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. 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.

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 \( p \) as the probe image and \( G \) as the gallery-set composed of \( N \) individuals with known identities, where \( g_i \in G \) corresponds to the \( i \)th subject in the gallery-set. Then, we can determine \( p \) identity (\( id \)) as

\[
\text{id} = \arg \max_i \text{sim}(\Psi(p), \Psi(g_i)), \tag{1}
\]

where \( \Psi \) corresponds to a feature extraction function and \( \text{sim}(\cdot, \cdot) \) 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).
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 non-linearly 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 of data, combining them in a super block and, then, performing a regression between the super block and the respective 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

$$a = X z \quad \text{and} \quad z = X^\top a. \quad (2)$$

Therefore, we have that we can update the scores as $a = XX^\top a$. Then, using a nonlinear transformation in matrix $X$, we obtain $\Phi$ 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 $s_q$) 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

$$\hat{y}_p = k_p^\top T(T^\top K_p T)^{-1} T^\top Y, \quad (3)$$

where $k_p$ is the kernel representation of the $j$th probe image ($p_j$) and $T \in \mathbb{R}^{n \times f}$ is a result of the concatenation of the scores $t \in \mathbb{R}^{n \times 1}$ computed for each factor in Algorithm 1. Similarly, we can compute the regression responses for a gallery sample ($\hat{y}_g$) using $K_p$ and the previous equation.
Algorithm 1: Kernel Multiblock PLS (Kernel MBPLS)

input : \( K_p, K_g, Y \) matrices and the integer \((f)\)

1 randomly initialize \( t \) and \( t_0 \)
2 for \( i=1 \) to \( f \) do
3 \hspace{1em} \( t_0 \leftarrow t \)
4 \hspace{1em} \( s_p = K_p t \)
5 \hspace{1em} \( s_g = K_g t \)
6 \hspace{1em} \( S \leftarrow [s_p, s_g] \)
7 \hspace{1em} \( w = S^\top \)
8 \hspace{1em} \( u = Sw \)
9 \hspace{1em} \( q = Y^\top u \)
10 \hspace{1em} \( t = Yq/q^\top q \)
11 end
12 \( K_p = K_p - tt^\top K_p \)
13 \( K_g = K_g - tt^\top K_g \)
14 \( Y = Y - tt^\top Y \)
15 end

III. KERNEL CROSS-VIEW COLLABORATIVE REPRESENTATION-BASED CLASSIFICATION

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

\[
s(i,j) = \frac{\hat{y}_i \cdot \hat{y}_j}{||\hat{y}_i|| \cdot ||\hat{y}_j||}. \tag{4}
\]

Algorithm 2: Kernel X-CRC.

input : Kernel matrices \((K_p, K_g)\)
output: Ranking list of gallery images \( R \)

1 Compute \( A_p, A_g, Q \) and \( W \) matrices
2 \( \beta_{g}^q \leftarrow W^{-1}P_{g}^{-1}, \beta_{g}^p \leftarrow W^{-1}A_g^{-1}A_p^{-1} \)
3 \( \beta_{p}^g \leftarrow Q^{-1}A_g^{-1}, \beta_{p}^g \leftarrow Q^{-1}A_g^{-1}A_p^{-1} \)
4 for \( p_j \in P \) do
5 \hspace{1em} for \( x_i \in X \) do
6 \hspace{2em} \( \alpha_x \leftarrow \beta_{g}^q k_{x_j}^q + \beta_{g}^p k_{x_j}^p \)
7 \hspace{2em} \( \alpha_p \leftarrow \beta_{g}^q k_{x_j}^q + \beta_{g}^p k_{x_j}^p \)
8 \hspace{2em} \( sim(i) \leftarrow \frac{\alpha_x \cdot \alpha_p}{||\alpha_x|| \cdot ||\alpha_p||} \)
9 \hspace{2em} \( R_j \leftarrow \text{sort}(sim, \text{descend}) \)
10 end
11 return \( R \)

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

IV. EXPERIMENTS

In this section, we perform a comprehensive evaluation of the proposed models 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
For a pair of probe and gallery images, we compute the coding vectors \( (\alpha_p, \alpha_g) \) that collaboratively represent them using training samples captured at their respective camera. A similarity term balances the trade-off between representativeness and similarity.

![Diagram of the proposed Kernel X-CRC method](image)

**Fig. 3.** Schematic representation of the proposed Kernel X-CRC method.

For a pair of probe and gallery images, we compute the coding vectors \( (\alpha_p, \alpha_g) \) that collaboratively represent them using training samples captured at their respective camera. A similarity term balances the trade-off between representativeness and similarity.

![Images of the same individual captured by different cameras](image)

**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.

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 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 the 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.

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 overfitting problem. Therefore, in the remaining experiments, we will focus on the GoG descriptor.

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.

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 result of the nonlinear and multitask formulation.

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>
<thead>
<tr>
<th>TABLE VII</th>
<th>INFLUENCE OF AVERAGE POOLING.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strategy</td>
<td>r = 1</td>
</tr>
<tr>
<td>No Pooling</td>
<td>79.7</td>
</tr>
<tr>
<td>Average</td>
<td>81.6</td>
</tr>
</tbody>
</table>

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 \( \Phi_p \) (i.e., probe camera) represents the camera 3, while the camera \( \Phi_g \) (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>
<thead>
<tr>
<th>TABLE VIII</th>
<th>RANK-1 MATCHING RATES FOR DIFFERENT GALLERY SETS.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Models</td>
<td>Market-1501</td>
</tr>
<tr>
<td></td>
<td>( G = (6) )</td>
</tr>
<tr>
<td>XQDA</td>
<td>76.4</td>
</tr>
<tr>
<td>Ker. X-CRC</td>
<td>78.0</td>
</tr>
</tbody>
</table>

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. 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>
<thead>
<tr>
<th>TABLE IX</th>
<th>TOP RANKED APPROACHES ON THE VIPeR DATASET.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Viper (p=316)</td>
</tr>
<tr>
<td></td>
<td>r = 1</td>
</tr>
<tr>
<td>Deep Ranking [34]</td>
<td>38.4</td>
</tr>
<tr>
<td>LOMO + XQDA [4]</td>
<td>40.0</td>
</tr>
<tr>
<td>MultiCNN [35]</td>
<td>47.8</td>
</tr>
<tr>
<td>GoG + XQDA [5]</td>
<td>48.2</td>
</tr>
<tr>
<td>Shangxuan et al. [30]</td>
<td>51.1</td>
</tr>
<tr>
<td>SCSP [36]</td>
<td>53.5</td>
</tr>
<tr>
<td>Kernel X-CRC</td>
<td>51.2</td>
</tr>
</tbody>
</table>

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 |

<table>
<thead>
<tr>
<th>TABLE X</th>
<th>TOP RANKED APPROACHES ON THE PRID450S DATASET.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>PRID450S (p=225)</td>
</tr>
<tr>
<td></td>
<td>r = 1</td>
</tr>
<tr>
<td>SCSP [36]</td>
<td>44.4</td>
</tr>
<tr>
<td>LOMO + XQDA [4]</td>
<td>61.4</td>
</tr>
<tr>
<td>Shangxuan et al. [30]</td>
<td>66.6</td>
</tr>
<tr>
<td>GoG + XQDA [5]</td>
<td>66.2</td>
</tr>
<tr>
<td>Kernel X-CRC</td>
<td>68.1</td>
</tr>
</tbody>
</table>

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


