [29], the QUML GRID dataset [30] and the CUHK03 dataset [31]. We partition each pedestrian image into several 10×10 sub-regions with an overlapping step of 5 pixels and extract the LOMO feature [6] for each sub-region. Several metric learning methods with the same LOMO feature are compared, and the state-of-the-art results are compared on three public datasets.

A. EXPERIMENTS ON VIPeR

VIPeR [29] is a widely used dataset for person re-identification. It contains 632 pairs of pedestrian images captured by two camera sensors in outdoor environment. All images in this dataset are scaled to 128×48 pixels. Following the widely used experiment protocol, we randomly divide 632 pairs of images evenly, including 316 pairs for training and the remaining 316 pairs for testing. The process is repeated 10 times to obtain an average performance.

We compare our approach with several metric learning methods, including ITML [22], LMNN [16], KISSME [24], XQDA [6], EquiDML [25], KRKISS [18] SSM [34] and MPML [37], and show the comparison results in Table 1. It should be noticed that the compared metric learning methods also employ the same LOMO features as the proposed method. The proposed nCRML achieves the best result. We also compare our approach with the state-of-the-art results reported on the VIPeR dataset. Table 2 shows the results, where we can observe that nCRML_CCM achieves the best result in all situations. These results prove the effectiveness of our method.

TABLE 1. Comparison with different metric learning methods with the same LOMO feature on the VIPeR dataset (P=316). The identification rates (%) at rank-1, rank-10 and rank-20 are listed.

Method	rank-1	rank-10	rank-20
ITML [22]	24.65	63.04	78.39
LMNN [16]	29.43	73.51	84.91
KISSME [24]	34.81	77.22	86.71
XQDA [6]	40.00	80.51	91.08
EquiDML [25]	40.92	81.68	91.80
KRKISS [18]	41.7	85.1	-
SSM [34]	42.22	83.54	92.82
MPML [37]	44.72	84.27	93.58
CRML	46.07	85.56	94.65
nCRML	**50.83**	**88.75**	**96.76**

TABLE 2. Comparison with the state-of-the-art results on the VIPeR dataset (P=316). The identification rates (%) at rank-1, rank-10 and rank-20 are listed.

Method	rank-1	rank-10	rank-20
PRDC [17]	15.66	53.86	70.09
KISSME [24]	19.60	62.20	77.00
MtMCML [36]	28.83	75.82	88.51
eSDC_ocsvm [19]	26.74	62.37	76.36
NFST [33]	42.28	82.94	92.06
LDML+ [26]	45.19	85.66	93.99
CRML	46.07	85.56	94.65
CRML_CCM	47.12	86.79	95.53
nCRML	50.83	88.75	96.76
nCRML_CCM	**52.20**	**89.81**	**97.78**

B. EXPERIMENTS ON QMUL GRID

The QUML GRID [30] is a challenging person re-identification dataset. It consists of 1275 person images captured by 8 disjoint visual sensors. Among them, there are 250 pedestrian image pairs, and each image pair is captured by different camera sensors. In addition, the remaining 775 pedestrian images do not belong to the above-mentioned 250 identities. In the experiment, the training set contains 125 image pairs, and the test set consists of 125 image pairs and the 775 additional images. We apply the average of 10 random trials as the accuracy.

TABLE 3. Comparison with different metric learning methods with the same LOMO feature on the QMUL GRID dataset (G=900). The identification rates (%) at rank-1, rank-10 and rank-20 are listed.

Method	rank-1	rank-10	rank-20
LDML [23]	8.16	22.24	27.36
ITML [22]	9.44	27.04	35.20
KISSME [24]	10.64	31.60	43.20
LMNN [16]	10.80	34.24	45.76
IGLML-XQDA [32]	18.80	44.08	55.52
MPML [37]	22.64	49.36	59.44
CRML	24.41	53.89	65.54
nCRML	**27.92**	**57.24**	**68.86**

Performance comparison of different metric learning methods with the same LOMO feature is presented in Table 3. We can observe that the proposed nCRML achieves the highest accuracy. Compared with Table 1, it can be seen that the GRID dataset is more challenging than the VIPeR dataset due to the GRID dataset with 8 disjoint camera sensors, while the VIPeR dataset only having two camera sensors. The success of nCRML proves that fully learning the contextual information in different low-dimensional subspaces could effectively handle the poor image conditions and complex viewpoint changes.

TABLE 4. Comparison with the state-of-the-art results on QMUL GRID (G=900). The identification rates (%) at rank-1, rank-10 and rank-20 are listed.

Method	rank-1	rank-10	rank-20
PRDC [17]	9.68	32.96	44.32
MRank-PRDC [35]	11.12	35.76	46.56
MRank-RankSVM [35]	12.24	36.32	46.56
MtMCML [36]	14.08	45.84	59.84
XQDA [6]	16.56	41.84	52.40
GOG_RGB [7]	22.80	52.31	64.10
CRML	24.41	53.89	65.54
CRML_CCM	25.43	54.79	66.52
nCRML	27.92	57.24	68.86
nCRML_CCM	**28.98**	**58.34**	**69.90**

Table 4 shows the results compared with the state-of-the-art results. In Table 4, our methods show quite better performance than the other existing algorithms. Especially, nCRML_CCM obtains 28.98% rank-1 identification rate, achieving a new state of the art. Compared with the second best one GOG_RGB, the improvement by nCRML_CCM is +6.18%, +6.03%, and +5.80% at rank-1, rank-10, and rank-20, respectively. The promising result indicates that the proposed algorithm could learn a robust metric to deal with such challenges on the GRID dataset.

C. EXPERIMENTS ON CUHK03

CUHK03 [31] is a large person re-identification dataset. It contains 13,164 pedestrian images of 1,360 pedestrians collected by six camera sensors. Each pedestrian is captured by two disjoint camera sensors and has 9.6 images in average. The CUHK03 dataset provides both pedestrian images manually labelled and detected with a pedestrian detector, which brings some part missing and misalignments in a realistic scene.

According to the experimental setting in [24], [31], [34], the dataset is randomly divided into a training set of 1,160 pedestrians and a test set of 100 pedestrians. All experiments are performed with 20 random partitions, and the average results are presented with the single-shot setting. The rank-1 accuracy of the state-of-the-art results in both the manually labelled setting and the detected setting are shown in Table 5. The proposed nCRML_CCM is +6.93% and +8.08% higher than the second best method EquiDML+LOMO [25] with the manually labelled setting and the detected setting, respectively. It proves the effectiveness of our method once again.

D. ABLATION STUDY

In this subsection, we implement the ablation study to perform in-depth analysis of the proposed method and the results are presented in Table 6. CRML_WDL represents the trans-

TABLE 5. Comparison with the state-of-the-art results at rank-1 (%) on CUHK03 with both manually labelled setting and detected setting (P=100).

Method	labelled	detected
ITML [22]	5.53	5.14
LMNN [16]	7.29	6.25
eSDC [19]	8.76	7.68
LDML [23]	13.51	10.92
KISSME [24]	14.17	11.70
XQDA+LOMO [6]	52.20	46.25
SSM+LOMO [34]	52.50	49.05
EquiDML+LOMO [25]	61.40	55.12
CRML	62.87	57.65
CRML_CCM	64.03	58.72
nCRML	67.26	62.18
nCRML_CCM	**68.33**	**63.20**

TABLE 6. Comparison of the proposed method with other ablation methods.

Method	rank-1	rank-10	rank-20
CRML_WDL	25.06	61.97	71.33
CRML_W	33.23	73.43	83.64
CRML ($\varepsilon = 0$)	39.61	80.54	89.77
CRML	46.07	85.56	94.65
nCRML_PCA	46.27	85.68	94.76
nCRML (n = 1)	47.68	85.96	94.89
CRML_CCM	47.12	86.79	95.53
nCRML	50.83	88.75	96.76
nCRML_CCM	**52.20**	**89.81**	**97.78**

formation matrix is initialized by PCA without discriminative learning. CRML_W indicates that we only utilize the second term in Eq. (5) to learn the transformation matrix, that is only the weak intra-region pairs and the extra-region pairs are utilized. CRML ($\varepsilon = 0$) represents that we only employ the first term in Eq. (5) to learn the transformation matrix, i.e., only the intra-region pairs and the extra-region pairs are utilized. nCRML_PCA illustrates that we replace RP with PCA to conduct the dimension reduction operation. nCRML (n = 1) denotes we only employ one time of RP operation. From Table 6, six conclusions can be drawn.

Firstly, CRML_W is inferior to CRML due to the absence of intra-region pairs which could provide the discriminative information in the transformation matrix learning. Secondly, compared with CRML ($\varepsilon = 0$), CRML improves +6.46% at rank-1, +5.02% at rank-10, and +4.88% at rank-20. The results verify the effectiveness of the weakly supervised information provided by weak intra-region pairs. Thirdly, compared with CRML_WDL, CRML greatly improves identification rates by about 20% in all situations. The results indicate that our method could learn the discriminative transformation matrix by maximizing the mutual information. Fourthly, nCRML achieves better results than CRML. This proves that RPs could avoid the dimensionality curse and fully learn the contextual information in different low-dimensional subspaces. Fifthly, nCRML (n = 1) achieves better performance than nCRML_PCA. This illustrates the RP operation is more effective than PCA. Finally, CRML and nCRML are improved with the help of CCM. It is because CCM could tackle the misalignment problem and seek the optimal match among the neighboring sub-regions in the process of similarity calculation.

E. INFLUENCE OF PARAMETERS

In this subsection, we analyze several important parameters on the VIPeR dataset and our experimental results have shown that the conclusions can be generalized to QMUL GRID and CUHK03 as well. To understand the influence of the coefficient ε in Eq. (5), we conduct experiments with different values of ε on VIPeR following the previous setting. Table 7 presents the rank-1 identification rate with different ε. It can be seen that when $\varepsilon = 0.1$, the proposed method achieves the best result.

TABLE 7. The rank-1 identification rate (%) on VIPeR with different ε.

ε	0	0.05	0.10	0.15	0.20
rank-1	39.61	43.53	**46.07**	44.20	42.74

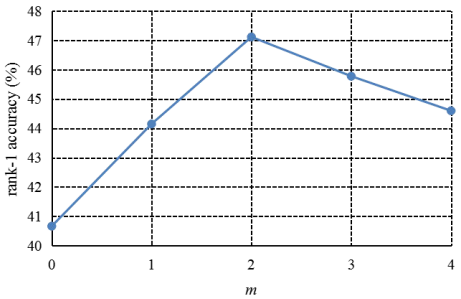

FIGURE 4. The rank-1 identification rate of the proposed CRML_CCM on VIPeR with different *m*.

The parameter m denotes the number of non-corresponding sub-region pairs belonging to the same pedestrian. The parameter m influences the number of weak intra-region pairs which could provide the contextual information for pedestrian images. The results with different m are shown in Fig. 4. When $m = 2$, we obtain the best result. This is because pedestrians usually have a large misalignment, and these misaligned pedestrians generally offset one or two sub-regions. When m is too large, it may bring in some noise so that the performance is weakened. When m is too small, it does not provide the enough contextual information in the matrix learning process.

The parameter c is the number of searching sub-regions in CCM. CCM could tackle the misalignment problem and search the most matched sub-region for each sub-region in a certain range. The results with different c are shown in Fig. 5. When c is too large, it may match the similar backgrounds, and when c is too small, CCM could not search the most matched sub-region due to the small search range. The proposed method achieves the best result when c is equal to 2.

In order to represent the pedestrian images, we extract the LOMO features for each sub-region. Then, RP is utilized to reduce the dimensionality to d. We analyze the number of

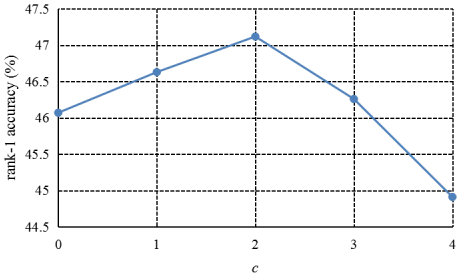

FIGURE 5. The rank-1 identification rate of the proposed CRML_CCM on VIPeR with different c.

RPs n and the reduced dimensionality d with the proposed nCRML on VIPeR. We set n to $\{0, 1, 2, 3, 4, 5\}$ ($d = s/3$). The rank-1 accuracies are $\{46.07\%, 47.68\%, 49.45\%, 50.83\%, 49.72\%, 47.83\%\}$. d is chosen from $\{2s/3, s/2, s/3, s/4, s/5\}$ ($n = 3$), and the rank-1 accuracies are $\{49.05\%, 49.34\%, 50.83\%, 48.71\%, 47.92\%\}$. Hence, $n = 3$ and $d = s/3$ yield superior accuracy.

TABLE 8. The rank-1 identification rate of the proposed CRML with different K on VIPeR.

K	1	2	3	5	7
rank-1	34.65	43.68	44.53	**46.07**	45.76

We comprehensively analyze the influence of the parameter K and the results are shown in Table 8 where we can see that the proposed CRML achieves the best result when $K = 5$. In addition, $K = 1$ means learning a global metric where the weak intra-region pairs and the contextual information are lost. So, it obtains a worse result than all other situations. When $K = 2$, it is the minimum value to ensure the existence of weak intra-region pairs, but the performance is significantly better than that of $K = 1$. Based on the above analysis, we can see that the weak intra-region pairs could learn the contextual information which is very important for the performance of person re-identification.

VI. CONCLUSION

In this paper, we have proposed the CRML to overcome appearance variances of pedestrian images in camera sensor networks. The CRML learns a discriminative transformation matrix using three kinds of sub-region pairs with the help of mutual information. We then employ the greedy axis rotation algorithm to optimize the transformation matrix. Furthermore, we have presented the nCRML to learn the contextual information in different low-dimensional subspaces and avoid the dimensionality curse. In the process of integration, we apply CCM to overcome the misalignment problem and locate the sub-region pairs with the maximum similarity in the neighborhood. Comprehensive experiments on three datasets have verified that the proposed method performs better than the state-of-the-art methods for person re-identification in camera sensor networks.

REFERENCES

[1] M. A. K. Azrag, T. A. A. Kadir, and A. S. Jaber, "Segment particle swarm optimization adoption for large-scale kinetic parameter identification of escherichia coli metabolic network model," *IEEE Access*, vol. 6, pp. 78622–78639, 2018.
[2] T. Si, Z. Zhang, and S. Liu, "Discrimination-aware integration for person re-identification in camera networks," *IEEE Access*, vol. 7, pp. 33107–33114, 2019.
[3] X. Liu, M. Jia, X. Zhang, and W. Lu, "A novel multichannel Internet of things based on dynamic spectrum sharing in 5G communication," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 5962–5970, Aug. 2019.
[4] W. Zheng, S. Gong, and T. Xiang, "Group association: Assisting re-identification by visual context," in *Person Re-Identification*, Jan. 2014, pp. 183–201.
[5] T. D'Orazio and G. Cicirelli, "People re-identification and tracking from multiple cameras: A review," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep./Oct. 2012, pp. 1601–1604.
[6] S. Liao, Y. Hu, X. Zhu, and S. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2197–2206.
[7] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1363–1372.
[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[9] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.
[10] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1367–1376.
[11] Z. Zhang, T. Si, and S. Liu, "Integration convolutional neural network for person re-identification in camera networks," *IEEE Access*, vol. 6, pp. 36887–36896, 2018.
[12] A. Li, D. Chen, Z. Wu, G. Sun, and K. Lin, "Self-supervised sparse coding scheme for image classification based on low rank representation," *PloS ONE*, vol. 13, no. 6, 2018, Art. no. e0199141.
[13] Z. Zhang, H. Zhang, and S. Liu, "Coarse-fine convolutional neural network for person re-identification in camera sensor networks," *IEEE Access*, vol. 7, pp. 65186–65194, 2019.
[14] Z. Zhang, D. Li, S. Liu, B. Xiao, and X. Cao, "Cross-domain ground-based cloud classification based on transfer of local features and discriminative metric learning," *Remote Sens.*, vol. 10, no. 1, p. 8, 2017.
[15] A. Li, Z. Wu, H. Lu, D. Chen, and G. Sun, "Collaborative self-regression method with nonlinear feature based on multi-task learning for image classification," *IEEE Access*, vol. 6, pp. 43513–43525, 2018.
[16] K. Q. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. 18th Int. Conf. Neural Inf. Process. Syst.*, Dec. 2005, pp. 1473–1480.
[17] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 649–656.
[18] C. Zhao, Y. Chen, X. Wang, W. K. Wong, D. Miao, and J. Lei, "Kernelized random KISS metric learning for person re-identification," *Neurocomputing*, vol. 275, pp. 403–417, Jan. 2018.
[19] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3586–3593.
[20] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1268–1277.
[21] J. Li, A. J. Ma, and P. C. Yuen, "Semi-supervised region metric learning for person re-identification," *Int. J. Comput. Vis.*, vol. 126, no. 8, pp. 855–874, 2018.
[22] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 209–216.
[23] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 498–505.
[24] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2288–2295.

[25] J. Wang, Z. Wang, C. Laing, C. Gao, and N. Sang, "Equidistance constrained metric learning for person re-identification," *Pattern Recognit.*, vol. 74, pp. 38–51, Feb. 2018.
[26] X. Yang, M. Wang, and D. Tao, "Person re-identification with metric learning using privileged information," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 791–805, Feb. 2018.
[27] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2855–2862.
[28] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.
[29] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveill. (PETS)*, Oct. 2007, pp. 1–7.
[30] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1988–1995.
[31] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.
[32] J. Zhang and X. Zhao, "Integrated global-local metric learning for person re-identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 596–604.
[33] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1239–1248.
[34] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3356–3365.
[35] C. C. Loy, C. Liu, and S. Gong, "Person re-identification by manifold ranking," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 3567–3571.
[36] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3656–3670, Aug. 2014.
[37] H.-M. Hu, W. Fang, B. Li, and Q. Tian, "An adaptive multi-projection metric learning for person re-identification across non-overlapping cameras," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2809–2821, Sep. 2019.

ZHONG ZHANG (M'14–SM'19) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently an Associate Professor with Tianjin Normal University, Tianjin, China. He has published more than 100 articles in international journals and conferences, such as Pattern Recognition, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, *Signal Processing* (Elsevier), CVPR, ICPR, and ICIP. His research interests include computer vision, signal processing, pattern recognition, and deep learning.

TONGZHEN SI is currently pursuing the master's degree with Tianjin Normal University, Tianjin, China. His research interests include computer vision, sensor networks, person re-identification, and deep learning.

MEIYAN HUANG is currently pursuing the master's degree with Tianjin Normal University, Tianjin, China. Her research interests include sensor networks, person re-identification, and deep learning.

SHUANG LIU (M'18–SM'19) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. She is currently an Associate Professor with Tianjin Normal University, Tianjin, China. Her research interests include computer vision and deep learning.