CORRESPONDÊNCIA ENTRE PESSOAS EM UMA REDE DE CÂMERAS DE VIGILÂNCIA

RAPHAEL FELIPE DE CARVALHO PRATES

CORRESPONDÊNCIA ENTRE PESSOAS EM UMA REDE DE CÂMERAS DE VIGILÂNCIA

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

Orientador: William Robson Schwartz

Belo Horizonte
Março de 2019

RAPHAEL FELIPE DE CARVALHO PRATES

CORRESPONDÊNCIA ENTRE PESSOAS EM UMA REDE DE CÂMERAS DE VIGILÂNCIA

Thesis presented to the Graduate Program in Departamento de Ciência da Computação of the Universidade Federal de Minas Gerais - Departamento de Ciência da Computação in partial fulfillment of the requirements for the degree of Doctor in Departamento de Ciência da Computação.

Advisor: William Robson Schwartz

Belo Horizonte
March 2019

Prates, Raphael Felipe de Carvalho

P912c Correspondência entre pessoas em uma rede de câmeras de vigilância / Raphael Felipe de Carvalho Prates. — Belo Horizonte, 2019
xix, 113 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais - Departamento de Ciência da Computação
Orientador: William Robson Schwartz

1. Computação - Teses. 2. Visão por computador. 3. Processamento de Imagens. 4. Aprendizado do computador. I. Título.

CDU 519.6*82.10 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Correspondência entre pessoas em uma rede de câmeras de vigilância

RAPHAEL FELIPE DE CARVALHO PRATES

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

William Robson Schwartz
Prof. William Robson Schwartz - Orientador
Departamento de Ciência da Computação - UFMG

Prof. Moacir Antonelli Ponti
Departamento de Ciência da Computação - USP

Prof. Adriano Alonso Veloso
Departamento de Ciência da Computação - UFMG

Prof. Erickson Rangel do Nascimento
Departamento de Ciência da Computação - UFMG

Prof. Guillermo Cámara Chávez
Departamento de Computação - UFOP

Belo Horizonte, 29 de Março de 2019.

Acknowledgments

I 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) and the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project). Part of the results presented in this dissertation were obtained through research on a project titled "HAR-HEALTH: Reconhecimento de Atividades Humanas associadas a Doenças Crônicas", sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91.

Resumo

O número de redes de câmeras de vigilância é cada vez maior como consequência da crescente preocupação com segurança. A grande quantidade de dados coletados demanda sistemas de vigilância inteligentes para extrair informações que sejam úteis aos oficiais de segurança. De forma a alcançar esse objetivo, esse sistema deve ser capaz de correlacionar as informações capturadas por diferentes câmeras de vigilância. Nesse cenário, a re-identificação de pessoas é de central importância para estabelecer uma identidade global para indivíduos capturados por diferentes câmeras usando apenas a aparência visual. No entanto, trata-se de uma tarefa desafiadora, uma vez que a mesma pessoa quando capturada por câmeras distintas sofre uma drástica mudança de aparência como consequência das variações no ponto-de-vista, iluminação e pose. Trabalhos recentes abordam a re-identificação de pessoas propondo descritores visuais robustos ou funções de correspondência entre câmeras, as quais são funções que aprendem a calcular a identidade correta de imagens capturadas por diferentes câmeras. Porém, a maior parte desses trabalhos é prejudicada por problemas como ambiguidade entre indivíduos, a escalabilidade e o número reduzido de imagens rotuladas no conjunto de treino. Nesta tese, abordamos o problema de correspondência de indivíduos entre câmeras de forma a tratar os problemas já mencionados e, portanto, obter melhores resultados. Especificamente, propomos duas direções: o aprendizado de subespaços e os modelos de identificação indireta. O primeiro aprende um subespaço comum que é escalável com respeito ao número de câmeras e robusto em relação à quantidade de imagens de treino disponíveis. Na identificação indireta, identificamos imagens de prova e galeria baseado na similaridade com as amostras de um conjunto de treino. Resultados experimentais validam ambas as abordagens no problema de re-identificação de pessoas considerando tanto apenas um par de câmeras como situações mais realísticas com múltiplas câmeras.

Palavras-chave: Visão Computacional, Vigilância Inteligente, Re-identificação de pessoas.

Abstract

The number of surveillance camera networks is increasing as a consequence of the escalation of the security concerns. The large amount of data collected demands intelligent surveillance systems to extract information that is useful to security officers. In order to achieve this goal, this system must be able to correlate information captured by different surveillance cameras. In this scenario, re-identification of people is of central importance in establishing a global identity for individuals captured by different cameras using only visual appearance. However, this is a challenging task, since the same person when captured by different cameras undergoes a drastic change of appearance as a consequence of the variations in the point of view, illumination and pose. Recent work addresses the person re-identification by proposing robust visual descriptors orcross-view matching functions, which are functions that learn to match images from different cameras. However, most of these works are impaired by problems such as ambiguity among individuals, scalability, and reduced number of labeled images in the training set. In this thesis, we address the problem of matching individuals between cameras in order to address the aforementioned problems and, therefore, obtain better results. Specifically, we propose two directions: the learning of subspaces and the models of indirect identification. The first learns a common subspace that is scalable with respect to the number of cameras and robust in relation to the amount of training images available. we match probe and gallery images indirectly by computing their similarities with training samples. Experimental results validate both approaches in the person re-identification problem considering both only one pair of cameras and more realistic situations with multiple cameras.

Palavras-chave: Computer Vision, Smart Surveillance, Person Re-Identification.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The interest in video surveillance has increased as consequence of the demand for public safety and the wide spread of surveillance camera networks in public (e.g., airports, universities campus and streets) and private places. These cameras collect a vast volume of data to be manually analyzed by law enforcement officers or used for forensic purposes. Nonetheless, there is no human labor to cope with the exponential growth of these data. Therefore, intelligent video surveillance aims at extracting useful information from these visual data to assist the security personnel. To accomplish that, multiple modules infer complex semantic information based on the detection, tracking and recognition of persons and objects in images and videos [Wang, 2013].

Multi-camera surveillance is a desired requirement in an intelligent video surveillance system as the data captured by a single camera is limited spatially and temporally. For instance, to monitor a wide area, as a university campus, we need to use hundreds of cameras that usually have non-overlapping fields-of-view that further increase the monitored area coverage. In this setting, how to automatically integrate the information from multiple cameras is an important issue. Person re-identification emerged as a possible solution that attained increasing attention of the research community in the last years [Zheng et al., 2016].

Person re-identification (Re-ID) provides the identity for an individual as he/she moves along surveillance cameras with non-overlapping field-of-views. Thus, it is possible to monitor activities in large areas with a reduced cost (small number of cameras). For instance, Re-ID allows the retrieval of videos of the same person in an entire camera network or the development of an online multi-camera tracking algorithm using the correspondence between individuals in multiple cameras.

A Re-ID system is composed of the person detection, tracking, feature extraction and cross-view matching [Apurva and Shah., 2014; Zheng et al., 2016], as illustrated in

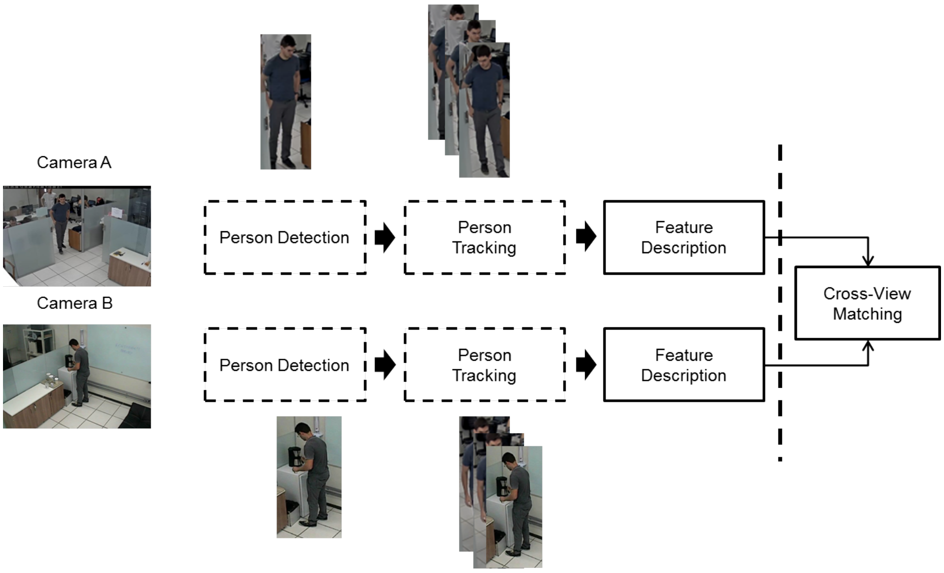

Figure 1.1: Schematic representation of the main steps in a single-shot person re-identification system. First, person detection and tracking are used in each camera to obtain the person locations (bounding boxes) at consecutive frames. Then, we extract the feature descriptor that will be used in the cross-view matching function to compute the similarity between individuals captured by these different cameras.

Figure 1.1. While in person detection and tracking we determine the person location at multiple frames in each camera, it is in the feature extraction and cross-view matching that we actually match images or videos captured by different cameras to assign the identity for an unknown individual.

In this dissertation, we focus on the single-shot and closed-set scenario, which is the most common setting in the person re-identification literature. In the single-shot, we have just have a single image captured from each individual at different and non-overlapping cameras. In addition, in the closed-set, we consider that the gallery-set (known identities) always includes the probe (unknown identity) and the number of individuals enrolled in the gallery is fixed.

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

$$id = \arg\max_{i} sim(\Psi(\mathbf{p}), \Psi(\mathbf{g}_i)), \tag{1.1}$$

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

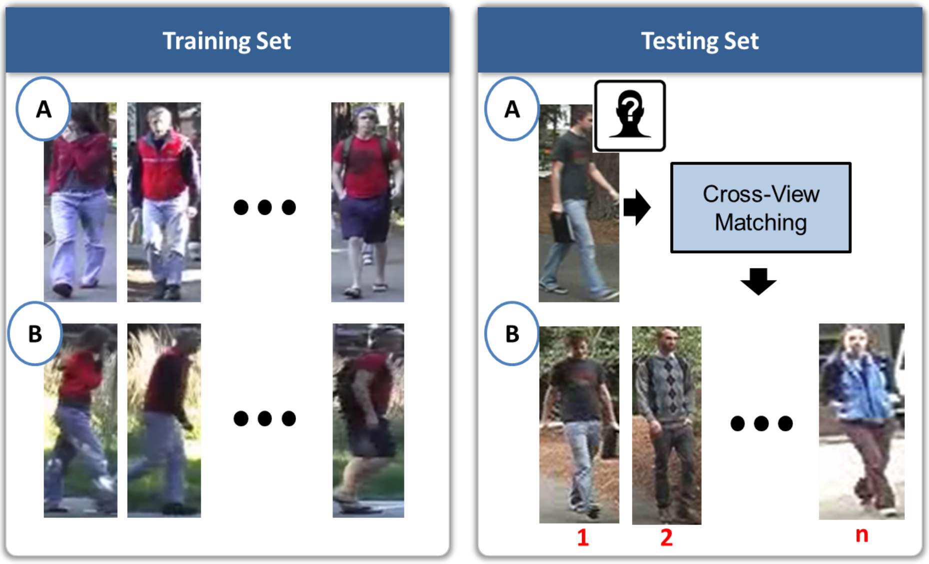

Figure 1.2: Training and Testing sets. In the training set the better configuration of feature descriptors and cross-view functions are learned using labeled individuals in a pair of cameras. Then, in the testing set, these functions are deployed using a disjoint subset of individuals.

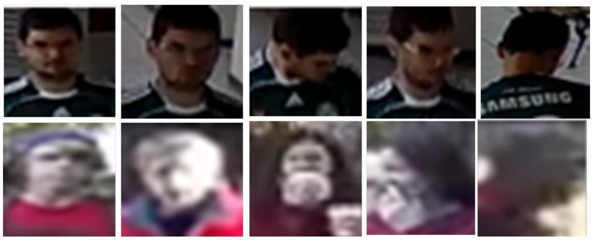

Figure 1.3: Faces from images captured by distinct surveillance cameras.

captured by different surveillance cameras. Then, these functions are deployed in a testing set whose identities are disjoint from the training set, as illustrated in Figure 1.2.

In a Re-ID system, the main challenge corresponds to rank the gallery-set based on the similarity with a probe (unknown identity) due to the low-resolution images provided by surveillance cameras where biometric cues such as face and iris are unreliable (see Fig. 1.3). In this case, the matching between probe and gallery images is based on appearance cues provided by clothes, mainly color and texture features. Therefore, it is prone to issues as ambiguity between individuals and the distinct camera conditions. The ambiguity is a result of the fact that individuals tend to dress similarly. Differently, the camera conditions are affected by different illumination conditions, camera's pose, background clutter and occlusion (see Figure 1.4). In this scenario, the design of feature descriptors that are simultaneously robust to different camera conditions and

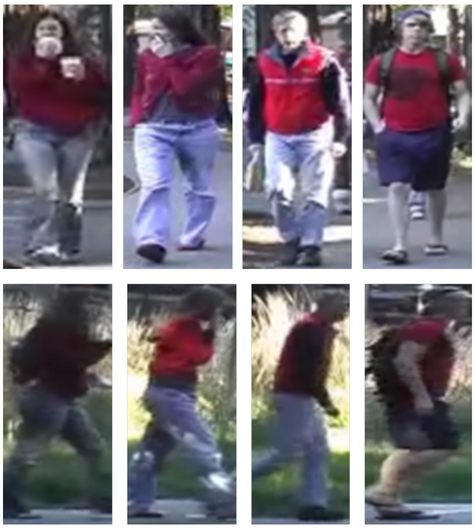

Figure 1.4: Images of the same person (columns) look more dissimilar than images of different persons as consequence of the high inter-person similarity and the different camera conditions (illumination, background, pose and camera viewpoints).

discriminative is an unattainable problem.

Deep learning-based methods are a possible solution to learn the feature representation and a matching function in a unique end-to-end framework. Nonetheless, experimental results obtained using deep learning approaches are still suboptimal. That can be explained by the laborious work of collecting and annotating images for each camera pair, which restricts the size of the person Re-ID datasets and prevents the generalization of the learned model. Therefore, there is still place for more traditional pipelines in person Re-ID that combine high-dimensional handcrafted and/or deeply learned descriptors with a cross-view matching function learned independently.

Another important issue that is still neglected by the person re-identification community is the scalability of the learned models with respect to the number of surveillance cameras (c). As an example, most of the approaches in literature consider a distinct model for each camera pair. In this scenario, we would need c(c-1)/2 models ($O(c^2)$), which is not scalable for real-world surveillance systems. Figure 1.5 illustrates both settings with the number of models learned. Therefore, how to design more scalable cross-view matching functions in such scenario is an important issue that we address in this dissertation.

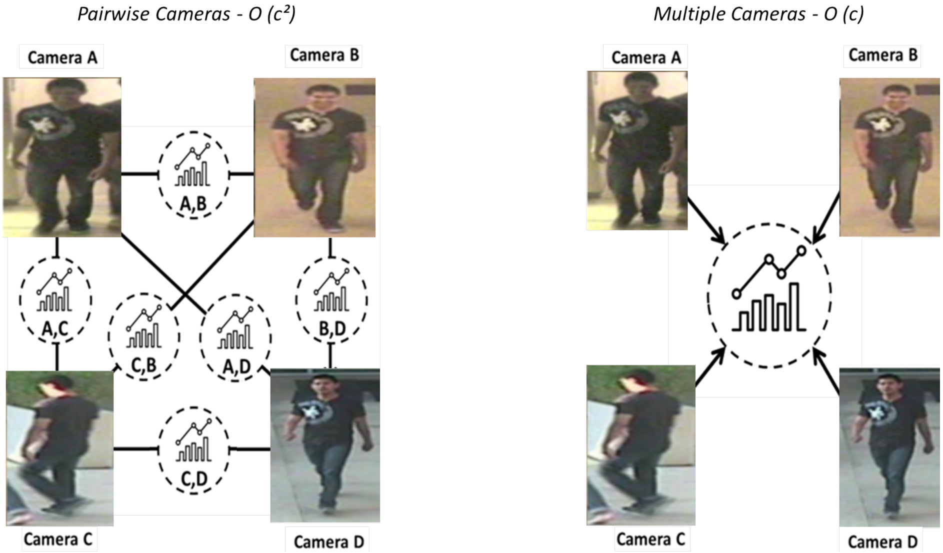

Figure 1.5: Pairwise and multiple cameras cross-view matching models. Notice that the pairwise learns a model for each pair of cameras. Differently, the multiple cameras learns a single for the entire camera network.

1.1 Motivation

As a consequence of the different camera conditions, feature descriptors extracted from an individual in a camera A will change drastically when the same person is seen again in a camera B. Thus, a simple Euclidean distance is ineffective to perform the matching between probe and gallery samples.

One alternative to tackle the camera transition problem consists in learning a transformation in the feature space that projects closer images of the same person independently of the camera conditions as illustrated in Figure 1.6. Nonetheless, the

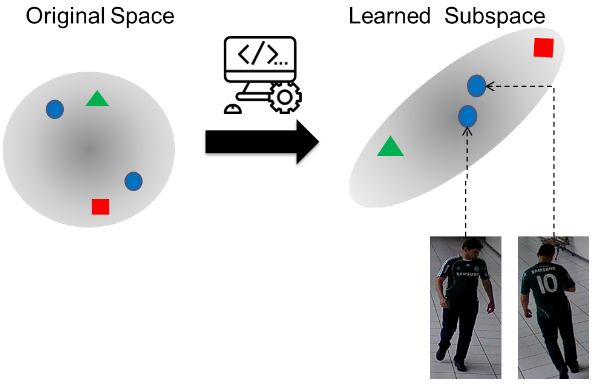

Figure 1.6: Subspace Learning. Images of the same person captured by different surveillance cameras are correlated in a common and low-dimensional subspace.

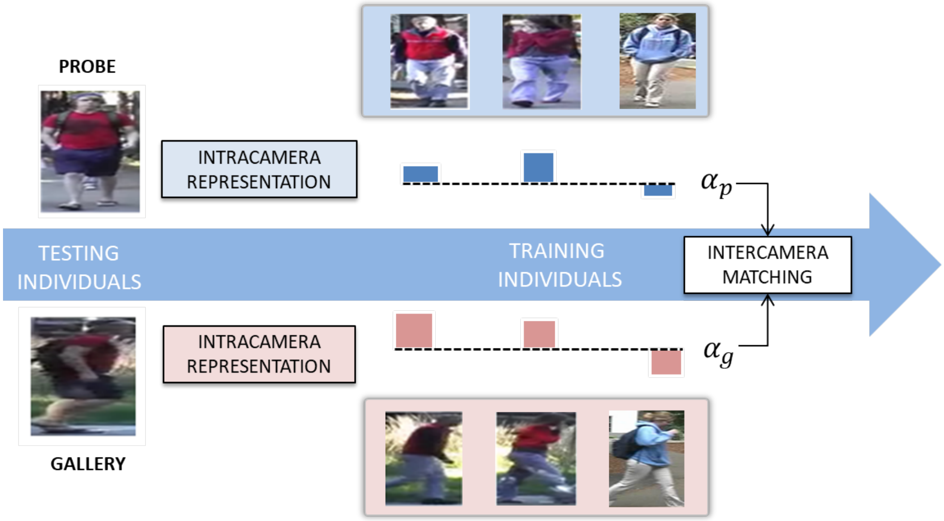

Figure 1.7: Schematic representation of the indirect matching approach. Training images of the same individuals are placed in a blue and red boxes to indicate probe and gallery cameras, respectively. First, we compute for probe and gallery samples an intra-camera representation (α_p and α_g) based on some similarity model. Then, using these representations, we indirectly perform the inter-camera matching between probe and gallery samples.

drastic appearance due to the distinct camera suggests a nonlinear transformation between cameras. Besides, these models need to deal with real-world settings where we have a huge number of surveillance cameras and a small number of labeled samples. Therefore, it motivates us to address the person Re-ID problem with nonlinear subspace learning models that can be learned with a small number of samples and are scalable with respect to the number of cameras.

A different solution is avoiding the camera transition by performing only intra-camera matching. For instance, we can compute the similarity of a probe image with a subset of individuals in the training set captured by the same camera. As these individuals are in training set, we also have their images in the gallery camera to compare with the samples in the gallery set. Then, these representations of similarity are used to indirectly compare probe and gallery images as illustrated in Figure 1.7.

One problem that also impacts the performance of person re-identification methods is the ambiguity between individual's clothes (i.e., individuals dressing uniforms). This problem can be reduced when additional information such pose and gender are available as attributes. These attributes can be manually annotated or automatically obtained using state-of-the-art approaches [Schumann and Stiefelhagen, 2017]. Figure 1.8 shows some examples of manually annotated attributes. Therefore, to address the ambiguity problem, it is important to enable the cross-view matching function to learn from attributes information.

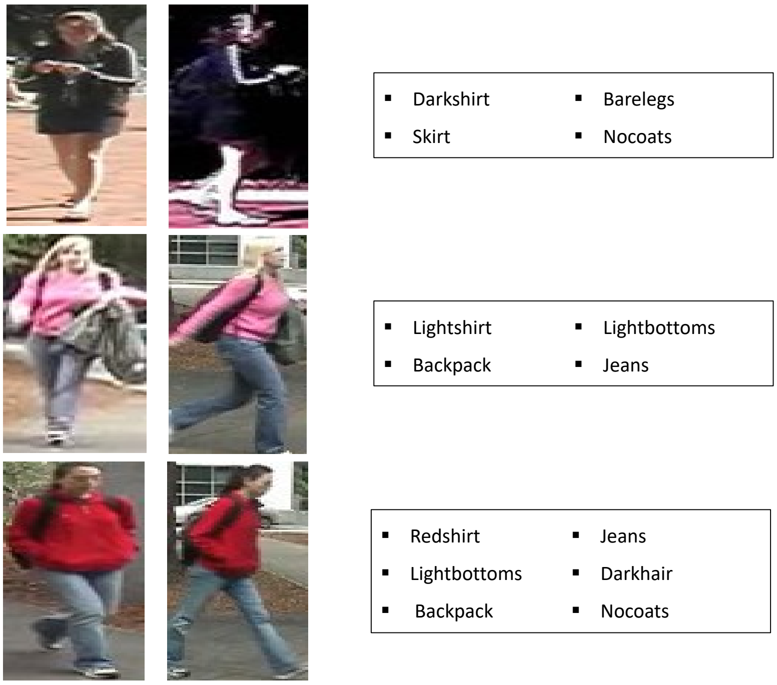

Figure 1.8: Manually labeled attributes for three individuals in VIPeR dataset.

In this work, we tackle some important issues in the person re-identification problem as the different camera conditions, the reduced number of training samples, the scalability with respect to the number of surveillance cameras and the ambiguity between individuals. Specifically, we show that subspace learning and the indirect matching strategies are successful to obtain feature representations that are robust to the different camera conditions. In addition, we present models that can be learned using datasets with small number of training samples, multiple cameras and additional labels (i.e. attributes). More importantly, we show that by using attribute and identity labels we can reduce the ambiguity between individual and boost the obtained results.

1.2 Hypotheses

In this dissertation, we assume two main hypotheses. The Subspace Learning hypothesis assumes that *“images of the same person when captured by different cameras share some subtle characteristics that can be captured by nonlinearly mapping feature descriptors from probe and gallery cameras into a common and low-dimensional subspace”*. In addition, the Indirect Matching hypothesis considers that *“images that are similar when captured by one camera will remain similar at a second camera and, therefore, it is possible to indirectly match images from different cameras using intra-camera similarity* *values"*. Based on these hypotheses, we elaborate some secondary hypotheses that are presented as follows.

Subspace Learning

An usual approach to match images captured by different surveillance cameras consists in learning either linear or nonlinear projections onto a common subspace that correlates feature descriptors extracted from different cameras. In fact, due to the nonlinear variations of feature descriptors across cameras, an improved matching performance is usually achieved using nonlinear subspace learning methods.

The main drawback of current subspace learning methods is that they are based on pairwise models and, therefore, are not scalable to real-world scenarios as the number of camera pairs grows quadratically with the number of cameras. Therefore, our hypothesis is that *"we can hierarchically model the nonlinear subspace learning problem in such a way that the number of learned projections matrices grows linearly with the number of surveillance cameras"*.

Besides, as a consequence of the small number of training samples (i.e., small-sample-size problem), these subspace learning methods present suboptimal results as they can not estimate the covariance matrices of the data with high accuracy. Therefore, another hypothesis is that *"we can learn nonlinear subspace models that do not depend on the estimation of covariance matrices and, therefore, are robust to the small-sample-size problem"*.

Indirect Matching

One possible and simple solution to tackle the camera transition problem consists in performing an indirect matching of individuals by using a subset of labeled persons that appeared in both cameras (i.e., training set). These methods assume that when individuals are similar in one camera, they will remain similar when captured by another, as illustrated in Figure 1.9. Then, using the intra-camera similarity, they are able to match persons captured by different cameras. In this scenario, one important issue is how to handle the variations on illumination, background clutter and occlusion that are present even when considering images captured by the same camera.

Regarding the indirect matching, our first hypothesis is that *"due to the ambiguity problem, we can find individuals in the training set - the prototypes - that are similar to any given sample (probe or gallery images). More importantly, as they are in the*

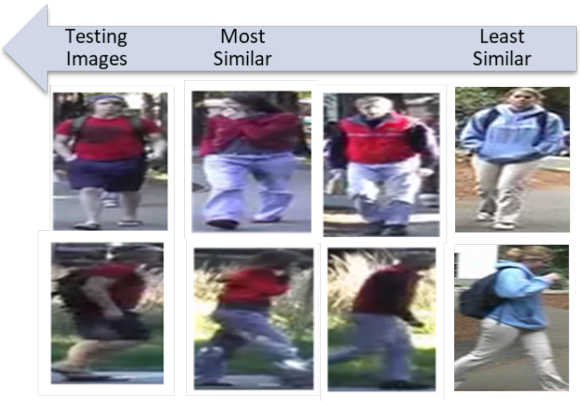

Figure 1.9: Indirect Matching. Notice that similar individuals captured by the first camera remain similar when comparing images captured by the second camera. These stable similarities is the main assumption of the Indirect Matching hypotheses.

training set, we can use them to learn discriminative models in the specific camera where these models will be deployed and, then, handle the camera transition problem".

Our second hypothesis considers that *"by learning a nonlinear regression model for each camera, we can compute an improved representation of probe and gallery images based on the similarity with the entire training set".* Notice that while in the first hypothesis we restrict our model to the prototypes subset, here we use all the available images in the training set.

Differently, our third hypothesis is that *"probe and gallery images can be represented as a linear combination of training samples captured at their respective cameras using a multi-task framework. More importantly, we claim that using the computed linear coefficients we can effectively match probe and gallery images".*

1.3 Objectives

In this work we address the person re-identification problem as a subspace learning or indirect matching problem. Specifically, we proposed six different methods with the following objectives: (1) Propose a model robust with respect to the number of training samples to reduce the burden of annotating persons identities at multiple cameras. (2) Dimish the camera transition problem using strategies as the nonlinear mapping of all cameras to a common subspace or indirectly matching samples from different cameras based on the similarity with samples in a training set. (3) Reduce the ambiguity problem by including additional labels, such as attributes, that boost the discriminative power of the learned model. (4) Work with models that are scalable with respect to the number of surveillance cameras.

1.4 Contributions

In this section, we present the main contributions of this dissertation. For a better exposition of these achievements, we categorized them in those related to the subspace learning hypotheses and those that correspond to the indirect matching hypotheses.

Subspace Learning

In this dissertation, we address person re-identification as the problem of learning a nonlinear common subspace where the direct matching between probe and gallery images is successful. Specifically, we propose three methods to learn projection matrices to the low-dimensional subspace: the *Kernel Hierarchical PCA*, the *Kernel PLS for Subspace Learning* and the *Kernel Multiblock PLS* .

The proposed methods are novel nonlinear extensions of common subspace learning models and possess the required characteristics for the person re-identification problem. For instance, *Kernel PLS for Subspace Learning* learns a common subspace the maximize the covariance between feature descriptors captured by different cameras and is robust to small-sample-size problem. Differently, *Kernel Hierarchical PCA* and *Kernel Multiblock PLS* learn a consensus directions between different cameras in a hierarchical formulation that grows linearly with the number of surveillance cameras. While the *Kernel Hierarchical PCA* only maximizes the covariance of the data, the *Kernel Multiblock PLS* performs a nonlinear regression between latent scores and response variables (e.g., identity labels) (see Chapter 3). Experimental results demonstrate that the nonlinear extension is important to reach improved results and, more importantly, these results are comparable to other common subspace learning models from literature that are not scalable and require the careful adjustment of regularization parameters.

Indirect Matching

The main drawback of nonlinear common subspace learning models resides in the direct matching between images captured by different cameras. We tackle this problem by indirectly matching probe and gallery images using labeled individuals in training set. We propose three methods to indirectly match probe and gallery images: *Prototypes-based Person Re-Identification*, *Cross-View Kernel PLS* and the *Kernel Cross-View Collaborative Representation based Classification.*

Our first attempt to indirectly match probe and gallery images was the *Prototypes-based Person Re-Identification* that learns discriminative models using a subset of similar individuals to probe and gallery images captured by their respective camera in the training set (prototypes subset). However, determining prototypes subset revealed to be a very challenging task, mainly for the smaller datasets. Thus, in the proposed *Cross-View Kernel PLS*, we use a nonlinear regression model to represent probe and gallery images based on the similarity with the entire training set. Similarly, in the novel *Kernel Cross-View Collaborative Representation based Classification*, we use a collaborative multi-task framework to adaptively represent probe and gallery images using coding vectors that balance the representativeness and discriminative power (see Chapter 4). The obtained experimental results show that our indirect matching strategy reaches interesting results. In fact, the proposed *Kernel Cross-View Collaborative Representation based Classification* achieved state-of-the-art results in the three datasets evaluated.

1.5 Outline

The chapters of this dissertation are organized as follows. Chapter 2 provides an overview of the main and most recent works in person re-identification. In Chapter 3, we present the proposed methods to approach person re-identification as a subspace learning problem. Then, Chapter 4 describes the proposed methods related to Indirect Matching of probe and gallery images. In Chapter 5, we present the experimental evaluation of the main parameters related to these methods in a pairwise camera setting and compare them against state-of-the-art approaches. Then, in Chapter 6, we show experimental results considering the multiple cameras scenario. Finally, Chapter 7 concludes with the contributions, limitations of this work and future research directions.

1.6 Publications

This dissertation has resulted so far in the seven publications listed in the following.

1. **Prates, Raphael Felipe**; Schwartz, William Robson. *Appearance-Based Person Re-Identification by Intra-Camera Discriminative Models and Rank Aggregation.* In: Biometrics (ICB), 2015 International Conference on. IEEE, 2015. p. 65-72.

2. **Prates, Raphael Felipe**; Schwartz, William Robson. *CBRA: Color-Based Ranking Aggregation for Person Re-Identification.* In: Image Processing (ICIP), 2015 IEEE International Conference on. IEEE, 2015. p. 1975-1979.

3. **Prates, Raphael Felipe**; Oliveira, Marina; Schwartz, William Robson. *Kernel Partial Least Squares for Person Re-Identification.* In: Advanced Video and

Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on. IEEE, 2016. p. 249-255

4. **Prates, Raphael Felipe**; Schwartz, William Robson. *Kernel Hierarchical PCA for Person Re-Identification.* In: 23th International Conference on Pattern Recognition, ICPR. 2016. p. 4-8.

5. **Prates, Raphael Felipe**; Dutra, Cristianne; Schwartz, William Robson. *Predominant Color Name Indexing Structure for Person Re-Identification.* In: Image Processing (ICIP), 2016 IEEE International Conference on. IEEE, 2016. p. 779-783.

6. **Prates, Raphael Felipe**; Schwartz, William Robson. *Kernel Multiblock Partial Least Squares for a Scalable and Multicamera Person Re-Identification System .* In: Journal of Electronic Imaging, Volume 27, Issue 3, 2018.

7. **Prates, Raphael Felipe**; Schwartz, William Robson. *Kernel Cross-View Collaborative Representation based Classification for Person Re-Identification.* In: Journal of Visual Communication and Image Representation, Volume 58, 2019.

Chapter 2

Related Works

At the beginning, person Re-ID was not consolidated as a research topic. Instead, the matching between individuals captured by different cameras occurred as a part of the multi-camera tracking system, usually combining appearance features with camera calibration and topology information [Wang, 2013]. For instance, the work proposed by Zajdel et al. [2005] is the first to use the term person re-identification and was based on a Bayesian model to relate appearance and spatio-temporal cues with label information. Actually, the work of Gheissari et al. [2006] is the first to re-identify individuals using only appearance features, therefore, establishing person re-identification as an independent research topic in the computer vision community [Zheng et al., 2016].

Initially, most of the works on person re-identification focused on the single-shot scenario. In this scenario, the camera transition problems are extreme - due to lack of available data - and impact more on the obtained experimental results. However, we can employ detection and tracking algorithms to capture multiple images or videos of the same individuals. Then, multi-shot and video-based methods have been proposed by the person re-identification community with improved results [Farenzena et al., 2010]. Nonetheless, these methods also have to deal with the higher intraclass variations than interclass that can be even more challenge when dealing with videos [You et al., 2016].

Appearance features extracted from still images or videos capture only short-term signatures, which restricts the applicability of person re-identification systems. To capture long-term information, some works employed complementary data as gait features [Liu et al., 2015b] or thermal and depth images [Mogelmose et al., 2013]. Gait information, which corresponds to the walking style of a person, is a long-term biometry that can be captured by low-resolution cameras without any suspect collaboration [Liu et al., 2015c]. However, gait methods commonly extract person's silhouette, which is very difficult in low-resolution cameras [Gao et al., 2016a]. Differently, thermal and depth information are captured by additional cameras that are seldom present at indoor or outdoor environments and have restricted applicability. Then, how to extract long-term signatures for persons in uncontrolled environments is still an open problem.

In this chapter, we review works proposed in the person re-identification literature. Notice that despite focusing on the single-shot setting, we also present the main works that consider different scenarios to provide a better overview of the literature. We divide these works based on the information used to perform person re-identification as single-shot (Section 2.1), multiple-shot (Section 2.2) and videos (Section 2.3). It is important to highlight that we provide a brief description of these methods. For a more detailed discussion, please refer to the person re-identification surveys [Apurva and Shah., 2014; Zheng et al., 2016].

2.1 Single-Shot Person Re-Identification

Single-shot person Re-ID consists in matching a person across cameras using a single image captured by each camera. These methods are motivated by the reduced frame-rate of conventional surveillance cameras that limits the number of frames available. Due to the small amount of collected data and the different camera conditions, single-shot person Re-ID is a very challenging task [Apurva and Shah., 2014]. Therefore, several works addressed single-shot person re-identification designing better feature descriptors (Section 2.1.2) or cross-view matching functions (Section 2.1.3). These steps are illustrated in Figure 1.1.

2.1.1 Labeled Samples

An important issue when dealing with person re-identification is the small number of labeled samples. It is a consequence of troublesome task of annotating a large number of images at multiple cameras. One attempt to tackle this problem that is worth mentioning corresponds to the generation of synthetic data. As an example, SOMAset [Barbosa et al., 2018] consists of synthetic data generated with different height, weight, body proportions, outfits and 250 different poses (see Figure 2.1). Similarly, in [Bak et al., 2018], the authors utilize synthetic images with multiple illumination conditions to increase the model's generalization to unseen illumination conditions. Nonetheless, these artificial images create a domain gap problem as they are only coarse approximations of real-world images absent of realism.

In this section, we discuss some recent works that addressed the small number of annotated samples problem using state-of-the-art techniques as Generative Adversarial

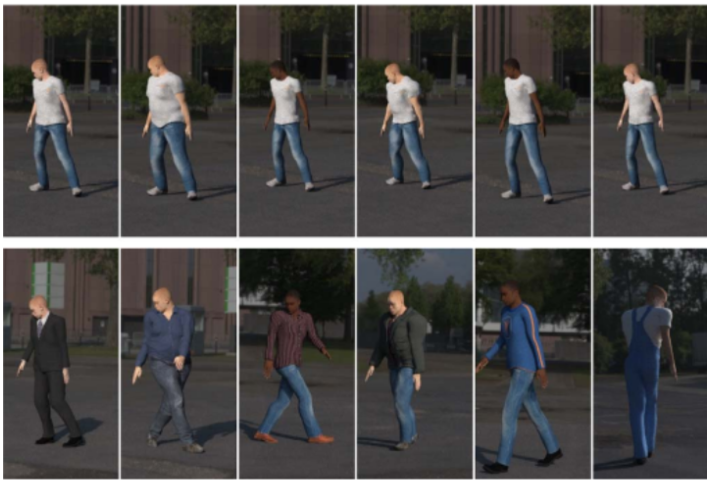

Figure 2.1: Synthetic images from SOMAset database. Illustration taken from [Barbosa et al., 2018]

Networks (GANs) [Siarohin et al., 2018; Zhong et al., 2018; Liu et al., 2018b; Zheng et al., 2017c; Deng et al., 2018; Bak et al., 2018] and Domain Adaptation [Fu et al., 2018; Geng et al., 2016; Li et al., 2018; Peng et al., 2016, 2018].

Zheng et al. [2017c] is the inaugural work using GANs to automatically generate images that boost person Re-ID models. In [Zhong et al., 2018], the authors improved the generative model using CycleGAN to transit between two cameras. Similarly, Bak et al. [2018] used CycleGAN to reduce the domain shift between synthetic and real-world images. Recently, more supervision has been deployed to generate more reliable samples. For instance, Siarohin et al. [2018] used pose information as input to generate images at a specific pose and Deng et al. [2018] employed a siamese network with a similarity preserving loss. Liu et al. [2018b] combined the pose with the class information to ensure that the generated sample has the target pose while preserving the identity.

Differently from GANs, Domain Adaptation uses large pre-existent annotated databases to learn representations that are transferred to an unlabeled dataset. An early strategy consists in modelling the problem as an asymmetric multi-task dictionary learning where specific and shared dictionaries are learned using iterative optimization algorithms with graph Laplacian regularisation terms [Peng et al., 2016, 2018]. In this way, the shared dictionaries transfer a representation from labeled to unlabeled samples. The main drawback of these methods is the fact that the feature descriptors and dictionaries are not learned end-to-end. To address this issue, Li et al. [2018] proposed an encoder-decoder architecture with shared and specific modules that are learned jointly using labeled and unlabeled samples. Differently, [Geng et al., 2016] uses a co-training strategy that alternates between subspace learning and CNN self-training to iteratively assigns pseudo-labels to the samples and updates the initial model, which was pre-trained in a large collection of labeled Re-ID images.

2.1.2 Feature Descriptors

In this section, we describe different feature descriptors that have been proposed in the literature exploiting both feature representation and body parts from where they will be extracted and matched [Satta, 2013]. These methods seek for a representation that remains stable in the presence of variations caused by the different camera conditions.

Regarding the body locations, the goal is to undertake the spatial misalignment between camera viewpoints. Figure 2.2 illustrates some of these body models proposed in literature. The simplest and most common strategy corresponds to equally divide the image in a fixed number of horizontal stripes [Wei-Shi et al., 2011], which is highly sensitive to the pedestrian detector errors. Thus, there have been some efforts to define body parts adaptively. Farenzena et al. [2010] used the symmetry of human body and Cheng et al. [2011] employed Pictorial Structures to automatically detect body regions. Differently, Zhao et al. [2013] discovered cross-view discriminative and robust image patches based on saliency information and Chen et al. [2016a] captured spatial distribution of patches between cameras to constraint the inter-camera matching.

Deep learning-based methods have also been proposed to deal with spatial misalignments between cameras and learn the feature representation in unique framework [Li et al., 2014; Varior et al., 2016; Zheng et al., 2018; Wang et al., 2018; Li et al., 2017]. In [Li et al., 2014], the authors handle with photometric and geometric transformations using a patch matching layer, while Varior et al. [2016] exploited the dependence between different spatial regions of the same image using Long Short-Term Memory (LSTM) architecture and a siamese network. More recently, Zheng et al. [2018] and Zheng et al. [2018] explored attentional mechanisms to highlight specific feature maps while handling the inter-camera misalignment problem, while Li et al. [2017] explicitly learn body parts segmentation using spatial transformers network.

A different perspective uses the person's pose information as input to construct a more reliable embedding representation. In [Jiang et al., 2018], the authors estimated the body orientation and parts location based on the pose information and proposed orientation-guided loss that pulls closer images of the same person with similar orientations. Likewise, Sarfraz et al. [2018] the authors employed the pose information as input in a model that learns view predictors and, consequently, a pose-sensitive embedding representation. Differently, Zheng et al. [2017a] construct the PoseBox representation

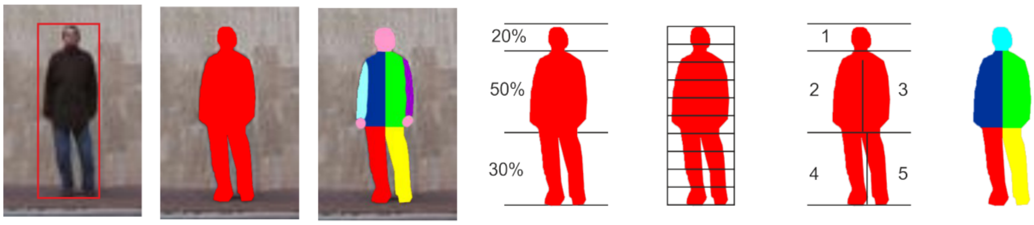

Figure 2.2: From left to right. First we have the pedestrian detection, foreground segmentation and the body segmentation. Then, different strategies of body segmentation are presented. The first two consider a fixed number of stripes, while in the last two we have five body parts obtained using human body symmetry [Farenzena et al., 2010]. Illustration taken from [Vezzani et al., 2013].

based on the pose and affine transformations.

After defining body locations, another important step is to represent these regions. For instance, Ma et al. [2012b] captured pixel information using local descriptor encoded by Fisher Vector, while Matsukawa et al. [2016] used a hierarchical Gaussian distribution to capture the mean and covariance from local patches. In [Ma et al., 2012a], biologically inspired features and covariance descriptors are combined to describe image regions. Differently, Yang et al. [2014] used a probabilistic model to relate color features to semantic color names and Liao et al. [2015] constructed a stable representation using horizontal occurrence of local features and max-pooling. Recently, Shangxuan et al. [2016] combined handcrafted feature descriptors with Convolutional Neural Network (CNN) features obtaining a more discriminative representation and the Mirror Representation [Chen et al., 2015] was proposed to obtain an augmented feature representation that aligns feature distributions across cameras.

In this work, we assume that it is not possible to properly handle the camera transition by directly matching feature descriptors captured by different cameras using a simple Euclidean distance. Therefore, we address the problem of learning a cross-view matching function to efficiently match probe and gallery images. In the next section, we present some works in literature that also addressed person re-identification learning cross-view matching functions.

2.1.3 Cross-View Matching

In the following paragraphs, we present the main attempts to learn a cross-view matching function to compute the similarity between images captured by different cameras. In this setting, feature descriptors are combined with a cross-view matching function learned using labeled and/or unlabeled images captured at each surveillance camera to obtain higher matching performance. In the following sections, we present these works divided based on their main contribution as metric learning, common subspace learning, information retrieval and unsupervised learning.

Metric Learning

Metric learning methods are motivated by the observation that the obtained results are suboptimal when using state-of-the-art feature descriptors with a simple Euclidean distance. Therefore, they use the pairwise (same or not-same) or triplet constraints to learn a distance function that is smaller between pairs of the same person than when considering different persons [Zheng et al., 2016]. The most commonly used metric is the generalization of Euclidean distance considering scaling and rotations of feature space - the Mahalanobis distance. In this formulation, the squared distance between vectors x_i and the x_j is compute as

$$d(x_i, x_j) = (x_i - x_j)^\top \mathbf{M}(x_i - x_j), \tag{2.1}$$

where $\mathbf{M}$ is a positive semidefinite matrix. In the following paragraphs, we explain different strategies to efficiently compute $\mathbf{M}$.

One of the first works to address person re-identification as a distance metric learning was proposed by Wei-Shi et al. [2011], where the authors maximize the probability of true pairs having smaller distances than wrong pairs using a costly optimization framework. Differently, in KISSME [Koestinger et al., 2012], the authors assume that the difference between samples are drawn from a Gaussian distribution to efficiently compute a Mahalanobis distance using log likelihood ratio test. Actually, KISSME is a two-step method that first needs to project the data into a low-dimensional subspace that eliminates dimension correlations to them properly learn the metric distance. Therefore, Liao et al. [2015] learn a better Mahalanobis matrix considering the subspace and distance learning in a unique framework. Differently, Yang et al. [2016] learn an improved distance function considering not only the differences between samples as well the commonness - sum of two samples - have a Gaussian distribution. However, these methods are based on strong assumptions about the distribution of the data, which are not always true.

One possible solution consists in considering more general metric learning approaches. For instance, in WARCA [Jose and Fleuret, 2016], the authors learn a global Mahalanobis distance in a low-dimensional subspace computed using orthonormal regularization. Differently, NLML [Huang et al., 2015] learns multiple local and nonlinear metric distances using neural networks and large margin optimization and

MLAPG [Liao and Li, 2015] used a logistic metric learning model with an asymmetric weighting strategy to address the imbalance between positive and negative samples.

Deep learning-based methods have been also explored in the distance metric learning for person re-identification. The idea consists in learning the feature representation and metric distance in an end-to-end framework. For instance, in the DeepML [Yi et al., 2014], the authors used a siamese network to match images captured by different cameras based on a simple cosine distance, while Zheng et al. [2017b] employed verification and identification losses. Similarly, Shi et al. [2015] applied the same architecture to learn a constrained Mahalanobis distance, while Wang et al. [2016] included a triplet loss function.

Triplet loss is widely used in the person re-identification problem. For instance, MultiCNN [Cheng et al., 2016] applied the triplet loss function with a margin to learn local and global feature representations and Ding et al. [2015] used triplet units in the deep relative distance comparison framework. Similarly, McLaughlin et al. [2016a] employed triplet loss jointly with pose, attributes and identity information in multi-task framework that regularizes parameters and avoids overfitting. In Hermans et al. [2017], the authors show that the triplet loss can be successful applied in person Re-ID with the proper selection of architectures and hard-negative mining. More recently, Chen et al. [2017] proposed a quadruplet loss with a larger inter-class variation and a smaller intra-class variation than the triplet loss. The main drawback of these methods is that they need a large amount of labeled data for each pair of cameras to learn a proper matching function, which is unrealistic in real-world scenarios.

Subspace Learning

Subspace learning methods have been widely employed in supervised Re-ID approaches. The idea is to compute a linear or nonlinear mapping function into a low-dimensional subspace where the matching between images captured by different cameras is successful even when using a simple cosine distance [Apurva and Shah., 2014; Zheng et al., 2016]. These methods are closely related to the metric learning approaches, as we can factorize the $\mathbf{M}$ ($\mathbf{M} = L^\top L$) and, then, Equation 2.1 corresponds to a simple euclidean distance in a learned common subspace.

The first work to address person re-identification as a subspace learning problem was proposed by An et al. [2013b], where the authors used Canonical Correlation Analysis (CCA) to learn projection matrices that maximize the correlation of feature descriptors captured by a camera pair. After that, Liao et al. [2015] learned a low-dimensional subspace using cross-view quadratic discriminant analysis (XQDA) and local Fisher discriminant analysis (LFDA) was used in [Pedagadi et al., 2013]. More importantly, in [Lisanti et al., 2014; Xiong et al., 2014], improved results are achieved using a nonlinear mapping before computing the projection matrices. For instance, Lisanti et al. [2014] proposed a nonlinear extension of CCA method - Kernel CCA (KCCA) - obtaining improved results in the evaluated datasets. In fact, the drastic appearance changes indicate a strong nonlinear behaviour of feature descriptors when captured by different cameras. Nonetheless, these methods learn a common subspace for each pair of cameras and, therefore, are not scalable to scenarios with many surveillance cameras. In addition, as they estimate covariance matrices using small number of samples, they need to include regularization parameters that are dataset specific and require a careful fine-tunning.

In this work, we also address person re-identification as a nonlinear subspace learning problem. Nonetheless, differently from previous works, we propose a hierarchical nonlinear subspace learning model that can handle with multiple cameras efficiently (Section 3.2). Furthermore, the proposed KPLS for subspace learning is robust to the small-sample-size problem and does not require any regularization.

Information Retrieval

There are also efforts to close the gap between person Re-ID and information retrieval. In this case, the goal is to explore strategies from information retrieval problems as re-ranking, ranking aggregation and hashing methods to improve the obtained ranking lists.

An et al. [2013a] address person re-identification as a re-ranking problem using detection of soft-biometric attributes (e.g, short hair) and Liu et al. [2013] use human strong and/or weak feedback in a post-rank optimization framework. To reduce the human effort, unsupervised post-rank optimization methods have been proposed using a least squares regression model with manifold regularization [Ma and Li, 2015] or content and context information [Garcia et al., 2015; Zhong et al., 2017]. For instance, Zhong et al. [2017] proposed a successful re-ranking approach based on the aggregation of content (i.e. feature representation) and context (i.e. nearest neighbors) distances.

Some works have proposed a combination of multiple experts at decision level using structural learning framework [Liu et al., 2015a; Paisitkriangkrai et al., 2015] or ranking aggregation [Prates and Schwartz, 2015]. More recently, Deep Ranking [Chen et al., 2016b] addressed Re-ID as a ranking problem to learn the feature representation and ranking function in a unique deep learning framework, while Liu et al. [2018c] proposed an end-to-end strategy that learns a discriminative binary coding using adversarial training to discriminate between real-valued and binary embedding.

Unsupervised Learning

The main drawback of the discussed methods is that they depend on the acquisition of labeled training data at each camera, which restricts the applicability in real-world scenario. Therefore, some works addressed person Re-ID as unsupervised problem.

In [Kodirov et al., 2015a], the authors iteratively learned a common dictionary representation and sparse coding for a pair of cameras using unlabeled training data. To include cross-view discriminative information, they used a graph Laplacian regularization term that keeps similar images close to each other in the learned dictionary representation. Differently, Kodirov et al. [2016] learned the graph representation and dictionary in a unique framework that uses ℓ_1-regularized graph to alleviate outliers problem and Lisanti et al. [2015a] used an iterative sparse representation method with adaptive weighting strategies to successfully rank gallery images based on the reconstruction error. More recently, Fan et al. [2018] proposed a clustering and sample selection algorithm that iteratively updates a CNN model based on an unlabeled database.

Despite these interesting works, the matching rates for unsupervised Re-ID are by a large margin inferior to the supervised scenarios. Therefore, in this work, we focus on the supervised person re-identification problem.

2.2 Multiple-Shot Person Re-Identification

In most of the real-world surveillance applications, we can collect multiple images of the same person when walking through the cameras field-of-view. In fact, by using multiple images of the same person in each surveillance camera, we can capture the same person in different poses and illumination conditions. Therefore, several works have been proposed in literature to transform the additional data provided by multiple images into a useful information to match individuals in a pair of cameras [Apurva and Shah., 2014; Zheng et al., 2016].

One of the main advantages of using multi-shot scenarios consists in the large amount of information available for each individual. Thus, some works employ all the images captured for each individual to perform person re-identification in supervised [Li et al., 2015b] or unsupervised scenarios [Khan and Brémond, 2016; Khan and Bremond, 2016]. For instance, Li et al. [2015b] use all the images available for each individual in training set to learn random forests in a low-dimensional subspace obtained by random projections. They showed that using all the available images of a given probe,

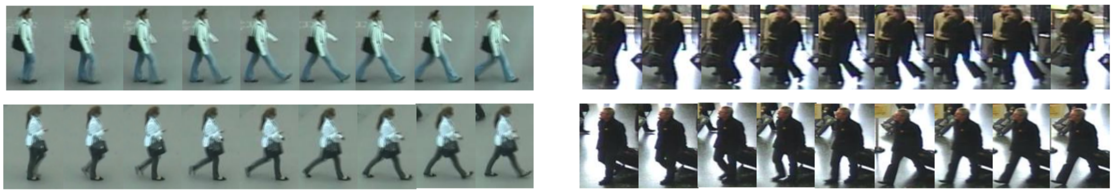

Figure 2.3: Ambiguity problem in multi-shot datasets. Each column shows images of different datasets, while in the rows we have different persons that are dressing similarly. Based on these images, we can notice that ambiguity is still a difficult problem when dealing with multiple images. Illustration taken from [You et al., 2016].

they are able to select the most discriminative trees and improve the obtained results. Differently, Khan and Bremond [2016] considered multiple images of the same person as positive image pairs to learn a metric distance in an unsupervised manner, while in Khan and Brémond [2016], the authors used a Multi Channel Appearance Mixture to model appearance variations inside each camera as a Gaussian Mixture Model (GMM). However, as these methods employed all the images available or just randomly selected some images, they are introducing noise into the matching models due to the high intra-camera appearance variability.

Some works tackled the appearance variability problem considering only a subset of images of the same person when captured by different cameras. Therefore, the main problem consists in finding a criterion to select the more reliable images for each camera pair. For instance, Karanam et al. [2015] restricted the number of gallery images used to represent a given probe image based on block-sparsity criterion. Then, they ranked the gallery images based on the reconstruction error. Li et al. [2015a] proposed an iterative procedure of hierarchical cluster and subspace learning that aims at selecting the most discriminative states of each individual before learning a metric distance, while Yang et al. [2011] performed key frame selection based on variations of color and pose information. Differently, in Harandi et al. [2012], the authors used similar and dissimilar pose images to learn two binary classifiers that boosted the recognition rate and Cho and Yoon [2016] computed camera intrinsic and extrinsic parameters to estimate four different poses (left, right, frontal and back) that are used when learning pose-aware matching functions.

The main drawback of multi-shot methods is that they require much more storage and computational cost when compared to single-shot scenario. Besides, the increased computational cost is not followed by a proportional gain in the matching performance. This occurs because problems that are present when dealing with single images, as the ambiguity between individuals, are not completely solved when using multiple images of the same individuals, as illustrated in Figure 2.3. In fact, You et al. [2016]

experimentally demonstrated that these problems become even harder to solve.

2.3 Video-Based Person Re-Identification

Similarly to the multi-shot, video-based person Re-ID also overcomes some common single-shot person Re-ID problems as pose, occlusion and illumination conditions. More importantly, video provides spatiotemporal cues helpful to identify individuals in low-resolution images and without the suspect collaboration. For instance, the gait information - the way a person walks - can be captured using spatiotemporal features and is important to match and align video sequences of the same person captured by different cameras [Liu et al., 2015b]. Therefore, in this section, we discuss the video-based person Re-ID problem focusing on the spatiotemporal information.

Early works on gait recognition are based on binary silhouette extraction to capture different stages of a walking cycle [Liu et al., 2015c]. Then, based on these cycles, these methods temporally align different video sequences in such a way that it is possible to describe and match videos based on gait information extracted from silhouette images or using the original RGB images. Examples of spatiotemporal descriptors include the Image Self-Similarity Plot [BenAbdelkader et al., 2004], Gait Energy Image [Han and Bhanu, 2006] and the Space-Time Interest Points [Kusakunniran, 2014]. Nonetheless, these descriptors are view-dependent and, then, their performance reduces when dealing with multiple cameras (e.g. person re-identification) [Bashir et al., 2010].

Gait matching across different cameras is one the main challenges to face when applying gait recognition methods in real-world applications [Bashir et al., 2010]. In one of the first works to tackle this problem, Bashir et al. [2010] estimated the camera viewpoint and maximized the correlation between gait-based feature descriptors extracted from different camera viewpoints in a common and low-dimensional subspace learned using Canonical Correlation Analysis (CCA). However, as they used only gait information, they discarded appearance features crucial to disambiguate individuals.

To combine gait and appearance information, Liu et al. [2015b] employed gait and appearance descriptors to learn a unique Metric Learning to Rank (MLR) model that explores the complementarity of gait and appearance information. Similarly, Kan Liu and Huang [2015] also used gait information to both discriminate and temporally align video sequences. However, instead of using the entire body information, they focused on spatially aligned cuboid regions (body-action units) from where they extract feature descriptors. Differently, Gao et al. [2016b] compute walking cycles by detecting displacements of lowest body parts regions estimated using superpixels. Thus, they could divide a walking cycle into segments and represent each segment using temporal pooling. Despite the interesting results, these methods depend on the correct estimative of the walking cycles, which is a very challenging task when we consider real-world scenarios with heavy occlusion and cluttered backgrounds.

Some works improved temporal alignment using complementary information or including a fragment selection criterion in the learned framework. For instance, Kawai et al. [2012] learn multiple matching models, one for each pose, and explored gait to both discriminate and synchronize video sequences. Differently, Wang et al. [2014] compute the fragments importance and the matching function in a unique multi-task framework that uses a pool of fragments. In Zhu et al. [2016], the authors add in the metric learning a criterion to minimize the intra-video covariance while discriminating between different videos. Similarly, You et al. [2016] learned discriminative features for cross-view matching using average pooling in a top-push distance-learning model.

Some works tackled the problem of learning a single representation for an entire video sequence using Recurrent Neural Network (RNN) architectures. Yan et al. [2016] used a simple pipeline to combine frame-wise handcrafted descriptors with dynamic information available in the entire sequence using Long Short-Term Memories (LSTMs). Differently, McLaughlin et al. [2016b] learned the feature representation and temporal information in a unique framework that combines CNNs, RNNs and siamese network. Since they use the image and optical flow information, they are able to correlate spatiotemporal features from different stages of a video sequence. Recently, Liu et al. [2018a] achieved improved results using a two-stream convolutional architecture that learns motion contextual information instead of using optical flow.

2.4 Complementary Information for Person Re-Identification

Only few works have addressed person re-identification using multimodal data (e.g. depth and thermal images). Multimodal data provide complementary information to appearance features that can be useful when matching images from different cameras [Mogelmose et al., 2013]. For instance, Pala et al. [2015] addressed person re-identification using a depth camera to extract nine anthropometric measures that are more robust to pose changes, such as the person's height. Differently, Mogelmose et al. [2013] proposed a tri-modal method to combine RGB, depth and thermal data, as illustrated in Figure 2.4. Considering calibrated thermal and RGB-D cameras, they used depth information to compute foreground mask from where RGB and thermal infor-

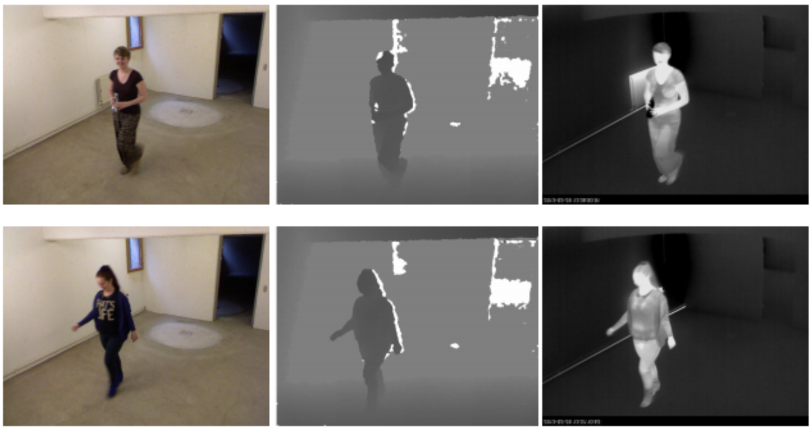

Figure 2.4: Complementary sensor data for person re-identification. Left, middle and right are the RGB, depth and thermal images, respectively. Based on these images, it is possible to observe that these sensors are capturing different information that can be combined to improve person re-identification results. Illustration taken from [Mogelmose et al., 2013].

mation are extracted. Despite the interesting results, these methods have a restricted applicability due to the combination of high cost and short-range sensors.

Chapter 3

Subspace Learning

As mentioned in the previous chapters, one approach to correctly match images captured by different surveillance cameras consists in learning a common, usually low-dimensional, subspace that correlates feature descriptors of individuals in training set captured by different surveillance cameras. We can roughly divide subspace learning literature in linear and nonlinear models. The first avoids the computation of costly kernel functions and, therefore, is more scalable. Nonetheless, improved results are reported when nonlinearly mapping the feature descriptors is performed before computing the linear projections. In fact, nonlinear subspace learning models are simple and straightforward extensions that can be efficiently computed using the "kernel trick" (see Figure 3.1).

In this section, we present the proposed nonlinear subspace learning models: the *Kernel PLS for Subspace Learning* (Section 3.1), the *Kernel Hierarchical PCA* (Section 3.2) and the *Kernel Multiblock PLS* (Section 3.3). Finally, Section 3.4 outlines the main aspects of these methods.

Notation. We use the following notation in the description. Bold lower-case letters denote column vectors and bold upper-case letters denote matrices (e.g., $\mathbf{a}$ and $\mathbf{A}$, respectively). We represent the ith image from probe and gallery cameras, as $\mathbf{p}_i$ and $\mathbf{g}_i \in \mathbb{R}^d$, respectively, where d denotes the dimension of the feature space. Without loss of generality, we assume that n testing images from probe camera constitute the probe set $\mathbf{P} \in \mathbb{R}^{n \times d}$ and n testing images from gallery camera represent the gallery set $\mathbf{G} \in \mathbb{R}^{n \times d}$. Similarly, the set of all n training images from probe and gallery cameras compose the matrices $\mathbf{X}_p$ and $\mathbf{X}_g \in \mathbb{R}^{n \times d}$, respectively.

When describing the nonlinear extensions, we use $\phi(.)$ to denote a nonlinear mapping function of input variables to a feature space $\mathcal{F}$, i.e., $\phi : x_i \in \mathbb{R}^d \rightarrow \phi(x_i) \in \mathcal{F}$

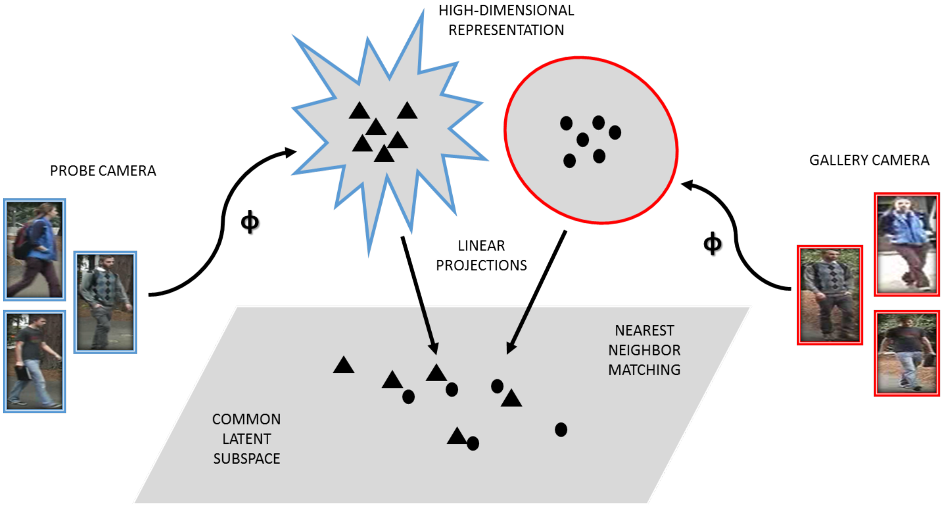

Figure 3.1: Subspace Learning. First, feature descriptors from training images captured from probe and gallery cameras are mapped to a high-dimensional feature space where linear projections matrices are learned using some criterion. Then, probe and gallery are projected onto the learned subspace and matched using the a simple cosine distance. Illustration from [Prates and Schwartz, 2016].

and, $\mathbf{\Phi}_p$ and $\mathbf{\Phi}_g$ are the resulting matrices after nonlinearly mapping $\mathbf{X}_p$ and $\mathbf{X}_g$, respectively. In this dissertation, we apply the "kernel trick" to avoid explicitly mapping the data into a high-dimensional space substituting the cross-product by $K = \mathbf{\Phi}\mathbf{\Phi}^\top$, where $K \in \mathbb{R}^{n\times n}$ is the *kernel Gram matrix*. Particularly, we define the *kernel Gram matrices* $\mathbf{K}_p$ as

$$\mathbf{K}_p = \begin{bmatrix} \mathbf{k}_0^0 & \mathbf{k}_1^0 & \cdots & \mathbf{k}_n^0 \\ \mathbf{k}_0^1 & \mathbf{k}_1^1 & \cdots & \mathbf{k}_n^1 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{k}_0^n & \mathbf{k}_1^n & \cdots & \mathbf{k}_n^n \end{bmatrix}, \tag{3.1}$$

where $\mathbf{k}_j^i \in \mathbb{R}$ represents the computation of $\phi(x_i)\phi(x_j)^\top$ and, x_i and x_j are training samples from the probe camera. Similarly, we compute the matrix $\mathbf{K}_g$ using training samples from the gallery camera. Finally, we define the row-vector representation $\mathbf{k}_i^g \in \mathbb{R}^{1\times n}$, whose element $\mathbf{k}_i^g(j)$ corresponds to the kernel function applied using feature descriptors of the ith gallery image $\mathbf{g}_i$ and the the jth training sample at the gallery camera ($\mathbf{x}_g^j \in \mathbf{X}_g$). Similarly, for the ith probe image $\mathbf{p}_i$, we compute its vector $\mathbf{k}_i^p \in \mathbb{R}^{1\times n}$ using training images at probe camera ($\mathbf{X}_p$).

3.1 Kernel PLS for Subspace Learning

In this section, we use an adaptation of the classical Partial Least Squares (PLS) to learn a low-dimensional subspace that maximizes the covariance between feature descriptors extracted from probe and gallery cameras [Rosipal and Krämer, 2006], which we coined as PLS for Subspace Learning.

PLS for Subspace Learning is a linear model and, improved results have been reported in literature when addressing appearance changes caused by different camera conditions using a nonlinear model [Lisanti et al., 2014; Xiong et al., 2014; Ma et al., 2012c]. Therefore, we propose a nonlinear extension of the PLS for Subspace Learning, the *Kernel PLS for Subspace Learning*. To the extent of our knowledge, we are the first work to propose a Kernel PLS for Subspace Learning.

In the next sections, we first present the PLS for Subspace Learning model (Section 3.1.1) to then introduce the necessary adaptations to obtain the proposed Kernel PLS for Subspace Learning (Section 3.1.2).

3.1.1 Partial Least Squares for Subspace Learning

PLS is a statistical method that computes the regression between independent ($\mathbf{X}$) and dependent ($\mathbf{Y}$) variables using score vectors $\mathbf{t}$ and $\mathbf{u}$ that are computed by means of weight vectors $\mathbf{w}$ and $\mathbf{c}$ such that

$$[cov(\mathbf{t}, \mathbf{u})]^2 = \max_{|\mathbf{w}|=|\mathbf{c}|=1} [cov(\mathbf{Xw}, \mathbf{Yc})]^2, \tag{3.2}$$

where $cov(\mathbf{t}, \mathbf{u})$ is the sample covariance between $\mathbf{t}$ and $\mathbf{u}$. This problem is solved by [Wold, 1985] using the nonlinear iterative partial least squares (NIPALS) method, a robust procedure to solve singular value decomposition problems.

Despite being widely used at chemometrics [Wold, 1985; Westerhuis et al., 1998], only recently PLS has attracted the attention of computer vision researches [Schwartz et al., 2009; Schwartz and Davis, 2009a]. For instance, Schwartz and Davis [2009a] used PLS regression models to weight high-dimensional feature descriptors based on their discriminative power in an one-against-all approach. However, they assume that subtle characteristics present in the gallery set will also be present in images captured by the probe camera, which disregards the camera transition problem.

In this section, we present the PLS for Subspace Learning model that is the PLS model indicated to deal with symmetric blocks of data (e.g. multiple camera-views) [Rosipal and Krämer, 2006]. Similarly to Canonical Correlation Analysis (CCA), PLS for Subspace Learning is more suitable for modeling relationship between blocks of variables in opposition to prediction purposes. However, differently from CCA, PLS for Subspace Learning is robust to the small-sample-size as it does not require the estimation of covariance matrices in the original feature space.

In Algorithm 1, we present the algorithm to obtain the latent factors for the PLS for Subspace Learning model. It is a simple modification of the conventional NIPALS algorithm that performs rank-one deflation of each block matrices using its associated score and loadings. Therefore, instead of performing regression, we are only discovering components on $\mathbf{X}_p$ that are also relevant for $\mathbf{X}_g$. To compute the projections into a low-dimensional subspace we use the approximation of $\mathbf{X}_p = \mathbf{T}\mathbf{W}_p^\top$ and the fact that matrices $\mathbf{T}$ and $\mathbf{W}_p$ are orthonormal to obtain the following equations

$$\begin{aligned} \mathbf{W}_p &= \mathbf{X}_p^\top \mathbf{T} \\ \mathbf{T} &= \mathbf{X}_p \mathbf{W}_p (\mathbf{W}_p^\top \mathbf{W}_p)^{-1}, \end{aligned} \tag{3.3}$$

and, combing both equations, we achieve

$$\mathbf{W}_p = \mathbf{X}_p^\top \mathbf{T} (\mathbf{T}^\top \mathbf{X}_p \mathbf{X}_p^\top \mathbf{T})^{-1}, \tag{3.4}$$

where matrix $\mathbf{T} \in \mathbb{R}^{n \times f}$ contains the computed score vectors ($\mathbf{t} \in \mathbb{R}^n$) for each factor in Algorithm 1 and the number of factors f is a positive integer. Similarly, we can compute the projection $\mathbf{W}_g \in \mathbb{R}^{d \times f}$ using the score vectors $\mathbf{u} \in \mathbb{R}^n$ and the matrix $\mathbf{X}_g$.

In the test stage, we use these projection matrices to compute the low-dimensional representation for the i*th* testing image from probe and gallery cameras as

$$\begin{aligned} \mathbf{t}_i^p &= \mathbf{p}_i \mathbf{W}_p \quad \text{and} \\ \mathbf{t}_i^g &= \mathbf{g}_i \mathbf{W}_g. \end{aligned} \tag{3.5}$$

We assume that the learned subspace handles the camera transition problem. Then, we perform the matching using the low-dimensional representation of probe ($\mathbf{t}_i^p$) and gallery images ($\mathbf{t}_i^g$) and a simple cosine distance.

In the next section, we present a straightforward modification of Algorithm 1 to tackle the nonlinearity of appearance changes that occurs in the person re-identification problem.

Algorithm 1: Partial Least Squares for Subspace Learning.

```
   input : $\mathbf{X}_p$, $\mathbf{X}_g$ matrices and the number of factors ($f$)
1  randomly initialize $\mathbf{u}$ and $\mathbf{u}_0$
2  for $i$=1 to $f$ do
3      while $\|\boldsymbol{u} - \boldsymbol{u}_0\| > \varepsilon$ do
4          $\mathbf{u}_0 \leftarrow \mathbf{u}$
5          $\mathbf{w} = \mathbf{X}_p^\top \mathbf{u}$
6          $\mathbf{t} = \mathbf{X}_p \mathbf{w}, \quad \mathbf{t} \leftarrow \frac{\mathbf{t}}{\|\mathbf{t}\|}$
7          $\mathbf{c} = \mathbf{X}_g^\top \mathbf{t}$
8          $\mathbf{u} = \mathbf{X}_g \mathbf{c}, \quad \mathbf{u} \leftarrow \frac{\mathbf{u}}{\|\mathbf{u}\|}$
9      end
10     $\mathbf{X}_p \leftarrow \mathbf{X}_p - \mathbf{t}\mathbf{t}^\top \mathbf{X}_p, \quad \mathbf{X}_g \leftarrow \mathbf{X}_g - \mathbf{u}\mathbf{u}^\top \mathbf{X}_g$
11 end
```

3.1.2 Proposed Method

In this section, we present the proposed Kernel PLS for Subspace Learning and provide the algorithm to compute the projection matrices.

The proposed Kernel PLS for Subspace Learning is closely related to the method proposed by Lisanti et al. [2014] that uses Kernel Canonical Correlation Analysis method (KCCA) in the person re-identification problem. However, instead of maximizing the correlation, the proposed method considers the covariance in the objective function. Rewriting Equation 3.2 to include the relation between correlation and covariance statistics, we obtain Equation 3.6. Therefore, the main advantage of the proposed method consists in the fact that maximizing the covariance we are optimizing the correlation as well capturing the variance at both input and output spaces [Sharma and Jacobs, 2011].

$$[cov(\mathbf{t}, \mathbf{u})]^2 = \max_{|\mathbf{w}|=|\mathbf{c}|=1} [var(\mathbf{X}\mathbf{w}) corr(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{c}) var(\mathbf{Y}\mathbf{c})]^2. \tag{3.6}$$

To nonlinearly correlate feature descriptors of the same person from probe and gallery cameras, we use matrices $\mathbf{K}_p$ and $\mathbf{K}_g$, which are computed using labeled image pairs from both cameras. Specifically, we combine rows 6 and 7 from Algorithm 1, obtaining

$$\begin{aligned} \mathbf{t} &= \mathbf{X}_p \mathbf{X}_p^\top \mathbf{u} \\ \mathbf{t} &= \mathbf{K}_p \mathbf{u}. \end{aligned} \tag{3.7}$$

Likewise, from rows 8 and 9, we compute the score $\mathbf{u}$. Thus, using Algorithm 2 we obtain the score vectors $\mathbf{T}$ and $\mathbf{U}$.

Algorithm 2: KPLS for Subspace Learning.

```
input : $\mathbf{K}_p$, $\mathbf{K}_g$ matrices and the number of factors ($f$)
randomly initialize $\mathbf{u}$ and $\mathbf{u}_0$
for $i$=1 to $f$ do
    while $\|\boldsymbol{u} - \boldsymbol{u}_0\| > \varepsilon$ do
        $\mathbf{u}_0 \leftarrow \mathbf{u}$
        $\mathbf{t} = \mathbf{K}_p\mathbf{u}, \quad \mathbf{t} \leftarrow \frac{\mathbf{t}}{\|\mathbf{t}\|}$
        $\mathbf{u} = \mathbf{K}_g\mathbf{t}, \quad \mathbf{u} \leftarrow \frac{\mathbf{u}}{\|\mathbf{u}\|}$
    end
    $\mathbf{K}_p \leftarrow (\mathbf{I} - \mathbf{t}\mathbf{t}^\top)\mathbf{K}_p(\mathbf{I} - \mathbf{t}\mathbf{t}^\top), \quad \mathbf{K}_g \leftarrow (\mathbf{I} - \mathbf{u}\mathbf{u}^\top)\mathbf{K}_g(\mathbf{I} - \mathbf{u}\mathbf{u}^\top)$
end
```

To project new samples onto the learned low-dimensional space, we define projection matrices $\mathbf{W}_p$ and $\mathbf{W}_g \in \mathbb{R}^{n \times d}$ that relate score vectors $\mathbf{T}$ and $\mathbf{U}$ with matrices $\mathbf{K}_p$ and $\mathbf{K}_g$, respectively, such that

$$\begin{aligned} \mathbf{T} &= \mathbf{K}_p\mathbf{W}_p \text{ and} \\ \mathbf{U} &= \mathbf{K}_g\mathbf{W}_g, \end{aligned} \tag{3.8}$$

where projection matrices $\mathbf{W}_p$ and $\mathbf{W}_g$ are computed as

$$\begin{aligned} \mathbf{W}_p &= \mathbf{T}(\mathbf{T}^\top\mathbf{K}_p\mathbf{T})^{-1} \text{ and} \\ \mathbf{W}_g &= \mathbf{U}(\mathbf{U}^\top\mathbf{K}_g\mathbf{U})^{-1}. \end{aligned} \tag{3.9}$$

For the i*th* gallery image ($\mathbf{g}_i$) and the j*th* probe image ($\mathbf{p}_j$), we first compute its kernel representation $\mathbf{k}_i^g$ and $\mathbf{k}_j^p$, respectively. Then, we obtain its low-dimensional representation as

$$\begin{aligned} \mathbf{t}_i^g &= \mathbf{k}_i^g\mathbf{W}_g \quad \text{and} \\ \mathbf{t}_j^p &= \mathbf{k}_j^p\mathbf{W}_p. \end{aligned} \tag{3.10}$$

Finally, we use the cosine between $\mathbf{t}_j^p$ and $\mathbf{t}_i^g$ as the similarity between $\mathbf{p}_j$ and $\mathbf{g}_i$.

3.2 Kernel Hierarchical PCA

Despite being widely employed in chemometrics and biochemical process monitoring [Westerhuis et al., 1998], Hierarchical PCA (HPCA) is not well known by the computer vision community. Therefore, we first introduce the Hierarchical PCA method (Section 3.2.1) and, then, present the novel Kernel HPCA method that efficiently computes the consensus projections into a common subspace (Section 3.2.2). It is important to highlight that due to the hierarchical formulation, these models learn a unique subspace for the entire camera network. It is crucial for realistic person re-identification systems as the number of pairwise models grow quadratically with the number of cameras.

3.2.1 Hierarchical PCA (HPCA)

Principal Component Analysis (PCA) is a statistical method for unsupervised dimensionality reduction that linearly project the data to a common latent subspace that explains most of the variance among samples [Schölkopf et al., 1998]. Classical applications of PCA in computer vision include eigenfaces [Turk and Pentland, 1991], eigengait [BenAbdelkader et al., 2001] and eigentracking [Black and Jepson, 1998]. However, while many computer vision problems present a strong nonlinear behavior, PCA is a linear model. Therefore, Schölkopf et al. [1998] proposed an extension of PCA that nonlinearly maps the data into a high-dimensional feature space and computes the principal components using a "kernel trick", which is known as Kernel PCA (KPCA).

Despite the success of PCA, there are few works that address person Re-ID problem using PCA-based methods [Martinel and Micheloni, 2014; Yang et al., 2011]. Martinel and Micheloni [2014] used PCA to learn a low-dimensional representation of image dissimilarities and, then, train a binary classifier using equivalence constraints. Differently, Yang et al. [2011] proposed an unsupervised approach that learns a low-dimensional representation for each gallery images using KPCA. In fact, the small number of samples in each class and the camera transition problem compose a challenging scenario to conventional PCA-based methods in the person Re-ID problem.

A possible solution to tackle the person Re-ID problem is to use multiblock multivariate models [Westerhuis et al., 1998]. These models have been employed when additional information is available for grouping variables in a meaningful blocks (e.g., different camera views). For instance, Hierarchical PCA (HPCA) [Westerhuis et al., 1998] is a multiblock extension of PCA that seeks for a consensus direction among all blocks. HPCA is useful when there is a meaningful division of data into blocks. A typical example corresponds to multiple measures of the same object (e.g., images of the same subject captured by multiple cameras).

To enable a better understanding, we present a two-block NIPALS algorithm and the description of this method. However, it is important to highlight that these methods can be easily adapted to handle multiple data blocks, not being limited to two. To the best of our knowledge, this is the first work to propose a common subspace learning model that can efficiently deal with more than two cameras simultaneously.

Algorithm 3 computes the HPCA model. It consists of the following stages. Let $\mathbf{X}_p$ and $\mathbf{X}_g$ be two blocks of data, HPCA starts with a super score $\mathbf{t} \in \mathbb{R}^n$, which is an initial guess of consensus and can be a column of these blocks. Then, this super score is regressed on blocks $\mathbf{X}_p$ and $\mathbf{X}_g$ to compute the loadings $\mathbf{w}_p$, $\mathbf{w}_g \in \mathbb{R}^m$. These loadings are employed to compute the block scores $\mathbf{s}_p$ and $\mathbf{s}_g \in \mathbb{R}^n$ for probe and gallery cameras, respectively. Then, the normalized scores constitute the super block $\mathbf{S} \in \mathbb{R}^{n\times 2}$. Finally, the super score $\mathbf{t}$ is regressed on the super block $\mathbf{S}$ to obtain the super weight $\mathbf{w} \in \mathbb{R}^m$, which is used to update the value of super score $\mathbf{t}$. This process repeats until the convergence of super score $\mathbf{t}$ to a predefined precision. Thus, while block scores ($\mathbf{s}_p$ and $\mathbf{s}_g$) are resultant of block variables, the super score $\mathbf{t}$ is derived using all variables. After the convergence, block variables $\mathbf{X}_p$ and $\mathbf{X}_g$ are deflated to obtain a new score vector at the next iterations until reaching the expected number of factors f, which is a parameter of the model.

Algorithm 3: Hierarchical PCA (HPCA)

```
input  : $\mathbf{X}_p$, $\mathbf{X}_g$ matrices and the number of factors ($f$)
randomly initialize $\mathbf{t}$ and $\mathbf{t}_0$
for $i$=1 to $f$ do
    while $\|\boldsymbol{t} - \boldsymbol{t}_0\| > \varepsilon$ do
        $\mathbf{t}_0 \leftarrow \mathbf{t}$
        $\mathbf{w}_p = \mathbf{X}_p^\top \mathbf{t}$
        $\mathbf{s}_p = \mathbf{X}_p \mathbf{w}_p \quad \mathbf{s}_p \leftarrow \frac{\mathbf{s}_p}{\|\mathbf{s}_p\|}$
        $\mathbf{w}_g = \mathbf{X}_g^\top \mathbf{t}$
        $\mathbf{s}_g = \mathbf{X}_g \mathbf{w}_g \quad \mathbf{s}_g \leftarrow \frac{\mathbf{s}_g}{\|\mathbf{s}_g\|}$
        $\mathbf{S} \leftarrow [s_p, s_g]$
        $\mathbf{w} = \mathbf{S}^\top \mathbf{t}$
        $\mathbf{t} = \mathbf{S}\mathbf{w} \quad \mathbf{t} \leftarrow \frac{\mathbf{t}}{\|\mathbf{t}\|}$
    end
    $\mathbf{X}_p = \mathbf{X}_p - \mathbf{t}\mathbf{w}_p^\top, \quad \mathbf{X}_g = \mathbf{X}_g - \mathbf{t}\mathbf{w}_g^\top$
end
```

In this dissertation, we assume that the learned common subspace deals with appearance changes caused by the data capture executed by different cameras. Therefore, it is possible to compute the similarity between individuals $\mathbf{g}_i$ and $\mathbf{p}_j$ using their respective super scores $\mathbf{t}_i^g$ and $\mathbf{t}_j^p$ in a simple nearest neighbor method. Similarly to the previous methods, the equation for computing this super score can be obtained using the orthonormal constraint and the approximation of $\mathbf{X}_p = \mathbf{T}\mathbf{W}_p$ as

$$\mathbf{t}_j^p = \mathbf{p}_j \mathbf{X}_p^\top \mathbf{T}(\mathbf{T}^\top \mathbf{X}_p \mathbf{X}_p^\top \mathbf{T})^{-1}, \tag{3.11}$$

where $\mathbf{T} \in \mathbb{R}^{n \times f}$ is constructed storing the computed super scores $\mathbf{t}$ for each factor. Similarly, we can compute the super scores $\mathbf{t}_i^g$ for a testing sample $\mathbf{g}_i$ using $\mathbf{X}_g$ in Equation 3.11. Then, the matching between probe and gallery images is based on the cosine between super scores $\mathbf{t}_j^p$ and $\mathbf{t}_i^g$.

3.2.2 Proposed Method

The complex appearance changes in images of the same individual captured by distinct cameras and the high matching performance of recent nonlinear person Re-ID models [Lisanti et al., 2014; Xiong et al., 2014; Ma et al., 2012c] suggest that we can reach improved results using nonlinear subspace learning models. Therefore, in this section, we present a kernel extension to HPCA model, which we call *Kernel HPCA*. To the best of our knowledge, this is the first kernel extension to the multiblock PCA method.

Kernel HPCA relates data blocks nonlinearly with principal components to obtain block scores ($\mathbf{s}_p$ and $\mathbf{s}_g$). Therefore, our proposed method captures high-order correlation between input variables to learn the common latent subspace. Furthermore, we present an efficient derivation of NIPALS algorithm to iteratively compute the principal components (or factors) of Kernel HPCA.

We assume the nonlinear transformation of the input variables from probe and gallery cameras (blocks) as the *kernel Gram matrices* $\mathbf{K}_p$ and $\mathbf{K}_g$, respectively. Thus, combining rows 6 and 7 from Algorithm 3 and applying the nonlinear transformation, we obtain that

$$\mathbf{s}_p = \mathbf{K}_p \mathbf{t}. \tag{3.12}$$

Likewise, we rewrite rows 8 and 9 from Algorithm 3 as

$$\mathbf{s}_g = \mathbf{K}_g \mathbf{t}. \tag{3.13}$$

Then, we derive a rank-one approximation of *kernel Gram matrix* $\mathbf{K}_p$ as

$$\mathbf{K}_p \approx \mathbf{t}\mathbf{t}^\top \mathbf{K}_p. \tag{3.14}$$

An analogous rank-one approximation to *kernel Gram matrix* $\mathbf{K}_g$ results in

$$\mathbf{K}_g \approx \mathbf{t}\mathbf{t}^\top \mathbf{K}_g. \tag{3.15}$$

Algorithm 4: Kernel Hierarchical PCA (Kernel HPCA).

```
input : $\mathbf{K}_p$, $\mathbf{K}_g$ and the number of factors ($f$)
randomly initialize $\mathbf{t}$ and $\mathbf{t}_0$
for $i$=1 to $f$ do
    while $\|\boldsymbol{t} - \boldsymbol{t}_0\| > \varepsilon$ do
        $\mathbf{t}_0 \leftarrow \mathbf{t}$
        $\mathbf{s}_p = \mathbf{K}_p\mathbf{t}$,   $\mathbf{s}_p \leftarrow \frac{\mathbf{s}_p}{\|\mathbf{s}_p\|}$
        $\mathbf{s}_g = \mathbf{K}_g\mathbf{t}$,   $\mathbf{s}_g \leftarrow \frac{\mathbf{s}_g}{\|\mathbf{s}_g\|}$
        $\mathbf{S} \leftarrow [s_p, s_g]$
        $\mathbf{w} = \mathbf{S}^\top\mathbf{t}$
        $\mathbf{t} = \mathbf{S}\mathbf{w}$   $\mathbf{t} \leftarrow \frac{\mathbf{t}}{\|\mathbf{t}\|}$
    end
    $\mathbf{K}_p \leftarrow \mathbf{K}_p - \mathbf{t}\mathbf{t}^\top\mathbf{K}_p$,   $\mathbf{K}_g \leftarrow \mathbf{K}_g - \mathbf{t}\mathbf{t}^\top\mathbf{K}_g$
end
```

Algorithm 4 presents our Kernel HPCA method. Kernel HPCA is based on the NIPALS algorithm for the computation of Hierarchical PCA with the required modifications to efficiently handle the nonlinear transformation of input variables. Similarly to HPCA, Kernel HPCA computes an orthonormal matrix $\mathbf{T} \in \mathbb{R}^{m \times f}$ whose columns store the super scores $\mathbf{t}$ obtained at each iteration. Thus, using the nonlinear mapping, we compute the super score $\mathbf{p}_j$ for the j*th* image from probe camera using its kernel representation $\mathbf{k}_j^p$ as

$$\mathbf{t}_j^p = \mathbf{k}_j^p \mathbf{T}(\mathbf{T}^\top \mathbf{K}_p \mathbf{T})^{-1}. \tag{3.16}$$

Identically, we compute the super score $\mathbf{g}_i$ for the i*th* image from gallery camera using its kernel representation $\mathbf{k}_i^g$ as

$$\mathbf{t}_i^g = \mathbf{k}_i^g \mathbf{T}(\mathbf{T}^\top \mathbf{K}_g \mathbf{T})^{-1}. \tag{3.17}$$

Similarly to HPCA model, we assume that the learned subspace handles the camera transition problem and the direct comparison between samples from different blocks results in a high performance when using their super scores. However, due the nonlinear transformation in the input variables, we believe that the common subspace learned with Kernel HPCA method is able to handle more complex feature transitions. Experimental results (see Section 5.2) corroborate our hypothesis with great improvement when compared with its counterpart - the HPCA model.

3.3 Kernel Multiblock Partial Least Squares

In this section, we discuss the Multiblock PLS (Section 3.3.1) and, then, present the novel Kernel Multiblock PLS (Section 3.3.2). Similarly to the HPCA, the Multiblock PLS also learns projections into a low-dimensional subspace using the input data divided into multiblocks. Differently, it simultaneously learns a regression between latent vectors and response variables (e.g. identity labels) in the latent space.

3.3.1 Multiblock PLS (MBPLS)

The multiblock Partial Least Squares (MBPLS) is commonly used to isolate the impact of different measures on the output in chemometrics [Brás et al., 2005] and process monitoring and control [MacGregor et al., 1994]. For instance, in [Brás et al., 2005], the authors analyse a chemical process using multiblock PLS that considers as input near-infrared and mid-infrared sensors. To the best of our knowledge MBPLS has not been applied in the computer vision or image processing problems. Therefore, in the next paragraphs, we introduce the MBPLS.

Multiblock PLS is useful when the independent variable ($\mathbf{X}$) can be divided into meaningful blocks, such as multiple measures from the same object [Westerhuis et al., 1998]. These blocks are useful to compute the individual contribution in the response and improve the learned latent representation. MBPLS accomplishes that by learning block-specific projections that drive the different blocks of data to a common score vector $\mathbf{t}$ and, therefore, correlates them in a common subspace. More importantly, the score vector $\mathbf{t}$ is iteratively updated based on the covariance between the response vector and the block scores. This process is repeated for each factor after the deflation of data and response matrices. In the following, we present a brief mathematical description of the MBPLS model. For a more detailed discussion, please refer to work [MacGregor et al., 1994].

Let matrices $\mathbf{X}_p$ and $\mathbf{X}_g \in \mathbb{R}^{n \times d}$ be our data ($\mathbf{X}$) divided into two blocks and $\mathbf{Y} \in \mathbb{R}^{n \times l}$ be the matrix with the regression responses. MBPLS starts with a random guess for the latent score $\mathbf{t} \in \mathbb{R}^n$. Then, this score is regressed on blocks $\mathbf{X}_p$ and $\mathbf{X}_g$ to compute the loadings $\mathbf{w}_p$ and $\mathbf{w}_g \in \mathbb{R}^d$. Block variables are multiplied by the estimated loading vectors to compute the block-specific score vectors $\mathbf{s}_p$ and $\mathbf{s}_g \in \mathbb{R}^n$, which are normalized to unit length and concatenated in a super block $\mathbf{S} \in \mathbb{R}^{n \times 2}$. As these score vectors are computed aiming to approximate the score vector $\mathbf{t}$, a regression between $\mathbf{S}$ and $\mathbf{t}$ is performed to obtain the weight vector $\mathbf{w} \in \mathbb{R}^2$, where $\mathbf{w}_i$ corresponds to the importance of the ith data block in the regression model, and a single score vector $\mathbf{u} \in \mathbb{R}^n$ that combines information from multiple blocks. Finally, $\mathbf{t}$ is updated to be the result of the regression of $\mathbf{Y}$ into $\mathbf{u}$. This process repeats until the convergence of score $\mathbf{t}$ to a predefined precision. After the convergence, the input variables are deflated to obtain a different score vector at each iteration. This process repeats f times, where f is the number of latent scores – the only parameter of the MBPLS model.

Algorithm 5: Multiblock PLS (MBPLS)

```
input  : $\mathbf{X}_p$, $\mathbf{X}_g$, $\mathbf{Y}$ matrices and the number of factors ($f$)
randomly initialize $\mathbf{t}$ and $\mathbf{t}_0$
for $i$=1 to $f$ do
    while $\|\boldsymbol{t} - \boldsymbol{t}_0\| > \varepsilon$ do
        $\mathbf{t}_0 \leftarrow \mathbf{t}$
        $\mathbf{w}_p = \mathbf{X}_p^\top \mathbf{t}$
        $\mathbf{s}_p = \mathbf{X}_p \mathbf{w}_p \quad \mathbf{s}_p \leftarrow \frac{\mathbf{s}_p}{\|\mathbf{s}_p\|}$
        $\mathbf{w}_g = \mathbf{X}_g^\top \mathbf{t}$
        $\mathbf{s}_g = \mathbf{X}_g \mathbf{w}_g \quad \mathbf{s}_g \leftarrow \frac{\mathbf{s}_g}{\|\mathbf{s}_g\|}$
        $\mathbf{S} \leftarrow [s_p, s_g]$
        $\mathbf{w} = \mathbf{S}^\top \mathbf{t}$
        $\mathbf{u} = \mathbf{S}\mathbf{w} \quad \mathbf{u} \leftarrow \frac{\mathbf{u}}{\|\mathbf{u}\|}$
        $\mathbf{q} = \mathbf{Y}^\top \mathbf{u}$
        $\mathbf{t} = \mathbf{Y}\mathbf{q}/\mathbf{q}^\top \mathbf{q}$
    end
    $\mathbf{X}_p = \mathbf{X}_p - \mathbf{t}\mathbf{t}^\top \mathbf{X}_p, \quad \mathbf{X}_g = \mathbf{X}_g - \mathbf{t}\mathbf{t}^\top \mathbf{X}_g, \quad \mathbf{Y} = \mathbf{Y} - \mathbf{t}\mathbf{t}^\top \mathbf{Y}$
end
```

Let $\mathbf{W}_p \in \mathbb{R}^{d \times f}$ be a projection matrix obtained by storing the loading vector $\mathbf{w}_p$ computed at each iteration of Algorithm 5. Similarly, $\mathbf{T} \in \mathbb{R}^{n \times f}$ is matrix of scores $\mathbf{t}$ obtained at all the f iterations. Due to the normalization and deflation steps of MBPLS, $\mathbf{T}$ matrix is orthonormal ($\mathbf{T}^\top \mathbf{T} = \mathbf{I}$). Therefore, similarly to the previous approaches, the closed-form solution of $\mathbf{W}_p$ is obtained as

$$\begin{aligned} \mathbf{X}_p &= \mathbf{T}\mathbf{W}_p^\top, \\ \mathbf{W}_p &= \mathbf{X}_p^\top \mathbf{T}. \end{aligned} \tag{3.18}$$

Thus, we can compute the regression coefficients ($\beta_p \in \mathbb{R}^{d \times l}$) according to Bennett et al. [2003] as

$$\beta_p = \mathbf{W}_p(\mathbf{T}^\top \mathbf{X}_p \mathbf{W}_p)^{-1}\mathbf{T}^\top \mathbf{Y}. \tag{3.19}$$

Finally, combining Equations 3.18 and 3.19, we obtain the regression response $\hat{\mathbf{y}}_p$ for $\hat{\mathbf{x}}_p$ as

$$\begin{aligned} \hat{\mathbf{y}}_p &= \hat{\mathbf{x}}_p \beta_p, \\ \hat{\mathbf{y}}_p &= \hat{\mathbf{x}}_p \mathbf{X}_p^\top \mathbf{T}(\mathbf{T}^\top \mathbf{X}_p \mathbf{X}_p^\top \mathbf{T})^{-1}\mathbf{T}^\top \mathbf{Y}. \end{aligned} \tag{3.20}$$

Similarly, we can compute the regression response $\hat{\mathbf{y}}_g$ for a sample from the gallery camera ($\hat{\mathbf{x}}_g$) using $\mathbf{X}_g$ and $\mathbf{T}$ based on the Equation 3.20.

3.3.2 Proposed Method

In this section, we describe the proposed Kernel MBPLS that nonlinearly relates data blocks 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 efficiently compute the regression coefficients of the Kernel MBPLS.

Considering the *kernel Gram matrices* $\mathbf{K}_p$ and $\mathbf{K}_g$ computed using training samples from probe and gallery cameras, respectively. Thus, combining rows 6 and 7 from Algorithm 5 to consider the nonlinear mapping, we obtain that

$$\mathbf{s}_p = \mathbf{K}_p \mathbf{t}. \tag{3.21}$$

Similarly, we rewrite the deflation equations from row 16, in Algorithm 5, multiplying both sides by $\mathbf{X}_p^\top$ and applying the nonlinear mapping to obtain the rank-1 deflation as

$$\mathbf{K}_p = \mathbf{K}_p - \mathbf{t}\mathbf{t}^\top \mathbf{K}_p. \tag{3.22}$$

Algorithm 6 presents our Kernel MBPLS method that efficiently handles the nonlinear mapping of the multiblock data. Thus, employing the “kernel trick” in Equation 3.20, we obtain

$$\hat{\mathbf{y}}_p = \mathbf{k}_j^p \mathbf{T}(\mathbf{T}^\top \mathbf{K}_p \mathbf{T})^{-1}\mathbf{T}^\top \mathbf{Y}, \tag{3.23}$$

Algorithm 6: Kernel Multiblock PLS (Kernel MBPLS)

```
input : $\mathbf{K}_p$, $\mathbf{K}_g$, $\mathbf{Y}$ matrices and the number of factors ($f$)
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
    $\mathbf{K}_p = \mathbf{K}_p - \mathbf{t}\mathbf{t}^\top\mathbf{K}_p, \quad \mathbf{K}_g = \mathbf{K}_g - \mathbf{t}\mathbf{t}^\top\mathbf{K}_g, \quad \mathbf{Y} = \mathbf{Y} - \mathbf{t}\mathbf{t}^\top\mathbf{Y}$
end
```

where $\mathbf{k}_j^p$ is the kernel representation of the j*th* probe image ($\mathbf{p}_j$).

The response matrix $\mathbf{Y}$ assumes the form

$$\mathbf{Y} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_C} & \mathbf{0}_{n_C} & \cdots & \mathbf{1}_{n_C} \end{bmatrix}, \tag{3.24}$$

where scalar C and n_i represents the number of classes and number of samples in class C_i, respectively. $\mathbf{1}_{n_i}$ and $\mathbf{0}_{n_i}$ denote a $n_i \times 1$ vector of ones or zeros, respectively. Notice that we do not perform the feature mapping in matrix $\mathbf{Y}$ as we aiming at predict its actual values.

In this work, we use the regression responses $\hat{\mathbf{y}}_i$ and $\hat{\mathbf{y}}_j$ as discriminative signature of samples from data blocks i and j (i.e., two distinct surveillance cameras). Specifically, the similarity between i and j ($s(i,j)$) is computed using the cosine similarity between $\hat{\mathbf{y}}_i$ and $\hat{\mathbf{y}}_j$ as

$$s(i,j) = \hat{\mathbf{y}}_i\hat{\mathbf{y}}_j / \parallel \hat{\mathbf{y}}_i \parallel\parallel \hat{\mathbf{y}}_j \parallel . \tag{3.25}$$

3.4 Final Remarks

In this chapter, we presented the three proposed approaches to tackle the person re-identification as a nonlinear subspace learning problem. These methods address the camera transition problem by projecting the data in a common subspace where a more robust and discriminative representation is obtained. Specifically, we presented the Kernel PLS for Subspace Learning, the Kernel Hierarchical PCA and the Kernel Multiblock PLS. In the followings paragraphs we summarize the main aspects of these methods.

The Kernel PLS for Subspace Learning learns a common and low-dimensional representation that maximizes the covariance between feature descriptors of the same individuals extracted from two non-overlapping cameras. It is based on the classical NIPALS formulation and, therefore, is limited to a pair cameras for model. Differently, the Kernel Hierarchical PCA employs a hierarchical formulation allowing that feature descriptors from multiple cameras to be integrated in a unique common subspace that scales linearly with the number of surveillance cameras.

Kernel Multiblock PLS is also a hierarchical model that scales linearly with the number of surveillance cameras. In addition, it allows the inclusion of additional labels, such as attributes, in a response $\mathbf{Y}$ and, therefore, better handles with the ambiguity problem.

Chapter 4

Indirect Matching

One of the main challenges addressed in the person re-identification is the drastic modification in the appearance of individuals as a result of the different camera acquisition conditions. In this chapter, we avoid the direct comparison between feature descriptors extracted from different cameras by using the proposed indirect matching strategy. Our main assumption is that two individuals whose images are similar when captured by a first camera will remain similar when captured by second camera. Thus, instead of directly comparing images obtained from different cameras, we compare them indirectly based on the similarity with individuals in a training set.

In the following sections, we present the three endeavors to indirectly matching probe and gallery images. The *Prototype-based Person Re-Identification* (Section 4.1), the *Cross-View Kernel PLS* (Section 4.2) and *Cross-View Kernel Collaborative Representation based Classification* (Section 4.3). When describing these methods, we used the same notation presented in Chapter 3. Finally, Section 4.4 summarizes the main characteristics of these approaches.

4.1 Prototype-Based Person Re-Identification

In this section, we present the *Prototype-based Person Re-Identification* method that indirectly matches probe and gallery images using prototypes. The main idea is that we can use the ambiguity between different persons to discover similar individuals for a given sample (probe or gallery image) in the training set, which we coined as the prototypes. Then, using these individuals, we can learn models at the specific cameras where the models will be deployed, bypassing the camera transition problem (see Figure 4.1). More importantly, we explore the close relation between person re-identification and

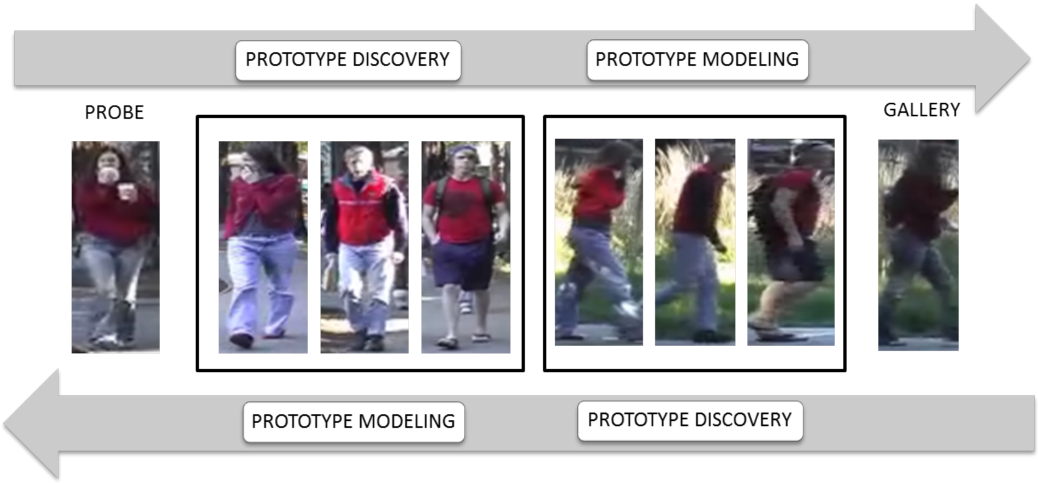

Figure 4.1: For a given sample (probe or gallery image), we compute the similarity with training samples captured by the same camera to determine the most similar individuals, which constitute the prototypes. Then, we use the class information to transit to the opposite camera where the models are learned and deployed.

information retrieval approaches and propose to combine complementary ranking lists using a ranking aggregation method.

The proposed prototype-based person re-identification model shares some aspects with methods from literature. Similarly to Liu et al. [2012], we also compute the feature importance for each probe and gallery images adaptively. However, instead of using an unsupervised approach, we use the class information to shift between cameras avoiding between cameras comparison. In An et al. [2013b], the authors also indirectly compare probe and gallery images using training samples. In fact, they learn a common subspace where they compare probe and gallery images using the similarity with all training images captured at the respective cameras. Differently, we use the idea of prototypes and our goal is to compute the feature importance adaptively, considering probe and gallery appearances. The idea of prototypes was explored in Guo et al. [2007] to handle with drastic appearance variations caused by non-overlapping cameras. Nonetheless, they deal with the problem of vehicle recognition and the prototypes are synthetic images instead of real images captured by each camera.

In the following sections, we present the proposed method in more details. Specifically, Section 4.1.1 presents the prototype discovery approach where we define a subset of similar individuals in training set for a specific sample image. Then, Section 4.1.2 presents the model learned to weight the feature descriptors accordingly to information present in the prototypes. Finally, Section 4.1.3 shows how complementary ranking lists are obtained starting from probe or gallery image.

4.1.1 Prototype Discovery

In short, the prototype discovery consists in computing the similarity between images captured by the same camera. These similarities are crucial as they provide the identity information that we can use to seamless move between different cameras in the training set. We accomplish that using a robust feature representation and a discriminative model to weight the feature descriptor accordingly with the importance. More specifically, we use well-known feature descriptors in person re-identification community to capture the appearance information and the one-shot similarity (OSS) [Wolf et al., 2009] to compute similarity between images.

In the OSS, the similarity between two images is computed using two discriminative model and a fixed subset of negatives images, denoted as background subset. Each model focuses on learning the discerning characteristics of a given sample when compared to the background. Algorithm 7 presents the OSS algorithm to compute the similarity between the $i th$ sample image (probe or gallery) and the $j th$ training image captured by the same camera, represented by $\mathbf{s}_i$ and $\mathbf{t}_j$, respectively. Notice that $\mathbf{s}_i$ ($\mathbf{t}_j$) is deployed in the model learned using the background and $\mathbf{t}_j$ ($\mathbf{s}_i$). Thus, the scores will be as higher as more similar are $\mathbf{s}_i$ and $\mathbf{t}_j$.

The background subset ($\mathbf{B}$) can be any set of images without intersection with the images that we want to compute the similarity. For instance, when computing the similarity between a probe image and the training images captured by the probe camera, we use as background subset training images captured by the gallery camera. In addition, we use the Partial Least Squares (PLS) regression as the discriminative models. PLS fits well in the person re-identification problem as it handles with the high-dimensional feature descriptor and the multicollinearity problem [Schwartz and Davis, 2009b].

Algorithm 7: One-Shot-Similarity (OSS)

```
input  : $\mathbf{s}_i$, $\mathbf{t}_j$ vectors and the background subset ($\mathbf{B}$)
$model_1 \leftarrow$ PLS ($\mathbf{s}_i$, $\mathbf{B}$),    $model_2 \leftarrow$ PLS ($\mathbf{t}_j$, $\mathbf{B}$)
$score_1 = model_1$ ($\mathbf{t}_j$),    $score_2 = model_2$ ($\mathbf{s}_i$)

$score \leftarrow \frac{score1+score2}{2}$
return $score$
```

After computing the similarity between a given sample and training images, we return the K most similar images in the training set and their respective identities, where K is a parameter of the model. It is important to highlight that using the identity information available for the training samples, we can obtain the prototypes subset at any camera.

4.1.2 Prototype Modeling

In this section, we present the prototype modeling that uses the prototypes to learn representative models in the cameras where the models will be deployed. For instance, for a probe image, we discover the prototypes in the probe camera but we use the class information to learn the subtle characteristics present in the prototypes using the training images at the gallery camera. Consequently, we can indirectly compute the similarity between probe and a gallery image using this prototype model.

Similarly to the prototype discovery, we also use the PLS regression model to weight the feature descriptor based on its discriminative power. However, instead of using as negative samples images captured by a different cameras (e.g., background subset) as in the OSS model, we use the training images captured by the same camera but not present in the prototype subset. Therefore, we can focus on the nuances that discriminate the prototypes from the remaining images captured in the same camera instead of artefacts that appear due to different camera conditions.

Once we learn the prototype model for a given sample, we can compute the similarity with any image captured by an opposite camera. For instance, when we learn the prototype model for a probe (gallery), we can use this model compute the similarity with the gallery (probe) images.

4.1.3 Sample Ranking

Previous section explains how to learn prototypes model for a given sample image. In this section, we show that probe and gallery images can be used to compute probe- and gallery-based ranking lists, respectively. Then, we present a ranking aggregation method to combine both ranking lists in an improved prototypes result.

The gallery-based ranking list is computed using as sample a gallery image. Thus, for each gallery image we learn a prototype model in the probe camera and then, we use this model to compute the similarity between a given probe and the respective gallery image. Differently, the probe-based ranking lists learns the prototypes model in the gallery camera and projects all the gallery images in this unique model to compute the similarity with the given probe. To compute the probe-based ranking list, we have to consider Figure 4.1 in direction left-to-right, while the right-to-left direction corresponds to the gallery-based ranking list. Notice that both ranking lists consist on the gallery images sorted in descending order of similarity with a given probe.

In this dissertation, we assume that probe- and gallery-based ranking lists have complementary information and combine their results using a ranking aggregation strategy. Specifically, we employ the Stuart ranking aggregation method [Stuart et al., 2003] that was proposed to handle lists of genes and is robust to noisy information and scalable, which are necessary requisites for the person Re-ID problem.

Finally, the obtained ranking list corresponds to an indirect matching of probe and gallery images using prototypes. However, the usual approach in in literature consists in directly matching probe and gallery images. For instance, KISSME [Koestinger et al., 2012] is a well-known metric learning method that computes a Mahalanobis distance based on a likelihood ratio test. We claim that due to the different strategies, the proposed prototype-based method is complementary to KISSME, and combine them using ranking aggregation. The obtained experimental results (see Section 5.4) confirm our expectation with improved results when combining both approaches.

4.2 Cross-View Kernel PLS

In Section 4.1, we presented our first attempt to indirectly handle the camera transition and weight the feature descriptors based on their discriminative power. However, the prototype-based approach depends on the discovery of a prototype subset, which is difficult due to variations of illumination and background that occur even in a fixed camera. In addition, the prototype modeling is based on a linear model, while state-of-the-art methods have shown improved working in a nonlinear feature space.

In this section, we present the *Cross-View Kernel PLS (X-KPLS)* method that computes a robust representation for a given sample using the entire training set at their respective camera based on a nonlinear regression model. More importantly, we learn the nonlinear regression model to map the feature descriptors into different vertices of a regular simplex, which ensures the best separability between classes and then, the model generalization [Vapnik, 1998]. Thus, we obtain a novel representation that is robust to the different camera conditions and has a dimensionality that depends uniquely on the number of images in the training set.

Indirectly matching probe and gallery images based on the similarity with the entire training set is not novel in person re-identification. An et al. [2013b] described probe and gallery images based on the similarity with training samples. They computed the similarity in a low-dimensional subspace that maximizes the correlation between

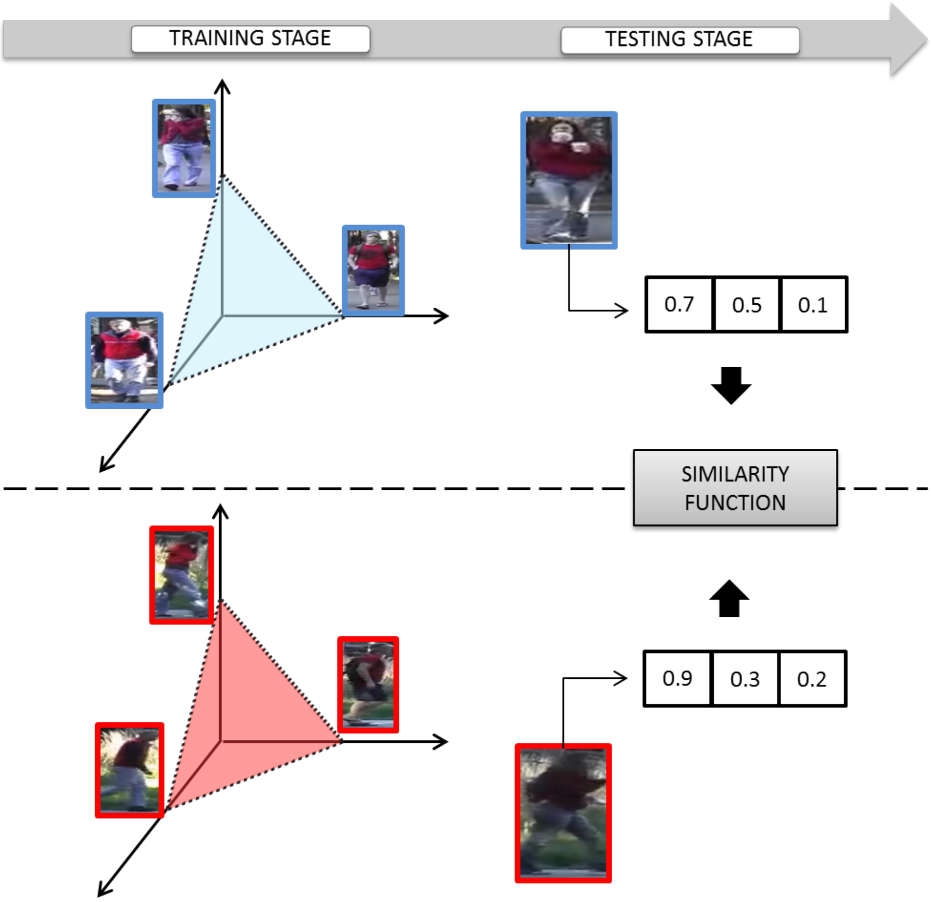

Figure 4.2: Schematic representation of the proposed X-KPLS model. In the training stage, we learn for each camera a nonlinear regression model that maps the feature descriptor extracted from training samples to vertices of a regular simplex. Then, in the testing stage, we apply probe and gallery images in the learned regression models to compare them using the regression responses and a similarity function.

images captured from different cameras. Despite the similar ideas, X-KPLS works in a nonlinear feature space and uses a regression model to compute the low-dimensional signatures of probe and gallery images.

In the following sections, we present the proposed X-KPLS method. First, Section 4.2.1 presents a brief introduction regarding the Kernel Partial Least Squares (KPLS) method. Then, we present the X-KPLS model that computes signatures for probe and gallery images as the regression responses computed by the camera-specific KPLS models in Section 4.2.2. Figure 4.2 illustrates the proposed X-KPLS method.

4.2.1 Kernel Partial Least Squares (KPLS)

In this section, we present a brief description of the Kernel PLS, which is a straightforward extension of the PLS regression model. In the following paragraphs, we consider the input feature descriptor and responses as the zero-mean matrices $\mathbf{X} \in \mathbb{R}^{n \times m}$ and $\mathbf{Y} \in \mathbb{R}^{n \times q}$, respectively.

The main idea in Kernel PLS consists in nonlinearly transform the input variables to a feature space $\mathcal{F}$, i.e, $\mathbf{\Phi} : x_i \in \mathbb{R}^m \to \mathbf{\Phi}(x_i) \in \mathcal{F}$, such that a linear regression in $\mathcal{F}$ corresponds to a nonlinear regression at the original space $\mathbb{R}^m$. However, instead of explicitly map the data into the high-dimensional space, the Kernel PLS model uses the "kernel trick" to substitute the cross-product by $K_x = \mathbf{\Phi}\mathbf{\Phi}^\top$, where $K_x \in \mathbb{R}^{n \times n}$ is the *kernel Gram matrix.*

Algorithm 8: Kernel Partial Least Squares (KPLS).

```
input : $\mathbf{K}_x$, Y matrices and the number of factors ($f$)
randomly initialize $\mathbf{u}$ and $\mathbf{u}_0$
for $i$=1 to $f$ do
    while $\|\boldsymbol{u} - \boldsymbol{u}_0\| > \varepsilon$ do
        $\mathbf{u}_0 \leftarrow \mathbf{u}$
        $\mathbf{t} = \mathbf{K}_x\mathbf{w}, \quad \mathbf{t} \leftarrow \frac{\mathbf{t}}{\|\mathbf{t}\|}$
        $\mathbf{c} = \mathbf{Y}^\top\mathbf{t}$
        $\mathbf{u} = \mathbf{Y}\mathbf{c}, \quad \mathbf{u} \leftarrow \frac{\mathbf{u}}{\|\mathbf{u}\|}$
    end
    $\mathbf{K}_x \leftarrow (\mathbf{I} - \mathbf{t}\mathbf{t}^\top)\mathbf{K}_x(\mathbf{I} - \mathbf{t}\mathbf{t}^\top)$
    $\mathbf{Y} \leftarrow \mathbf{Y} - \mathbf{t}\mathbf{t}^\top\mathbf{Y}$
end
```

Algorithm 8 presents the Kernel PLS method proposed in Rosipal and Trejo [2002]. Note that they do not perform a nonlinear map from $\mathbf{Y}$ onto a high-dimensional space, since they want to predict $\mathbf{Y}$ from $\mathbf{K}_x$. For instance, in classification problems, the response matrix $\mathbf{Y}$ assumes the form

$$\mathbf{Y} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_C} & \mathbf{0}_{n_C} & \cdots & \mathbf{1}_{n_C} \end{bmatrix}, \tag{4.1}$$

where scalar C represents the number of classes and n_i indicates the number of samples in class C_i. $\mathbf{1}_{n_i}$ and $\mathbf{0}_{n_i}$ denote a $n_i \times 1$ vector of ones or zeros, respectively. In fact, this strategy also has the advantage of promoting the best separability of training data and generalization ability of the learned model [Vapnik, 1998].

Finally, KPLS predicts the responses as $\mathbf{Y} \approx \Phi\mathbf{B}$. Thus, for the training and testing data, we obtain the responses $\hat{\mathbf{Y}}$ and $\hat{\mathbf{Y}}_t \in \mathbb{R}^n$, respectively, as

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{T}\mathbf{T}^\top\mathbf{Y} \text{ and} \\ \hat{\mathbf{Y}}_t &= \mathbf{k}_t\mathbf{U}(\mathbf{T}^\top\mathbf{K}_x\mathbf{U})^{-1}\mathbf{T}^\top\mathbf{Y},\end{aligned} \tag{4.2}$$

where matrices $\mathbf{T}$ and $\mathbf{U} \in \mathbb{R}^{n\times f}$ contain the computed score vectors from Algortihm 8, and $\mathbf{k}_t \in \mathbb{R}^n$ consists of the computation of $\Phi(\mathbf{x})\boldsymbol{\Phi}^\top$, which corresponds to the kernel function applied using a given sample $\mathbf{x}$ and the training images $\mathbf{X}$.

4.2.2 Proposed Method

In this section, we present the proposed X-KPLS method that learns a Kernel PLS regression model for each camera. These regression models map feature descriptors of the same class to the same vertex of a regular simplex, which is only possible due to the class information available in the training set. Thus, when computing the regression response for probe and gallery images, we can use the models learned at probe and gallery cameras, respectively. Figure 4.2 schematically represents the X-KPLS method.

X-KPLS represents the i*th* probe image ($\mathbf{p}_i$) and the j*th* gallery image ($\mathbf{g}_j$) using training images at their respective cameras, $\mathbf{X}_p$ and $\mathbf{X}_g$, respectively, based on camera-specific KPLS models. To achieve that, we first use Algorithm 8 to learn two nonlinear regression models - one for probe and the other for gallery camera. These models are learned for the probe and gallery using as output ($\mathbf{Y}$) a response matrix defined accordingly to Equation 4.1, and the input matrix ($\mathbf{X}$) consists of the feature descriptors extracted from training samples at their respective cameras (see Algorithm 9).

In the testing stage, to represent a probe image ($\mathbf{p}_i$), we first define the matrix $\mathbf{K}_p$, whose element i, j is the result of a kernel function using as input the feature descriptors $\mathbf{x}_p^i$ and $\mathbf{x}_p^j$. Then, using Algorithm 9, we compute the probe's signature $\hat{\mathbf{Y}}_p^i$ accordingly to Equation 4.2, where matrices $\mathbf{T}$ and $\mathbf{U}$ are obtained using Algorithm 8. Similarly, we compute the signature for j*th* gallery image $\hat{\mathbf{Y}}_g^j$ using the training images at camera gallery camera ($\mathbf{X}_g$). Notice that these signatures have values proportional to the similarity between probe and gallery images with training samples at their respective cameras. Therefore, use signatures $\hat{\mathbf{Y}}_t^{\mathbf{p}_i}$ and $\hat{\mathbf{Y}}_t^{\mathbf{g}_j}$ to indirectly compare the probe image $\mathbf{p}_i$ and gallery image $\mathbf{g}_j$ using the cosine similarity.

Algorithm 9: Cross-View Kernel PLS (X-KPLS)

```
input : $\mathbf{X}_p$, $\mathbf{X}_g$ are input matrices and, $\mathbf{k}_i^p$, $\mathbf{k}_j^g$ are row-vectors
$model_p \leftarrow$ KPLS $(\mathbf{X}_p, \mathbf{Y})$,    $model_g \leftarrow$ KPLS $(\mathbf{X}_g, \mathbf{Y})$
$\hat{\mathbf{Y}}_p^i = model_p\ (\mathbf{k}_i^p)$,    $\hat{\mathbf{Y}}_g^j = model_g\ (\mathbf{k}_j^g)$
$score \leftarrow \frac{\hat{\mathbf{Y}}_p^i \hat{\mathbf{Y}}_g^j}{\|\hat{\mathbf{Y}}_p^i\| \|\hat{\mathbf{Y}}_g^j\|}$
return $score$
```

4.3 Kernel Cross-View Collaborative Representation based Classification

In this section, we propose a novel method to address the person Re-ID problem using a nonlinear supervised Collaborative Representation based Classification (CRC) framework, the *Kernel Cross-View Collaborative Representation based Classification (Kernel X-CRC)*. *Kernel X-CRC* efficiently represents each pair probe **p** and gallery **g** images collaboratively using its camera-view specific training samples $\mathbf{X}_p$ and $\mathbf{X}_g$, respectively.

Some works also investigate the person re-identification problem using sparse or collaborative representations [Lisanti et al., 2015b; Zeng et al., 2015; Karanam et al., 2015; Harandi et al., 2012; Kodirov et al., 2015b]. *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 occurs in Karanam et al. [2015]. Differently from dictionary learning-based approaches [Harandi et al., 2012; Kodirov et al., 2015b], 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 [Lisanti et al., 2015b; Karanam et al., 2015; Harandi et al., 2012; Kodirov et al., 2015b], we efficiently model the strong nonlinear transition of features between cameras achieving an analytical solution.

To the best of our knowledge, this is the first work addressing the person re-identification problem as a multi-task collaborative representation problem. Furthermore, it is important to emphasize that, even though employed to the person re-identification problem, the proposed approach provides a general framework to other multi-view computer vision problems, such as the matching of multi-modal biometrics.

In the following sections, we first discuss the Collaborative Representation based Classification in the context of person re-identification (Section 4.3.1) and, then, we present the novel *Kernel X-CRC* method (Section 4.3.2) that computes the similarity between probe and gallery images based on its respective coding vectors.

4.3.1 Collaborative Representation based Classification

In this section, we present the Collaborative Representation based Classification (CRC) method and show a simple adaptation to consider the person re-identification problem. Then, we motivate the proposed *Kernel X-CRC* method.

Due to the lack of available samples or the high cost of collecting and annotating a large number images for each individual captured by different cameras, conventional person re-identification datasets contain only hundreds of individuals. In this scenario, we have to deal with the small-sample-size problem. It is also a common problem in the face identification task that has been addressed coding a query (probe) sample as sparse (SRC [Wright et al., 2010]) or collaborative (CRC [Zhang et al., 2012]) combination of the gallery samples. Thus, the ith probe image ($\mathbf{p}_i$) is represented using a gallery-set $\mathbf{G}$ with proper regularization term as

$$\min_{\alpha} \parallel \mathbf{p}_i - \mathbf{G}\boldsymbol{\alpha} \parallel_2^2 + \lambda \parallel \boldsymbol{\alpha} \parallel_r, \tag{4.3}$$

where λ is a scalar and $\boldsymbol{\alpha}$ is the sparse ($r = 1$) or the collaborative ($r = 2$) coding vector [Zhang et al., 2012; Wright et al., 2010]. Then, the probe image is assigned to the class that results in the smallest reconstruction error.

Equation 4.3 directly reconstructs a probe image using gallery samples, which is only reasonable when probe and gallery images are feature descriptors from the same modality (e.g. same cameras). Therefore, this formulation will result in poor matching performance when directly applied in person re-identification as consequence of the camera transition problem. Therefore, in this work, we consider the problem of adapting this model to handle the person re-identification scenario. Specifically, we focus on the collaborative representation problem as it presents an analytical solution.

A straightforward attempt to adapt the collaborative representation to the person re-identification problem would be to compute the collaborative representation coefficients (coding vectors) $\boldsymbol{\alpha}_p$ and $\boldsymbol{\alpha}_g$ by solving the camera-specific optimization problems

$$\min_{\boldsymbol{\alpha}_p} \parallel \mathbf{p}_i - \mathbf{X}_p\boldsymbol{\alpha}_p \parallel_2^2 + \lambda \parallel \boldsymbol{\alpha}_p \parallel_2^2 \text{ and } \min_{\boldsymbol{\alpha}_g} \parallel \mathbf{g}_j - \mathbf{X}_g\boldsymbol{\alpha}_g \parallel_2^2 + \lambda \parallel \boldsymbol{\alpha}_g \parallel_2^2, \tag{4.4}$$

where each linear regularized model represents the probe or the gallery image using the respective images in the training set ($\mathbf{X}_p$ or $\mathbf{X}_g$). Therefore, similar coding coefficients would be expected when $\mathbf{p}_i$ and $\mathbf{g}_j$ correspond to the same individual acquired from different cameras ($i = j$). Then, the matching between probe and gallery images occurs indirectly using the coding vectors ($\boldsymbol{\alpha}_p$ and $\boldsymbol{\alpha}_g$). Note that $\mathbf{X}_g$ and $\mathbf{X}_p$ can be any representation of the training images able to balance discriminative power and

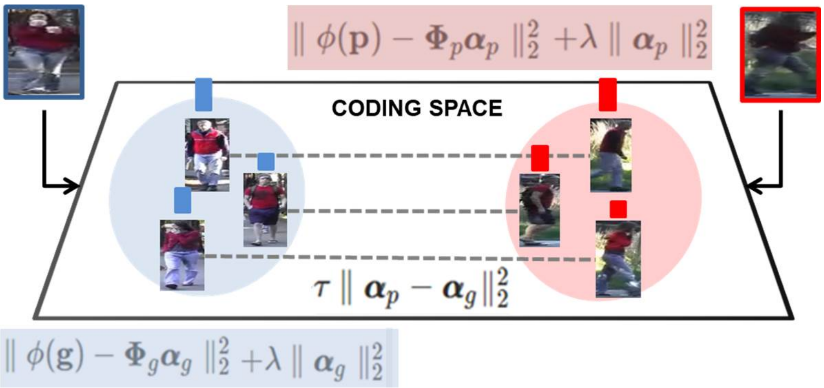
$\| \phi(\mathbf{p}) - \mathbf{\Phi}_p \boldsymbol{\alpha}_p \|_2^2 + \lambda \| \boldsymbol{\alpha}_p \|_2^2$
CODING SPACE
$\tau \| \boldsymbol{\alpha}_p - \boldsymbol{\alpha}_g \|_2^2$
$\| \phi(\mathbf{g}) - \mathbf{\Phi}_g \boldsymbol{\alpha}_g \|_2^2 + \lambda \| \boldsymbol{\alpha}_g \|_2^2$

Figure 4.3: 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. Illustration from [Prates and Schwartz, 2018].

robustness to camera transition. In fact, we show in the experimental results that we reach improved matching rates when working in a learned common subspaces.

Equation 4.4 considers probe and gallery cameras independently while elements $\boldsymbol{\alpha}_g^i$ and $\boldsymbol{\alpha}_p^i$ are the coding vectors of ith individual in training set. Therefore, a better model should consider not only the representativeness in each camera, but also the similarity between $\boldsymbol{\alpha}_p$ and $\boldsymbol{\alpha}_g$ in a unique multi-task framework. In addition, better results have been reported working in a nonlinear feature space [Lisanti et al., 2014].

4.3.2 Proposed Approach

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* $\| \boldsymbol{\alpha}_p - \boldsymbol{\alpha}_g \|_2^2$ in our multi-task formulation that balances representativeness and similarity, as illustrated in Figure 4.3.

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

$$\begin{aligned} \min_{\boldsymbol{\alpha}_g, \boldsymbol{\alpha}_p} & \| \phi(\mathbf{p}) - \mathbf{\Phi}_p \boldsymbol{\alpha}_p \|_2^2 + \| \phi(\mathbf{g}) - \mathbf{\Phi}_g \boldsymbol{\alpha}_g \|_2^2 \\ & + \lambda \| \boldsymbol{\alpha}_p \|_2^2 + \lambda \| \boldsymbol{\alpha}_g \|_2^2 + \tau \| \boldsymbol{\alpha}_p - \boldsymbol{\alpha}_g \|_2^2, \end{aligned} \tag{4.5}$$

where $\phi(.)$ is a nonlinear function and, $\mathbf{\Phi}_g$ and $\mathbf{\Phi}_p$ are resulting nonlinear mapping of $\mathbf{X}_g$ and $\mathbf{X}_p$, respectively. that we analytically derived with respect to $\boldsymbol{\alpha}_p$ and $\boldsymbol{\alpha}_g$

obtaining

$$\boldsymbol{\alpha}_p = \mathbf{A}_p^{-1}\boldsymbol{\alpha}_g + \mathbf{A}_p^{-1}\boldsymbol{\Phi}_p^\top\phi(\mathbf{p}) \text{ and } \boldsymbol{\alpha}_g = \mathbf{A}_g^{-1}\boldsymbol{\alpha}_p + \mathbf{A}_g^{-1}\boldsymbol{\Phi}_g^\top\phi(\mathbf{g}), \tag{4.6}$$

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

$$\mathbf{A}_p = \boldsymbol{\Phi}_p^\top\boldsymbol{\Phi}_p + (\lambda+\tau)\mathbf{I} \text{ and } \mathbf{A}_g = \boldsymbol{\Phi}_g^\top\boldsymbol{\Phi}_g + (\lambda+\tau)\mathbf{I}. \tag{4.7}$$

Note that the Equations in 4.6 are interdependent. Therefore, replacing $\boldsymbol{\alpha}_g$ and isolating $\boldsymbol{\alpha}_p$, we obtain

$$\boldsymbol{\alpha}_p = \tau\mathbf{Q}^{-1}\mathbf{A}_p^{-1}\mathbf{A}_g^{-1}\boldsymbol{\Phi}_g^\top\phi(\mathbf{g}) + \mathbf{Q}^{-1}\mathbf{A}_p^{-1}\boldsymbol{\Phi}_p^\top\phi(\mathbf{p}) \tag{4.8}$$

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

$$\mathbf{Q} = \mathbf{I} - \tau^2\mathbf{A}_p^{-1}\mathbf{A}_g^{-1}. \tag{4.9}$$

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

$$\boldsymbol{\alpha}_g = \tau\mathbf{W}^{-1}\mathbf{A}_g^{-1}\mathbf{A}_p^{-1}\boldsymbol{\Phi}_p^\top\phi(\mathbf{p}) + \mathbf{W}^{-1}\mathbf{A}_g^{-1}\boldsymbol{\Phi}_g^\top\phi(\mathbf{g}) \tag{4.10}$$

with $\mathbf{W}$ computed as

$$\mathbf{W} = \mathbf{I} - \tau^2\mathbf{A}_g^{-1}\mathbf{A}_p^{-1}. \tag{4.11}$$

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 its respective row vectors $\mathbf{k}^g$ and $\mathbf{k}^p$, respectively. Then, the similarity between a pair of probe $\mathbf{p}$ and gallery $\mathbf{g}$ is computed by the cosine similarity between $\boldsymbol{\alpha}_p$ and $\boldsymbol{\alpha}_g$, as described in Algorithm 10.

Algorithm 10: *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 using the above equations
$\boldsymbol{\beta}_g^g \leftarrow \mathbf{W}^{-1}\mathbf{P}_g^{-1}$, $\boldsymbol{\beta}_g^p \leftarrow \tau\mathbf{W}^{-1}\mathbf{A}_g^{-1}\mathbf{A}_p^{-1}$
$\boldsymbol{\beta}_p^p \leftarrow \mathbf{Q}^{-1}\mathbf{A}_p^{-1}$, $\boldsymbol{\beta}_p^g \leftarrow \tau\mathbf{Q}^{-1}\mathbf{A}_p^{-1}\mathbf{A}_g^{-1}$
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 $\boldsymbol{R}$
```

4.4 Final Remarks

In this chapter, we present the proposed methods that address the camera transition problem by avoiding the direct comparison between images captured from distinct cameras. It is accomplished by indirectly matching probe and gallery images based on the similarity with training samples. Specifically, we proposed the Prototypes-based Person Re-identification, the Cross-view Kernel PLS and the Kernel Cross-view Collaborative Representation based Classification. In the followings paragraphs, we discuss the main aspects of each method.

The Prototypes-based Person Re-Identification assumes a training set with enough samples to find a subset of similar individuals for a given sample, the prototypes. The prototypes subset is used to transit between cameras and, therefore, learn a discriminative model at the camera where the model will be deployed. Complementary models are learned when considering probe or gallery sample to compute the prototype subset. Finally, these models are combined using ranking a ranking aggregation approach. The main drawback of this method is the need of a large and diverse subset of training samples to discover similar prototypes for every probe/gallery sample. Differently, the X-KPLS and Kernel X-CRC avoid this issue using the entire training set.

The X-KPLS matches probe and gallery samples based on a similarity representation learned using the entire training set. The key idea is consists in using the training samples at each camera to learn a nonlinear regression model that maps from feature descriptors to vertices of a regular simplex (i.e. identity indexes). Thus, we have a model that computes a regression a response at each specific camera that is proportional to the similarity with the training samples. Besides, it scales linearly with the number of surveillance cameras. Then, the matching between probe and gallery cameras is performed by projecting them at probe and gallery regression models and computing the similarity between the responses. Nonetheless, it has the drawback that camera-specific models are learned independently of each other, which we tackle in the Kernel X-CRC.

The Kernel X-CRC also uses a similarity representation computed using the training set. This representation is learned using a collaborative representation framework that allows the representation of probe and gallery samples as a nonlinear combination of training samples at their respective cameras. Besides, it includes a similarity term to force a trade-off between similarity and representativeness. This optimization is solved simultaneous for a pair of cameras, which permits the flow of information between cameras. The main limitation of this method is that it must be computed for each pair of cameras.

Chapter 5

Experimental Results - Pairwise Cameras

In this chapter, 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 three camera pairwise datasets (VIPeR, PRID450S and CUHK01). To further enrich the discussion, we add attributes labels in the analysis of the regression-based approaches (Kernel MBPLS and Cross-view Kernel PLS). These attributes, which are obtained from work [Layne et al., 2014], are detailed in Table 5.1.

To set the parameters of the proposed methods, we use the common strategy in the person re-identification literature of using a validation and testing set composed by ten random partitions of images in training and probe/gallery subsets. These partitions have an equal number of samples, with the only exception of CUHK01 that has 971 unique identities. In this case, we used 486 individuals in the training set and the remaining 485 in the probe/gallery set.

When presenting results in tables, we report the mean *rank-k* matching rate in ten distinct 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. Besides, we report between parenthesis the standard deviation to assess the stability of the different methods.

In the following, we first present the three datasets evaluated and the employed feature descriptors. Then, in the following sections, we present the parameters setting of each method proposed. Finally, Section 5.7 shows the comparison of the proposed methods against state-of-the-art approaches.

Table 5.1: Manually annotated attributes in VIPeR dataset.

Redshirt	Blueshirt	Lightshirt	Darkshirt	Greenshirt
Nocoats	Sidebag	Darksidebag	Colourbottoms	Darkbottoms
Lightbottoms	Satchel	Barelegs	Shorts	Jeans
Male	Skirt	Patterned	Midhair	Darkhair
Bald	HandBag	Backpack		

Datasets

We selected three well-known person re-identification datasets to assess the performance of state-of-the-art approaches. These datasets are the VIPeR [Gray and Tao, 2008], PRID450S [Roth et al., 2014] and CUHK01 [Li et al., 2012]. These datasets consider the classical person re-identification scenario where we have images of the same individual captured by two non-overlapping surveillance cameras.

VIPeR is one of the most widely used dataset for person re-identification due to the challenging imposed by the image resolution, illumination, pose and background conditions. In fact, these drastic appearance changes make this dataset difficult even for humans. Then, a more controlled dataset - PRID450S - was released with 450 individuals captured by two non-overlapping and static surveillance cameras. However, these datasets consider only few individuals and images, which limits the application of data intensive approaches (e.g. deep learning). To tackle these problems, in 2012, CUHK01 was proposed with more individuals and multiple images from each individual (multi-shot scenario). More recently, multi-cameras and multi-shot datasets have been proposed to close the gap between the researches and real-world applications as we will discuss in Chapter 6. In the following paragraphs, we present each dataset and its main aspects.

VIPeR dataset. VIPeR [Gray and Tao, 2008][1] contains 632 labelled image pairs captured by two different outdoor cameras located in an academic environment where each subject appears once in each camera (single-shot scenario). The main difficulties occur due viewpoint changes, illumination and image quality. For instance, most of the image pairs show viewpoint change larger than 90 degrees. All images are normalized to 128 $\times$ 48 pixels for evaluations. Figure 5.1 illustrates some image pairs from VIPeR.

[1] Available at: https://vision.soe.ucsc.edu/projects

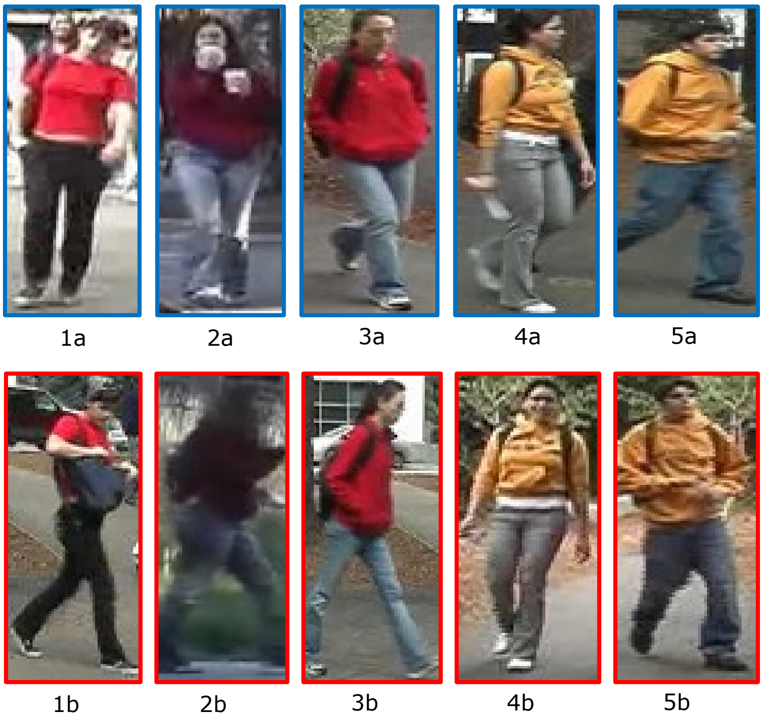

Figure 5.1: Example of images captured at camera A (first row) and B (last row), in VIPeR dataset. Individuals 4 and 5 are similar at camera A and B. However, it is not common. For instance, individuals 1a, 2a and 3a are similar at camera A, while at camera B they look quite different as a consequence of self-occlusion (1b) and illumination changes (2b).

PRID450S dataset. PRID450S [Roth et al., 2014][2] consists of 450 images pairs of pedestrians captured by two non-overlapping cameras. Each subject appears in single image at each camera (single-shot). The main challenges are related to changes in viewpoint, pose as well as significant differences in background and illumination (see Fig. 5.2).

CUHK01 dataset. CUHK01 [Li et al., 2012][3] dataset captures two disjoint camera-view images for each person in a campus environment. It contains 971 persons with two images from each camera-view (multi-shot scenario) that are normalized to 160 × 60 pixels for evaluations. Camera A captures the frontal view or back view, while camera B captures the side view of pedestrians (see Fig. 5.3).

Feature Descriptors

As designing a feature descriptor is not the focus of this dissertation, we evaluate the proposed methods using different feature descriptors from person re-identification

[2] Available at: http://lrs.icg.tugraz.at/download.php

[3] Available at: http://www.ee.cuhk.edu.hk/rzhao/

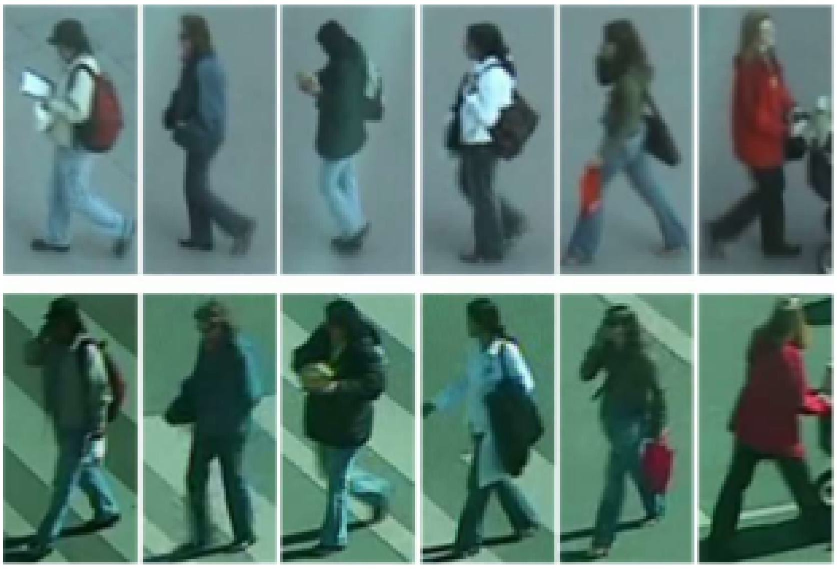

Figure 5.2: Image pairs of the same individual captured by different cameras from PRID450S dataset. Observe that the illumination and pose change person's visual appearance.

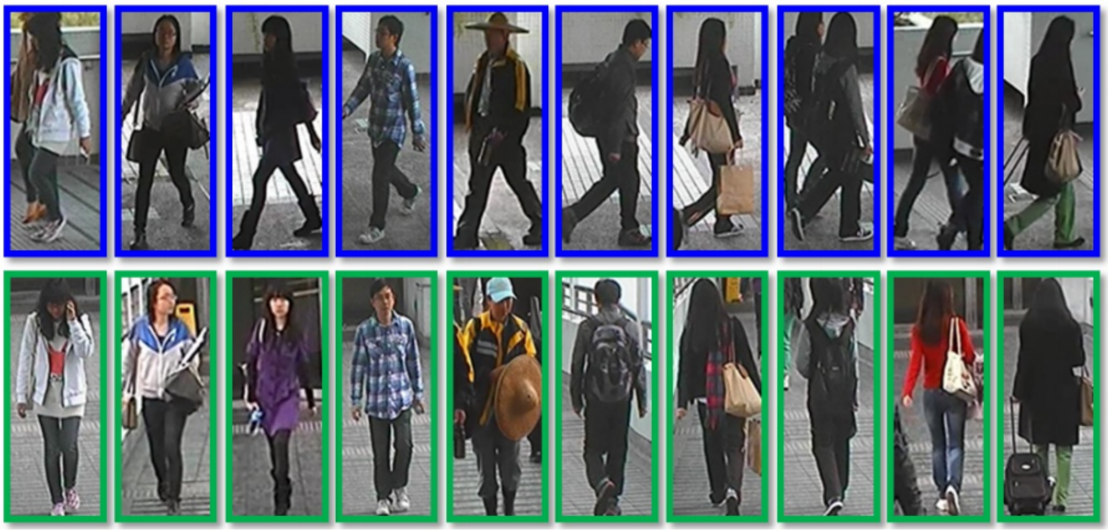

Figure 5.3: Image pairs of the same individual captured by different cameras from CUHK01 dataset. Notice that similarly to VIPeR, we also have different illumination conditions, pose and occlusion problems. Illustration taken from [Li et al., 2012].

literature. To accomplish that, we select two widely used descriptors: the Hierarchical Gaussian descriptor (GoG) [Matsukawa et al., 2016] and the combination of handcrafted and deeply learned features (LOMO + CNN) [Shangxuan et al., 2016]. In the following, we present a brief overview of these feature descriptors and, then, the kernel function that we use in the remaining experiments - Radial Basis Function (RBF) Kernel.

LOMO + CNN. LOMO [Liao et al., 2015] consists of color and texture descriptors extracted from multiple scales and horizontal stripes. Differently, the CNN features are automatically learned from image pixels and, to avoid overfitting, are constrained by previously extracted handcrafted descriptors using a feature fusion layer [Shangxuan et al., 2016]. The final descriptor is a simple high-dimensional concatenation of LOMO and CNN features.

GoG. GoG descriptor assumes a Gaussian distribution to capture texture and color information of local patches using covariance and mean statistics, respectively. Then, these Gaussian distributions are learned hierarchically from patches to image regions to compose the final Hierarchical Gaussian descriptor.

When mapping the data to a nonlinear feature space, we used the radial basis function kernel, or RBF kernel. The RBF kernel on feature vectors $\mathbf{x}$ and $\mathbf{y}$ is computed as

$$K(\mathbf{x}, \mathbf{y}) = exp(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}), \tag{5.1}$$

where σ is a parameter that we empirically set for each feature descriptor and dataset.

5.1 Kernel PLS for Subspace Learning

In this section, we consider parameters directly related to the proposed model. First, we present experimental results considering the already mentioned feature descriptors from literature. Then, we consider distinct values of sigma and latent factors. Finally, we compare the proposed kernel extension with its counterpart linear model - PLS for Subspace Learning - to demonstrate the performance gain due to the nonlinear mapping.

Feature Descriptors. Tables 5.2 and 5.3 present the obtained experimental results of the proposed approach with different feature descriptors in VIPeR and PRID450S datasets, respectively. Table 5.2 shows that the best performing feature descriptor is the GoG, achieving 37.6% of *rank-1* matching rate. Similarly, we observed improved results for GoG descriptor in the CUHK01 dataset. Nonetheless, as shown in Table 5.3, LOMO+CNN outperformed the GoG descriptor in PRID450S dataset. We credit this improvement to the distinct characteristics of each dataset that can benefit one or another feature representation.

Table 5.2: Matching rates for different ranking positions in the VIPeR dataset using the proposed Kernel PLS for Subspace Learning with different feature descriptors.

Descriptors	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
GoG	**37.6 (1.9)**	**72.1 (2.1)**	**84.0 (1.6)**	**89.4 (1.2)**	**92.4 (0.8)**
LOMO + CNN	29.9 (2.1)	61.4 (1.7)	73.7 (1.6)	79.3 (1.4)	83.7 (1.4)

Table 5.3: Matching rates for different ranking positions in the PRID450S dataset using the proposed Kernel PLS for Subspace Learning with different feature descriptors.

Descriptors	PRID450S (p=225)				
	r = 1	r = 5	r = 10	r = 15	r = 20
LOMO + CNN	**47.0 (2.4)**	**73.6 (1.8)**	81.8 (2.5)	86.2 (2.1)	89.0 (1.2)
GoG	39.6 (3.2)	70.9 (3.9)	**81.9 (2.8)**	**88.6 (2.2)**	**92.5 (2.1)**

Parameter σ. Table 5.4 shows the obtained *rank-1* matching rates using the proposed Kernel PLS for Subspace Learning with different values of sigma and the VIPeR dataset. Based on these results, 2.0 is the best value of sigma and therefore, is the value that we set for the following experiments.

Table 5.4: Rank-1 matching rates in the VIPeR dataset using the Kernel PLS for Subspace Learning with different values of parameter sigma.

Sigma	VIPeR (p=316)				
	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$	$\sigma = 5$
Rank-1	31.1 (2.5)	**37.6 (1.9)**	36.3 (2.5)	35.3 (2.4)	34.2 (2.8)

Latent Factors. Table 5.5 presents the *rank-1* matching rate for different number of latent factors, which is the parameter f in Algorithm 2. Based on these results, we conclude that the results stabilize after 200 components. Therefore, in the remaining experiments, we will employ 200 components for VIPeR. Likewise, we set the number of latent factors for PRID450S and CUHK01 as 200 and 300, respectively.

Nonlinear Mapping. Table 5.6 presents the experimental results using the Linear and the Kernel PLS model for Subspace Learning in VIPeR dataset. The obtained

Table 5.5: Rank-1 matching rates in the VIPeR dataset using the Kernel PLS for Subspace Learning with different numbers of latent factors.

Factors	VIPeR (p=316)				
	$f = 50$	$f = 100$	$f = 200$	$f = 250$	$f = 300$
Rank-1	31.1 (1.9)	35.0 (2.1)	**37.6 (1.9)**	37.4 (2.4)	37.8 (2.9)

experimental results show that the proposed nonlinear extension increases the *rank-1* matching rate in 12.1 percentage points. Figure 5.4 shows some ranking results for Kernel PLS and PLS for Subspace Learning models. These results agree with the quantitative results presented in Table 5.6 as for some individuals only the nonlinear model is able to return the correct gallery image between the top ten individuals.

Table 5.6: Matching rates for different ranking positions in the VIPeR dataset of PLS and Kernel PLS methods for Subspace Learning.

Methods	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
Kernel PLS	**37.6 (1.9)**	**72.1 (2.1)**	**84.0 (1.6)**	**89.4 (1.2)**	**92.4 (0.8)**
PLS	25.5 (1.8)	55.9 (4.0)	70.8 (2.8)	79.6 (2.1)	85.7 (1.2)

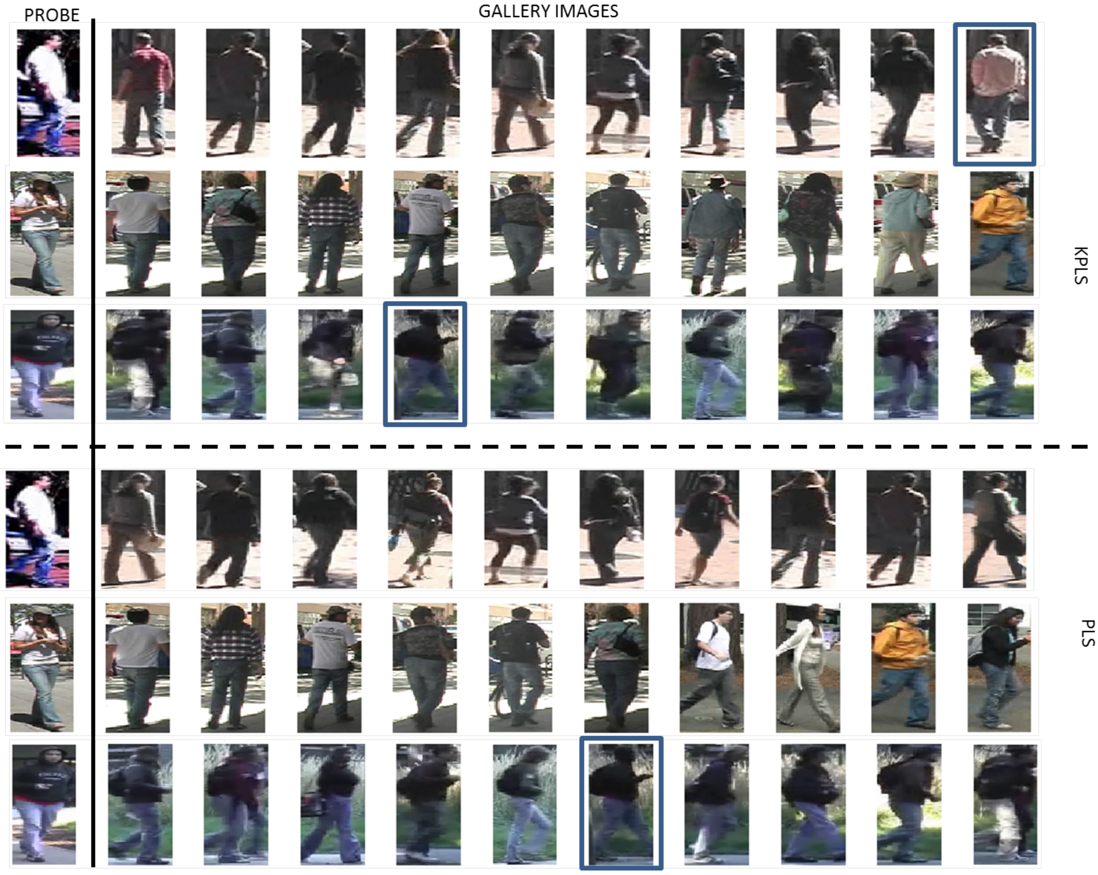

Figure 5.4: First ten individuals in gallery set ranked accordingly with the similarity with probe images using the Kernel PLS (first three rows) and the PLS (last three rows) for Subspace Learning. Individuals surrounded by a blue box correspond to the correct match for the probe image in the gallery set.

5.2 Kernel Hierarchical PCA

As described in Section 3.2, the proposed Kernel HPCA iteratively computes a low-dimensional representation that corresponds to a consensus between blocks of variables. In this section, we evaluate four parameters directly related with the performance of the proposed method: the feature descriptors, the values of sigma, the number of latent factors and the influence of the nonlinear mapping.

Feature Descriptors. In this experiment, we determine the matching rates of the proposed method with the evaluated feature descriptors. The obtained experimental results using VIPeR dataset are presented in Table 5.7. Based on these results, it is possible to notice that GoG descriptor outperforms the LOMO + CNN by a large margin. Similar results were also observed in CUHK01 dataset. However, we observed an improved *rank-1* matching rate for LOMO + CNN when compared against GoG descriptor in PRID450S, as shown in Table 5.8. Therefore, in the remaining experiments, we will employ GoG for VIPeR and CUHK01 datasets, and LOMO + CNN for PRID450S dataset.

Table 5.7: Matching rates for different ranking positions in the VIPeR dataset using the proposed Kernel HPCA method with different feature descriptors.

Descriptors	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
GoG	**39.0 (2.4)**	**74.3 (1.8)**	**86.0 (1.1)**	**92.0 (1.1)**	**94.4 (0.7)**
LOMO + CNN	33.3 (1.9)	67.2 (2.4)	79.4 (1.5)	84.7 (1.8)	88.4 (1.6)

Table 5.8: Matching rates for different ranking positions in the PRID450S dataset using the proposed Kernel HPCA method with different feature descriptors.

Descriptors	PRID450S (p=225)				
	r = 1	r = 5	r = 10	r = 15	r = 20
LOMO + CNN	**48.6 (3.5)**	74.5 (2.7)	83.3 (2.2)	87.6 (2.0)	90.5 (1.5)
GoG	48.0 (3.3)	**77.6 (2.3)**	**87.6 (2.1)**	**91.8 (2.0)**	**94.3 (1.8)**

Parameter σ. Table 5.9 presents the *rank-1* matching rates for distinct values of sigma in VIPeR dataset, which impacts in the output of the RBF-kernel computation. For these results, we notice that the best value for sigma is 2.0.

Table 5.9: Rank-1 matching rates in the VIPeR dataset using the Kernel HPCA method with different values of parameter sigma.

Sigma	VIPeR (p=316)				
	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$	$\sigma = 5$
Rank-1	31.0 (2.6)	**39.0 (2.4)**	37.8 (2.3)	36.6 (1.7)	35.8 (1.6)

Latent Factors. We also evaluated the influence of the number of latent factors in the experimental results, which is the parameter f from Algorithm 4. The obtained experimental results are presented in Table 5.10. Based on these results, it is possible to conclude that better experimental results are obtained using 200 components. Notice that the number of components is limited by the number of training samples available, which is 316 for VIPeR dataset. Similarly, we set the number of latent factors as 250 and 150 for CUHK01 and PRID450S datasets, respectively.

Nonlinear Mapping. Table 5.11 presents the obtained experimental results using the HPCA and the novel Kernel HPCA models. Based on these results, it is possible to conclude that the nonlinear mapping is responsible for a great improvement in the

Table 5.10: Rank-1 matching rates in the VIPeR dataset using the Kernel HPCA with different numbers of latent factors.

Factors	VIPeR (p=316)				
	$f = 50$	$f = 100$	$f = 200$	$f = 250$	$f = 300$
Rank-1	35.6 (2.4)	38.0 (2.2)	**39.0 (2.4)**	38.6 (2.1)	38.3 (2.1)

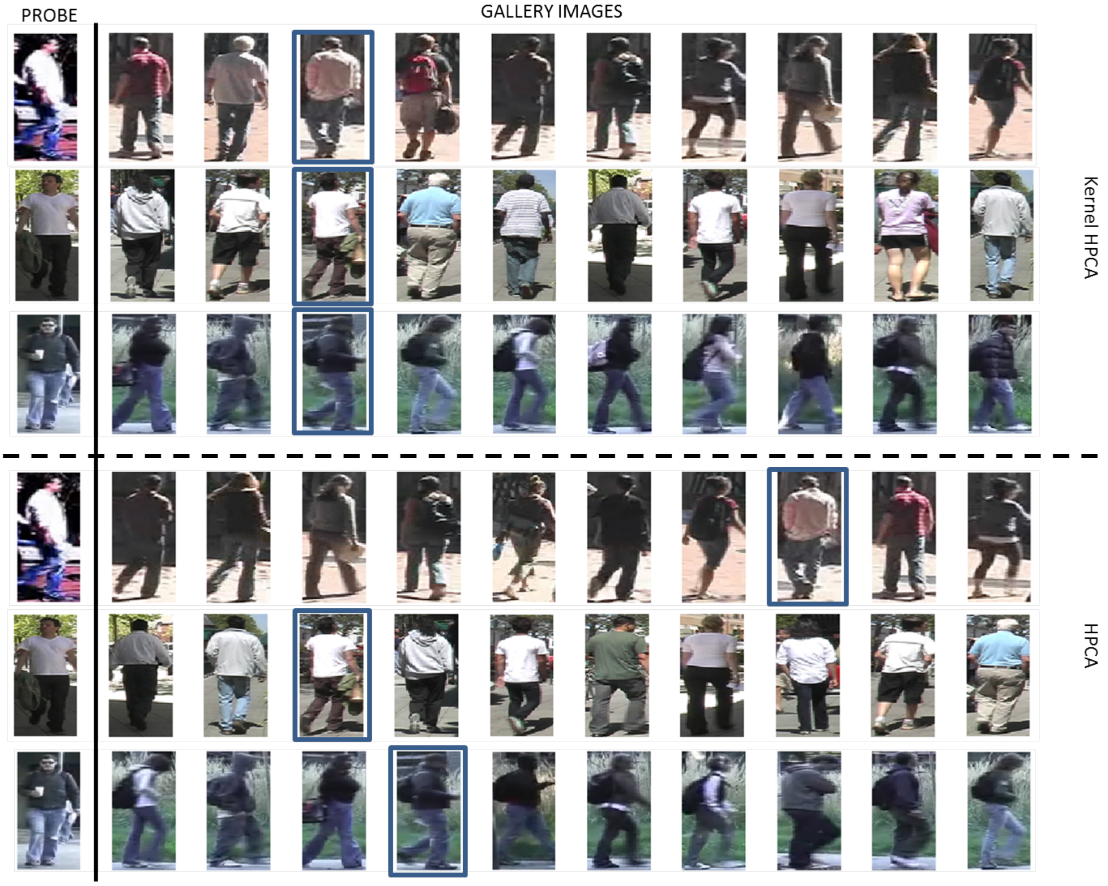

Figure 5.5: The first ten individuals in gallery set ranked accordingly with the similarity with probe images using the Kernel HPCA (first three rows) and the HPCA (last three rows) models. Individuals surrounded by a blue box correspond to the correct match for the probe image in the gallery set.

matching rates. For instance, the *rank-1* increases in almost five percentage points. Figure 5.5 shows qualitative results that support these conclusions.

Table 5.11: Matching rates for different ranking positions in the VIPeR dataset of HPCA and Kernel HPCA methods.

Methods	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
Kernel HPCA	**39.0 (2.4)**	**74.3 (1.8)**	**86.0 (1.1)**	**92.0 (1.1)**	**94.4 (0.7)**
HPCA	34.6 (2.2)	70.9 (1.5)	82.8 (1.6)	89.0 (0.9)	92.1 (0.7)

5.3 Kernel Multiblock PLS

In this section, we focus on discovering the best configuration of parameters for the proposed Kernel Multiblock PLS. Specifically, we evaluate different settings for the feature descriptors, values of sigma, number of latent factors and the influence in the obtained results of the nonlinear mapping. Finally, we perform an experiment assess the impact on the experimental results when adding manually labeled attributes in the response matrix **Y**.

Feature Descriptors. Table 5.12 presents the experimental results obtained using LOMO + CNN and GoG descriptors in VIPeR dataset. Similarly, Table 5.13 presents these results for PRID450S. We can notice that in both datasets GoG presents higher results than the LOMO + CNN. For instance, the GoG descriptor is 5.0 percentage points higher for the *rank-1* in VIPeR dataset. Based on these improved results, we set the feature descriptor as the GoG in the remaining experiments.

Table 5.12: GoG and LOMO + CNN descriptors evaluated using the proposed Kernel MBPLS in the VIPeR dataset.

Descriptors	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
GoG	**38.8 (2.3)**	**74.2 (1.4)**	**85.9 (1.3)**	**92.1 (1.0)**	**94.4 (0.9)**
LOMO + CNN	33.8 (2.1)	67.1 (2.1)	79.7 (1.5)	84.7 (1.5)	88.0 (1.5)

Table 5.13: GoG and LOMO+CNN descriptors evaluated using the proposed Kernel MBPLS in the PRID450S dataset.

Descriptors	PRID450S (p=225)				
	r = 1	r = 5	r = 10	r = 15	r = 20
GoG	**46.6 (2.8)**	**77.1 (3.0)**	**87.0 (1.4)**	**92.2 (1.5)**	**94.6 (1.8)**
LOMO + CNN	44.2 (2.0)	73.9 (1.9)	82.4 (2.1)	87.6 (1.5)	89.9 (1.6)

Parameter σ. Table 5.14 shows the obtained *rank-1* matching rates using the proposed Kernel MBPLS method with different values of sigma and the VIPeR dataset. These results demonstrate that the best value of sigma is 2.0, which is the value that we set in the following experiments.

Latent Factors. We evaluate different number of latent factors using the Kernel MBPLS and the VIPeR dataset. These results are shown in Table 5.15. For these results, it is possible to see that the best trade-off between high matching rates and

Table 5.14: Evaluation of different values of sigma using the Kernel Multiblock PLS method in the VIPeR dataset.

Sigma	VIPeR (p=316)				
	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$	$\sigma = 5$
Rank-1	31.1 (2.4)	**38.8 (2.3)**	37.8 (2.4)	36.3 (1.9)	35.7 (1.7)

computational cost is achieved when using 200 factors. Thus, we set the number of factors as 200 in the following tests.

Table 5.15: Evaluation of different number of factors using the Kernel Multiblock PLS in the VIPeR dataset.

Factors	VIPeR (p=316)				
	$f = 50$	$f = 100$	$f = 200$	$f = 250$	$f = 300$
Rank-1	35.7 (2.2)	38.2 (2.1)	**38.8 (2.3)**	38.8 (1.9)	38.4 (2.0)

Nonlinear Mapping. One important assumption of the proposed Kernel MBPLS is the positive impact of the nonlinear mapping in the results. To evaluate that, we compared the Kernel MBPLS with its linear counterpart, the MBPLS model. The obtained experimental results are presented in Table 5.16. Based on these results, we can observe improved results for all the ranking positions when using the Kernel MBPLS. Figure 5.6 corroborates with these results showing that for all probe images analyzed the Kernel MBPLS provided better ranking lists than the MBPLS.

Table 5.16: Matching rates for different ranking positions in VIPeR dataset when using the MBPLS and the proposed Kernel MBPLS.

Methods	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
Kernel MBPLS	**38.8 (2.3)**	**74.2 (1.4)**	**85.9 (1.3)**	**92.1 (1.0)**	**94.4 (0.9)**
MBPLS	33.8 (2.1)	66.4 (1.7)	78.9 (1.4)	85.0 (1.1)	88.3 (0.7)

Attributes. Kernel MBPLS is originally a nonlinear regression model and therefore, it suits to the inclusion of different information in the response matrix $\mathbf{Y}$. In Table 5.17, we present experiments regarding the use of identity, attribute and identity + attribute labels. For these results, we can notice that the addition of attribute labels results in an improved performance of the Kernel MBPLS. It can be explained to the better separability of the data in the low-dimensional subspace that is achieved when using attribute labels.

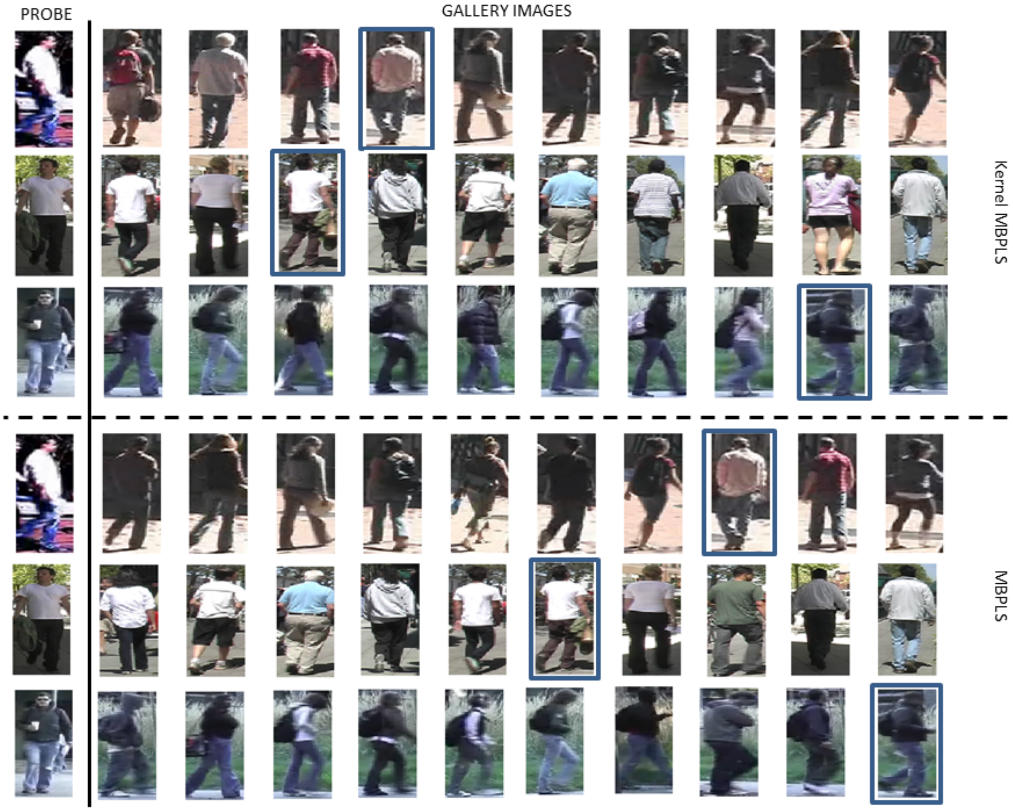

Figure 5.6: The top ten individuals in gallery set ranked accordingly with the similarity with probe images using the Kernel MBPLS (first three rows) and MBPLS (last three rows) regression models. Individuals surrounded by a blue box correspond to the correct match for the probe image in the gallery set.

Table 5.17: Evaluation of the proposed Kernel MBPLS in the VIPeR dataset using the Identity, Attributes and Identity + Attributes settings.

Method	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
Ident.	38.8 (2.3)	74.2 (1.4)	85.9 (1.3)	92.1 (1.0)	94.4 (0.9)
Attr.	21.5 (1.8)	48.6 (3.2)	61.5 (2.8)	69.0 (3.1)	74.4 (2.9)
Ident. + Attr.	**41.0 (1.5)**	**76.7 (2.4)**	**87.9 (0.9)**	**93.0 (0.9)**	**95.0 (0.5)**

5.4 Prototype-Based Person Re-Identification

As described in Section 4.1, the Prototype Discovery stage computes the similarity between a given sample image and the training images captured by the same camera using One-Shot Similarity (OSS) measure. The OSS measure depends on the choice of the discriminative model and the feature representation. We fixed the discriminative model as the PLS regression model and analyzed the influence of the different feature descriptors. We observed that the number of components (latent factors) used in the PLS model has a small impact in the obtained experimental results. Thus, we maintained it equals one.

Regarding the remaining stages, we evaluated the impact of the prototypes size on the obtained experimental results. Then, we show that probe- and gallery-based ranking lists are complementary and then, can be aggregated to achieve better results. Finally, we demonstrate that the proposed indirect matching using prototypes model is complementary to a well-known metric learning approach method that directly matches probe and gallery images.

Feature Descriptors. We first evaluated different feature descriptors on the VIPeR dataset. To reduce the computational cost, we first projected these high-dimensional feature descriptors into a low-dimensional feature space computed using Principal Component Analysis (PCA) method, which we fixed the dimensionality as 80 components. In addition, we report the obtained experimental results obtained after the aggregation of probe- and gallery-based ranking lists. According to Table 5.18, improved results were achieved when setting the descriptor as the GoG. These results are consistent with the obtained experimental results in PRID450S and CUHK01. Therefore, we focus on the GoG descriptor in the remaining experiments at these datasets.

Table 5.18: Matching rates for different ranking positions in the VIPeR dataset using the prototypes-based approach with different feature descriptors.

Descriptors	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
GoG	**26.3 (1.1)**	**57.4 (0.9)**	**72.8 (0.4)**	**81.3 (0.5)**	**86.4 (0.3)**
LOMO + CNN	23.7 (1.5)	56.8 (1.8)	73.3 (1.0)	82.3 (0.8)	87.7 (0.9)

PCA Projection. In Table 5.19, we measured the influence of the projection into a low-dimensional subspace on the obtained experimental results. Based on these results, it is possible to conclude that the low-dimensional representation improves the obtained results. We credit this performance gain to the reduction of the noise variations as we keep only the significant information. More importantly, as we apply the PCA using data from both cameras, we claim that it already focus on more stable features. Thus, in the remaining experiments we will use the GoG representation projected into a low-dimensional subspace computed using PCA.

Prototypes Size. An important parameter of the proposed model consists in the per-

Table 5.19: Matching rates for different ranking positions on the VIPeR dataset using the original GoG feature representation and a low-dimensional representation computed using PCA.

Descriptors	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
GoG + PCA	**26.3 (1.1)**	**57.4 (0.9)**	**72.8 (0.4)**	**81.3 (0.5)**	**86.4 (0.3)**
GoG	16.8 (0.7)	47.2 (0.9)	62.2 (1.2)	72.5 (0.5)	81.2 (0.8)

centage of individuals from training set used at each prototype subset - the parameter K. Using a small number of individuals, we take the risk of having similar individuals in different classes (positive and negative) when learning the prototypes model. Differently, if we have a huge number of individuals in the prototypes, it becomes too heterogeneous that no subtle characteristic is learned. Therefore, the parameter K must be carefully adjusted for each dataset.

Table 5.20 presents the obtained experimental results for different values of parameter K. Based on these experimental results, we conclude that the percentage of individuals that results in the best performance is 20%, which for the VIPeR dataset corresponds to approximately 63 individuals. We also performed an experimental evaluation of the parameter K in PRID450S and CUHK01 datasets reaching the values of 30% and 15%, respectively.

Table 5.20: Matching rates for different ranking positions on the VIPeR dataset using different percentage of prototypes subset.

Prototypes Size	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
10%	21.5 (2.1)	55.5 (2.6)	70.0 (1.5)	79.4 (1.9)	86.1 (0.4)
20%	**26.3 (1.1)**	**57.4 (0.9)**	**72.8 (0.4)**	**81.3 (0.5)**	**86.4 (0.3)**
30%	24.4 (1.5)	56.2 (1.2)	74.8 (1.0)	81.3 (1.0)	86.1 (0.6)

Aggregation Strategy. In this experiment, we validate our assumption that combining probe- and gallery-based ranking lists we obtain an improved ranking of gallery images. We tackle this problem using the Stuart ranking aggregation method [Stuart et al., 2003]. Figure 5.7 presents some ranking results for probe- and gallery-based methods. Based on these qualitative results, we can confirm that these ranking lists are capturing complementary information. For instance, some individuals are only correctly ranked when we consider both list, such as the first and the third individuals. Therefore, it is expected an improvement when aggregating these ranking lists.

Table 5.21 presents the obtained experimental results for the probe- and gallery-based results, and the final result reached aggregating both ranking lists. In fact, these results corroborate with our assumption. For instance, we obtained an improvement in the *rank-1* of 2.6 percentage points when compare to the second best result.

Table 5.21: Matching rates for different ranking positions on the VIPeR dataset for the probe-based, gallery-based and the aggregation strategies.

Strategy	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
Aggregated	**26.3 (1.1)**	**57.4 (0.9)**	**72.8 (0.4)**	**81.3 (0.5)**	**86.4 (0.3)**
Probe-based	23.7 (2.0)	51.6 (1.8)	68.5 (1.0)	78.3 (0.7)	83.7 (0.5)
Gallery-based	22.9 (2.5)	56.0 (1.3)	70.0 (1.2)	77.7 (1.0)	83.1 (0.9)

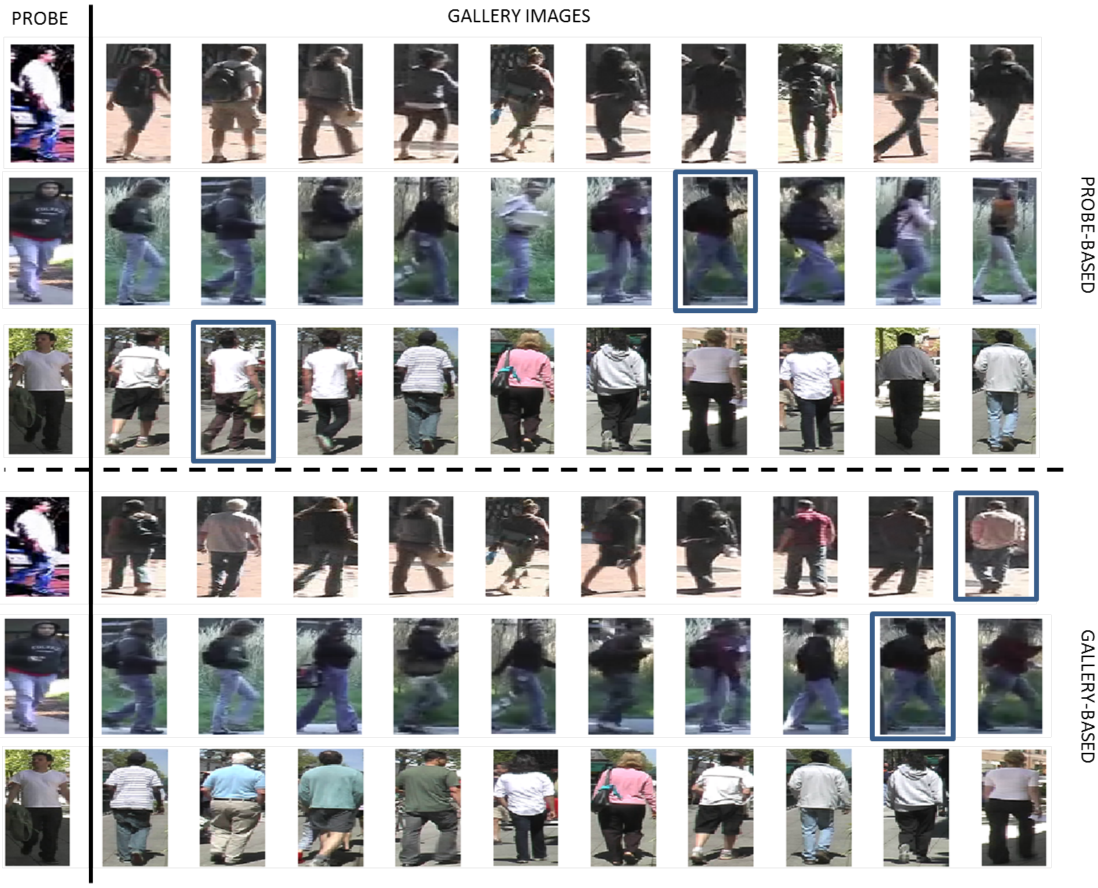

Figure 5.7: The top ten individuals in gallery set ranked accordingly with the similarity with probe images using probe-based (first three rows) and gallery-based (last three rows) ranking lists. Individuals surrounded by a blue box correspond to the correct match for the probe image in the gallery set.

Direct and Indirect Matching. The proposed method consists in an indirect matching of probe and gallery images using prototypes. However, a common approach in literature corresponds to directly compare probe and gallery images using a learned metric distance. Due to the different strategies, it is expected that the obtained results with the direct and indirect matching models are different and probably complementary. Therefore, we combined the obtained ranking list of KISSME metric learning [Koestinger et al., 2012] with the resulting ranking list of the proposed prototypes model using Stuart ranking aggregation method.

Table 5.22 shows a substantial improvement in both KISSME and the proposed Prototypes-based method when combined using ranking aggregation. For instance, the *rank-1* improved in 6.5 percentage points when compared to the second best performing approach.

Table 5.22: Matching rates for different ranking positions in the VIPeR dataset for the prototypes, KISSME and the final aggregation of direct and indirect matching models.

Strategy	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
Final	**35.0 (1.9)**	**71.5 (1.6)**	**84.3 (1.9)**	**91.0 (1.2)**	**94.3 (1.0)**
KISSME	28.5 (1.3)	64.7 (1.5)	81.3 (1.0)	88.3 (1.0)	91.8 (0.7)
Prototypes	26.3 (1.1)	57.4 (0.9)	72.8 (0.4)	81.3 (0.5)	86.4 (0.3)

5.5 Cross-View Kernel PLS (X-KPLS)

In this section, we evaluate the influence of the main parameters related to the X-KPLS model. First, we assess the performance of the proposed approach using the evaluated descriptors and distinct values of the parameter sigma. Then, we define the best number of latent factors. Finally, we present the performance gain due to the nonlinear modeling.

Feature Descriptors. In this experiment, we compare the performance of the evaluated feature descriptors on the proposed X-KPLS model using VIPeR dataset. Table 5.23 presents the obtained experimental results. Based on these results, we conclude that the best performing feature descriptor is the GoG, which also holds for PRID450S and CUHK01 datasets. Therefore, in the remaining experiments and datasets, we will use the GoG descriptor.

Parameter σ. In Table 5.24, we present the *rank-1* matching rates using the proposed Cross-view Kernel PLS method with different values of sigma and the VIPeR dataset.

Table 5.23: Matching rates for different ranking positions in the VIPeR dataset using the X-KPLS approach with different feature descriptors.

Descriptors	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
GoG	**38.7 (2.0)**	**74.1 (1.8)**	**86.2 (1.3)**	**91.5 (1.3)**	**94.3 (1.0)**
LOMO + CNN	34.9 (1.2)	69.0 (1.9)	81.2 (1.5)	86.7 (1.3)	90.3 (1.3)

According to these results, the best value of sigma is 2.0, which is the value that we set for the following experiments.

Table 5.24: Rank-1 matching rates in the VIPER dataset using the X-KPLS method with different values of parameter sigma.

Sigma	VIPeR (p=316)				
	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$	$\sigma = 5$
Rank-1	30.4 (2.0)	**38.7 (2.0)**	38.1 (3.1)	37.7 (3.0)	36.9 (3.7)

Latent Factors Table 5.25 shows the obtained experimental results using X-KPLS and different number of latent factors - the parameter f from Algorithm 8. Based on these results, we conclude that using 200 components we reach the best results with a reduced computational cost. We conducted similar experiments in PRID450S and CUHK01 datasets to set the number of latent factors as 150 and 300, respectively.

Table 5.25: Rank-1 matching rates in the VIPeR dataset using the X-KPLS with different numbers of latent factors.

Factors	VIPeR (p=316)				
	$f = 50$	$f = 100$	$f = 200$	$f = 250$	$f = 300$
rank-1	31.9 (1.7)	36.7 (2.0)	**38.7 (2.0)**	38.5 (2.1)	38.7 (1.6)

Nonlinear Mapping. Table 5.26 presents the obtained experimental results using the proposed model with linear (PLS) and nonlinear (Kernel PLS) regression models. Based on these results, it is possible to notice the improvement due to the nonlinear mapping. For instance, the *rank-1* increases 3.9 percentage points when using the Kernel PLS model. Figure 5.8 presents some qualitative results comparing the X-KPLS and PLS models. Observe that X-KPLS provides a higher ranking position for the corresponding gallery image in the two first probe images. Differently, the third gallery is not ranked between the top most similar individuals, which we relate to the drastic appearance change caused by the different pose and illumination conditions.

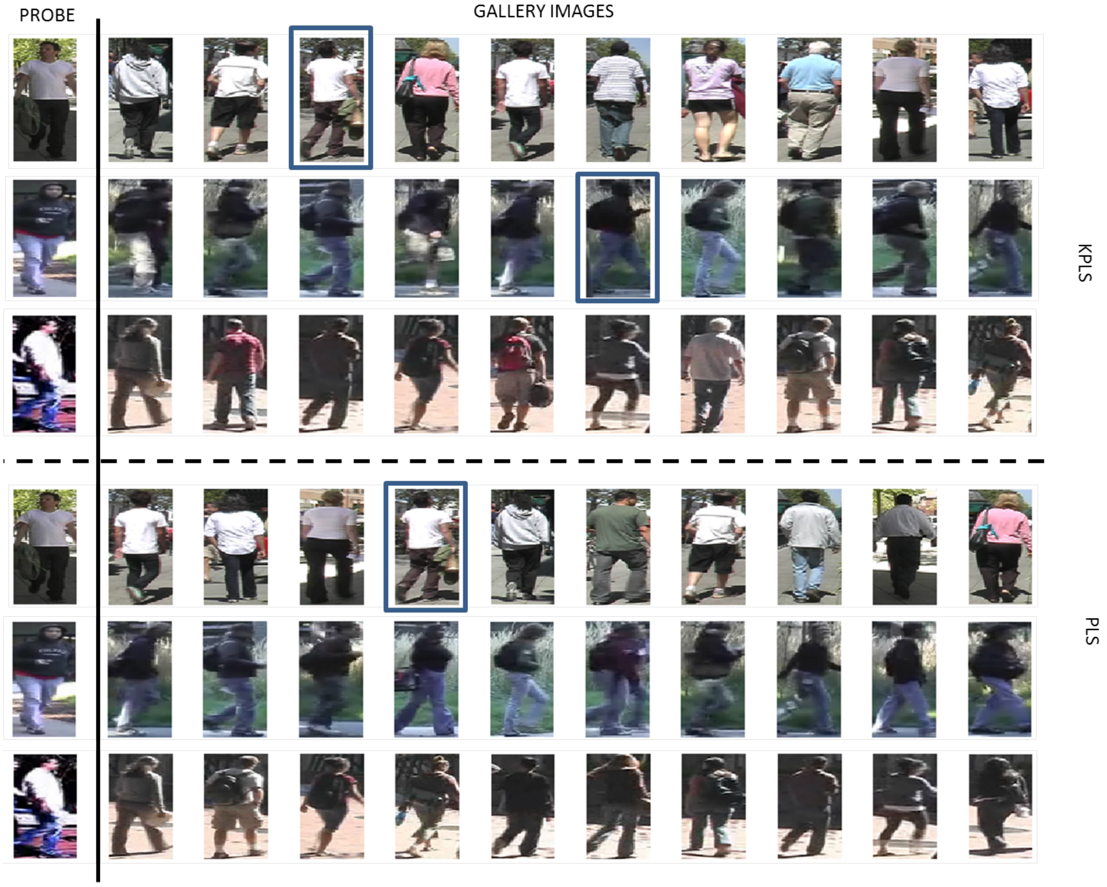

Figure 5.8: The top ten individuals in gallery set ranked accordingly with the similarity with probe images using Cross-View KPLS (first three rows) and PLS (last three rows) regression models. Individuals surrounded by a blue box correspond to the correct match for the probe image in the gallery set.

Table 5.26: Matching rates for different ranking positions in the VIPeR dataset using the X-KPLS with linear and nonlinear regression models.

Method	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
Kernel PLS	**38.7 (2.0)**	**74.1 (1.8)**	**86.2 (1.3)**	**91.5 (1.3)**	**94.3 (1.0)**
PLS	34.8 (2.4)	70.3 (1.4)	83.2 (1.4)	88.7 (0.9)	92.0 (0.6)

Attributes. The Cross-view KPLS model learns a nonlinear regression model for each camera. In Table 5.27 experiments, we consider the addition of attributes label in the response matrix **Y**. Similarly to the previous experiments, the attributes when combined with the identity labels boost the matching rates for all the ranking positions.

Table 5.27: Matching rates for different ranking positions in the VIPeR dataset using the X-KPLS with linear and nonlinear regression models.

Method	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
Ident.	38.7 (2.0)	74.1 (1.8)	86.2 (1.3)	91.5 (1.3)	94.3 (1.0)
Attr.	21.6 (1.6)	48.5 (3.3)	61.6 (2.9)	69.0 (3.2)	74.6 (3.0)
Ident. + Attr.	**39.5 (2.6)**	**76.0 (1.8)**	**87.4 (1.0)**	**92.1 (0.8)**	**94.8 (0.8)**

5.6 Kernel Cross-View Collaborative Representation based Classification

In the following sections, we evaluate the main parameters related to Kernel X-CRC model. First, we present the matching rates for different feature descriptors. Then, we assess the performance of the proposed method with different parameters values and, linear and nonlinear kernel functions. Finally, we analyze the performance improvement reached by working in a common and low-dimensional feature space.

Feature Descriptors. Table 5.28 presents the obtained experimental results, in VIPeR dataset, for the analyzed feature descriptors. Based on these results, we can notice that GoG achieves improved results when compared to LOMO+CNN. Similarly, we also observed improved results when using GoG descriptor in PRID450S and CUHK01 datasets. Therefore, we set GoG as the feature descriptor.

Table 5.28: Matching rates for different ranking positions in the VIPeR dataset using the proposed Kernel X-CRC with different feature descriptors.

Descriptors	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
GoG	**51.6 (2.0)**	**64.1 (2.3)**	**71.9 (2.7)**	**76.7 (2.4)**	**80.4 (2.4)**
LOMO + CNN	45.5 (2.1)	76.4 (2.4)	86.7 (1.5)	91.3 (0.8)	94.4 (0.7)

Parameter σ. Table 5.29 shows the obtained *rank-1* matching rates for different values of sigma in CUHK01 dataset. Based on these results, we set the value of sigma as 2.0, which is the value that we set for the following experiments.

Table 5.29: Rank-1 matching rates for different values of σ in the VIPeR dataset.

Parameter σ	VIPeR (p=316)			
	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$
rank-1	51.4 (2.4)	**51.6 (2.7)**	**51.6 (2.0)**	50.5 (2.1)

Parameter λ. In this experiment, we evaluate the proposed method with different values of λ when the parameter τ is fixed as 1.0. The parameter λ, from Equation 3.8, is the regularization parameter for the values of probe and gallery coding vectors (α_p and α_g). Table 5.30 presents the obtained experimental results for different values of λ in VIPeR dataset. Based on these results, we set λ as 0.01 for the following experiments.

Table 5.30: Rank-1 matching rates for different values of λ in the VIPeR dataset.

Parameter λ	VIPeR (p=316)			
	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 10$
rank-1	**51.6 (2.0)**	50.6 (2.8)	47.8 (2.1)	46.4 (2.4)

Parameter τ. In this experiment, we evaluate the impact on the experimental results of different values of τ, which is the parameter that multiplies the similarity term in Equation 3.8. For these experiments, we maintained the parameter λ fixed accordingly to the values defined in the previous experiment. Table 5.31 shows the obtained *rank-1* matching rates for different values of τ. These results show that τ equals 1.0 corresponds to the best trade-off between reconstruction and similarity terms and, therefore, the highest *rank-1* matching rate.

Table 5.31: Rank-1 matching rates for different values of τ in the VIPeR dataset.

Parameter τ	VIPeR (p=316)			
	$\tau = 0.01$	$\tau = 0.1$	$\tau = 1.0$	$\tau = 10$
rank-1	45.1 (2.4)	50.1 (2.1)	**51.6 (2.0)**	48.8 (1.6)

Nonlinear Mapping. We also evaluated the effect of the nonlinear mapping in the obtained experimental results. To accomplish that, we evaluated the proposed method using the RBF and linear kernel function. Table 5.32 presents the obtained experimental results using both kernel functions. Based on these results, we can conclude that the nonlinear mapping is important to improve the experimental results.

Subspace Learning. In Table 5.33, we compare the Kernel X-CRC using GoG and low-dimensional representations obtained using methods from literature, such as KCCA [Lisanti et al., 2014], XQDA [Liao et al., 2015] and MLAPG [Liao and Li, 2015]. Regarding these results, we conclude that a performance gain occurs only when dealing with XQDA and MLAPG methods, which consider same and not-same image pairs when learning the low-dimensional representation. For instance, using XQDA we

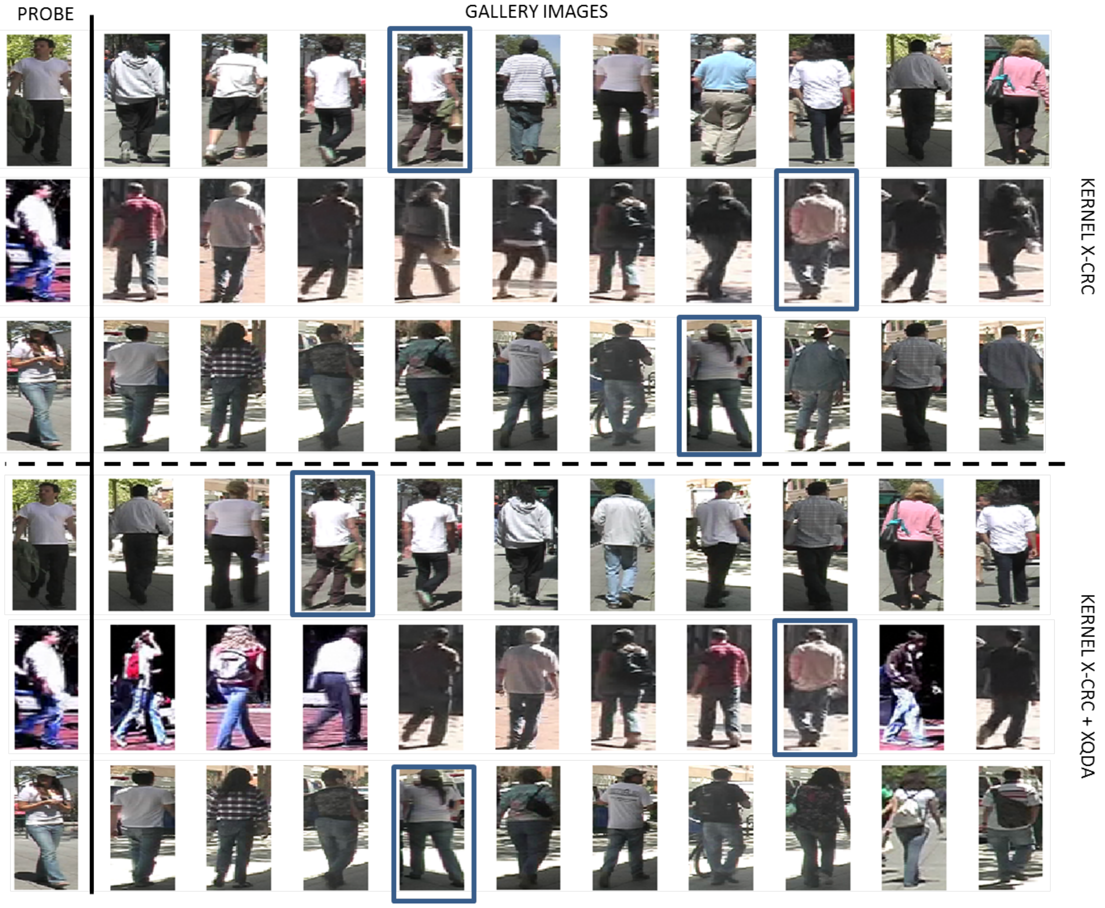

Figure 5.9: The top ten individuals in gallery set ranked accordingly with the similarity with probe images using Kernel X-CRC and the original GoG descriptor (first three rows) or the low-dimensional representation learned using XQDA (last three rows) regression models. Individuals surrounded by a blue box correspond to the correct match for the probe image in the gallery set.

Table 5.32: Matching rates for different ranking positions in the VIPeR dataset using Kernel X-CRC with nonlinear and linear kernel functions.

Kernel	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
RBF	**51.6 (2.0)**	**64.1 (2.3)**	**71.9 (2.7)**	**76.7 (2.4)**	**80.4 (2.4)**
Linear	49.6 (1.8)	77.5 (2.4)	87.9 (1.6)	91.9 (1.5)	94.2 (1.0)

improved the *rank-1* matching rate in more than 5.0 percentage points. Differently, KCCA computes a common subspace that only considers image pairs of the same individuals in the objective function, and, therefore, reduces the obtained experimental results. Therefore, in the remaining experiments, we will focus on the proposed Kernel X-CRC method using GoG descriptor projected into the XQDA feature space.

Figure 5.9 shows some ranking results for the Kernel X-CRC using only GoG and the low-dimensional representation computed using XQDA. Based on these results, it is possible to observe that Kernel X-CRC ranks the gallery images between the ten most similar individuals when using the GoG feature descriptor. Nonetheless, when using the representation obtained using XQDA, it is able to better discriminate between these individuals, improving the ranking position of the corresponding gallery image.

Table 5.33: Matching rates for different ranking positions in the VIPeR dataset using *Kernel X-CRC* in the original and low-dimensional feature spaces.

Descriptors	VIPeR (p=316)				
	r = 1	r = 5	r = 10	r = 15	r = 20
GoG + XQDA	**51.6 (2.7)**	**80.5 (2.0)**	89.6 (1.5)	93.4 (1.2)	95.4 (0.8)
GoG + MLAPG	48.0 (2.0)	78.1 (1.1)	87.1 (1.2)	91.4 (0.8)	93.9 (0.7)
GoG	46.3 (1.5)	79.4 (1.9)	89.3 (1.4)	93.3 (1.1)	95.6 (1.1)
GoG + KCCA	42.9 (1.0)	79.0 (0.7)	**89.7 (1.3)**	**94.0 (1.0)**	**96.0 (1.9)**
GoG + PCA	46.5 (1.6)	79.6 (2.4)	89.6 (1.2)	93.3 (1.0)	95.7 (1.2)

5.7 State-of-the-art Comparisons

In this section, we compare the proposed approaches for subspace learning and indirect matching with a large number of methods from state-of-the-art in the VIPeR , PRID450S and CUHK01 datasets. In the following tables, we report different methods from person re-identification literature ranked according to the *rank-1* matching rate. In addition, we also present the obtained experimental results with the methods proposed in lowest part of the table. As the methods from literature did not report the standard deviation, we just present these values for our methods. The only exception is the XQDA [Liao et al., 2015] that has the code online available.

VIPeR

Table 5.34 presents the matching rates from different approaches based on metric learning [Jose and Fleuret, 2016; Liao and Li, 2015; Huang et al., 2015], subspace learning [Lisanti et al., 2014; Liao et al., 2015; Zhang et al., 2016a; Chen et al., 2015] and deep learning [Shangxuan et al., 2016; Cheng et al., 2016; Chen et al., 2016b].

Regarding the results in Table 5.34, we can conclude that the Kernel MBPLS presents better results than the Kernel PLS for Subspace Learning model and the Kernel HPCA. We attribute these results to nonlinear regression to the identity + attribute labels, which allows a better separation of the data in the learned subspace and therefore, better generalization. Notice that all the proposed subspace learning methods outperform the classical KCCA model. We relate this improvement to the better feature representation and to the variance information that is considered only in the proposed model. Table 5.34 also shows that the proposed subspace learning models are outperformed by related approaches from literature as the Null Space [Zhang et al., 2016a] and KMFA [Chen et al., 2015] that also handle person re-identification as a common subspace learning problem. Nonetheless, these methods require the fine-tuning of regularization parameters and are not scalable to scenarios with multiple surveillance cameras. Otherwise, both models proposed do not depend on the regularization of covariance matrices and, the Kernel HPCA and the Kernel MBPLS efficiently deal with multiple surveillance cameras.

The indirect matching methods perform well in the person re-identification problem, as shown in Table 5.34, with the lowest results being reported by the prototype-based approach. These results are consistent with our observation that discovering the prototypes subset is a difficult task. Thus, X-KPLS greatly improves the obtained matching rates as consequence of using the entire training set instead of a subset of individuals. However, the Kernel X-CRC that achieves state-of-the-art results as consequence of the adaptive matching of probe and gallery images using a multi-task framework. Zhang et al. [2016b] also employs a specific model for each pair of probe and gallery images. Nonetheless, their models are obtained using a mapping function to relate feature descriptors to model parameters, which is very challenging for small datasets such as VIPeR.

Based on Table 5.34, the Kernel X-CRC results are close to the baseline that also uses the XQDA and GoG descriptor when considered the mean and the standard deviation. Nonetheless, as these measures are obtained using the same partitions, a better statistical analysis can be obtained using a paired t-test. In fact, performing this study, we observed that there is a strong evidence that the Kernel X-CRC is superior to the XQDA + GoG with a confidence interval of 95%. In fact, Kernel X-CRC results are only lower than the SCSP [Chen et al., 2016a] that combines global and local spatial matching models. However, SCSP [Chen et al., 2016a] did not provide the partition used in their experiments and, therefore, the direct comparison between the obtained results is not possible.

Figure 5.10 presents the top ten gallery images ranked accordingly the similarity with the same probe image, where each row corresponds to the application of a proposed similarity model. These results show that Kernel X-CRC model better discriminates the correct gallery image. For instance, the prototype-based and the KPLS for Subspace Learning did not return the correct image between the top most similar individuals.

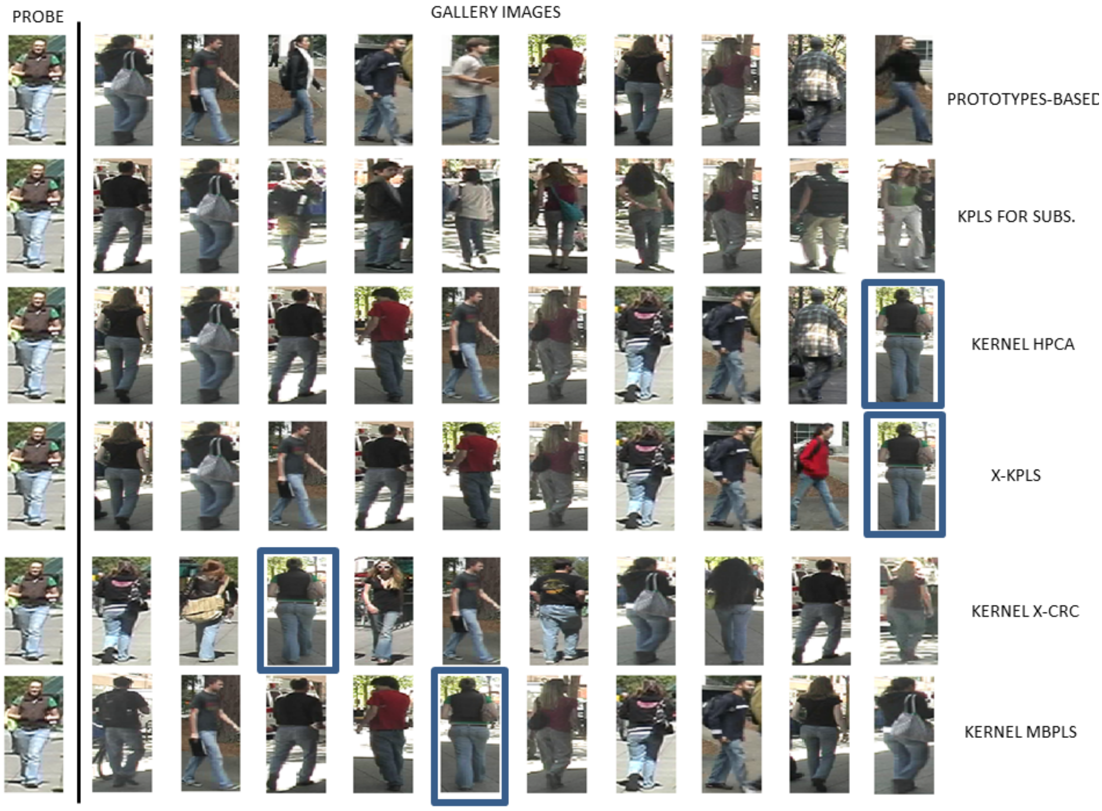

Figure 5.10: The first ten individuals in gallery set ranked accordingly with the similarity with probe images using the proposed approaches. Individuals surrounded by a blue box correspond to the correct match for the probe image in the gallery set.

PRID450S

In Table 5.35, we present the matching rates for different methods that address person re-identification using PRID450S dataset. Based on these results, Kernel MBPLS method reaches improved results when compared to Kernel PLS for Subspace Learning and Kernel HPCA, which corroborates with our assumption that nonlinear regression models perform better in person re-identification problem. However, both methods are outperformed by subspace learning models from literature (XQDA and KMFA) that consider same and not-same constraints and, therefore, are able to learn more discriminative subspaces.

As in the previous section, X-KPLS obtained results superior to the prototypes-based approach due to the representation using the entire training set instead of just a subset of individuals. More importantly, Kernel X-CRC reaches improved results when compared to the proposed methods, and the highest *rank-1* matching rate when compare to the methods from literature. As matter of fact, we performed a paired t-test that shows with 95% of confidence that the ranking results from Kernel X-CRC are

Method	Viper (p=316)				
	r = 1	r = 5	r = 10	r = 20	r = 30
KCCA [Lisanti et al., 2014]	37.0	-	85.0	93.0	-
Deep Ranking [Chen et al., 2016b]	38.4	69.2	81.3	90.4	94.1
LOMO + XQDA [Liao et al., 2015]	40.0	68.0	80.5	91.1	95.5
WARCA [Jose and Fleuret, 2016]	40.2	68.2	80.7	91.1	-
MLAPG [Liao and Li, 2015]	40.7	-	82.3	92.4	-
NLML [Huang et al., 2015]	42.3	71.0	85.2	94.2	-
Null Space [Zhang et al., 2016a]	42.3	71.5	82.9	92.1	-
Zhang et al. [2016b]	42.7	-	84.3	91.9	-
Mirror + KMFA [Chen et al., 2015]	43.0	75.8	87.3	94.8	-
Paisitkriangkrai et al. [2015]	45.9	77.5	88.9	95.8	-
MultiCNN [Cheng et al., 2016]	47.8	74.7	84.8	91.1	94.3
GoG + XQDA [Matsukawa et al., 2016]	48.2 (3.0)	77.3 (1.5)	87.6 (1.1)	91.5 (0.9)	-
Shangxuan et al. [2016]	51.1	81.0	91.4	96.9	-
SCSP [Chen et al., 2016a]	**53.5**	**82.6**	**91.5**	96.6	-
Kernel X-CRC	51.2 (2.4)	79.9 (2.2)	89.9 (1.5)	95.5 (1.0)	-
Kernel MBPLS	41.1 (1.7)	76.7 (2.6)	88.3 (2.0)	94.8 (1.1)	-
Kernel HPCA	39.1 (2.0)	75.3 (2.6)	86.9 (1.8)	94.2 (1.3)	-
X-KPLS	38.8 (2.2)	74.0 (2.3)	86.0 (1.4)	97.0 (0.8)	-
KPLS Subs.	38.1 (2.2)	72.8 (2.6)	84.0 (2.2)	92.5 (1.3)	-
Prototypes	33.7 (2.9)	70.3 (2.0)	84.0 (1.9)	97.0 (1.4)	-

Table 5.34: Top ranked approaches on the VIPer dataset.

superior to our baseline (XQDA + GoG). We attribute this improvement to the better representation obtained using GoG descriptor with XQDA [Matsukawa et al., 2016] and the nonlinear computation of coding vectors. It is also important to highlight that some methods that achieved interesting results in VIPeR dataset did not performed well in PRID450S due to small number of training samples, such as WARCA [Jose and Fleuret, 2016] and SCSP [Chen et al., 2016a]. Differently, the proposed Kernel X-CRC achieves state-of-the-art results in both datasets.

CUHK01

Table 5.36 presents the reported experimental results for different methods in literature that used the CUHK01 dataset considering the single-shot scenario. In addition, we also present the obtained experimental results of the proposed methods. Based on these results, we can notice that, similarly to the previous datasets evaluated, the Kernel MBPLS model performs better than KPLS for Subspace Learning and Kernel HPCA. Nonetheless, those methods are still worst than other subspace learning models from literature (XQDA and KMFA). As mentioned before, these methods include equivalence

Method	PRID450S (p=225)				
	r = 1	r = 5	r = 10	r = 20	r = 30
WARCA [Jose and Fleuret, 2016]	24.6	55.5	70.3	85.0	92.0
SCSP [Chen et al., 2016a]	44.4	71.6	82.2	89.8	93.3
Mirror + KMFA [Chen et al., 2015]	55.4	79.3	87.8	91.6	-
Zhang et al. [2016b]	60.5	-	88.6	93.6	-
LOMO + XQDA [Liao et al., 2015]	61.4	-	90.8	95.3	-
Shangxuan et al. [2016]	66.6	86.8	92.8	96.9	-
GoG + XQDA [Matsukawa et al., 2016]	66.2 (2.0)	87.8 (1.5)	92.6 (1.2)	95.2 (0.9)	-
Kernel X-CRC	**68.1 (1.8)**	**90.7 (1.4)**	**95.0 (1.3)**	**97.6 (0.8)**	-
Kernel MBPLS	47.3 (3.0)	75.4 (2.4)	86.0 (1.7)	94.2 (1.8)	-
Kernel HPCA	46.9 (3.2)	75.7 (2.4)	86.2 (1.7)	94.2 (1.9)	-
KPLS Subs.	46.4 (1.7)	72.8 (2.6)	84.0 (2.2)	92.5 (1.3)	-
X-KPLS	46.3 (2.9)	75.4 (1.4)	86.4 (1.8)	94.4 (2.0)	-
Prototypes	34.5 (3.6)	67.9 (2.0)	80.4 (1.9)	95.4 (1.6)	-

Table 5.35: Top ranked approaches on the PRID450S dataset.

constraints in the learned subspace and, therefore, are more discriminative. However, they do not address issues that are considered in the Kernel HPCA and Kernel MBPLS, such as the scalability and small-sample-size problem.

Table 5.36 shows that, as in the previous datasets, the indirect matching models perform well in the person re-identification problem. More importantly, Kernel X-CRC reached 62.2% in the *rank-1* matching rate, which is only inferior to the WARCA [Jose and Fleuret, 2016] model. Actually, WARCA and MultiCNN models seem to heavily depend on the number of training samples available as they performed considerable better in CUHK01 dataset. Differently, due to the collaborative representation approach, the proposed Kernel X-CRC performs well in both small and large scale datasets. Besides, as these methods used different partitions of the data, the direct comparison is not fair.

Method	CUHK01 (p=485)				
	r = 1	r = 5	r = 10	r = 20	r = 30
Mirror + KMFA [Chen et al., 2015]	40.4	64.6	75.3	84.1	-
Paisitkriangkrai et al. [2015]	53.4	76.4	84.4	90.5	-
MultiCNN [Cheng et al., 2016]	53.7	84.3	**91.0**	**96.3**	**98.3**
GoG + XQDA [Matsukawa et al., 2016]	56.2 (1.0)	78.8 (1.4)	85.7 (1.4)	89.1 (1.3)	91.8 (1.2)
WARCA [Jose and Fleuret, 2016]	**65.6**	**85.3**	90.5	95.0	-
Kernel X-CRC	62.2 (1.8)	83.0 (2.1)	89.3 (1.2)	92.12 (1.1)	94.2 (0.9)
X-KPLS	44.0 (2.5)	70.0 (2.0)	79.1 (1.5)	86.7 (1.3)	-
Kernel MBPLS	40.4 (2.6)	65.3 (2.1)	74.9 (1.2)	83.3 (1.0)	-
Kernel HPCA	39.9 (2.5)	65.7 (1.7)	75.1 (1.7)	83.1 (1.6)	-
Prototypes	35.2 (2.1)	62.1 (1.3)	70.0 (1.2)	78.7 (1.1)	-
KPLS Subs.	32.5 (2.0)	54.6 (1.7)	64.4 (2.0)	73.3 (1.8)	-

Table 5.36: Top ranked approaches on the CUHK01 dataset.

Chapter 6

Experimental Results - Multiple Cameras

In this chapter, we present the experimental results with different parameters and configurations considering the multiple cameras datasets WARD [Martinel and Micheloni, 2012] and RAID [Das et al., 2014]. These datasets contain more than two cameras and are useful to analyze person re-identification methods that consider multiple cameras model instead of the traditional camera pairwise models. Pairwise models capture nuances of a specific camera pair but are not scalable with the number of surveillance cameras. For instance, in a network with c connected cameras, we have $c(c-1)/2$ pair of cameras. One can argue that is possible to use spatio-temporal reasoning to cluster cameras in subsets [Loy et al., 2010]. Nonetheless, we would still have dozens of cameras clusters in a realistic setting. Therefore, we claim that is crucial to tackle the camera scalability problem when addressing the person re-identification problem.

In the following sections, we present experimental results for the three proposed multiple camera methods: Kernel Hierarchical PCA (Section 6.1), Kernel Multiblock PLS (Section 6.2) and the Cross-view Kernel PLS (Section 6.3). While the first learns a subspace that correlates samples of the same individual captured in the camera network, the last two are regression-based approaches and, therefore, include multiple information in the response matrix $\mathbf{Y}$. Particularly, we consider identity, attributes and the concatenation of identity and attributes labels. To accomplish that we manually labelled 24 attributes for each individual in the WARD dataset, which are detailed in Table 6.1. Despite being manually labeled, these attributes could be automatically predicted using some state-of-the-art method [Schumann and Stiefelhagen, 2017]. Figure 6.1 shows some examples of annotated attributes for three subjects in the WARD dataset.

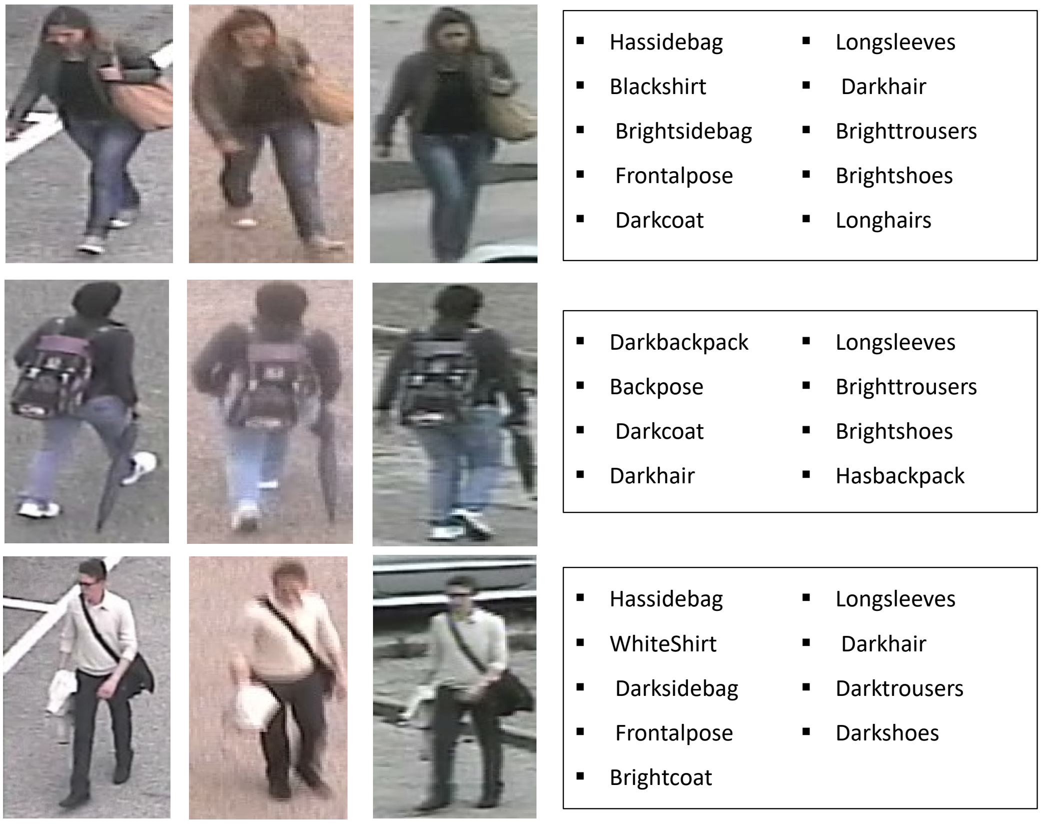

Figure 6.1: Manually labeled attributes for three individuals in the WARD dataset.

Table 6.1: Manually annotated attributes and pose information in WARD dataset.

Colorshirt	Blackshirt	Whiteshirt	Grayshirt	Strippedshirt
Stampedshirt	Hassidebag	Darksidebag	Brightsidebag	Hasbackpack
Darkbackpack	Brightbackpack	Frontalpose	Backpose	Darkcoat
Brightcoat	Darkhair	Brighthair	Blacktrousers	Brighttrousers
Darkshoes	Brightshoes	Longhairs	Longsleeves	

Despite the different strategies for training (pairwise or multiple cameras models), we keep the same testing protocol for a fair a comparison with state-of-the-art methods. Specifically, we select one camera as probe and other camera as gallery when computing the matching rates. Furthermore, when comparing different approaches, we use the mean matching rate for all the camera pairs as a performance measure. These mean matching rates are computed after running experiments on ten distinct partitions for each pair of cameras.

In the following, we describe the two multi-cameras dataset and feature descriptors evaluated.

Datasets

We consider the multiple cameras setting in these experiments, which is a more realistic scenario. Particularly, we selected two widely used multi-camera re-identification datasets: WARD [Martinel and Micheloni, 2012] and RAID [Das et al., 2014]. The WARD considers three surveillance cameras in an outdoor environment with small variations in the person's pose (see Fig. 6.2). Differently, the RAID dataset is a four-cameras dataset with indoor and outdoor cameras and large variations in pose and illumination (see Fig. 6.3). The following paragraphs details these datasets.

WARD dataset. WARD [Martinel and Micheloni, 2012] consists of a collection of 4786 images of 70 individuals captured by three non-overlapping cameras, where each individual has multiple images at each camera. The main challenges are related to the strong variations in the image resolution and illumination conditions.

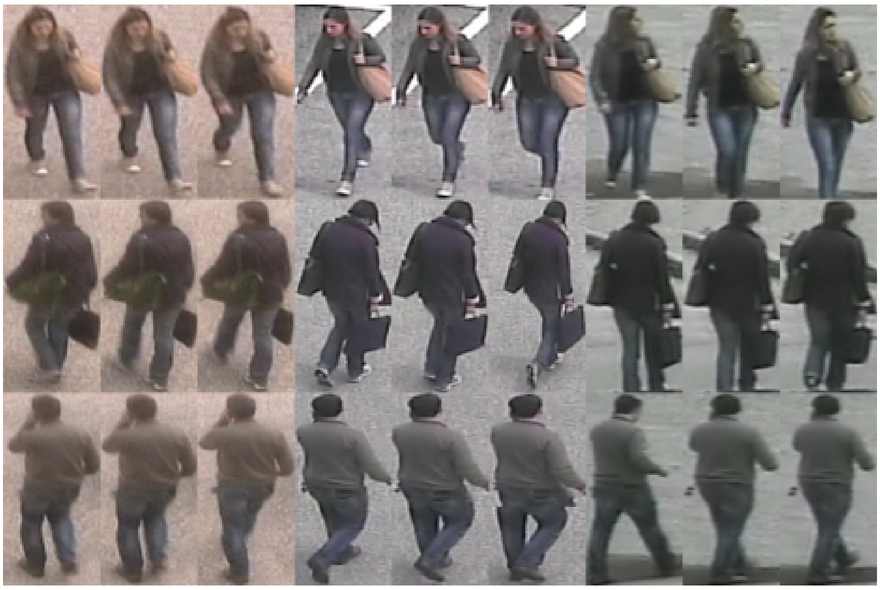

Figure 6.2: WARD dataset. Each row corresponds to images of the same individual captured by three surveillance cameras (i.e., each triplet corresponds to a distinct camera).

RAID dataset. RAID [Das et al., 2014] dataset is composed by 43 persons asked to walk through four surveillance cameras resulting in a total of 6920 annotated images. To assess the impact of different illumination conditions, the authors used two indoors and two outdoors cameras.

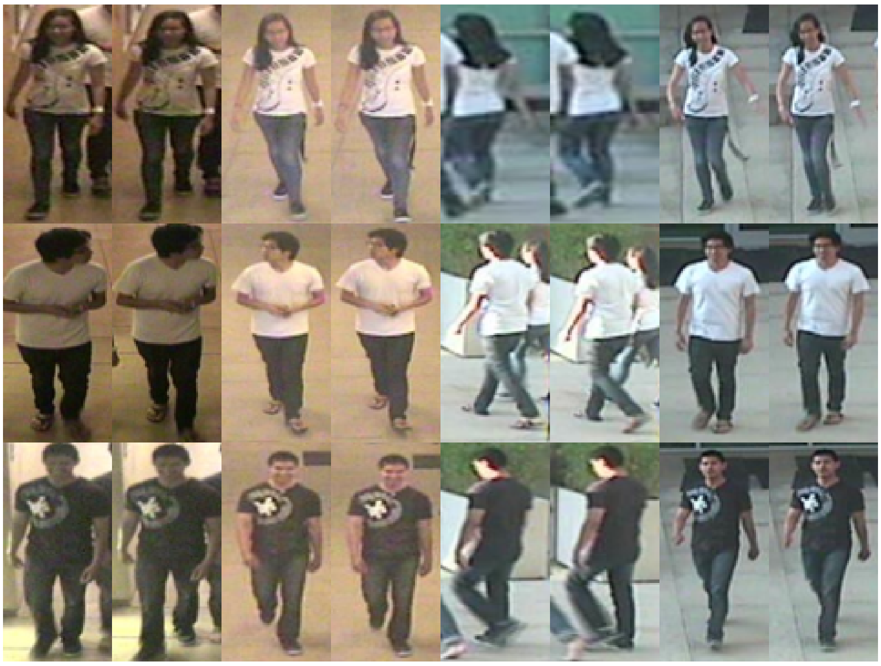

Figure 6.3: RAID dataset. Each row corresponds to images of the same individual captured by four surveillance cameras (i.e., each pair of images corresponds to a distinct camera).

Feature Descriptors

As the proposal of novel feature descriptors is not the focus of this work, we compared two available feature descriptors in the person re-identification literature. Specifically, we employed the GoG [Matsukawa et al., 2016] and the LOMO [Liao et al., 2015] descriptors, which we extracted using the online available code provided by the authors. To more information about the features and kernel functions, refer to Chapter 5.

Notice that we used in these experiments the LOMO alone instead of the fusion of LOMO and CNN features (LOMO + CNN) [Shangxuan et al., 2016]. The main reason for that is the fact that the authors did not make online available the code or the features for WARD and RAID datasets.

In the following sections, we present experiments considering the distinct methods proposed and these features descriptors.

6.1 Kernel Hierarchical PCA

In this section, we approach the multiple cameras person re-identification problem using the Kernel HPCA model. We consider the different parameters of the proposed method (Section 6.1.1) and the comparison between multiple and pairwise cameras (Section 6.1.2). It is important to highlight that the Kernel HPCA does not consider the response matrix **Y** when computing the low-dimensional representation and, therefore, we did not perform experiments including the attributes information.

6.1.1 Parameters Setting

In the following experiments, we consider different parameters that impact in the matching rate of the proposed Kernel HPCA model. In particular, these parameters are the feature descriptors, the distinct values of sigma (σ) and the number of latent factors.

Feature Descriptors. We performed experiments using the Kernel HPCA and the two feature descriptors, LOMO and GoG. Table 6.2 and 6.3 present the obtained matching rates for different ranking positions using WARD and RAID datasets, respectively. Similarly to previous experiments, GoG outperformed the LOMO by a large margin. For instance, the *rank-1* is 4.9 and 8.6 percentage points higher in WARD and RAID datasets, respectively. Thus, in the remaining experiments, we will employ the GoG descriptor.

Table 6.2: GoG and LOMO descriptors evaluated using the proposed Kernel HPCA in the WARD dataset.

Descriptors	WARD (p=35)				
	r = 1	r = 2	r = 3	r = 4	r = 5
GoG	73.1 (6.0)	83.1 (5.6)	88.9 (4.6)	92.2 (3.9)	94.0 (2.8)
LOMO	67.2 (7.9)	80.0 (8.6)	86.4 (6.4)	90.4 (4.8)	92.2 (4.0)

Table 6.3: GoG and LOMO descriptors evaluated using the proposed Kernel HPCA in the RAID dataset.

Descriptors	RAID (p=20)				
	r = 1	r = 2	r = 3	r = 4	r = 5
GoG	76.8 (9.3)	88.5 (6.5)	93.5 (4.9)	95.9 (3.9)	97.8 (2.3)
LOMO	68.2 (8.8)	83.3 (5.4)	89.8 (4.9)	93.0 (4.2)	95.5 (3.6)

Parameter σ. Table 6.4 and 6.5 present the *rank-1* matching rate for distinct values of sigma in WARD and RAID dataset, which is the parameter related to the RBF-kernel computation. Based on these results, it is possible to observe that suboptimal results are obtained using sigma as 0.1 and 0.5. Differently, best results are obtained for values closer to 1. Furthermore, we noticed that for values higher than 1.0 the experimental results did not present high variations. Based on these results, we set sigma as 3.0 and 1.0 for WARD and RAID datasets, respectively.

Table 6.4: Evaluation of different values of sigma using the Kernel HPCA method in the WARD dataset.

Sigma	WARD (p=35)			
	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 3$
Rank-1	4.1 (1.4)	53.2 (5.3)	70.9 (7.3)	**73.1 (6.0)**

Table 6.5: Evaluation of different values of sigma using the Kernel HPCA method in the RAID dataset.

Sigma	RAID (p=20)			
	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 3$
Rank-1	9.1 (2.3)	43.2 (4.9)	**76.8 (9.3)**	76.2 (8.7)

Latent Factors. Table 6.6 and 6.7 show the *rank-1* matching rates using WARD and RAID datasets with different number of factors in the Kernel HPCA model. For both datasets, we observed that adding too many components degrades the performance while increasing the computational cost. Thus, we achieved the best results using 30 and 20 components for WARD and RAID, respectively. We set these values for the remaining experiments.

Table 6.6: Evaluation of different number of factors using the Kernel HPCA in the WARD dataset.

Factors	WARD (p=35)			
	$f = 10$	$f = 20$	$f = 30$	$f = 50$
Rank-1	56.0 (7.8)	68.6 (6.7)	**73.1 (6.0)**	54.2 (5.3)

Table 6.7: Evaluation of different number of factors using the Kernel HPCA in the RAID dataset.

Factors	RAID (p=20)			
	$f = 10$	$f = 20$	$f = 30$	$f = 50$
Rank-1	72.9 (9.8)	**76.8 (9.3)**	74.4 (7.2)	67.5 (9.0)

6.1.2 Multiple Cameras

In this section, we compare the Kernel HPCA in a multiple camera setting, where a single model is learned for the entire camera network, with the pairwise counterpart that learns a model specific for each camera pair. It is important to emphasize that the multiple cameras model has the advantage of being scalable, but has to deal with all cameras at once. Differently, the pairwise learns subtle characteristics present at each camera pair and is limited to a scenario with only few cameras.

Table 6.8 presents the mean matching rate for different ranking positions when considering all possible combination of probe and gallery cameras. It is a common metric when assessing the overall performance of the proposed method in the surveillance system. Based on these results, it is possible to conclude that the proposed Kernel HPCA in a multiple camera setting is not just scalable, but also more discriminative than the pairwise model. We attribute these results to a better generalization of the learned low-dimensional representation when considering more cameras. Table 6.9 details these results for each pair of cameras. Notice that the multiple cameras outperforms the pairwise model for all camera pairs.

Table 6.8: Matching rates for multiple and pairwise cameras using the Kernel HPCA models in WARD dataset. These results are detailed for each pair of cameras.

Training Strategy	r = 1	r = 2	r = 3
Multiple Cameras	73.1 (6.0)	83.1 (5.6)	88.9 (4.6)
Pairwise Cameras	69.5 (6.5)	81.6 (5.3)	87.0 (4.0)

Table 6.9: Matching rates for multiple and pairwise cameras using the Kernel HPCA models in WARD dataset. These results are the mean matching rates when considering all the camera pairs.

Training Strategy	Probe/Gallery	r = 1	r = 2	r = 3
Multiple Cameras	A/B	81.4 (5.8)	89.7 (5.7)	94.0 (4.4)
	B/C	71.1 (6.5)	81.7 (5.7)	88.3 (4.6)
	A/C	66.6 (5.6)	77.7 (5.2)	84.3 (4.9)
Pairwise Cameras	A/B	75.1 (7.4)	88.6 (4.0)	92.3 (4.1)
	B/C	67.7 (4.7)	77.1 (7.3)	83.7 (5.2)
	A/C	65.7 (7.3)	77.1 (7.3)	83.7 (5.2)

Table 6.10 presents a similar analysis for the RAID dataset. In this case, better results were obtained using the pairwise model. In fact, the RAID dataset shows large variations of pose and illumination conditions than the WARD dataset, which can be better captured when considered each pair independently instead of using a unique that has to capture all the data variation at once.

Table 6.10: Matching rates for multiple and pairwise cameras Kernel HPCA models in RAID dataset.

Training Strategy	r = 1	r = 2	r = 3
Multiple Cameras	76.8 (9.3)	88.5 (6.5)	93.5 (4.9)
Pairwise Cameras	80.0 (7.3)	89.9 (4.8)	94.9 (3.4)

6.2 Kernel Multiblock PLS

In this section, we consider the proposed Kernel Multiblock PLS in a multiple camera setting. We first evaluate the different parameters of the proposed method (Section 6.2.1) and, then, we present the comparison between multiple cameras and pairwise settings (Section 6.2.2). While in these first experiments we use only identity information as the regression matrix **Y**, Section 6.2.3 considers the impact on the matching as consequence of the addition of attributes labels.

6.2.1 Parameters Setting

In this section, we perform experiments considering different parameters that impact in the matching rate of the proposed Kernel Multiblock PLS. Specifically, we evaluate the the feature descriptors, distinct values of sigma and the number of latent factors. In these experiments, we set the matrix **Y** using only the identity information.

Feature Descriptors. Table 6.11 presents the experimental results obtained using the proposed Kernel MBPLS using two widely used feature descriptors (GoG and LOMO) in the WARD dataset. These results demonstrate the superiority of the GoG when compared to LOMO. Similarly, we observe improved results in the RAID dataset when using GoG (see Table 6.12). Based on these improved results, we set the feature descriptor as the GoG for the remaining experiments.

Table 6.11: GoG and LOMO descriptors evaluated using the proposed Kernel MBPLS in the WARD dataset.

Descriptors	WARD (p=35)				
	r = 1	r = 2	r = 3	r = 4	r = 5
GoG	**76.5 (5.1)**	**86.8 (5.8)**	**92.1 (4.7)**	**94.5 (3.9)**	**95.8 (3.6)**
LOMO	69.8 (6.0)	84.0 (5.2)	91.1 (3.2)	94.3 (2.4)	95.7 (2.8)

Parameter σ. Tables 6.13 and 6.14 present the experimental results using the proposed Kernel MBPLS method with different values of sigma in WARD and RAID

Table 6.12: GoG and LOMO descriptors evaluated using the proposed Kernel MBPLS in the RAID dataset.

Descriptors	RAID (p=20)				
	r = 1	r = 2	r = 3	r = 4	r = 5
GoG	**78.4 (7.8)**	**89.7 (6.3)**	**94.4 (3.8)**	**97.1 (3.1)**	**98.4 (2.4)**
LOMO	71.4 (8.2)	85.8 (6.7)	92.3 (4.6)	95.5 (3.3)	96.9 (2.9)

datasets, respectively. These results demonstrate that the performance declines for values lower than 1.0. More importantly, the best results are achieved for sigma equals 3 and 1 for WARD and RAID datasets, respectively. We set sigma using these values in the remaining experiments.

Table 6.13: Evaluation of different values of sigma (σ) using the Kernel MBPLS in the WARD dataset.

Sigma	WARD (p=35)			
	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 3$
Rank-1	7.1 (3.4)	60.2 (7.3)	74.7 (6.4)	**79.0 (5.6)**

Table 6.14: Evaluation of different values of sigma (σ) using the Kernel MBPLS in the RAID dataset.

Sigma	RAID (p=20)			
	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 3$
Rank-1	10.6 (6.0)	71.3 (8.9)	**78.4 (7.8)**	78.2 (8.4)

Latent Factors. Table 6.15 and 6.16 show the obtained experimental results when using the Kernel MBPLS model with distinct number of factors in WARD and RAID datasets, respectively. Based on these results, we can notice that using a small number of components (i.e. 10 components) or large number (i.e. 70 components) we obtain suboptimal results. In fact, the best performance is achieved when using 50 and 30 components for WARD and RAID datasets, respectively.

6.2.2 Multiple Cameras

In this section, we assess the robustness of the proposed Kernel MBPLS to the addition of multiple surveillance cameras. To accomplish that, we compare the matching rates

Table 6.15: Evaluation of different number of factors using the Kernel MBPLS in the WARD dataset.

Factors	WARD (p=35)			
	$f = 10$	$f = 30$	$f = 50$	$f = 70$
Rank-1	53.9 (6.6)	75.3 (6.0)	**76.5 (5.1)**	63.9 (20.9)

Table 6.16: Evaluation of different number of factors using the Kernel MBPLS in the RAID dataset.

Factors	RAID (p=20)			
	$f = 10$	$f = 30$	$f = 50$	$f = 70$
Rank-1	71.3 (9.8)	**78.4 (7.8)**	75.3 (6.8)	75.0 (8.3)

obtained in a pairwise (i.e., a specific model for each pair of cameras) and multiple cameras (i.e., a single model for the camera network) settings with the Kernel MBPLS.

Table 6.17 presents the obtained matching rates for different ranking positions when considering each camera pair in the WARD dataset. Differently, Table 6.18 uses the mean matching rate to summarizes the results obtained for each camera pair in RAID dataset. Based on these results, it is possible to observe that the Kernel MBPLS is robust to the addition of cameras as there is just a small reduction in the matching rates when compared to the pairwise setting.

Table 6.17: Matching rates for multiple and pairwise cameras using the Kernel MBPLS models in WARD dataset. These results are detailed for each pair of cameras.

Training Strategy	Probe/Gallery	r = 1	r = 2	r = 3
Multiple Cameras	A/B	**83.7 (5.0)**	92.9 (5.9)	96.0 (5.3)
	B/C	74.9 (5.0)	85.7 (5.5)	92.0 (4.4)
	A/C	70.9 (5.2)	81.7 (5.9)	88.3 (4.4)
Pairwise Cameras	A/B	83.4 (6.3)	**93.1 (5.9)**	**96.3 (4.5)**
	B/C	78.0 (5.9)	87.1 (5.4)	93.7 (4.0)
	A/C	74.9 (6.8)	86.3 (5.8)	91.7 (3.9)

6.2.3 Attributes

Table 6.19 presents the obtained experimental results in the WARD dataset with and without attributes information. For these results, it is possible to conclude that the attributes produces worst results than the identity information when both are used

Table 6.18: Matching rates for multiple and pairwise cameras using the Kernel MBPLS models in RAID dataset. These results are the mean matching rates when considering all the camera pairs.

Training Strategy	r = 1	r = 2	r = 3
Pairwise Cameras	**83.1 (7.0)**	**93.5 (4.4)**	**97.0 (2.7)**
Multiple Cameras	78.4 (7.8)	89.7 (6.3)	94.4 (3.8)

isolated. It can be explained by the low-resolution images, different illumination conditions and the subjectivity of some attributes (e.g., dark or bright hair). More importantly, identity and attribute are complementary information. For instance, when using identity + attributes we improve 4.8 percentage points in the *rank-1* when compared to the identity only result.

Table 6.19: Evaluation of the proposed Kernel MBPLS in the WARD dataset using the Identity, Attributes and Identity + Attributes settings.

Method	WARD (p=35)				
	r = 1	r = 5	r = 10	r = 15	r = 20
Ident.	76.5 (5.1)	86.8 (5.8)	92.1 (4.7)	94.5 (3.9)	95.8 (3.6)
Attr.	67.4 (8.3)	81.2 (7.8)	86.5 (6.5)	90.8 (5.5)	93.0 (4.6)
Ident. + Attr.	**81.3 (6.4)**	**90.3 (5.4)**	**94.2 (4.8)**	**96.6 (2.9)**	**97.6 (2.6)**

6.3 Cross-view Kernel PLS (X-KPLS)

In this section, we consider the proposed Cross-view Kernel PLS (X-KPLS) in a multiple camera setting. Differently from the previous approaches (i.e., Kernel MBPLS and Kernel HPCA) that learned projections to a low-dimensional space jointly, X-KPLS learns them independently using non-linear regression models. Nonetheless, as a unique projection is learned for each camera, this model also scales with the number of surveillance cameras.

In the following, we first evaluate the different parameters of the proposed method (Section 6.3.1) and, then, we present experiments regarding the inclusion of attributes information (Section 6.3.2). Notice that we did not compare the pairwise and multiple camera models as the X-KPLS does not learn projections jointly and, therefore, there is no difference between these settings.

6.3.1 Parameters Setting

In this section, we consider the different parameters that impact in the matching rate of the proposed X-KPLS model. In particular, these parameters are the feature descriptors, the distinct values of sigma and the number of latent factors of the nonlinear regression model. In these experiments, we define the responses matrix $\mathbf{Y}$ using only the identity information.

Feature Descriptors. We evaluate the LOMO and GoG descriptors using the X-KPLS model in WARD and RAID datasets. Table 6.20 presents the matching rates for different ranking positions in WARD, while Table 6.21 shows the results in RAID. Based on these results, it is possible to determine that the GoG outperforms the LOMO for all ranking positions and in both datasets. Therefore, in the following experiments, we will use the GoG representation.

Table 6.20: GoG and LOMO descriptors evaluated using the proposed X-KPLS in the WARD dataset.

Descriptors	WARD (p=35)				
	r = 1	r = 2	r = 3	r = 4	r = 5
GoG	**76.4 (7.0)**	**86.1 (6.3)**	**91.0 (4.6)**	**93.7 (3.8)**	**95.6 (3.1)**
LOMO	70.5 (6.1)	83.4 (5.5)	88.0 (5.0)	92.7 (3.8)	94.6 (3.1)

Table 6.21: GoG and LOMO descriptors evaluated using the proposed X-KPLS in the RAID dataset.

Descriptors	RAID (p=20)				
	r = 1	r = 2	r = 3	r = 4	r = 5
GoG	**79.5 (8.4)**	**89.3 (5.9)**	**94.9 (4.0)**	**97.3 (3.4)**	**98.8 (2.3)**
LOMO	72.4 (7.9)	86.3 (6.1)	92.4 (4.5)	96.1 (3.1)	97.3 (2.9)

Parameter σ. We evaluate different values of sigma when computing the RBF kernel in WARD and RAID dataset. The obtained *rank-1* matching rates are presented in Table 6.22 and Table 6.23, respectively. Similarly to the previous methods, we observed a decline in the matching rates for values smaller than 1.0, while higher values are achieved when using sigma values equal or higher than 1.0. Based on these experiments, we define sigma equals 2.0 and 1.0 for WARD and RAID, respectively.

Latent Factors. We evaluate the number of factors used in the X-KPLS method. It is important to emphasize that one model is learned for each camera and that we set the same number of components for all models. Table 6.24 presents the number of factors

Table 6.22: WARD dataset Kernel XPLS with different values of parameter sigma.

Sigma	WARD (p=35)			
	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$
Rank-1	56.1 (6.7)	72.0 (7.5)	**76.4 (7.0)**	76.1 (7.0)

Table 6.23: RAID dataset Kernel XKPLS with different values of parameter sigma.

Sigma	RAID (p=20)			
	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$
Rank-1	70.9 (9.0)	**79.5 (8.4)**	79.3 (7.4)	79.3 (7.4)

for WARD dataset, while, in Table 6.25, we show these results in the RAID dataset. Observe that using few or too many components results in suboptimal results. It can be related to the overfitting or underfitting problems, respectively. More importantly, we find that best *rank-1* matching rates are achieved when using 30 components in both datasets.

Table 6.24: Evaluation of different number of factors using the X-KPLS in the WARD dataset.

Factors	WARD (p=35)			
	$f = 10$	$f = 20$	$f = 30$	$f = 50$
Rank-1	51.8 (6.3)	71.3 (7.8)	**76.4 (7.0)**	76.0 (6.1)

Table 6.25: Evaluation of different number of factors using the X-KPLS in the RAID dataset.

Factors	RAID (p=20)			
	$f = 10$	$f = 20$	$f = 30$	$f = 50$
Rank-1	67.4 (9.4)	77.3 (8.4)	**79.5 (8.4)**	79.5 (8.0)

6.3.2 Attributes

In this section, we evaluate the proposed X-KPLS considering three different settings: identity only, attributes only and attributes + identity information. To accomplish that, we used the manually labeled attributes information in WARD dataset, as already mentioned. As shown in Table 6.26, we observed improved results due to the better low-dimensional representation of the data when using attributes complementary to the identity information.

Table 6.26: Evaluation of the proposed Kernel X-KPLS in the WARD dataset using the Identity, Attributes and Identity + Attributes settings.

Method	WARD (p=35)				
	r = 1	r = 5	r = 10	r = 15	r = 20
Ident.	76.4 (7.0)	86.1 (6.3)	91.0 (4.6)	93.7 (3.8)	95.6 (3.1)
Attr.	66.7 (9.0)	80.9 (6.4)	87.5 (5.8)	91.0 (5.8)	92.5 (5.9)
Ident. + Attr.	**80.9 (6.6)**	**91.8 (5.4)**	**95.0 (5.0)**	**96.6 (3.8)**	**97.9 (2.6)**

6.4 State-of-the-art Comparisons

In this section, we compare the proposed methods with state-of-the-art approaches in the WARD and RAID datasets. To the best of our knowledge, there is no other methods in literature that address person re-identification as a multiple camera problem. Therefore, we compare the proposed multi-cameras methods with pairwise models. Table 6.27 shows these methods categorized in metric learning and subspace learning approaches. In addition, Table 6.27 presents that the multi-camera models grow linearly with the number of cameras, while the pairwise model are not scalable as we need to learn a different model for each one of the c(c-1)/2 pair of cameras.

When considering the proposed methods, we used the best configuration of parameters discussed in the previous sections. In addition, in the regression-based methods (Kernel MBPLS and X-KPLS), we used the identity and attributes labels as response matrix $\mathbf{Y}$.

Table 6.27: State-of-the-art approaches categorized in pairwise and multiple cameras models. In addition, we represent the methods complexity as function of the number of cameras c.

Strategy	Models	Methods	Complexity
Pairwise	Metric Learning	XQDA, KISSME, MLAPG	$O(c^2)$
	Subspace Learning	CCA, Kernel CCA	$O(c^2)$
Multi-camera	Subspace Learning	**Kernel HPCA**, **Kernel MBPLS**, **Cross-view Kernel PLS**	$O(c)$

WARD

Table 6.28 presents the *rank-1* matching rates of the methods analyzed in the different camera pairs of WARD dataset. These methods are grouped based on the scalability constraints in pairwise and multiple cameras models.

Regarding the multiple cameras models, we observe a superior performance of the Kernel MBPLS and the X-KPLS when compared to the Kernel HPCA. It can be explained by the better separability of the data when using identity and attributes labels in the response matrix **Y**. Between the Kernel MBPLS and X-KPLS, the latter presents slightly better performance, which is not statistically significant when we consider the standard deviations. In fact, both methods have a lot in common with the main difference being the factor that the X-KPLS computes each camera projection independently while in the Kernel MBPLS they are computed jointly in a single framework.

The pairwise models demonstrate that metric learning approaches present superior performance when compared to the subspace learning methods (CCA and KCCA). It is an indicative that working with similarities and dissimilarities is better than with correlation. As matter of fact, XQDA and KISSME outperforms the proposed methods in most of the cameras but have the disadvantage of being pairwise models.

Table 6.28: Mean *Rank-1* matching rate for different approaches on WARD dataset.

Models	Methods	Probe/Gallery Cameras		
		A/B	A/C	B/C
Pairwise	XQDA	88.3 (6.1)	82.6 (4.6)	**89.4 (4.7)**
	MLAPG	72.0 (7.1)	67.6 (6.3)	76.9 (7.3)
	KISSME	88.6 (5.4)	**84.3 (6.1)**	89.1 (3.5)
	CCA	80.3 (8.0)	62.9 (5.0)	70.0 (8.0)
	KCCA	82.6 (7.2)	65.4 (3.4)	70.9 (6.7)
Multiple	**Ker. HPCA**	81.1 (6.2)	64.0 (4.3)	71.7 (8.0)
	Ker. MBPLS	86.6 (5.2)	77.1 (4.5)	83.7 (5.6)
	X-KPLS	**88.9 (3.7)**	78.3 (4.1)	84.9 (6.0)

RAID

Table 6.29 presents the *rank-1* matching rates for each pair of cameras in the RAID datasets. These methods are divided in two subsets, the first subset consists of methods that are pairwise, while the second subset is composed by the proposed multiple camera models.

Considering the multiple camera models, we can observe that the Kernel MBPLS and X-KPLS have a mean value slightly superior to the Kernel HPCA, corroborating with the results observed in the RAID dataset. Nonetheless, it is not a strong evidence due to the large variation of the obtained results in all methods compared.

Similarly to the RAID dataset, in the pairwise models, we can see improved results in the metric learning methods (XQDA, KISSME and MLAPG) when compared to the subspace learning approaches (CCA and KCCA). In particular, the XQDA outperforms the remaining approaches by a large margin. This superior performance when compared to the multiple camera models can be related to the camera pair nuances that are learned when dealing with pairwise models. Nonetheless, while we have to train six XQDA models (i.e., one for each pair of cameras), we just need to learn a single multiple camera model.

Table 6.29: State-of-the-art approaches compared in the RAID dataset. The methods are categorized in pairwise (first lines) and multiple camera (last lines) models to better distinguish the scalability constraints.

Method	Probe/Gallery Cameras					
	A/B	A/C	A/D	B/C	B/D	C/D
XQDA	**99.5 (1.7)**	**80.5 (8.2)**	**93.2 (3.6)**	**81.6 (8.3)**	**97.9 (2.7)**	**90.0 (4.6)**
KISSME	**99.5 (1.7)**	70.5 (7.5)	85.8 (7.9)	77.4 (5.6)	95.8 (4.8)	82.6 (2.5)
MLAPG	98.9 (2.2)	74.2 (7.2)	82.6 (11.4)	77.4 (9.9)	95.3 (3.9)	81.6 (4.5)
CCA	87.4 (7.9)	66.8 (6.6)	76.8 (5.7)	71.0 (7.9)	83.7 (9.1)	83.2 (9.2)
KCCA	87.4 (6.7)	67.4 (6.9)	80.5 (9.3)	72.1 (7.5)	86.3 (5.7)	80.0 (8.5)
Ker. HPCA	87.0 (8.6)	65.5 (9.3)	79.5 (2.8)	70.5 (7.2)	92.0 (6.3)	81.0 (5.2)
Ker. MBPLS	95.0 (5.3)	66.5 (7.8)	81.0 (8.1)	72.5 (7.9)	91.5 (7.1)	81.0 (5.2)
X-KPLS	94.0 (6.1)	66.5 (7.8)	81.0 (7.7)	70.0 (8.2)	91.5 (7.8)	81.0 (7.7)

Chapter 7

Conclusions

In this chapter, we present a summary of the the main contributions of this work. Specifically, we discuss the cross-view matching models proposed regarding its main advantages and drawbacks (Section 7.1). Finally, we discuss some future works that have potential to tackle open issues in the person re-identification problem (Section 7.2).

7.1 Summary

In this dissertation, we address the person re-identification problem tackling some important issues that are still neglected by the person re-identification community as the different camera conditions, the ambiguity between individual's clothes and the scalability with respect to the number of surveillance cameras. To accomplish that, we proposed six approaches that handle the problem as common subspace learning or an indirect matching problem. In the following paragraphs, we highlight the main contributions of this work.

We tackle the camera transition problem using a common subspace learning strategy. In this way, we learn a common subspace that highlights features that are stable to the different camera transitions. Specifically, we proposed three common subspace learning models: the Kernel PLS for Subspace Learning, the Kernel Hierarchical PCA and the Kernel Multiblock PLS. These models have the advantage of being nonlinear and robust to datasets with a reduced number of training samples. In addition, the hierarchical formulations of the Kernel HPCA and the Kernel MBPLS allows a linear growth of the number of models with respect to the number of cameras.

Experimental results obtained using the proposed subspace learning models validate the proposed strategies: nonlinear modelling and hierarchical formulation. For instance, when using the nonlinear mapping, we noticed an improvement in the *rank-1* positions of all methods. In addition, the experiments using multiple cameras demonstrate that the models performance is only slightly diminished when more cameras are added during the training. Besides, superior results where observed when compared to classical subspace learning methods from literature. Nonetheless, these results are still lower than methods based on metric learning, which are constrained to a pair of cameras (e.g., XQDA).

A different solution to the camera transition consists of not performing the direct comparison between feature descriptors from distinct cameras. Instead, the comparison between probe and gallery samples occurs indirectly by using the similarity with training samples. Three proposed approaches explore this line of research: the Prototype-Based Person Re-identification, the Cross-view Kernel PLS and the Kernel Cross-view Collaborative Representation based Classification. Experimental results corroborate our claims with improved results when compared to baseline approaches, mainly for the Kernel X-CRC method. In fact, the experiments demonstrate that key strategies as the nonlinear mapping, the multitask formulation and the initial projection in a common subspace boost the Kernel X-CRC performance.

Finally, we also tackle the ambiguity between individuals wearing similar clothes by introducing attributes information that guide a better separation between individuals. Specifically, we include attributes labels in the nonlinear regression models Kernel MBPLS and X-KPLS and demonstrate improved results in a pairwise (VIPeR) and multicameras (WARD) database.

7.2 Future Works

In this work, we tackle the person re-identification problem in datasets with a reduced number of training samples. We focused on this scenario as the human labor necessary for the annotation of individuals identities in large cameras network is extremely high. In this scenario, deep learning approaches obtain suboptimal results as a consequence of the lack of generalization of such models. Therefore, we tackle the camera transition, ambiguity between individuals and scalability problems using subspace learning and indirect matching approaches that are learned using handcrafted descriptors.

Similar problems are present in different research areas where the cost of obtaining or annotating samples is high. For instance, when using multimodal biometrics, we usually have available only small collections due to the high cost of collecting the data and the privacy concerns. Therefore, as future works, we intend to evaluate the proposed methods in such settings.

Bibliography

An, L., Chen, X., Kafai, M., Yang, S., and Bhanu, B. (2013a). Improving person re-identification by soft biometrics based reranking. In *Distributed Smart Cameras (ICDSC), 2013 Seventh International Conference on*, pages 1–6.

An, L., Kafai, M., Yang, S., and Bhanu, B. (2013b). Reference-based person re-identification. In *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, pages 244–249.

Apurva, B.-G. and Shah., S. K. (2014). A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270 – 286. ISSN 0262-8856.

Bak, S., Carr, P., and Lalonde, J.-F. (2018). Domain adaptation through synthesis for unsupervised person re-identification. *arXiv preprint arXiv:1804.10094*.

Barbosa, I. B., Cristani, M., Caputo, B., Rognhaugen, A., and Theoharis, T. (2018). Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *Computer Vision and Image Understanding*, 167:50--62.

Bashir, K., Xiang, T., and Gong, S. (2010). Cross-view gait recognition using correlation strength. In *Proceedings of the British Machine Vision Conference*, pages 109.1--109.11. BMVA Press.

BenAbdelkader, C., Cutler, R., Nanda, H., and Davis, L. (2001). Eigengait: Motion-based recognition of people using image self-similarity. In *Audio-and Video-Based Biometric Person Authentication*, pages 284--294. Springer.

BenAbdelkader, C., Cutler, R. G., and Davis, L. S. (2004). Gait recognition using image self-similarity. *EURASIP Journal on Applied Signal Processing*, 2004:572--585.

Bennett, K., Bennett, K., Embrechts, M., and Embrechts, M. (2003). An optimization perspective on kernel partial least squares regression. In *Advances in Learning Theory: Methods, Models and Applications.*

Black, M. J. and Jepson, A. D. (1998). Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63--84.

Brás, L. P., Bernardino, S. A., Lopes, J. A., and Menezes, J. C. (2005). Multiblock pls as an approach to compare and combine nir and mir spectra in calibrations of soybean flour. *Chemometrics and Intelligent Laboratory Systems*, 75(1):91--99.

Chen, D., Yuan, Z., Chen, B., and Zheng, N. (2016a). Similarity learning with spatial constraints for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1268--1277.

Chen, S.-Z., Guo, C.-C., and Lai, J.-H. (2016b). Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353--2367.

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

Chen, Y.-C., Zheng, W.-S., and Lai, J. (2015). Mirror representation for modeling view-specific transform in person re-identification. In *International Conference on Artificial Intelligence*, pages 3402--3408. AAAI Press.

Cheng, D., Gong, Y., Zhou, S., Wang, J., and Zheng, N. (2016). 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*, pages 1335--1344.

Cheng, D. S., Cristani, M., Stoppa, M., Bazzani, L., and Murino, V. (2011). Custom pictorial structures for re-identification. In *British Machine Vision Conference (BMVC)*.

Cho, Y.-J. and Yoon, K.-J. (2016). Improving person re-identification via pose-aware multi-shot matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Das, A., Chakraborty, A., and Roy-Chowdhury, A. K. (2014). Consistent re-identification in a camera network. In *European Conference on Computer Vision*, volume 8690 of *Lecture Notes in Computer Science*, pages 330--345. Springer.

Deng, W., Zheng, L., Kang, G., Yang, Y., Ye, Q., and Jiao, J. (2018). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 6.

Ding, S., Lin, L., Wang, G., and Chao, H. (2015). Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993--3003.

Fan, H., Zheng, L., Yan, C., and Yang, Y. (2018). Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):83.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360--2367. IEEE.

Fu, Y., Wei, Y., Wang, G., Li, J., Zhou, X., Shi, H., and Huang, T. (2018). One shot domain adaptation for person re-identification. *arXiv preprint arXiv:1811.10144*.

Gao, C., Wang, J., Liu, L., Yu, J. G., and Sang, N. (2016a). Temporally aligned pooling representation for video-based person re-identification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 4284–4288.

Gao, C., Wang, J., Liu, L., Yu, J.-G., and Sang, N. (2016b). Temporally aligned pooling representation for video-based person re-identification. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 4284--4288. IEEE.

Garcia, J., Martinel, N., Micheloni, C., and Gardel, A. (2015). Person re-identification ranking optimisation by discriminant context information analysis. In *International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 13-16 December, 2015*, pages 1305--1313.

Geng, M., Wang, Y., Xiang, T., and Tian, Y. (2016). Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*.

Gheissari, N., Sebastian, T. B., and Hartley, R. (2006). Person reidentification using spatiotemporal appearance. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1528--1535. IEEE.

Gray, D. and Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262--275. Springer.

Guo, Y., Shan, Y., Sawhney, H., and Kumar, R. (2007). Peet: Prototype embedding and embedding transition for matching vehicles over disparate viewpoints. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1--8. IEEE.

Han, J. and Bhanu, B. (2006). Individual recognition using gait energy image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(2):316--322.

Harandi, M. T., Sanderson, C., Hartley, R., and Lovell, B. C. (2012). Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach. In *Computer Vision–ECCV 2012*, pages 216--229. Springer.

Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Huang, S., Lu, J., Zhou, J., and Jain, A. K. (2015). Nonlinear local metric learning for person re-identification. *arXiv preprint arXiv:1511.05169*.

Jiang, N., Liu, J., Sun, C., Wang, Y., Zhou, Z., and Wu, W. (2018). Orientation-guided similarity learning for person re-identification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2056--2061. IEEE.

Jose, C. and Fleuret, F. (2016). Scalable metric learning via weighted approximate rank component analysis. *arXiv preprint arXiv:1603.00370*.

Kan Liu, Bingpeng Ma, W. Z. and Huang, R. (2015). A spatio-temporal appearance representation for video-based pedestrian re-identification. In *International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 13-16 December, 2015*, pages 3810--3818.

Karanam, S., Li, Y., and Radke, R. (2015). Sparse re-id: Block sparsity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 33--40.

Kawai, R., Makihara, Y., Hua, C., Iwama, H., and Yagi, Y. (2012). Person re-identification using view-dependent score-level fusion of gait and color features. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2694--2697. IEEE.

Khan, F. M. and Brémond, F. (2016). Person re-identification for real-world surveillance systems. *CoRR*, abs/1607.05975.

Khan, F. M. and Bremond, F. (2016). Unsupervised data association for metric learning in the context of multi-shot person re-identification. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*.

Kodirov, E., Xiang, T., and Gong, S. (2015a). Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *BMVC*, volume 3, page 8.

Kodirov, E., Xiang, T., and Gong, S. (2015b). Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 44.1–44.12. BMVA Press.

Kodirov, E., Xiang, T., Zhenyong, F., and Gong, S. (2016). Person re-identification by unsupervised ℓ_1 graph learning. In *Computer Vision–ECCV 2016*, pages 178--195. Springer.

Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P., and Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kusakunniran, W. (2014). Recognizing gaits on spatio-temporal feature domain. *Information Forensics and Security, IEEE Transactions on*, 9(9):1416–1423.

Layne, R., Hospedales, T. M., and Gong, S. (2014). Attributes-based re-identification. In *Person Re-Identification*, pages 93--117. Springer.

Li, D., Chen, X., Zhang, Z., and Huang, K. (2017). Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 384--393.

Li, W., Zhao, R., and Wang, X. (2012). Human reidentification with transferred metric learning. In *Asian Conference on Computer Vision*, pages 31--44. Springer.

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152--159.

Li, Y., Wu, Z., Karanam, S., and Radke, R. J. (2015a). Multi-shot human re-identification using adaptive fisher discriminant analysis. In *Proceedings of the British Machine Vision Conference*, pages 1--12.

Li, Y., Wu, Z., and Radke, R. J. (2015b). Multi-shot re-identification with random-projection-based random forests. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 373--380. IEEE.

Li, Y.-J., Yang, F.-E., Liu, Y.-C., Yeh, Y.-Y., Du, X., and Wang, Y.-C. F. (2018). Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. *arXiv preprint arXiv:1804.09347*.

Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197--2206.

Liao, S. and Li, S. Z. (2015). Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3685--3693.

Lisanti, G., Masi, I., Bagdanov, A., and Del Bimbo, A. (2015a). Person re-identification by iterative re-weighted sparse ranking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(8):1629–1642.

Lisanti, G., Masi, I., Bagdanov, A., and Del Bimbo, A. (2015b). Person re-identification by iterative re-weighted sparse ranking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(8):1629–1642.

Lisanti, G., Masi, I., and Del Bimbo, A. (2014). Matching people across camera views using kernel canonical correlation analysis. In *Proceedings of the International Conference on Distributed Smart Cameras*, ICDSC '14, pages 10:1--10:6, New York, NY, USA. ACM.

Liu, C., Gong, S., Loy, C. C., and Lin, X. (2012). Person re-identification: What features are important? In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 391--401. Springer.

Liu, C., Loy, C., Gong, S., and Wang, G. (2013). Pop: Person re-identification post-rank optimisation. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 441–448.

Liu, H., Jie, Z., Jayashree, K., Qi, M., Jiang, J., Yan, S., and Feng, J. (2018a). Video-based person re-identification with accumulative motion context. *IEEE transactions on circuits and systems for video technology*, 28(10):2788--2802.

Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., and Hu, J. (2018b). Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099--4108.

Liu, X., Wang, H., Wu, Y., Yang, J., and Yang, M.-H. (2015a). An ensemble color model for human re-identification. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 868–875.

Liu, Z., Qin, J., Li, A., Wang, Y., and Van Gool, L. (2018c). Adversarial binary coding for efficient person re-identification. *arXiv preprint arXiv:1803.10914*.

Liu, Z., Zhang, Z., Wu, Q., and Wang, Y. (2015b). Enhancing person re-identification by integrating gait biometric. *Neurocomputing*, 168:1144--1156.

Liu, Z., Zhang, Z., Wu, Q., and Wang, Y. (2015c). Enhancing person re-identification by integrating gait biometric. *Neurocomputing*, 168:1144--1156.

Loy, C. C., Xiang, T., and Gong, S. (2010). Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90(1):106--129.

Ma, A. and Li, P. (2015). Query based adaptive re-ranking for person re-identification. In *Computer Vision – ACCV 2014*, volume 9007 of *Lecture Notes in Computer Science*, pages 397–412. Springer International Publishing.

Ma, B., Su, Y., and Jurie, F. (2012a). Bicov: a novel image representation for person re-identification and face verification. In *Proceedings of the British Machine Vision Conference*, pages 57.1--57.11. BMVA Press.

Ma, B., Su, Y., and Jurie, F. (2012b). Local descriptors encoded by fisher vectors for person re-identification. In *Computer Vision - ECCV 2012. Workshops and Demonstrations*, volume 7583 of *Lecture Notes in Computer Science*, pages 413--422. Springer Berlin Heidelberg.

Ma, B., Su, Y., and Jurie, F. (2012c). Local descriptors encoded by fisher vectors for person re-identification. In *European Conference on Computer Vision (ECCV)*, pages 413--422. Springer.

MacGregor, J. F., Jaeckle, C., Kiparissides, C., and Koutoudi, M. (1994). Process monitoring and diagnosis by multiblock pls methods. *AIChE Journal*, 40(5):826--838.

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

Martinel, N. and Micheloni, C. (2014). Person re-identification by modelling principal component analysis coefficients of image dissimilarities. *Electronics Letters*, 50(14):1000--1001.

Matsukawa, T., Okabe, T., Suzuki, E., and Sato, Y. (2016). Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1363--1372.

McLaughlin, N., del Rincon, J. M., and Miller, P. (2016a). Person re-identification using deep convnets with multi-task learning. *IEEE Transactions on Circuits and Systems for Video Technology*.

McLaughlin, N., Martinez del Rincon, J., and Miller, P. (2016b). Recurrent convolutional network for video-based person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mogelmose, A., Bahnsen, C., Moeslund, T., Clapes, A., and Escalera, S. (2013). Tri-modal person re-identification with rgb, depth and thermal features. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 301 -- 307.

Paisitkriangkrai, S., Shen, C., and van den Hengel, A. (2015). Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846--1855.

Pala, F., Satta, R., Fumera, G., and Roli, F. (2015). Multi-modal person re-identification using rgb-d cameras. *Circuits and Systems for Video Technology, IEEE Transactions on*, PP(99):1–1. ISSN 1051-8215.

Pedagadi, S., Orwell, J., Velastin, S., and Boghossian, B. (2013). Local fisher discriminant analysis for pedestrian re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peng, P., Tian, Y., Xiang, T., Wang, Y., Pontil, M., and Huang, T. (2018). Joint semantic and latent attribute modelling for cross-class transfer learning. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1625--1638.

Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., and Tian, Y. (2016). Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1306--1315.

Prates, R. and Schwartz, W. R. (2016). Kernel hierarchical pca for person re-identification. In *23th International Conference on Pattern Recognition, ICPR 2016, Cancun, MEXICO, December 4-8, 2016.*

Prates, R. and Schwartz, W. R. (2018). Kernel cross-view collaborative representation based classification for person re-identification. *Journal of Visual Communication and Image Representation.*

Prates, R. F. and Schwartz, W. R. (2015). Cbra: Color-based ranking aggregation for person re-identification. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 1975--1979. IEEE.

Rosipal, R. and Krämer, N. (2006). Overview and recent advances in partial least squares. In *Subspace, latent structure and feature selection*, pages 34--51. Springer.

Rosipal, R. and Trejo, L. (2002). Kernel partial least squares regression in reproducing kernel hilbert space. *The Journal of Machine Learning Research*, 2:97--123.

Roth, P. M., Hirzer, M., Köstinger, M., Beleznai, C., and Bischof, H. (2014). Mahalanobis distance learning for person re-identification. In *Person Re-Identification*, pages 247--267. Springer.

Sarfraz, M. S., Schumann, A., Eberle, A., and Stiefelhagen, R. (2018). A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proc. CVPR*, pages 420--429.

Satta, R. (2013). Appearance descriptors for person re-identification: a comprehensive review. *CoRR*, abs/1307.5748.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299--1319.

Schumann, A. and Stiefelhagen, R. (2017). Person re-identification by deep learning attribute-complementary information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20--28.

Schwartz, W. and Davis, L. (2009a). Learning discriminative appearance-based models using partial least squares. pages 322–329. ISSN 1550-1834.

Schwartz, W. and Davis, L. (2009b). Learning discriminative appearance-based models using partial least squares. In *Computer Graphics and Image Processing, 2009 XXII Brazilian Symposium on*, pages 322–329. ISSN 1550-1834.

Schwartz, W. R., Kembhavi, A., Harwood, D., and Davis, L. S. (2009). Human detection using partial least squares analysis. In *Computer vision, 2009 IEEE 12th international conference on*, pages 24--31. IEEE.

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

Sharma, A. and Jacobs, D. W. (2011). Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 593--600. IEEE.

Shi, H., Zhu, X., Liao, S., Lei, Z., Yang, Y., and Li, S. Z. (2015). Constrained deep metric learning for person re-identification. *arXiv preprint arXiv:1511.07545.*

Siarohin, A., Sangineto, E., Lathuilière, S., and Sebe, N. (2018). Deformable gans for pose-based human image generation. In *CVPR 2018-Computer Vision and Pattern Recognition.*

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *science*, 302(5643):249--255.

Turk, M. A. and Pentland, A. P. (1991). Face recognition using eigenfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586--591.

Vapnik, V. N. (1998). *Statistical learning theory.* Adaptive and learning systems for signal processing, communications, and control. Wiley, New York. ISBN 0-471-03003-1.

Varior, R. R., Shuai, B., Lu, J., Xu, D., and Wang, G. (2016). A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135--153. Springer.

Vezzani, R., Baltieri, D., and Cucchiara, R. (2013). People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 46(2):29.

Wang, C., Zhang, Q., Huang, C., Liu, W., and Wang, X. (2018). Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 365--381.

Wang, G., Lin, L., Ding, S., Li, Y., and Wang, Q. (2016). Dari: Distance metric and representation integration for person verification. *arXiv preprint arXiv:1604.04377*.

Wang, T., Gong, S., Zhu, X., and Wang, S. (2014). Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688--703. Springer.

Wang, X. (2013). Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3--19.

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

Westerhuis, J. A., Kourti, T., and MacGregor, J. F. (1998). Analysis of multiblock and hierarchical pca and pls models. *Journal of chemometrics*, 12(5):301--321.

Wold, H. (1985). *Encyclopedia of Statistical Sciences*, volume 6. John Wiley & Sons.

Wolf, L., Hassner, T., and Taigman, Y. (2009). The one-shot similarity kernel. In *International Conference on Computer Vision (ICCV)*.

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T. S., and Yan, S. (2010). Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031--1044.

Xiong, F., Gou, M., Camps, O., and Sznaier, M. (2014). Person re-identification using kernel-based metric learning methods. In *European Conference on Computer Vision (ECCV)*, pages 1--16. Springer.

Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., and Yang, X. (2016). Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, pages 701--716. Springer.

Yang, J., Shi, Z., and Vela, P. (2011). Person reidentification by kernel pca based appearance learning. In *CRV*.

Yang, Y., Liao, S., Lei, Z., and Li, S. Z. (2016). Large scale similarity learning using similar pairs for person verification. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., and Li, S. Z. (2014). Salient color names for person re-identification. In *Computer Vision–ECCV 2014*, pages 536--551.

Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 34--39. IEEE.

You, J., Wu, A., Li, X., and Zheng, W.-S. (2016). Top-push video-based person re-identification. *arXiv preprint arXiv:1604.08683*.

Zajdel, W., Zivkovic, Z., and Krose, B. J. A. (2005). Keeping track of humans: Have i seen this person before? In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 2081–2086. ISSN 1050-4729.

Zeng, M., Wu, Z., Tian, C., Zhang, L., and Hu, L. (2015). Efficient person re-identification by hybrid spatiogram and covariance descriptor. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 48--56.

Zhang, L., Xiang, T., and Gong, S. (2016a). Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1239--1248.

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

Zhang, Y., Li, B., Lu, H., Irie, A., and Ruan, X. (2016b). Sample-specific svm learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, R., Ouyang, W., and Wang, X. (2013). Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586--3593.

Zheng, L., Huang, Y., Lu, H., and Yang, Y. (2017a). Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*.

Zheng, L., Yang, Y., and Hauptmann, A. G. (2016). Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*.

Zheng, Z., Zheng, L., and Yang, Y. (2017b). A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13.

Zheng, Z., Zheng, L., and Yang, Y. (2017c). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 3.

Zheng, Z., Zheng, L., and Yang, Y. (2018). Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhong, Z., Zheng, L., Cao, D., and Li, S. (2017). Re-ranking person re-identification with k-reciprocal encoding. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3652--3661. IEEE.

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

Zhu, X., Jing, X.-Y., Wu, F., and Feng, H. (2016). Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In *International Joint Conference on Artificial Intelligence*.