Matching People Across Surveillance Cameras

Raphael Prates
Smart Sense Laboratory
Department of Computer Science
Universidade Federal de Minas Gerais, Brazil
prates@dcc.ufmg.br

William Robson Schwartz
Smart Sense Laboratory
Department of Computer Science
Universidade Federal de Minas Gerais, Brazil
william@dcc.ufmg.br

***Abstract*—This work addresses the person re-identification problem, which consists on matching images of individuals captured by multiple non-overlapping surveillance cameras. Works from literature tackle this problem extracting characteristics that are robust to different poses and illumination conditions, and matching functions to assign the correct identity for individuals. More scalable and accurate matching functions is the focus of this work[1]. Specifically, we propose two matching methods: the Kernel MBPLS and the Kernel X-CRC. The Kernel MBPLS is a nonlinear regression model that is scalable with respect to the number of cameras and allows the inclusion of additional labelled information (e.g., attributes). Differently, the Kernel X-CRC is a nonlinear and multitask matching function that can be used jointly with subspace learning approaches to boost the matching rates. We present an extensive experimental evaluation of both approaches in four datasets (VIPeR, PRID450S, WARD and Market-1501). Experimental results demonstrate that the Kernel MBPLS and the Kernel X-CRC outperforms approaches from literature. Furthermore, we show that the Kernel X-CRC can be successfully applied in large-scale datasets.**

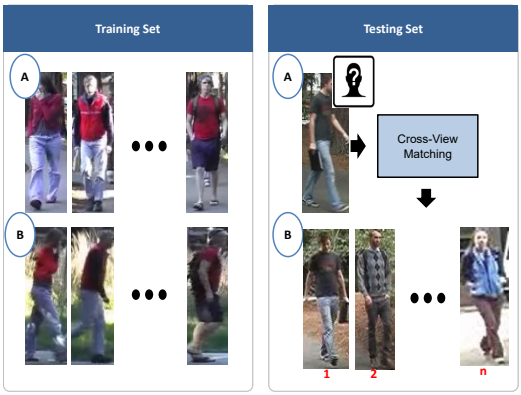

Fig. 1. Training and test sets. In the training set the better configuration of feature descriptors and cross-view functions are learned using labeled individuals in a pair of cameras. Then, in the test set, these functions are deployed using a disjoint subset of individuals.

I. Introduction

Person re-identification (Re-ID) consists in establishing correspondences between pedestrian images captured by multiple non-overlapping cameras. The goal in a person re-identification system is to look for previous occurrences of a probe image in the gallery-sets of all cameras connected in a surveillance camera network. Re-ID is important to provide a broad view of the people's behavior to the security personnel and has attracted the researches attention in the past years [2].

Formally, let us consider $\mathbf{p}$ as the probe image and $\mathbf{G}$ as the gallery-set composed of N individuals with known identities, where $\mathbf{g}_i \in \mathbf{G}$ corresponds to the ith subject in the gallery-set. Then, we can determine $\mathbf{p}$ identity (id) as

$$id = \arg\max_i sim(\Psi(\mathbf{p}), \Psi(\mathbf{g}_i)), \quad (1)$$

where Ψ corresponds to a feature extraction function and $sim(\cdot,\cdot)$ is some cross-view matching function. In a supervised setting, the cross-view matching and feature extraction functions are learned using a training set, which consists of labeled individuals captured by different surveillance cameras. Then, these functions are deployed in a test set whose identities are disjoint from the training set, as illustrated in Figure 1.

Due to the ambiguity between individuals and the computational cost required to match pedestrian images in the entire surveillance network, the person re-identification is a challenging task. In fact, the same subject when captured by different surveillance cameras might look more dissimilar than different subjects as consequence of the variations in the camera capture conditions (e.g., camera viewpoint and illumination) and the person's pose [3]. In addition, dozens of cameras would be necessary to monitor just hundreds of individuals in medium-sized environment. Therefore, the scalability of the person re-identification system with respect of the number of cameras is an important and still overlooked issue for real-world applications.

The person re-identification literature have addressed the ambiguity between individuals with novel features descriptors [4]–[6], which are more robust to the different camera conditions, and with camera pairwise matching models [7]–[13]. The latter is an interesting solution since these models capture specific transitions of feature descriptors for a single pair of cameras, such as the variations due to different camera viewpoints and illumination conditions. One widely used approach is the cross-view quadratic discriminant analysis (XQDA) [4] that learns a common and low-dimensional representation and a matching function.

Despite outperforming results, the pairwise matching of probe and gallery cameras is not suitable for a real-world scenario. In a surveillance camera network with c connected cameras, we can have $c(c-1)/2$ pairs of camera (see Figure 2),

[1]This work corresponds to a PhD thesis [1]

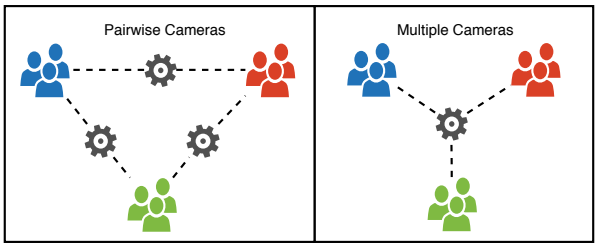

Fig. 2. Pairwise and multiple cameras cross-view matching models, which are represented using gear icons. Notice that the pairwise learns a model for each pair of cameras. Differently, the multiple cameras learns a single for the entire camera network.

which is not scalable for most of the real-world applications with thousands of surveillance cameras (e.g. an airport).

In this work, we tackle both problems: the scalability with respect to the number of surveillance cameras and the ambiguity between individuals. Regarding the scalability problem, we proposed the Kernel Multiblock Partial Least Squares (Kernel MBPLS) that considers multiple sources of data when projecting the data onto a low-dimensional subspace that correlates the input data with the responses. To reduce the ambiguity between individuals, we propose the Kernel Cross-view Collaborative Representation based Classification (Kernel X-CRC) that collaboratively represents test samples as a nonlinear combination of training samples.

Kernel HPCA [7] is the unique approach in literature that also addresses the scalability problem in person re-identification. However, while the Kernel HPCA only learns a common subspace, the Kernel MBPLS also performs a regression that imposes a better separation of the data in the learned subspace. Some works also investigate the person re-identification problem using sparse or collaborative representations [6], [14]–[17]. *Kernel X-CRC* has some key advantages when compared to the these methods. For instance, *Kernel X-CRC* is a general method that does not assume a block structure in the coefficients representation as required in [15]. Differently from dictionary learning-based approaches [16], [17], our work represents probe and gallery images using training samples, which avoids solving costly optimization problems without sacrificing the matching rate. More importantly, different from previous works [6], [15]–[17], we efficiently model the strong nonlinear transition of features between cameras achieving an analytical solution.

Experimental results in the WARD, which is a multiple cameras dataset, demonstrate that Kernel MBPLS is not only scalable but also surpasses subspace learning approaches from literature. Similarly, experimental results show that the proposed Kernel X-CRC outperforms approaches from literature when considering VIPeR and PRID450S datasets. Besides, we demonstrate that the Kernel X-CRC can be successfully adjusted to work in Market-1501, a large-scale dataset.

The predominant contributions of this work for the person re-identification problem are: (i) a common subspace learning that combines scalability with respect to the number of cameras with higher matching rates, (ii) a nonlinear and multitask matching function that boosts the matching rates of subspace learning approaches from literature and (iii) an extensive experimental evaluation and discussion of the employed strategies in four person re-identification datasets.

The methods described in this work correspond to the main achievements during the doctorate research [18], [19]. Besides these approaches, we also proposed additional methods for subspace learning that maximizes the covariance between a pair of cameras (Kernel PLS [20]) or multiple cameras (Kernel HPCA [21]). We also tackled the person re-identification problem by comparing probe and gallery with a fixed subset of individuals [22], using inverted indexing lists [23] and as a ranking aggregation problem [24].

II. Kernel Multiblock Partial Least Squares

In this section, we describe the proposed Kernel MBPLS that nonlinearly relates data blocks (i.e., images from different surveillance cameras) and responses in a learned latent space. Despite its simplicity, the proposed method captures high-order correlations between input variables and responses. Furthermore, we show how to compute the regression coefficients of the Kernel MBPLS efficiently.

The overall idea of the proposed Kernel Multiblock PLS consists on computing scores for each block (s_i) of data, combining them in a super block (S) and, then, performing a regression between the super block and the responsive matrix Y. In the following paragraphs, we give a more detailed description of the proposed Kernel MBPLS that is presented in Algorithm 1. This algorithm is a modification of the classic NIPALS algorithm [25] that decomposes a matrix X into orthonormal scores (a) and loadings (z) as $X = az^\top$, such that we can represent a and z as

$$a = Xz \quad \text{and} \quad z = X^\top a. \tag{2}$$

Therefore, we have that we can update the scores as $a = XX^\top a$. Then, using a nonlinear transformation in matrix X, we obtain Φ and using the "kernel trick" to substitute the cross-product $\Phi\Phi^\top$ by the *kernel Gram matrix* $K \in \mathbb{R}^{n\times n}$, we have that $a = Ka$. From lines 4 to 7 of Algorithm 1, we can notice that the block-scores (s_p and s_g) are updated in this manner to construct the super block S. Then, in lines 8 and 9, we represent the super block as $S = uw^\top$ and, in lines 10 and 11, we represent the response matrix as $Y = uq^\top$. Notice that as we use the same scores for both S and Y, we correlate them in the learned low-dimensional subspace. Finally, after the convergence, the input matrices are deflated (lines 13 and 14) and the process continues until the number of factors (f) has been reached.

Finally, we can compute the regression response as

$$\hat{\mathbf{y}}_p = \mathbf{k}_j^p \mathbf{T}(\mathbf{T}^\top \mathbf{K}_p \mathbf{T})^{-1}\mathbf{T}^\top \mathbf{Y}, \tag{3}$$

where $\mathbf{k}_j^p$ is the kernel representation of the *jth* probe image ($\mathbf{p}_j$) and $\mathbf{T} \in \mathbb{R}^{n\times f}$ is a result of the concatenation of the scores $\mathbf{t}_i \in \mathbb{R}^{n\times 1}$ computed for each factor in Algorithm 1. Similarly, we can compute the regression responses for a gallery sample ($\hat{\mathbf{y}}_g$) using $\mathbf{K}_p$ and the previous equation.

Algorithm 1: Kernel Multiblock PLS (Kernel MBPLS)

input : $\mathbf{K}_p$, $\mathbf{K}_g$, $\mathbf{Y}$ matrices and the integer (f)

randomly initialize $\mathbf{t}$ and $\mathbf{t}_0$
for *i=1* **to** f **do**
  while $\|\boldsymbol{t} - \boldsymbol{t}_0\| > \varepsilon$ **do**
    $\mathbf{t}_0 \leftarrow \mathbf{t}$
    $\mathbf{s}_p = \mathbf{K}_p\mathbf{t} \quad \mathbf{s}_p \leftarrow \frac{\mathbf{s}_p}{\|\mathbf{s}_p\|}$
    $\mathbf{s}_g = \mathbf{K}_g\mathbf{t} \quad \mathbf{s}_g \leftarrow \frac{\mathbf{s}_g}{\|\mathbf{s}_g\|}$
    $\mathbf{S} \leftarrow [s_p, s_g]$
    $\mathbf{w} = \mathbf{S}^\top\mathbf{t}$
    $\mathbf{u} = \mathbf{S}\mathbf{w} \quad \mathbf{u} \leftarrow \frac{\mathbf{u}}{\|\mathbf{u}\|}$
    $\mathbf{q} = \mathbf{Y}^\top\mathbf{u}$
    $\mathbf{t} = \mathbf{Y}\mathbf{q}/\mathbf{q}^\top\mathbf{q}$
  end
  $\mathbf{K}_p = \mathbf{K}_p - \mathbf{t}\mathbf{t}^\top\mathbf{K}_p, \quad \mathbf{K}_g = \mathbf{K}_g - \mathbf{t}\mathbf{t}^\top\mathbf{K}_g,$
  $\mathbf{Y} = \mathbf{Y} - \mathbf{t}\mathbf{t}^\top\mathbf{Y}$
end

In this work, we assume that when comparing test samples measured from data blocks i and j, we can use the regression responses $\hat{\mathbf{y}}_i$ and $\hat{\mathbf{y}}_j$ as discriminative signatures. Specifically, the matching between i and j ($s(i,j)$) is computed using the cosine similarity between $\hat{\mathbf{y}}_i$ and $\hat{\mathbf{y}}_j$ as

$$s(i,j) = \hat{\mathbf{y}}_i\hat{\mathbf{y}}_j / \parallel \hat{\mathbf{y}}_i \parallel\parallel \hat{\mathbf{y}}_j \parallel . \tag{4}$$

III. Kernel Cross-view Collaborative Representation-based Classification

In this section, we present the Kernel Cross-View Collaborative Representation based Classification (Kernel X-CRC) that efficiently represents each pair probe $\mathbf{p}$ and gallery $\mathbf{g}$ images collaboratively using its camera-view specific training samples $\mathbf{X}_p$ and $\mathbf{X}_g$, respectively.

Considering as related tasks the representation of probe and gallery images using training images from their respective cameras, the proposed Kernel X-CRC model simultaneously estimates $\boldsymbol{\alpha}_g$ and $\boldsymbol{\alpha}_p$ in a multi-task collaborative representation framework. Thus, we aim at estimating the most similar coding vectors $\boldsymbol{\alpha}_g$ and $\boldsymbol{\alpha}_p$ that simultaneously describe probe and gallery subjects. To compute these coding vectors, we introduce a *similarity term* $\parallel \boldsymbol{\alpha}_p - \boldsymbol{\alpha}_g \parallel_2^2$ that balances representativeness and similarity, as illustrated in Figure 3.

The proposed Kernel X-CRC model results in the following optimization problem

$$\begin{aligned}\min_{\boldsymbol{\alpha}_g,\boldsymbol{\alpha}_p} \parallel \phi(\mathbf{p}) - \boldsymbol{\Phi}_p\boldsymbol{\alpha}_p \parallel_2^2 + \parallel \phi(\mathbf{g}) - \boldsymbol{\Phi}_g\boldsymbol{\alpha}_g \parallel_2^2 \\ +\lambda \parallel \boldsymbol{\alpha}_p \parallel_2^2 +\lambda \parallel \boldsymbol{\alpha}_g \parallel_2^2 +\tau \parallel \boldsymbol{\alpha}_p - \boldsymbol{\alpha}_g \parallel_2^2,\end{aligned} \tag{5}$$

where $\phi(.)$ is a nonlinear function and, $\boldsymbol{\Phi}_g$ and $\boldsymbol{\Phi}_p$ are resulting nonlinear mapping of $\mathbf{X}_g$ and $\mathbf{X}_p$, respectively. Analytically deriving Equation 5 with respect to $\boldsymbol{\alpha}_p$ and $\boldsymbol{\alpha}_g$, we obtain

$$\begin{aligned}\boldsymbol{\alpha}_p = \mathbf{A}_p^{-1}\boldsymbol{\alpha}_g + \mathbf{A}_p^{-1}\boldsymbol{\Phi}_p^\top\phi(\mathbf{p}) \text{ and} \\ \boldsymbol{\alpha}_g = \mathbf{A}_g^{-1}\boldsymbol{\alpha}_p + \mathbf{A}_g^{-1}\boldsymbol{\Phi}_g^\top\phi(\mathbf{g}),\end{aligned} \tag{6}$$

where projections matrices $\mathbf{A}_p$ and $\mathbf{A}_g$ are given by

$$\begin{aligned}\mathbf{A}_p = \boldsymbol{\Phi}_p^\top\boldsymbol{\Phi}_p + (\lambda+\tau)\mathbf{I} \text{ and} \\ \mathbf{A}_g = \boldsymbol{\Phi}_g^\top\boldsymbol{\Phi}_g + (\lambda+\tau)\mathbf{I}.\end{aligned} \tag{7}$$

Note that Equations in 6 are interdependent. Therefore, replacing $\boldsymbol{\alpha}_g$ and isolating $\boldsymbol{\alpha}_p$, we obtain

$$\boldsymbol{\alpha}_p = \tau\mathbf{Q}^{-1}\mathbf{A}_p^{-1}\mathbf{A}_g^{-1}\boldsymbol{\Phi}_g^\top\phi(\mathbf{g}) + \mathbf{Q}^{-1}\mathbf{A}_p^{-1}\boldsymbol{\Phi}_p^\top\phi(\mathbf{p}) \tag{8}$$

with projection matrix $\mathbf{Q}$ corresponding to

$$\mathbf{Q} = \mathbf{I} - \tau^2\mathbf{A}_p^{-1}\mathbf{A}_g^{-1}. \tag{9}$$

Similarly, we can compute the coding vector $\boldsymbol{\alpha}_g$ as

$$\boldsymbol{\alpha}_g = \tau\mathbf{W}^{-1}\mathbf{A}_g^{-1}\mathbf{A}_p^{-1}\boldsymbol{\Phi}_p^\top\phi(\mathbf{p}) + \mathbf{W}^{-1}\mathbf{A}_g^{-1}\boldsymbol{\Phi}_g^\top\phi(\mathbf{g}) \tag{10}$$

with $\mathbf{W}$ computed as

$$\mathbf{W} = \mathbf{I} - \tau^2\mathbf{A}_g^{-1}\mathbf{A}_p^{-1}. \tag{11}$$

To avoid explicitly mapping of data to a high-dimensional space, we can use the "kernel trick" substituting cross-product $\boldsymbol{\Phi}_g^\top\boldsymbol{\Phi}_g$ and $\boldsymbol{\Phi}_p^\top\boldsymbol{\Phi}_p$ by the *kernel Gram matrix* $\mathbf{K}_g$ and $\mathbf{K}_p \in \mathbb{R}^{n\times n}$, respectively. Furthermore, we replace $\boldsymbol{\Phi}_g^\top\phi(\mathbf{g})$ and $\boldsymbol{\Phi}_p^\top\phi(\mathbf{p})$ by their respective row vectors $\mathbf{k}^g$ and $\mathbf{k}^p$.

Algorithm 2: *Kernel X-CRC.*

input : Kernel matrices ($\mathbf{K}_g$ and $\mathbf{K}_p$)
output: Ranking list of gallery images $\mathbf{R}$

Compute $\boldsymbol{A}_g$, $\boldsymbol{A}_p$, $\boldsymbol{Q}$ and $\boldsymbol{W}$ matrices
$\boldsymbol{\beta}_g^g \leftarrow \mathbf{W}^{-1}\mathbf{P}_g^{-1}$, $\boldsymbol{\beta}_g^p \leftarrow \tau\mathbf{W}^{-1}\mathbf{A}_g^{-1}\mathbf{A}_p^{-1}$
$\boldsymbol{\beta}_p^p \leftarrow \mathbf{Q}^{-1}\mathbf{A}_p^{-1}$, $\boldsymbol{\beta}_p^g \leftarrow \tau\mathbf{Q}^{-1}\mathbf{A}_p^{-1}\mathbf{A}_g^{-1}$
for $\boldsymbol{p}_j \in \boldsymbol{P}$ **do**
  for $\boldsymbol{x}_i \in \boldsymbol{X}$ **do**
    $\boldsymbol{\alpha}_x \leftarrow \boldsymbol{\beta}_g^g\mathbf{k}_i^g + \boldsymbol{\beta}_g^p\mathbf{k}_j^p$, $\boldsymbol{\alpha}_p \leftarrow \boldsymbol{\beta}_p^g\mathbf{k}_i^g + \boldsymbol{\beta}_p^p\mathbf{k}_j^p$
    $sim(i) \leftarrow \frac{\boldsymbol{\alpha}_g^\top\boldsymbol{\alpha}_p}{\|\boldsymbol{\alpha}_g\|\|\boldsymbol{\alpha}_p\|}$
  end
  $\mathbf{R}_j \leftarrow sort(sim, descend)$
end
return $\boldsymbol{R}$

Due to the multi-task learning framework, a pair of probe ($\mathbf{p}$) and gallery images ($\mathbf{g}$) will compute $\boldsymbol{\alpha}_p$ and $\boldsymbol{\alpha}_g$ that balances the representativeness in each camera with the similarity between coding vectors. This balance will only result in a similar coding vector if $\mathbf{p}$ corresponds to the respective gallery image of $\mathbf{g}$. Therefore, we match probe and gallery using the cosine similarity between $\boldsymbol{\alpha}_p$ and $\boldsymbol{\alpha}_g$, as described in Algorithm 2.

IV. Experiments

In this section, we perform a comprehensive evaluation of the proposed methods assessing the influence of different parameters in the obtained experimental results and providing a broad comparison with state-of-the-art approaches in two camera pairwise and single-shot datasets (VIPeR [26], PRID450S [27]) and two multiple cameras and multi-shot datasets (Market-1501 [28] and WARD [29]). In the following

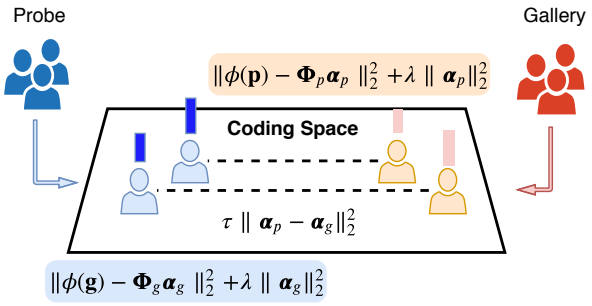

Fig. 3. Schematic representation of the proposed *Kernel X-CRC* method. For a pair of probe and gallery images, we compute the coding vectors (α_p and α_g) that collaboratively represent them using training samples captured at their respective camera. A similarity term balances the trade-off between representativeness and similarity.

Fig. 4. Images of the same individual (columns) captured by different cameras (rows) in the employed person re-identification datasets. From left to right: PRID450S, VIPER, Market-1501 and WARD datasets.

paragraphs, we first present the four datasets evaluated and the training and evaluation protocols. Then, in the remaining sections, we present the ablation study and state-of-the-art comparisons for Kernel MBPLS (Section IV-A) and Kernel X-CRC approaches (Section IV-B).

1) Datasets: VIPER contains 632 labelled pairs captured by two outdoor cameras in very challenging conditions. PRID450S has 450 labelled image pairs capture by two surveillance cameras. Differently, Market-1501 contains 32,668 images of 1,501 individuals captured from six surveillance cameras, where the same individual can have multiple images at the same camera (i.e., multi-shot). Similarly, the WARD is a multi-shot dataset consisting of 4,786 images of 70 individuals from three surveillance cameras. Figure 4 shows examples of individuals from the four datasets evaluated.

2) Training and Evaluation Protocols: To set the parameters of the proposed methods, we use the common strategy in the person re-identification literature of using a validation and test set composed by ten random partitions of images in training and probe/gallery subsets with equal number of identities [4], [5]. When presenting results in tables, we report the mean *rank-k* matching rate in ten distinct

TABLE I
ABLATION STUDY ON THE WARD DATASET.

Features		Labels		Scenario		
GoG	LOMO	ID	Attr	Multi	Pair	*Rank-1*
✓		✓		✓		76.5
	✓	✓		✓		69.8
✓		✓			✓	78.8
✓			✓	✓		67.4
✓		✓	✓	✓		81.3

partitions of the data, which consists of the percentage of individuals correctly identified when considering the *top-k* ranking positions, a widely employed metric to compare Re-ID approaches.

3) Feature Descriptors: As designing a feature descriptor is not the focus of this work, we evaluate the proposed methods using two widely used descriptors: the Hierarchical Gaussian descriptor (GoG) [5] and the LOMO [4]. LOMO consists of color and texture descriptors extracted from multiple scales and horizontal stripes that are summarized using an horizontal pooling operation. GoG is a hierarchical Gaussian distribution of covariance and mean statistics that captures texture and color information of image patches.

A. Kernel MBPLS

1) Ablation Study: In this section, we present the experimental results using the proposed Kernel MBPLS with different strategies and using the WARD dataset. Specifically, we evaluate state-of-the-art feature descriptors as the GoG [5] and LOMO [4], labelled information as identity and attributes, and the multiple cameras (multi) and camera pairwise (pair) scenarios. To consider the attributes information, we manually labelled 24 attributes for each individual in the WARD dataset.

Table I presents the obtained mean *rank-1* for all camera pairs. For instance, the first row shows the *rank-1* when using the GoG feature descriptor, the identity information to construct the matrix Y and a multiple cameras setting, which means that a unique model is learned for the entire system. Based on these results, it is possible to conclude that the highest results are obtained when considering identity and attribute labels due to its complementary (i.e., line 5). Besides, the GoG descriptor results in a large gain when compared to the LOMO (i.e., line 1 and 2). Finally, the experiment using a pairwise approach - one model learned for each camera pair - presents a small improvement when compared to the same configuration using a multiple camera approach (i.e., lines 1 and 3) as a consequence of learning subtle variations that occur in a camera pair. However, it has the huge disadvantage of growing quadratically with the number of cameras.

2) State-of-the-art Comparisons: Table II presents the obtained *rank-1* results for the different camera pairs when using the proposed Kernel MBPLS and other subspace learning approaches from the literature, such as the Kernel CCA (KCCA), CCA and the Kernel HPCA. These methods are divided in

pairwise and multiple cameras models, while the first learns a model for each pair of cameras, the latter learns a single model for the entire camera system. Kernel MBPLS presents superior results, which can be related to the nonlinear regression that better separates samples in the learned subspace. Thus, the Kernel MBPLS is scalable and presents higher matching rates than similar approaches from literature.

TABLE II
MEAN *Rank-1* MATCHING RATE ON WARD DATASET.

Models	Methods	Probe/Gallery		
		A/B	A/C	B/C
Pairwise	CCA	80.3	62.9	70.0
	KCCA	82.6	65.4	70.9
Multiple	Ker. HPCA	81.1	64.0	71.7
	Ker. MBPLS	**86.6**	**77.1**	**83.7**

B. Kernel X-CRC

1) Feature Descriptors: Table III presents the experimental results considering distinct feature descriptors in the VIPeR dataset with the proposed Kernel X-CRC model. Specifically, we evaluated the WHOS [6], the LOMO+CNN [30] and the GoG [5] descriptors. While the WHOS is a simple concatenation of shape, texture and color descriptors extracted at different resolutions, the LOMO+CNN consists of a combination of the LOMO descriptor with a feature representation learned using a deep learning model and trained on the VIPeR dataset. According to the results, we observed that the direct application of deep learning approaches results in worst results when compared to the GoG descriptor. It can be related to the small number of samples, which favours the overfiting problem. Therefore, in the remaining experiments, we will focus on the GoG descriptor.

TABLE III
FEATURE DESCRIPTORS EVALUATION ON THE VIPER DATASET.

Feature Descriptor	Viper (p=316)		
	r = 1	r = 5	r = 10
WHOS	43.0	74.0	84.4
LOMO+CNN	46.8	77.3	88.7
GoG	**51.4**	**81.2**	**89.7**

2) Subspace Learning Approaches: One important component in the Kernel X-CRC method is the projection in a low-dimensional subspace as it is responsible to reduce the computational cost while improving the matching rates. Table IV compares subspace learning approaches from literature when using the as matching function the proposed Kernel X-CRC or a simple cosine distance. According to the results, Kernel X-CRC represents a large margin gain for all approaches, which we credit to the nonlinear formulation, collaborative representation and the multitask approach. In the following experiments, we study the influence of each component.

3) Influence of the XQDA: To further understand the separated contribution of the Kernel X-CRC and the XQDA [4] model, we compare the experimental results using the XQDA and different matching functions, such as the cosine distance, Mahalanobis distance and the MLAPG [31] methods. According to the results showed in Table V, all matching functions represent an improvement in the matching rates when compared to the cosine distance. More importantly, between all methods evaluated, the Kernel X-CRC is the one with highest results, which demonstrates its contribution as a matching function.

TABLE IV
SUBSPACE EVALUATION ON THE VIPER DATASET.

Method	Viper (p=316)		
	XQDA [4]	MLAPG [31]	KCCA [7]
Cosine	45.1	41.4	40.5
Kernel X-CRC	**51.4**	**47.6**	**45.7**

TABLE V
DIFFERENT MATCHING FUNCTIONS IN THE XQDA SUBSPACE.

Method	Viper (p=316)		
	r = 1	r = 5	r = 10
XQDA + Cosine	45.1	74.2	84.9
XQDA + Mahalanobis	46.2	74.7	85.6
XQDA + MLAPG	47.6	76.8	86.6
XQDA + Kernel X-CRC	**51.4**	**81.2**	**89.7**

4) Ablation Study: Table VI presents the obtained results using the baselines SRC [32] and the CRC [33], and the different strategies employed in the Kernel X-CRC formulation. Out of these results, we notice that the linear kernel diminishes the *rank-1* in 1.0 percentage point and that without the multi-task formulation the *rank-1* reduces in 2.3 percentage points. Furthermore, Table VI demonstrates a great improvement in the Kernel X-CRC when compared to classical approaches as a results of the nonlinear and multitask formulation.

TABLE VI
RESULTS OF THE BASELINE APPROACHES ON THE VIPER DATASET.

Approach	Viper (p=316)		
	r = 1	r = 5	r = 10
SRC	39.8	65.2	74.9
CRC	47.6	77.6	86.0
Linear Kernel	50.4	79.9	88.4
Without Multi-task	49.1	79.7	88.6
Kernel X-CRC	**51.4**	**81.2**	**89.7**

5) Large-Scale and Multiple Cameras Datasets: In the following experiments, we used the Market-1501 due to the large number of samples and multiple cameras. As feature descriptor, we employed the deep learning representation computed using the ResNet50 model, as proposed in [2]. The Kernel X-CRC computes a kernel function that grows quadratically with the number of samples. In large-scale person re-identification datasets, each individual has multiple samples that are obtained using some tracking algorithm - which we coined tracklets. To make the Kernel X-CRC suitable in this setting, we can compute the average of the feature representation in the entire tracklet. In this way, we limit the complexity of the kernel computation to the number of individuals instead of the number of samples. Table VII presents the comparison

between using the average pooling and considering all the samples. From these results, we can notice that the average pooling not only reduces the computational complexity, but also increases the matching rate.

TABLE VII
INFLUENCE OF AVERAGE POOLING.

Strategy	Market-1501		
	r = 1	r = 2	r = 3
No Pooling	79.7	85.5	88.6
Average	**81.6**	**87.1**	**89.5**

Despite the fact that Equation 5 considers a pair of cameras, the proposed Kernel X-CRC is not limited to the camera pairwise scenario. To demonstrate that, we propose a simple experiment where the matrices $\mathbf{\Phi}_p$ (i.e., probe camera) represents the camera 3, while the camera $\mathbf{\Phi}_g$ (i.e., gallery camera) represents different subsets of cameras. Table VIII presents the obtained experiments for this setting. For instance, the first column represents the *rank-1* when matching camera 3 with camera 6 (pairwise setting), and the last column consists on matching camera 3 with all camera except the camera 3. Based on these results, we can observe that the proposed Kernel X-CRC is not only suitable for a multiple cameras, but also improves the performance when more cameras are available. Furthermore, the Kernel X-CRC outperforms the XQDA method in all settings.

TABLE VIII
RANK-1 MATCHING RATES FOR DIFFERENT GALLERY SETS.

Models	Market-1501			
	G = (6)	G = (6,4)	G = (6,5)	G = G/3
XQDA	76.4	75.6	86.2	89.8
Ker. X-CRC	**78.0**	**78.6**	**87.8**	**90.8**

6) State-of-the-art Comparisons: Table IX presents the matching rates of different approaches considering the VIPeR dataset. These methods are based on metric learning [4], [5], subspace learning [18], deep learning [30], [34], [35] and the SCSP [36], which imposes spatial constraints when matching samples. Based on these results, the proposed Kernel X-CRC outperforms the subspace and deep learning approaches. In fact, the small number of samples results in a challenging scenario for deep learning methods. Besides, the subspace learning approaches employ a simple cosine distance, while the Kernel X-CRC uses a nonlinear and multitask matching function that boosts the results. Finally, the SCSP [36] shows the highest matching rates, which can be associated to the spatial constraints that are successfully captured in VIPeR dataset as the majority of the individuals appear in a frontal view at one camera and side view at the other.

Table X shows the obtained experimental results in PRID450S dataset. Similarly to the VIPeR dataset, the proposed Kernel X-CRC surpassed the subspace learning and deep learning approaches demonstrating that the obtained results are consistent when considering different datasets.

TABLE IX
TOP RANKED APPROACHES ON THE VIPER DATASET.

Method	Viper (p=316)		
	r = 1	r = 5	r = 10
Deep Ranking [34]	38.4	69.2	81.3
LOMO + XQDA [4]	40.0	68.0	80.5
MultiCNN [35]	47.8	74.7	84.8
GoG + XQDA [5]	48.2	77.3	87.6
Shangxuan *et al.* [30]	51.1	81.0	91.4
SCSP [36]	**53.5**	**82.6**	**91.5**
Kernel X-CRC	51.2	79.9	89.9

Differently, the SCSP [36] reached the smallest matching rates between the methods evaluated on PRID450S, which we credit to the challenging pose variations that occur in this dataset.

TABLE X
TOP RANKED APPROACHES ON THE PRID450S DATASET.

Method	PRID450S (p=225)		
	r = 1	r = 5	r = 10
SCSP [36]	44.4	71.6	82.2
LOMO + XQDA [4]	61.4	-	90.8
Shangxuan *et al.* [30]	66.6	86.8	92.8
GoG + XQDA [5]	66.2	87.8	92.6
Kernel X-CRC	**68.1**	**90.7**	**95.0**

V. CONCLUSIONS

In this work, we addressed the person re-identification problem using two approaches that are extensively evaluated in four datasets from literature that diversify in number of samples and cameras. The Kernel MBPLS is a nonlinear regression model that projects samples onto a common and low-dimensional subspace. Experimental results demonstrate the Kernel MBPLS is not only scalable with respect to the number of cameras, but also presents superior results when compared to similar approaches from literature. Besides, we show an improvement in the matching rates when additional labels are included, such as attributes. The Kernel X-CRC indirectly matches samples using a multitask formulation. Experimental results demonstrate its superiority as a matching function for subspace learning approaches from literature. In fact, we achieved the highest and the second highest *rank-1* in PRID450S and VIPeR datasets, respectively. Besides, we demonstrate that the Kernel X-CRC can be successfully adapted to large-scale and multiple cameras datasets.

ACKNOWLEDGMENTS

The authors would like to thank the National Council for Scientific and Technological Development – CNPq (Grants 311053/2016-5 and 438629/2018-3), the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00540-17), the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project). This study was financed in part by the Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

[1] R. Prates and W. R. Schwartz, "Matching people across surveillance cameras," Ph.D. dissertation, Department of Computer Science, UFMG, Belo Horizonte, Brazil, April 2019.

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

[3] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets," *arXiv preprint arXiv:1605.09653*, 2016.

[4] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.

[5] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1363–1372.

[6] G. Lisanti, I. Masi, A. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 8, pp. 1629–1642, Aug 2015.

[7] G. Lisanti, I. Masi, and A. Del Bimbo, "Matching people across camera views using kernel canonical correlation analysis," in *Proceedings of the International Conference on Distributed Smart Cameras*, ser. ICDSC '14. New York, NY, USA: ACM, 2014, pp. 10:1–10:6.

[8] R. Prates, M. Oliveira, and W. R. Schwartz, "Kernel partial least squares for person re-identification," in *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2016.

[9] Z. Wei-Shi, G. Shaogang, and X. Tao, "Person re-identification by probabilistic relative distance comparison," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 203–208.

[10] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *European Conference on Computer Vision*. Springer, 2016, pp. 135–153.

[11] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 1–16.

[12] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1239–1248.

[13] C. Jose and F. Fleuret, "Scalable metric learning via weighted approximate rank component analysis," *arXiv preprint arXiv:1603.00370*, 2016.

[14] M. Zeng, Z. Wu, C. Tian, L. Zhang, and L. Hu, "Efficient person re-identification by hybrid spatiogram and covariance descriptor," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 48–56.

[15] S. Karanam, Y. Li, and R. Radke, "Sparse re-id: Block sparsity for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 33–40.

[16] M. T. Harandi, C. Sanderson, R. Hartley, and B. C. Lovell, "Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 216–229.

[17] E. Kodirov, T. Xiang, and S. Gong, "Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification," in *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, September 2015, pp. 44.1–44.12.

[18] R. Prates and W. R. Schwartz, "Kernel multiblock partial least squares for a scalable and multicamera person reidentification system," *Journal of Electronic Imaging*, vol. 27, no. 3, p. 033041, 2018.

[19] ——, "Kernel cross-view collaborative representation based classification for person re-identification," *Journal of Visual Communication and Image Representation*, 2018.

[20] R. Prates, M. Oliveira, and W. R. Schwartz, "Kernel partial least squares for person re-identification," in *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2016.

[21] R. Prates and W. R. Schwartz, "Kernel hierarchical pca for person re-identification," in *23th International Conference on Pattern Recognition, ICPR 2016, Cancun, MEXICO, December 4-8, 2016.*, 2016.

[22] ——, "Appearance-based person re-identification by intra-camera discriminative models and rank aggregation," in *International Conference on Biometrics, ICB 2015, Phuket, Thailand, 19-22 May, 2015*, 2015, pp. 65–72.

[23] C. D. Raphael Prates and W. R. Schwartz, "Predominant color name indexing structure for person re-identification," in *2016 IEEE International Conference on Image Processing (ICIP)*. Springer, 2016, pp. 779–783.

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

[25] H. Wold, *Encyclopedia of Statistical Sciences*. John Wiley & Sons, 1985, vol. 6.

[26] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European conference on computer vision*. Springer, 2008, pp. 262–275.

[27] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*. Springer, 2014, pp. 247–267.

[28] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Computer Vision, IEEE International Conference on*, 2015.

[29] N. Martinel and C. Micheloni, "Re-identify people in wide area camera network," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 31–36.

[30] W. Shangxuan, C. Ying-Cong, L. Xiang, Y. Jin-Jie, and Z. Wei-Shi, "An enhanced deep feature representation for person re-identification," in *WACV2016: IEEE Winter Conference on Applications of Computer Vision.*, March 2016.

[31] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3685–3693.

[32] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.

[33] L. Zhang, M. Yang, X. Feng, Y. Ma, and D. Zhang, "Collaborative representation based classification for face recognition," *arXiv preprint arXiv:1204.2358*, 2012.

[34] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2353–2367, 2016.

[35] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1335–1344.

[36] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1268–1277.