UNIVERSITY OF CALIFORNIA
RIVERSIDE

Real-World Person Identification

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

by

Le An

December 2014

Dissertation Committee:

Dr. Bir Bhanu, Chairperson
Dr. Matthew J. Barth
Dr. Subir Ghosh

The Dissertation of Le An is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

Upon the completion of this work, I own my gratitude to a great number of people. I would first like to express my deepest gratitude to Dr. Bir Bhanu, for his support and guidance as my advisor during my PhD study. I would like to thank my committee members, Dr. Matthew J. Barth and Dr. Subir Ghosh for their constructive comments to improve this work. I would also like to thank Dr. Gerardo Beni and Dr. Yingbo Hua who were in my oral qualifying exam committee and gave me insightful feedback and suggestions.

I am grateful to Dr. Mehran Kafai, Dr. Songfan Yang, Dr. Ninad Thakoor, with whom I collaborated extensively and shared a lot of inspirational ideas. I am also grateful to Suresh Kumar, as my colleague and friend, who had brought enormous optimism to all of us. I am indebted and grateful to my wife and colleague, Xiaojing Chen, for her companion and support to my research and all other aspects in my life.

I would like to thank Zhixing Jin, Linan Feng, Dr. Yu Sun, Dr. Yiming Li, Dr. Alberto Cruz and all of my other current and former colleagues, friends at University of California, Riverside and at other places who have offered me help, support, and joy. I would also like to thank all of my family members who have been supportive in many aspects during the completion of this work.

The materials of some chapters in this dissertation have appeared in "Face image super-resolution using 2D CCA" © 2014 Elsevier by L. An *et al.* and "Dynamic Bayesian network for unconstrained face recognition in surveillance camera networks" © 2013 the IEEE by L. An *et al.*

I dedicate this Dissertation,

To my father, Jiyuan An, and my mother, Shuhua Wu,

for their unconditional love, support, sacrifice, and encouragement.

To my wife, Xiaojing Chen, for her support, understanding, and endless love.

Without you, I would not have gone so far.

ABSTRACT OF THE DISSERTATION

Real-World Person Identification

by

Le An

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, December 2014
Dr. Bir Bhanu, Chairperson

Person Identification or recognition has been receiving broad interests and it is highly desirable in applications such as security monitoring, authentication, etc. In order to recognize a person, different traits, including fingerprint, face, and gait, can be used. Among these possible traits, face and body are preferred since they can be acquired without the person's cooperation. In controlled environment, recognition is less challenging with well posed subject in high resolution. However, in real-world scenarios, where the image of a person exhibits variations in pose, illumination, and resolution, standard pattern recognition methods may fail. Driven by the necessity for person identification in real-world, we have proposed several identification methods. Specifically, we have developed a face image super-resolution method as a pre-processing step to improve the face recognition accuracy. In addition, to recognize person in a surveillance setting with multiple cameras, we have developed an algorithm that utilizes multiple cameras for face recognition by encoding the person-specific dynamics with a dynamic Bayesian network. In case the face of a person cannot be reliably acquired, identifying person by body appearance is preferred. To this end, we have proposed two methods to identify person in multiple surveillance cameras, using a novel reference descriptor and a sparse

representation, respectively. To validate the proposed method in this dissertation, we have conducted extensive results on publicly available datasets. Results show that each of the aforementioned method achieves state-of-the-art performance in various person identification tasks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Face is frequently used to identify person. Given an image of a face, the image resolution plays an important role when it is to be recognized. In Chapter 3, a face super-resolution method using two-dimensional canonical correlation analysis (2D CCA) is presented. A detail compensation step is followed to add high-frequency components to the reconstructed high-resolution face. Unlike most of the previous research on face super-resolution algorithms that first transform the images into vectors, in our approach the relationship between the high-resolution and the low-resolution face images are maintained in their original 2D representation. In addition, rather than approximating the entire face, different parts of a face image are super-resolved separately to better preserve the local structure. The proposed method is compared with various state-of-the-art super-resolution algorithms using multiple evaluation criteria including face recognition performance. Results on publicly available datasets show that the proposed method super-resolves high quality face images which are very close to the ground-truth and performance gain is not dataset dependent. The method is very efficient both in the training and testing phases compared to the other approaches.

The demand for robust face recognition in real-world surveillance cameras is increasing due to the needs of practical applications such as security and surveillance. Although face recognition has been studied extensively in the literature, achieving good performance in surveillance videos with unconstrained faces is inherently difficult. During the image acquisition process, the non-cooperative subjects appear in arbitrary poses and resolutions in different lighting conditions, together with noise and blurriness of images. In addition, multiple cameras are usually distributed in a camera network and different cameras often capture a subject in different views. In Chapter 4, we aim at tackling this unconstrained face recognition problem and utilizing multiple cameras to improve the recognition accuracy using a probabilistic approach. We propose a Dynamic Bayesian Network (DBN) to incorporate the information from different cameras as well as the temporal clues from frames in a video sequence. The proposed method is tested on a public surveillance video dataset with a three-camera setup. We compare our method to different benchmark classifiers with various feature descriptors. The results demonstrate that by modeling the face in a dynamic manner the recognition performance in a multi-camera network is improved over the other classifiers with various feature descriptors and the recognition result is better than using any of the single camera.

Normally a face of a person cannot be acquired easily in real-world surveillance cameras due to arbitrary human pose, illumination, etc. In this case, we aim at identifying persons across non-overlapping cameras, which is known as person re-identification that matches people at different time and location. Re-identifying people is of great importance in crucial applications such as wide-area surveillance and visual tracking. Due to the appearance variations in pose, illumination, and occlusion in different camera views, person re-identification is inherently difficult. To address these challenges, a

reference-based method is proposed in Chapter 5 for person re-identification across different cameras. Instead of directly matching people by their appearance, the matching is conducted in reference space where the descriptor for a person is translated from the original color or texture descriptors to similarity measures between this person and the exemplars in the reference set. A subspace is learned in which the correlations of the reference data from different cameras are maximized using Regularized Canonical Correlation Analysis (RCCA). For re-identification, the gallery data and the probe data are projected into this RCCA subspace and the reference descriptors (RDs) of the gallery and probe are generated by computing the similarity between them and the reference data. The identity of a probe is determined by comparing the RD of the probe and the RDs of the gallery. A re-ranking step is added to further improve the results using a saliency-based matching scheme. Experiments on publicly available datasets show that the proposed method outperforms the state-of-the-art approaches.

In addition, we developed a L2 regularized sparse-representation based person re-identification framework in Chapter 6. Specifically, to address this multi-view matching problem, we first learn a subspace in which the goal is to maximize the correlation between data from different cameras but corresponding to the same people. Given a probe from one camera view, we represent it using a sparse representation from a jointly learned coupled dictionary in the learned subspace. The L1 induced sparse representation is regularized by an L2 regularization term. The introduction of L2 regularization allows learning a sparse representation while maintaining the stability of the sparse coefficients. To compute the matching scores between probe and gallery, their L2 regularized sparse representations are used with a modified cosine similarity measure. Experimental results with extensive comparisons on publicly available datasets demonstrate that the proposed method outperforms the state-of-the-art methods and using L2 regularized

sparse representation (L1+L2) more accurate matching is achieved compared to using the L1 regularization or L2 regularization only.

Each chapter in this dissertation stands alone as a complete description of each aforementioned method. Before we dive into details of individual methods, related work is presented in Chapter 2.

Chapter 2

Related Work

In this chapter, we present related work in face super-resolution, face recognition, and person re-identification in order to overview relevant literatures before we discuss details of our own methods in these topics.

2.1 Face Super-Resolution

Face is commonly used to recognize humans. In real-world applications such as video surveillance, detected faces are often of low-resolution, which makes the recognition task difficult. Face image super-resolution, also referred as face hallucination, is a natural solution to solve this problem. Although in some work super-resolution and recognition are handled simultaneously without generating high-resolution images [63], it is still desirable to obtain a super-resolved face image from low-resolution feed in case where examination or validation by human is required.

In the past several decades, various super-resolution methods have been proposed. Based on the input, those methods can be categorized into two classes. The methods in the first class take advantage of multiple images of the same scene and

reconstruct HR images by aggregating information from all the LR images via motion estimation or registration [42, 136, 194]. However, these methods strongly rely on accurate motion information. The methods for super-resolution in the second class are based on a single image. This class of methods has received a lot of attention recently [33, 144, 49]. Some learning based methods try to model the relationship between the LR face images and HR face images [45, 175]. An example of a learning-based approach is [147]. Recently, a joint learning approach is proposed by Gao *et al.* [48] in which two projection matrices are trained simultaneously and the original LR and HR feature spaces are mapped onto a unified feature subspace. This produces improved results compared to the neighbor-embedding based methods. A sparse neighbor selection scheme is proposed in [47] for image SR and achieves *state-of-the-art* results.

Beyond the generic super-resolution algorithms, specific approaches for certain kind of images such as face images have been proposed [71, 100, 106, 193]. Due to the highly structured shape of a human face, more accurate face super-resolution can be achieved by learning this structural information from an appropriate training dataset. In [155], a semi-coupled dictionary learning model is proposed with application to super-resolution and face synthesis. In [100], a two-step face hallucination approach is developed by first globally modeling the face with a Gaussian assumption and then locally refining the face using path-based nonparametric Markov random field (MRF). Inspired by this work, a number of two-step face hallucination approaches have been proposed [71, 106, 193].

As indicated in [25], the downsampling process maintains the structure in the image manifolds. Furthermore, previous research has shown that face images reside on a non-linear manifold which is linear and smooth locally and it is commonly assumed that the manifolds of LR face images and HR face images have similar local structure [62].

As a consequence, in recent years, manifold learning has been explored by researchers to hallucinate face images under the assumption that the manifolds of LR and HR images have similar local neighborhood structures [82, 159, 193]. These methods directly work in the subspaces of LR and HR images using standard subspace techniques such as the Principal Component Analysis (PCA). Compared to performing reconstruction in the original input feature space, the reconstruction is more meaningful and reliable in the subspaces. In a different manner, Ma *et al.* [113] bypassed the necessity for subspace learning by reconstructing each small patch of a face image separately. A fast face super-resolution method was proposed recently by substituting the nonlinear mapping with multiple local linear transformations [73]. Instead of super-resolving in the image domain, a feature-domain super-resolution framework is proposed in [124] for face recognition.

Among the manifold learning approaches, canonical correlation analysis has been widely adopted in recent years. Canonical correlation analysis (denoted as 1D CCA in this chapter) was first introduced in [67]. It is a multivariate statistical model to analyze the correlation between two sets of variables. 1D CCA finds linear combinations of the variables in each set that have maximum correlations. The model that 1D CCA delivers is a high dimensional relationship between two sets of variables with a few pairs of canonical variables. There are different generalizations of 1D CCA such as kernel CCA [60], locality preserving CCA (LPCCA) [143], and neural network based CCA [51].

There have been many applications utilizing 1D CCA and its variants. In [178] a 2D-3D face matching method is proposed using 1D CCA to capture the relationship between 2D face images and 3D face data for recognition. Recently kernelized CCA has been applied in facial expression recognition [191]. Xu *et al.* [169] proposed a multimodal

recognition scheme with ear and profile face using kernel CCA. Beyond the face domain, 1D CCA is also broadly applied. In [30] data fusion and group analysis of biomedical data is performed using 1D CCA. In [125] 1D CCA is used to analyze remotely sensed data in a geographic information system. The handwritten character recognition is also formulated in a framework with 1D CCA [142]. Recently, Li *et al.* [89] used 1D CCA to maximize the intra-individual correlations for face recognition at different poses.

In terms of super-resolution, Huang *et al.* [71] proposed a face hallucination method based on 1D CCA to determine a coherent subspace in which the correlation between the LR and HR images is maximized. The face images are first vectorized and projected to PCA subspace, then 1D CCA is applied to enhance the correlation of the HR and LR image projections. This approach tries to find the consistency between the LR and HR face images and is able to generate more realistic face images compared to the previous work in [100, 193]. Recently a vehicle logo super-resolution method using 1D CCA is reported in [7] to improve the vehicle recognition accuracy.

2.2 Face Recognition

With more face images becoming available from various sources, many face related processing tools are demanded, such as face clustering [183, 181], face retrieval [79] and face recognition [5]. For real-world face recognition, the following factors are usually considered: pose variation, illumination change, facial expression change, and misalignment.

To tackle the pose variations, Arashloo *et al.* [15] proposed a Markov random field based image matching for pose-invariant face recognition. Prabhu *et al.* [132] constructed a 3D model for each subject in the database using a single 2D image,

then the synthesized face images at different poses were generated from the 3D model for matching. Li *et al.* [91] evaluated the probability that two faces have the same underlying identity cause for matching faces instead using a distance-based method, achieving better results than current methods for face recognition with varying poses. Li *et al.* [90] proposed a probabilistic elastic matching method to handle pose variation. In this model, a Gaussian mixture model (GMM) was used to capture the spacial-appearance distribution of the faces in the training set and an SVM classifier is used for face verification.

To mitigate the illumination change, Hussain *et al.* [74] proposed a new face representation called Local Quantized Patterns (LQP), which is robust to illumination variations and using this feature representation improved performance on the challenging Labeled Faces in the Wild (LFW) [70] dataset was achieved. Li *et al.* [93] used near infrared images for face recognition such that visible illumination changes were bypassed. Tan *et al.* [145] proposed an efficient illumination normalization technique using Gamma correction, difference of Gaussian filtering, masking and contrast equalization. This normalization step boosted the recognition performance on several benchmark datasets.

For face recognition with expressions, a popular solution is to reproduce the neutral faces from the faces with expressions for matching. This is opposite to the goal of facial expression analysis in which facial expressions are supposed to be retained [14, 177, 176]. Nagesh *et al.* [121] proposed that the images of the same subject with different expressions can be viewed as an ensemble of inter-correlated signals and the sparsity accounts for the variation in expressions. In light of this observation, the holistic face image and the facial expression image were generated for face recognition. Hsieh *et al.* [68] removed the expression from a given face by using optical flow computed from the input face with respect to a neutral face.

When a face is mis-aligned and sent to a recognition system, the recognition performance may degrade significantly [138]. In real-world scenarios, aligning faces may be difficult, if not impossible. To handel this problem, Liao *et al.* [99] proposed to use an arbitrary patch of a face image for recognition with an alignment-free face representation. Wang *et al.* [171] proposed a misalignment-robust face recognition method by inferring the spatial misalignment parameters in a trained subspace. Cui *et al.* [31] tackled misalignment by extracting sparse codes of position-free patches within each spatial block in the image. A pairwise-constrained multiple metric learning was proposed to integrate the face descriptors from all blocks.

Recently there have been approaches that try to take care of these aforementioned factors simultaneously. Wolf *et al.* [163] combined multiple face representations and background statistics to improve face recognition in unconstrained environment. Berg *et al.* [22] utilized an external set of faces for identity-preserving alignment and identity classifier learning. A collection of the classifiers is able to discriminate the subjects whose faces are captured in the wild. Müller *et al.* [120] separated the learning of the invariance from learning new instances of individuals. A set of examples called model was used to learn the invariance and new instances were compared by rank list similarity. When a face needs to be compared with a large database, linear search is no longer an affordable approach. Kafai *et al.* [79] proposed a hashing method using discrete cosine transform (DCT) for face retrieval and close to linear search performance were reported. Chen *et al.* [26] utilized multi-level features for large-scale face image retrieval and it was shown that multi-level features outperformed single-level features.

To recognize faces from videos, in general there are two principles: using 2D images from video sequences directly, or generating a 3D face model to cope with pose variation.

Within the 2D-based methods, normally the faces are first extracted from the video frames manually or using an automated face detector [152]. Subsequently, either all the face images or only the exemplar face images are used for the recognition task. An appearance manifold was built in [86] to represent each pose by an affine plane to cope with the pose variations in video sequences. In [107] a Hidden Markov Model (HMM) was used for video-based face recognition. In this model the temporal characteristics were analyzed over time. Stallkamp *et al.* [141] presented a real-time video based face identification system using a local appearance-based model and multiple frame weighting schemes. In [105] the face recognition in video was tackled by exploiting the spatial and temporal information based on Bayesian keyframe learning and nonparametric discriminant embedding. Recently, Biswas *et al.* [23] proposed a learning-based likelihood measurement to match high-resolution frontal view gallery images with probe images from surveillance videos. Wong *et al.* [164] proposed a patch-based image quality assessment method to select a subset of the "best" face images from the video sequences to improve the recognition performance. In [32], the video-based face recognition was converted to the problem of matching two image sets from different video sequences and it needs an independent reference set to align the images sets to be matched. An *et al.* [8, 5, 10] proposed to align faces with pose variations from different frames to a frontal face template. These aligned faces were then averaged to generate a single face representation for the video data with rectified pose.

In an effort to recognize faces using more than one camera, some prior work has been done. Xie *et al.* [167] trained a reliability measure and it was used to select the most reliable camera for recognition. In [61] a cylinder head model was built to track and fuse face recognition results from different cameras. These approaches were tested on videos taken in controlled environment with higher resolution than typical surveillance video

data. For application in surveillance cameras, a person re-identification method was proposed in [19] which depends on the robustness of the face tracker. A face recognition framework for mass transport security surveillance was proposed in [108].

In 3D-based approaches, the 3D face models are either computed or captured directly with a 3D scanner. Xu *et al.* [170] developed a framework using 3D face models for pose and illumination invariant face recognition from video sequences by integrating the effects of motion and lighting changes. In [129], the system used the images in the video as probe to compare with the 2D projection of the gallery 3D model. Liao *et al.* [98] used a single image for each individual in the gallery set to construct a 3D model to synthesize various face views. The 3D based methods are in general computationally expensive. Furthermore, a 3D model is difficult to be constructed from low-resolution videos, thus, the application of 3D models in surveillance cameras is limited. A recent survey of video based face recognition can be found in [137].

2.3 Person Re-Identification

Two major directions to tackle person re-identification are to extract invariant feature representations and to learn specialized distance metrics across different camera views. In feature-driven approaches, robust features which are invariant to change in pose and illumination conditions are studied. Cheng *et al.* [28] adopted pictorial structures to localize the human parts and search part-to-part correspondences to match subjects. Farenzena *et al.* [44] extracted features accounting for the overall chromatic content, the spatial arrangement and the presence of recurrent local motifs to match individuals with appearance variation. Bak*et al.* [16] learned a model in a covariance metric space to select features based on the idea that different regions for each subject

should be matched specifically. Gray *et al.* [53] used AdaBoost to select the most discriminative features instead of handcrafted features. Prosser *et al.* [133] formulated the re-identification as a relative ranking problem instead of an absolute scoring problem. Hirzer *et al.* [65] proposed a two-step method by first using a descriptive model to obtain an initial ranking, which was refined in the second step by a discriminative model with human feedback. Kviatkovsky *et al.* [84] discovered the color intra-distribution structure and showed that this structure was invariant under certain illumination changes and could be combined with the covariance descriptor for person re-identification. Yang *et al.* [179] proposed a salient color names based color descriptor which can guarantee that a higher probability will be assigned to the color name which is nearer to the color. This color descriptor can be computed efficiently in advance. Ma *et al.* [112] used both biologically inspired features and covariance descriptors to handle background and illumination variations. Martinel *et al.* [115] presented an appearance-based approach by computing a novel discriminative signature from multiple local features.

Beyond low-level features, semantic features have been explored for improved re-identification results. Mining semantic attributes or concepts from image or video has been studied extensively in multimedia domain [117, 118]. For person re-identification, Kuo *et al.* [83] applied semantic color names to describe an image of a person instead of using color histograms for better stability. Layne *et al.* [85] proposed mid-level semantic attributes to describe person for the purpose of re-identification. An *et al.* [11] used biometric attributes such as gender from images to re-rank the initial re-identification results from low-level features. Zhao *et al.* [187, 186] proposed to use salient features for person re-identification. The saliency was estimated using unsupervised learning and was combined with existing methods (*e.g.*, [44]) to improve the recognition performance. Liu *et al.* [104] proposed a post-rank optimization method which allowed a human-in-

the-loop to select negative samples. This improved the performance gain over 30% and as compared to the exhaustive search, the time efficiency significantly improved. Liu *et al.* [102] provided extensive study of feature importance for person re-identification and proposed a method for on-the-fly mining of feature. For person re-identification on mobile devices, Vernier [150] *et al.* introduced a client-server system which improved the re-identification performance over time with reduced computation time. Zhao *et al.* [188] learned discriminative mid-level filters from automatically discovered patch clusters and those filters were able to identify specific visual patterns in order to distinguish different persons. Zhang *et al.* [182] proposed a framework which can leverage heterogeneous contextual information such as gait together with facial features to identify person from low-quality video surveillance data.

On the other hand, researchers are also investigating in robust matching methods. Hirzer *et al.* [66] proposed a relaxed pairwise learned metric (RPLM) based on Mahalanobis distance learning which took advantages of the structure of the data with reduced computational cost. It achieved state-of-the-art results with simple feature descriptors. Köstinger *et al.* [81] proposed a simple yet effective method to learn the distance metric called KISS metric from a statistical inference perspective. Tao *et al.* [146] extended the KISS metric by introducing regularization to robustly estimate covariance matrices against the instability in calculating the inverse of a covariance matrix from a small size training set. Zheng *et al.* [190] formulated re-identification as a relative distance comparison problem. It maximized the likelihood such that the distance between a pair of images of the same person is smaller than a pair of images of different people. Liu *et al.* [103] incorporated attribute information into the framework of [190] to further improve the re-identification results by feature weighting. Li *et al.* [94] jointly partitioned the image spaces of two camera views into different configurations based

on the similarity of cross-view transforms. Image pairs with similar transforms were projected to a common feature space for matching. Li *et al.* [96] proposed a filter pairing neural network in which misalignment, pose difference, occlusions and background clutter were jointed handled with the help of abundant data. Xiong *et al.* [168] applied multiple kernel-based metrics in conjunction with histogram-based features and showed improvement over state-of-the-art on several datasets.

Standard metric learning techniques such as Large Margin Nearest Neighbor (LMNN) [162], Information Theoretic Metric Learning (ITML) [35], and Logistic Discriminant Metric Learning (LDML) [55] were also applied to person re-identification. Dikmen *et al.* [37] developed a variant of LMNN by introducing a reject option to the unfamiliar matches (LMNN-R) and achieved improved results. Martinel *et al.* [116] extracted multiple features from image pairs and obtained a so-called distance feature vector. The re-identification was achieved by classifying this distance feature vector using a trained binary classifier. Pedagadi *et al.* [130] used local Fisher discriminant analysis (LFDA) to reduce feature dimensionality for person re-identification. It outperformed other metric learning-based methods. Mignon *et al.* [119] proposed pairwise constrained component analysis (PCCA) to learn a low-dimensional mapping in which distances between data points complied with a set of sparse training pairwise constraints. Loy *et al.* [109] reported a manifold ranking approach in which the probe information was propagated along the data manifold in an unsupervised manner. It showed that the performance of existing metric learning based methods could be significantly improved by integrating the manifold ranking. Comprehensive survey on person re-identification can be found in [151, 39, 20, 50].

Chapter 3

Face Image Super-Resolution using 2D CCA

3.1 Introduction

In real-world scenarios, faces of high resolution is what is desired but difficult to acquire. For instance, the faces captured by surveillance cameras or image sensors on mobile devices are often with low resolution. Using these face images in face recognition algorithms may yield poor recognition accuracy. Thus, it is beneficial to first enhance face image quality through image super-resolution techniques.

The low-resolution (LR) images can be considered as being generated by the imaging process where the original high-resolution (HR) images undergo blurring and downsampling [128]. Usually noise is introduced to further degrade the image quality. The purpose of super-resolution is to reverse this imaging process in order to recover the high-resolution images from the low-resolution observations.

However, the 1D CCA was not designed specifically for the image data. To fit the image data into 1D CCA formulation, the image has to be first converted into a

1D vector. On the other hand, an image is inherently represented in a 2D matrix. The appearance of an image becomes obsolete when reshaped into a vector. To tackle this problem, 2-dimensional CCA has been proposed and it is specifically suitable for image analysis [87]. 2D CCA is formulated in such a manner that it takes two sets of images and explores their relations directly without the necessity to first vectorize each image.

For face image super-resolution (SR), as a common routine for the data representation in manifold based SR methods, the face images are first reshaped into vectors and then super-resolution is performed. The reshaped vectors have large dimensions. For computational feasibility, PCA is applied [71, 193]. However, in this reshaping process the intrinsic 2D spatial structure information of face is erased. In this chapter, inspired by the 2D CCA techniques [87], a two-step 2D CCA based face super-resolution approach is developed that can preserve the intrinsic 2D spatial structure of face images in the super-resolution process. In the *first step* the HR face is reconstructed. Since the reconstructed face is not rich in facial details, we apply a high-frequency detail mask to the reconstructed faces in the *second step*. Figure 3.1 shows the system diagram of the proposed approach.

More specifically, during the training in the *first step*, the learned projection matrices by 2D CCA are able to project the HR face images and LR face images into a subspace where their correlation is maximized. When a testing LR face image is provided, the optimal combination of its K nearest neighbors in the LR training set is found in this subspace. Due to the structural similarity between the HR and LR face images, we are then able to reconstruct a HR face image using the same K nearest neighbors in the HR training set. In the *second step*, a high-frequency detail mask is generated using the neighborhood derived in the first step and added to the reconstructed face image

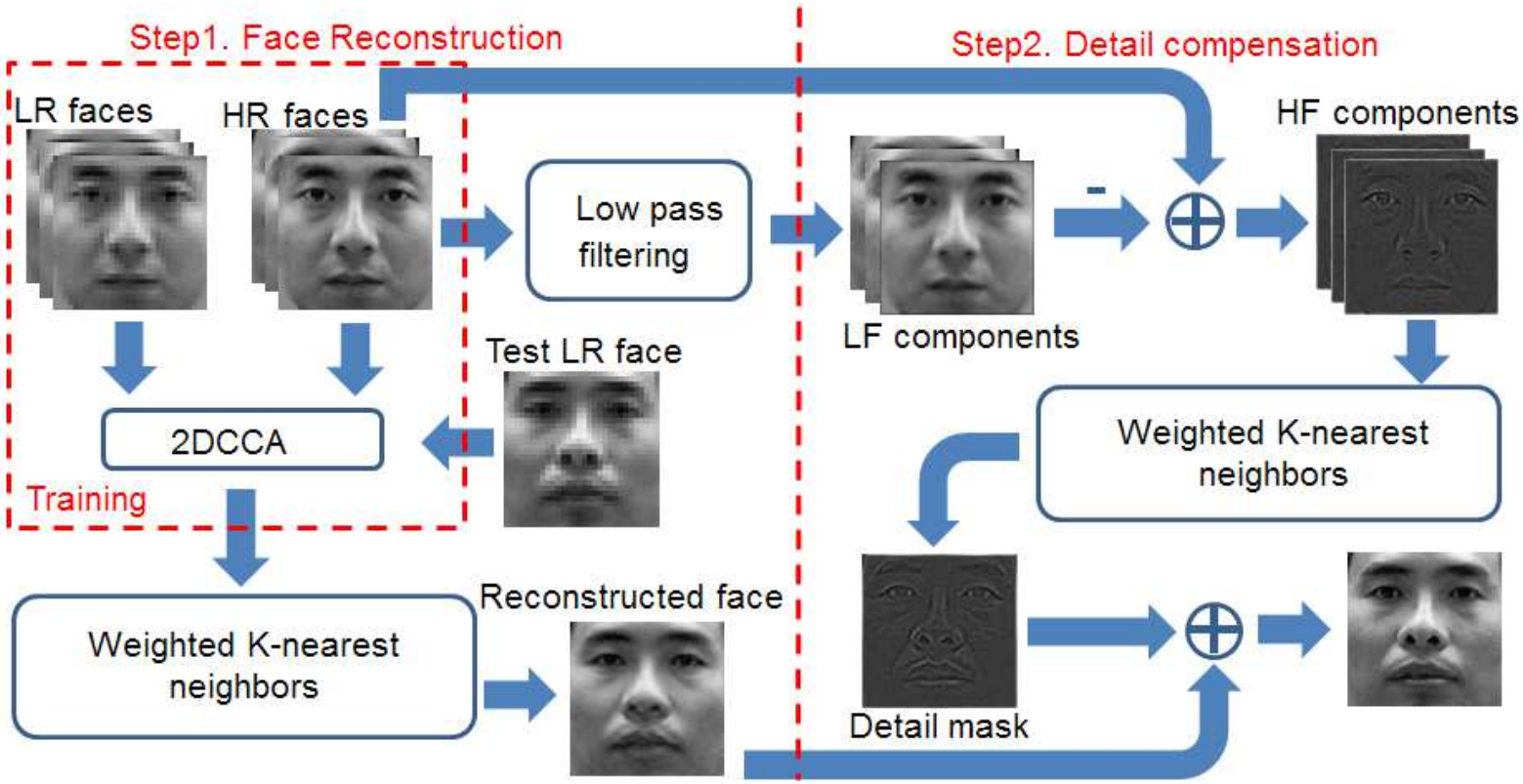

Figure 3.1: The system diagram of the proposed face super-resolution algorithm. The training set contains HR and LR face images and they are projected to a coherent subspace by 2D CCA in which the correlation between the HR face images and the LR face images is maximized. Given an input LR image, it is first projected into this subspace using the projection matrices obtained in the training step. K weighted nearest neighbors in the LR training set are found that reconstruct the input LR face with minimum error. The same neighborhood is applied to the HR training set to generate the super-resolved image. A detail compensation step is followed by reconstructing a high-frequency mask from the high-frequency components. The final output is the sum of the reconstructed face image and the detail mask.

to yield the final output HR image. Inspired by the locality based method [153], we process different parts of a face independently instead of a holistic approach to further improve the output quality.

3.1.1 Contributions of This Chapter

Compared to the previous work for face image SR including the 1D CCA based method in [71], the contributions of this chapter are as follows:

1. This is the first chapter that explores 2D CCA for face image super-resolution. To the authors' best knowledge this has not been done before using 2D CCA. The proposed method demonstrates superior performance compared to the state-of-the-art super-resolution methods [175, 71, 113, 49] (see Figure 3.7).

2. The proposed 2D CCA super-resolution algorithm works directly on the original 2D face image representation. The method is computationally efficient and it achieves the best performance compared to the other methods [175, 71, 113, 49] (see Table 3.3).

3. Thorough experiments are conducted to validate the approach, both quantitatively and qualitatively, using comprehensive metrics including reference based metrics (PSNR, SSIM, SVD [140]) and non-reference based metric (DM [34]). Cross-dataset validation is also performed (see Figure 3.8). Results from the experiments show that the approach is not datasets or image dependent, which is crucial from practical considerations (see Figure 3.7). In addition, a recognition task using the super-resolved images by the proposed method lead to the highest recognition accuracy compared to the other methods (see Table 3.2).

In the rest of this chapter, Section 3.2 provides mathematics for 1D CCA and 2D CCA. Section 3.3 presents the proposed face super-resolution algorithm. The experimental results and comparisons are given in Section 3.4. Finally, Section 3.5 concludes this chapter.

3.2 1D and 2D CCA

3.2.1 1D CCA Formulation

1D CCA was first introduced in [67]. 1D CCA finds basis for two sets of random variables such that the correlation between the projections of these two sets of random variables is maximized. Given two centered (zero mean) datasets, $X = \{x_i \in \mathbb{R}^m, i = 1, 2, ..., N\}$ and $Y = \{y_i \in \mathbb{R}^n, i = 1, 2, ..., N\}$, 1D CCA aims at obtaining two basis vectors $W_X \in \mathbb{R}^m$ and $W_Y \in \mathbb{R}^n$ such that the correlation coefficient ρ of $W_X^T X$ and $W_Y^T Y$ is maximized. The objective function to be maximized is given by

$$\begin{aligned} \rho &= \frac{Cov(W_X^T X, W_Y^T Y)}{\sqrt{Var(W_X^T X)}\sqrt{Var(W_Y^T Y)}} \\ &= \frac{W_X^T C_{XY} W_Y}{\sqrt{W_X^T C_{XX} W_X W_Y^T C_{YY} W_Y}} \end{aligned} \tag{3.1}$$

where C_{XX} and C_{YY} is the autocovariance matrix of X and Y. C_{XY} denotes the covariance matrix of X and Y.

Equivalently, the 1D CCA can be formulated as a constrained optimization problem by

$$\underset{W_X, W_Y}{\operatorname{argmax}} W_X^T C_{XY} W_Y \tag{3.2}$$

subject to $W_X^T C_{XX} W_X = 1$ and $W_Y^T C_{YY} W_Y = 1$.

3.2.2 2D CCA Formulation

For some data types such as image, the data representation is inherently two-dimensional. Thus, it is desirable to analyze data in the original 2D space without reshaping the data into 1D vectors. Motivated by 2D Principal Component Analysis (2D PCA) [173], 2D CCA was recently developed in [87]. Given two centered datasets, $X = \{x_i \in \mathbb{R}^{m_x \times n_x}, i = 1, 2, ..., N\}$ and $Y = \{y_i \in \mathbb{R}^{m_y \times n_y}, i = 1, 2, ..., N\}$, 2D CCA seeks two left projection matrices $L_X \in \mathbb{R}^{m_x \times d_1}$ and $L_Y \in \mathbb{R}^{m_y \times d_1}$ and two right projection matrices $R_X \in \mathbb{R}^{n_x \times d_2}$ and $R_Y \in \mathbb{R}^{n_y \times d_2}$ such that the correlation coefficient ρ between the two projected datasets $L_X^T X R_X$ and $L_Y^T Y R_Y$ is maximized. ρ is given by

$$\rho = \frac{Cov(L_X^T X R_X, L_Y^T Y R_Y)}{\sqrt{Var(L_X^T X R_X)}\sqrt{Var(L_Y^T Y R_Y)}} \tag{3.3}$$

ρ can be written in two parts as

$$\rho_L = \frac{L_X^T C_{XY}^R L_Y}{\sqrt{L_X^T C_{XX}^R L_X L_Y^T C_{YY}^R L_Y}} \tag{3.4}$$

$$\rho_R = \frac{R_X^T C_{XY}^L R_Y}{\sqrt{R_X^T C_{XX}^L R_X R_Y^T C_{YY}^L R_Y}} \tag{3.5}$$

where C_{XX}^R is the autocovariance matrix of XR_X, C_{YY}^R is the autocovariance matrix of YR_Y, and C_{XY}^R is the covariance matrix of XR_X and YR_Y. Similarly, C_{XX}^L is the covariance matrix of $L_X^T X$, C_{YY}^L is the covariance matrix of $L_Y^T Y$, and C_{XY}^L is the covariance matrix of $L_X^T X$ and $L_Y^T Y$. The equivalent constrained problem for 2D CCA is

$$\underset{L_X, L_Y, R_X, R_Y}{\operatorname{argmax}} \; Cov(L_X^T X R_X, L_Y^T Y R_Y) \tag{3.6}$$

subject to $Var(L_X^T X R_X) = 1$ and $Var(L_Y^T Y R_Y) = 1$.

3.2.3 Difference in Solving 1D CCA and 2D CCA

Note that the optimization required for 1D CCA in (3.2) and 2D CCA in (3.6) are different. Using Lagrange multiplier, the solution of the optimization problem for 1D CCA is equivalent to the solution of the following generalized eigenvalue problems

$$\begin{aligned} C_{XY}W_Y &= \lambda C_{XX}W_X \\ C_{YX}W_X &= \lambda C_{YY}W_Y \end{aligned} \tag{3.7}$$

where $C_{YX} = C_{XY}^T$. However, the generalized eigenvalue problem for 2D CCA is different, it involves the following two sets of equations

$$\begin{aligned} C_{XY}^R L_Y &= \lambda C_{XX}^R L_X \\ C_{YX}^R L_X &= \lambda C_{YY}^R L_Y \end{aligned} \tag{3.8}$$

$$\begin{aligned} C_{XY}^L R_Y &= \lambda C_{XX}^L R_X \\ C_{YX}^L R_X &= \lambda C_{YY}^L R_Y \end{aligned} \tag{3.9}$$

The projection matrices L_X, L_Y and R_X, R_Y are solved in an iterative manner. At each iteration, to obtain the updated L_X and L_Y, R_X and R_Y are fixed, and L_X and L_Y are obtained by computing the d_1 largest generalized eigenvectors in (3.8). Similarly, to obtain the updated R_X and R_Y, L_X and L_Y are fixed, and R_X and R_Y are obtained by computing the d_2 largest generalized eigenvectors in (3.9). This process continues until convergence when the updates from the last iteration to the current iteration become very small. In our experiments, L_X, L_Y and R_X, R_Y converge in a few iterations.

3.3 2D CCA for Face Super-resolution

The proposed face super-resolution approach consists of two steps: the *first step* is face reconstruction and the *second step* is detail compensation which further refines a face reconstructed in the first step since a reconstructed face through manifold learning often does not contain sufficient details.

3.3.1 Face Reconstruction

There are two key parts for face reconstruction using 2D CCA: training and reconstruction. During the training, a 2D CCA model is learned. For reconstruction, the learned model is used to construct HR faces from the LR input. It is to be noted that there exists no publication for 2D CCA for face super-resolution and face is an important structure with enormous number of applications. Further, we will see in Section 3.4 that 2D CCA based approach provides the best performance compared to almost all the recently published papers on super-resolution.

3.3.1.1 Training

In the training, 2D CCA is applied to find the left and right projection matrices that project the HR and LR images into a subspace in which the correlation between the projections is maximized. Given the HR training set $X = \{x_i \in \mathbb{R}^{m_x \times n_x}, i = 1, 2, ..., N\}$ and the corresponding LR training set $Y = \{y_i \in \mathbb{R}^{m_y \times n_y}, i = 1, 2, ..., N\}$, the mean faces μ_X and μ_Y are subtracted to obtain the centered datasets $\hat{X}$ and $\hat{Y}$, respectively.

The left transforms $L_{\hat{X}}$ and $L_{\hat{Y}}$ and the right transforms $R_{\hat{X}}$ and $R_{\hat{Y}}$ are obtained by maximizing

$$\rho = \frac{Cov(L_{\hat{X}}^T \hat{X} R_{\hat{X}}, L_{\hat{Y}}^T \hat{Y} R_{\hat{Y}})}{\sqrt{Var(L_{\hat{X}}^T \hat{X} R_{\hat{X}})}\sqrt{Var(L_{\hat{Y}}^T \hat{Y} R_{\hat{Y}})}} \tag{3.10}$$

The image datasets $\hat{X}$ and $\hat{Y}$ are now transformed to $P_X = L_{\hat{X}}^T \hat{X} R_{\hat{X}}$ and $P_Y = L_{\hat{Y}}^T \hat{Y} R_{\hat{Y}}$.

3.3.1.2 Reconstruction

In order to perform super-resolution, a LR image i_{LR} is provided. The LR image is projected to the subspace by

$$P_i^{LR} = L_{\hat{Y}}^T (i_{LR} - \mu_Y) R_{\hat{Y}} \tag{3.11}$$

We assume that P_i^{LR} can be reconstructed by a linear combination of its K nearest neighbors in P_Y and the coefficients w_j's are obtained by minimizing the reconstruction error given by

$$\underset{\{w_j\}_{j=1}^K}{\operatorname{argmin}} \left\| P_i^{LR} - \sum_{j=1}^{K} w_j P_{Y_j} \right\|_F \tag{3.12}$$

subject to the constraint $\sum_{j=1}^K w_j = 1$. P_{Y_j} denotes a sample in the LR dataset, and $\|\cdot\|_F$ calculates the Frobenius norm. The details on solving this constrained least square problem can be found in [134].

After obtaining the reconstruction weights $\{w_j\}_{j=1}^K$, the projection of the desired HR image i_{HR} in the 2D CCA space is reconstructed by

$$P_i^{HR} = \sum_{j=1}^{K} w_j P_{X_j} \tag{3.13}$$

where P_{X_j} is the HR version corresponding to P_{Y_j}.

Similar to (3.11), P_i^{HR} is related to i_{HR} by

$$P_i^{HR} = L_{\hat{X}}^T (i_{HR} - \mu_X) R_{\hat{X}} \tag{3.14}$$

so i_{HR} is derived as

$$i_{HR} = L_{\hat{X}}^{T\,\dagger} P_i^{HR} R_{\hat{X}}^{\dagger} + \mu_X \tag{3.15}$$

where † denotes the Moore-Penrose pseudoinverse operation since $L_{\hat{X}}^T$ and $R_{\hat{X}}$ are not directly invertible.

The super-resolution approach for the whole face mentioned above is based on the rationale that the same neighborhoods are preserved in both HR dataset and LR dataset. Instead of generating a model for the whole face, we divide a face into three parts from top to bottom: eyes part, nose part, and mouth part (for the aligned face images, each part is taken within a predefined region). The same super-resolution procedure is applied directly with the only difference being that the training LR and HR image pairs and the input LR image are now certain parts of the face. The partitioning improves the global reconstruction precision by refining local reconstruction separately. The final result is obtained by stitching the three independently reconstructed parts together as shown in Figure 3.2. We average the pixels on the boundaries from different parts to generate a smooth output.

3.3.2 Detail Compensation

During face reconstruction, the projection of the face data into a subspace inevitably loses some information and this is often observed as the lack of high-frequency details. Furthermore, the neighborhood reconstruction itself is essentially an averaging process which further smooths the reconstruction results. To alleviate this problem, we

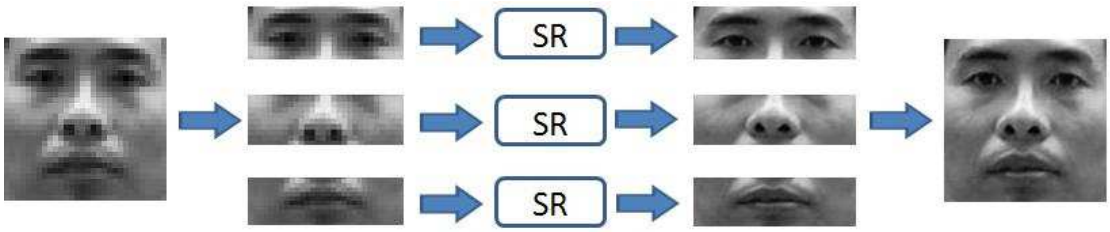

Figure 3.2: A Face is divided into three parts corresponding to eyes region, nose region, and mouth region. Super-resolution is performed separately for each part and the outputs are merged together to form the high-resolution output.

add a detail compensation step in order to generate faces with high-frequency details.

For a HR image x_j in the training set X, a Gaussian filter is applied, effectively as a low-pass filter. The output $\tilde{x}_j$ is a blurred version of the original image that mainly contains the low-frequency components (LF). By subtracting the low-passed image $\tilde{x}_j$ from the original image x_j, an image h_j that contains mainly high-frequency (HF) components is generated.

The reconstruction weights are computed during the training as in (3.12) to generate a HF compensation mask i_{comp} by

$$i_{comp} = \sum_{j=1}^{K} w_j h_j \tag{3.16}$$

where $\{h_j\}_{j=1}^{K}$ corresponds to the same neighborhood as that of $\{P_{X_j}\}_{j=1}^{K}$ and $\{P_{Y_j}\}_{j=1}^{K}$, and K is the number of chosen nearest neighbors. The final output is given by

$$i_{out} = i_{HR} + i_{comp} \tag{3.17}$$

Similar as it was done in face reconstruction, each HF image is divided into three parts and the three sets of calculated weights are applied to generate the three reconstructed HF masks. The HF masks are then combined together to form a detail compensation mask for the whole face.

3.4 Experiments

3.4.1 Experimental Protocols

3.4.1.1 Face Datasets

We evaluate our method on the CAS-PEAL-R1 dataset [46] which contains face images of 1040 individuals. We use frontal face images of these individuals with normal lighting and neutral expression. All images are cropped and geometrically normalized with the locations of the eyes and mouths fixed to form the the HR images of size 120×120 and the LR images of size 30×30. Thus, the magnification factor is 4 in our experiments. We selected images of 940 individuals for training and the rest of the images are used for testing. As studied and pointed out in [71] as well as found out empirically in our experiments, the larger the neighborhood size K is, the better the results of neighborhood reconstruction are. Therefore, in our experiments all the training images are used in the K nearest neighbor reconstruction. To generate the LF images, a Gaussian filter with $\sigma = 0.8$ is applied with a 5×5 mask. This low-pass filtering of the image simulates the formation of the LR image from HR image. The selection of the parameters for the Gaussian filter is similar to the settings in the previous work for image SR [38, 172, 4].

In addition, we also use CUHK student dataset [160] which contains 188 subjects with frontal faces to test our algorithm. The images in this dataset are cropped and aligned in the same manner. Figure 3.3 shows some sample images from both datasets.

For training, the HR images are divided into three parts of size 40×120, 48×120, and 40×120. The middle part has 4 rows of pixels at top and at bottom overlapping with the upper part and the lower part. The corresponding LR images are

Figure 3.3: Sample images from two datasets: CAS-PEAL-R1 dataset [46] (top) and CUHK student dataset [160] (bottom).

divided to 10×30, 12×30, and 10×30. For testing, we concatenate the super-resolved parts and average the overlapping pixels. By averaging the overlapping regions the final output is smooth and consistent (see Figure 3.7). We choose $d_1 = d_2 = 30$ in projection matrices empirically to maintain the reconstruction accuracy in an efficient manner. The projection matrices converge in about 10 iterations during optimization.

3.4.1.2 Methods Compared

We compare our results to four state-of-the-art methods: artifact free super-resolution method using iterative curve based interpolation (ICBI) [49], sparse representation based super-resolution (SPR) [175], the position-patch based method (PP) [113], and the 1D CCA based face reconstruction method (1D CCA) [71]. Among these methods, 1D CCA and PP are specially designed to super-resolve face images while ICBI and SPR are the super-resolution algorithms for generic images. For ICBI, SPR and PP we used the default settings provided by the authors of these papers [49, 175, 113].

We use the same training set in 1D CCA as in our method [71]. For face reconstruction in PP, the input HR-LR pairs come from our training set. Note that we do not compare with [124] since it is not directly related to image super-resolution and we do not compare with [155] since it is primarily designed for face image-sketch synthesis.

3.4.1.3 Metrics for Quantitative Evaluation

We calculate the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [161] scores for the super-resolved face images. In addition, we calculate the distortion measure (DM) [34] that evaluates the distortion from the original image in the frequency domain, in which other image quality measures usually do not work. The DM is calculated by

$$DM = \int_0^{f_{max}} \left[1 - DTF(\frac{f_r}{f_N})\right] CSF(f_r) df_r \tag{3.18}$$

where f_r is the radial frequency, f_N is the Nyquist frequency and f_{max} is a predefined value. DTF is a distortion transfer function and CSF is contrast sensitivity function which approximates the human visual system (HVS).

Furthermore, we also apply a recently introduced SVD-based quality measure [140] in our experiments. The SVD-based image quality measure tries to mimic a human viewer by measuring different distortion types at different levels. In this metric first a graphical measure is calculated for each image block of size $n \times n$ by

$$D_i = \sqrt{\sum_{i=1}^{n} (s_i - \hat{s}_i)^2} \tag{3.19}$$

where s_i is the singular value of the original block and $\hat{s}_i$ is the singular value of the distorted block. The global measure is obtained by

$$M_{SVD} = \frac{\sum_{i=1}^{\frac{k}{n}\times\frac{k}{n}} |D_i - D_{mid}|}{\frac{k}{n}\times\frac{k}{n}} \tag{3.20}$$

where D_{mid} is the median of the sorted D_i and $k \times k$ is the size of the image.

3.4.2 Experimental Results

3.4.2.1 Effect of Part-Based SR

As aforementioned, instead of performing SR on the whole face, we divide the face into three parts and reconstruct each part individually. The final output merges the three parts together. Figure 3.4 shows the difference by subtracting the reconstructed faces from the original HR images for some sample images using the holistic method and the part-based method.

Examining the difference between the original images and the proposed part based SR results, we find that reconstructing the entire face using one trained model brings more error especially in the regions of eyes, nose and mouth. By specializing the trained model for a specific part of a face, the output is closer to the ground-truth with less distortions.

3.4.2.2 Effect of Detail Compensation

In Figure 3.5 we examine the Fourier transform on the face images after face reconstruction and detail compensation. The magnitude of the Fourier transform is drawn as the heat map where the magnitude decreases from red to blue. After face reconstruction, the low-frequency components are dominating, as can be seen in the center of the heat map. The magnitude decreases from the center towards the corners of the heat map. This agrees with our visual impression that the reconstructed faces

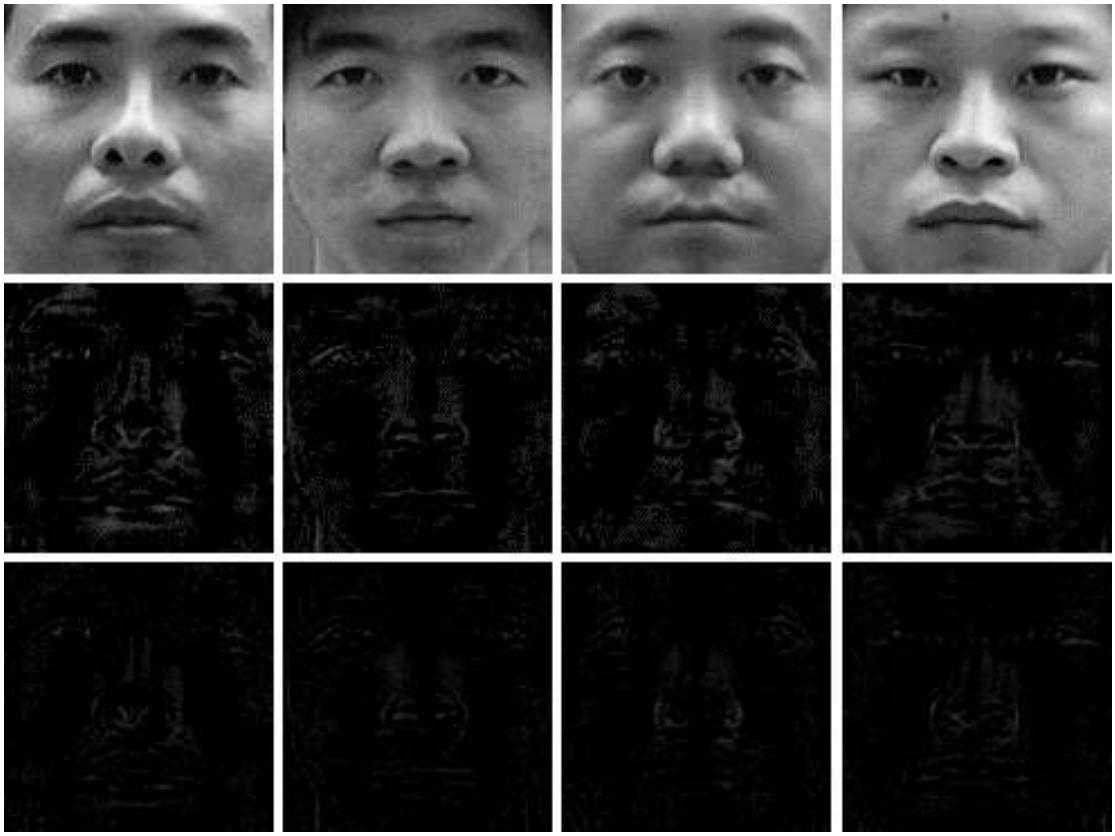

Figure 3.4: Original HR images (top), residue between the original images and whole face based SR results (middle), residue between the original images and part based SR results (bottom).

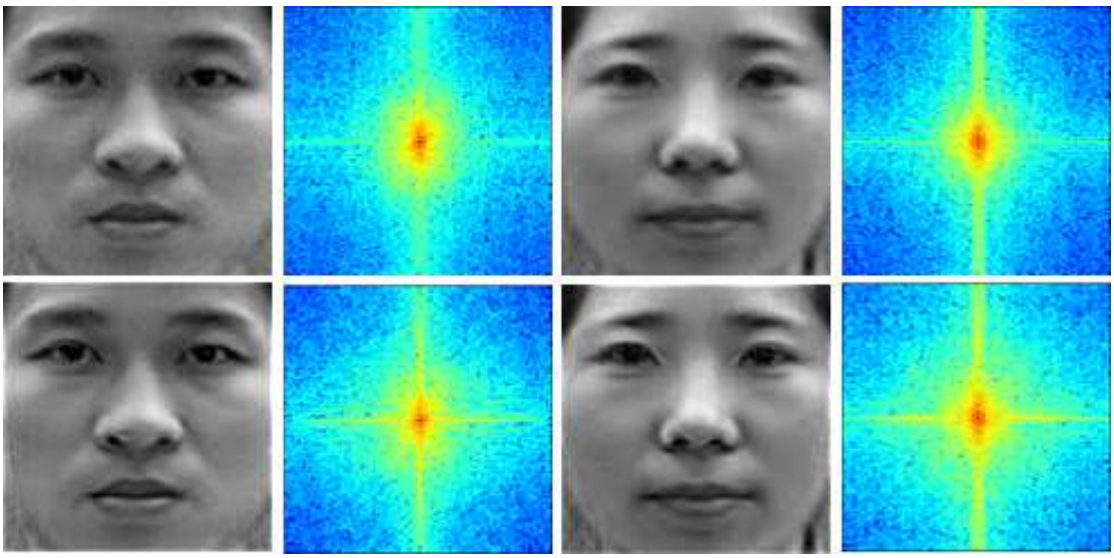

Figure 3.5: Effects of detail compensation. (Top) Reconstructed faces by 2D CCA. (Bottom) Results after detail compensation. The heat map to the right of the image shows the magnitude of its Fourier transform.

are not sufficiently sharp. After detail compensation, the magnitude is increased in the high-frequency components. Thus, the face images look sharper with less blurriness and more details.

Table 3.1 shows the effect of part-based SR and detail compensation. The performance gain occurs from holistic face SR to part-based SR for all the three evaluation metrics (PSNR, SSIM, and DM). Further improvement is achieved by detail compensation. The scores of SVD [140] are similar in the comparisons. The quantitative results indicate that both the part-based SR and detail compensation help to improve the output image quality. These quantitative results correspond to the visual observations from Figure 3.4 and Figure 3.5.

Table 3.1: Evaluation of the effects of part-based SR and detail compensation. The evaluation metrics include PSNR, SSIM [161], DM [34] and SVD [140]. For all the metrics, the higher score is better.

	Holistic Face	Part-Based	Detail Comp.
PSNR(dB)	32.45	34.46	34.89
SSIM	0.843	0.884	0.885
DM(dB)	35.38	35.87	38.25
SVD	0.667	0.672	0.668

3.4.2.3 Effect of Projection Dimension

The effect of the dimension $d1$ and $d2$ in the projection matrices is examined. As shown in Figure 3.6, when $d1$ and $d2$ are small, the reconstruction is not accurate. The choice of $d1 = d2 = 30$ generates good results. As $d1$ and $d2$ become much larger, the differences in the outputs are not easily noticeable while the computation and memory expenses increase. As a result, we choose $d1 = d2 = 30$ in our experiments as the dimension of the left and right projection matrices.

3.4.2.4 Comparison with 1D CCA Based Method

In the 1D CCA based method [71] the images are first converted into vectors. From Figure 3.7(e) we can see that although 1D CCA is able to super-resolve the faces, the output images suffer from distortions. This inaccurate reconstruction visually causes the super-resolved image deviate from the ground-truth image. In other words, the output face images look different from the actual subjects (see the distortions on the subjects' noses, eyes, mouths, and chins). This may degrade the performance of latter

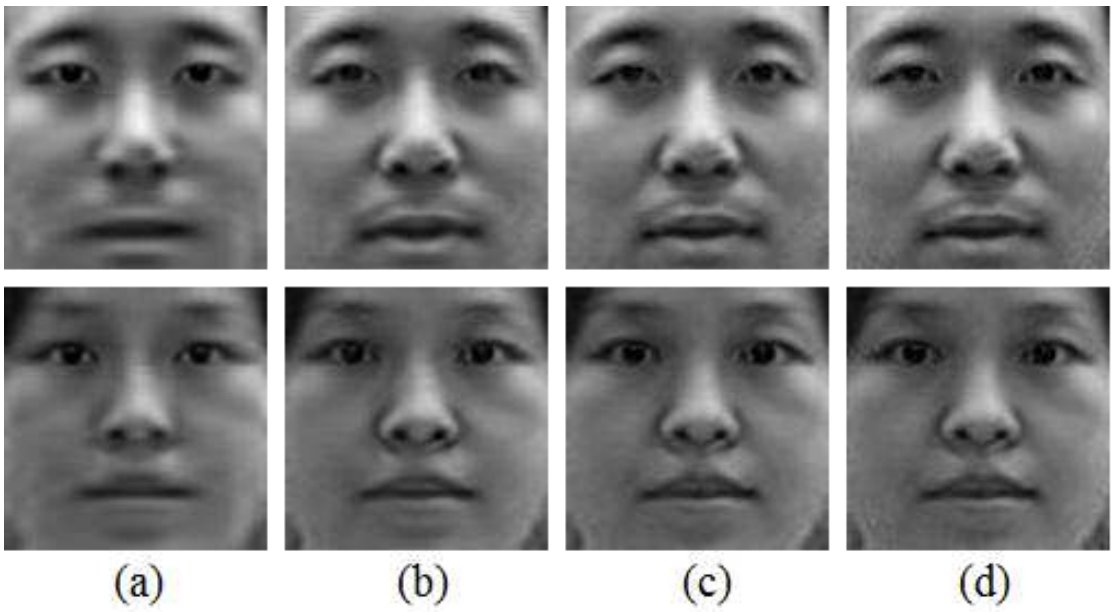

Figure 3.6: Effects of dimension of the projection matrices. (a) $d = 10$. (b) $d = 20$. (c) $d = 30$. (d) $d = 40$.

processing steps such as face recognition (see Table 3.2). The reason for the distortions is that the images are first reshaped into 1D vectors and then the relationships in the 1D CCA subspace are explored. However, since the data are intrinsically 2D structured, this reshaping process would inevitably discard the 2D spatial information in the original data representation.

In the proposed 2D CCA based methods, those distortions are significantly reduced (see Figure 3.7(f)) with respect to the ground-truth (see Figure 3.7(g)). It is evident that by bypassing the image vectorization, the output image is better reconstructed in term of its underlying structure.

Figure 3.8 shows the box plots for the quality measures by PSNR, SSIM, DM, and SVD. Box plot is a non-parametric display of differences between groups of numerical data. The proposed method outperforms the 1D CCA based method in all of the reference based and non-reference based metrics. Especially, 1D CCA yields poor scores

Figure 3.7: The super-resolution Results. Top three rows are from CAS-PEAL-R1 dataset [46] and bottom three rows are from CUHK dataset [160]. (a) Low-resolution images (enlarged by pixel replication). (b) Results by ICBI [49]. (c) Results by SPR [175]. (d) Results by PP [113]. (e) Results by 1D CCA [71]. (f) Results by the proposed method. (g) Original high-resolution images.

for SSIM as a similarity measure, since its results are distorted from the ground-truth. Compared with 1D CCA based method, both the reconstruction error and artifacts are reduced by our method, thus leading to better quantitative scores. In [71] the 1D CCA based method outperforms some of the representative face super-resolution methods including [100], a common baseline method for face super-resolution. On the other hand, the proposed method outperforms the method in [71] (see Figure 3.8). Therefore, the proposed method is better than [100].

3.4.2.5 Comparison with Other Methods

We compare the proposed approach with some state-of-the-art methods. Figure 3.7 shows sample results with different methods. We summarize the findings as follows:

- The results by ICBI [49] (see Figure 3.7(b)) do not contain sufficient details and the blurriness in the output is not removed. The interpolation based method is not able to reconstruct facial details.

- SPR based method [175] tackles the SR problem from the perspective of compressed sensing. It is based on the assumption that the sparse representation can be recovered correctly from the downsampled signal. As shown in Figure 3.7(c), the results contain more detail and the faces are reconstructed properly. However the staircase noise is noticeable along the curved edges.

- PP based method [113] divides the face image into many small connected patches and then uses neighbor embedding to super-resolve each patch separately. The final results are constructed by stitching the small patches together. As can be seen from the results (see Figure 3.7(d)), the general structure of the faces is well

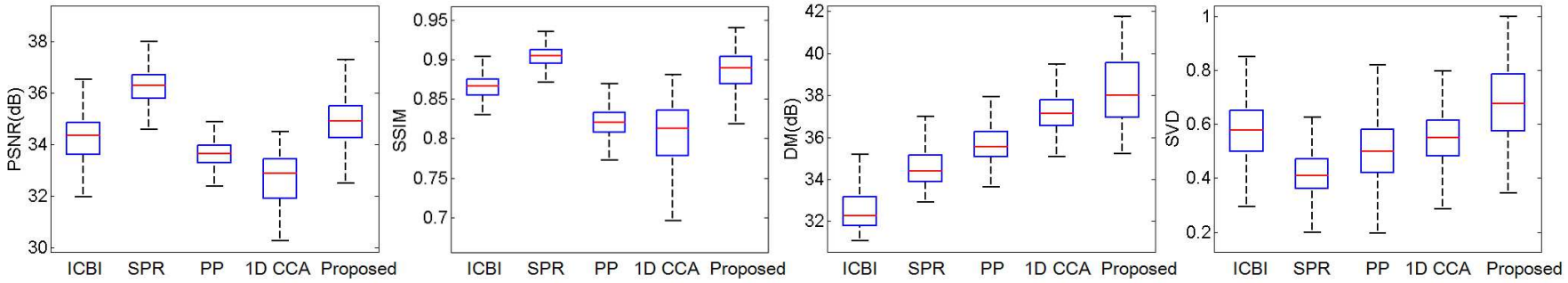

Figure 3.8: The box plots of the proposed method and four state-of-the-art methods: ICBI [49], SPR [175], PP [113], and 1D CCA [71]. The red bar indicate the median of the results. Above median is the upper quartile and maximum value while below median is the lower quartile and minimum value. Metrics from left to right: PSNR, SSIM [161], DM [34] and SVD [140]. For all the metrics, the higher score is better.

maintained. The blockness artifacts are explicitly visible due to the reconstruction in a local manner without the global refinement.

- The 2D spatial structure of the LR faces is well maintained in the 2D CCA subspace since the image data are fed directly in the optimization process without being converted into vectors. Figure 3.7(f) shows the final results by our method. In this case, there is more resemblance between the output and the original images. The detail compensation further improves the results by adding more details to the faces, i.e. the contours of the eyes and mouths. Compared to the original HR face images, the final outputs of the proposed algorithm are realistic without explicit artifacts, confirming that the 2D structural information is well maintained by using 2D CCA method.

The training images are from the CAS dataset only. When applied the trained model to the CUHK dataset, which is taken under different illumination conditions (see Figure 3.3), the outputs are still satisfactory. Thus, it is evident that the trained model is not dataset dependent and this merit makes our method generalizable.

The proposed method outperforms PP [113], and ICBI [49] in PSNR and SSIM. For PSNR the results by SPR [175] are higher than the proposed method. However, this contradicts with our visual examination. In fact, as indicated in [161], PSNR itself does not translate the visual quality to scores faithfully. For SSIM, our method is comparable to SPR [175]. For DM and SVD our method yields the best results, which means in these aspects the proposed method is able to generate highest quality outputs with minimum artifacts and distortion. The results by DM and SVD are more coherent with the visual quality on the super-resolved images from Figure 3.7.

Table 3.2: Recognition accuracy using super-resolved images.

Method	ICBI [49]	SPR [175]	PP [113]	1D CCA [71]	Step 1	Proposed
Accuracy	75%	97.22%	87.5%	64.58%	98.96%	99.31%

3.4.2.6 Results on Real World Data

We test the proposed method on a real world image with LR faces. We manually extracted the faces from the picture and aligned them as in the previous experiments. Some samples of the super-resolved face images are shown in Figure 3.9. These subjects are not present in the database we used for training and testing above. Note that the quality of the input image is significantly worse than the quality of the images from CAS-PEAL-R1 or CUHK datasets due to noise, blurriness and artifacts caused by compression. Still our algorithm is able to generate reasonably good results.

3.4.3 Effect of Super-Resolution on Recognition

In order to evaluate the effects of super-resolution to face recognition, we conduct a recognition experiment with LBP-based face recognition [1] as the baseline method. The 100 subjects from CAS-PEAL-R1 dataset [46] and 188 subjects from CUHK dataset [160] used for testing in the above experiments are now combined to form a dataset of 288 subjects. The gallery set contains HR image for those subjects and the query data are the super-resolved images using different methods. Table 3.2 shows the recognition accuracy.

From the recognition rates we can see that the images generated by the proposed methods lead to better recognition result compared to the other methods. The

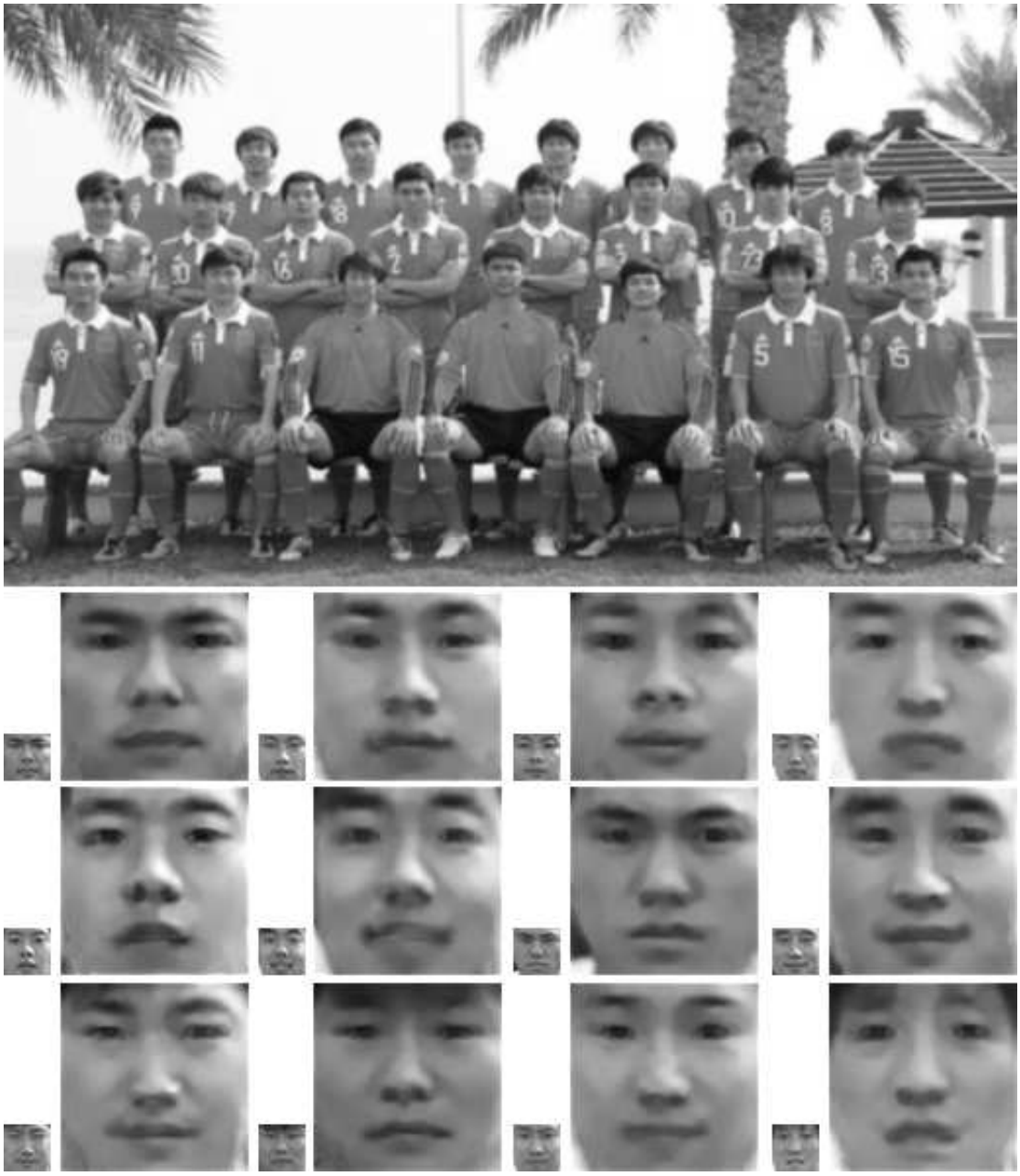

Figure 3.9: Results on a real world image. (Top) Original image. (Bottom) Some extracted LR faces (small images) and the super-resolved faces (large images).

result by 1D CCA [71] is less competitive since the generated faces are distorted from the ground-truth, which adversely affects the recognition performance. The recognition rate using 2D CCA only (step 1) without detail compensation is 98.96%, which is better compared to other methods. By further compensating the details, the highest recognition rate of 99.31% is achieved.

We would expect that by using more sophisticated features and classifiers the recognition rates using images by different super-resolution methods would also increase. However, given parameter settings in the baseline face recognition method, the 2D CCA super-resolved faces result the best recognition rate. This is in agreement with the visual quality assessment in Figure 3.7 in which the faces are faithfully reconstructed using 2D CCA .

3.4.4 Computational Complexity

One of the advantages of our method is the simplicity of computation. The computational complexity depends on the solution to solve the generalized eigenvalue problem to solve ((3.8) and (3.9)). Many eigenvalue solvers can be used. For instance, the Arnoldi iteration [148] is an efficient and popular algorithm. Given a generalized eigenvalue problem $Ax = \lambda BX$, suppose the matrix size of A and B be N by N, then the computational complexity using Arnoldi iteration is $O(dN^2 + d^2N)$, where d is the number of significant eigenvalues. In (3.8) and (3.9) for the solution of 2D CCA, the sizes of covariance matrices corresponding to A and B are small (i.e., they are of the order of image width or height). This implies smaller N in $O(dN^2 + d^2N)$, thus 2D CCA is computationally efficient. On the other hand, in 1D CCA, since images are first reshaped to vectors, the corresponding covariance matrices in (3.7) are much larger (i.e., they are of the order of total number of pixels in an image), which lead to more

Table 3.3: Comparison of training time (in seconds) and average time to super-resolve a face image.

Method	ICBI [49]	SPR [175]	PP [113]	1D CCA [71]	Proposed
Training	–	351.86s	–	103.36s	64.30s
Testing	0.087s	39.23s	112.81s	0.48s	1.38s

expensive 1D CCA as compared to the 2D CCA.

Among the methods compared in this chapter, SPR [175], 1D CCA [71], and the proposed 2D CCA based method require a training process. Note that in PP [113] the HR-LR pairs are used to reconstruct the testing image and no explicit model or subspace is formed via training, thus, we do not consider it here as a learning based approach that requires training. All the programs are implemented in MATLAB and were executed on a desktop with a 2.4 GHz CPU and 3 GB of RAM. The implementation of our method is not optimized. Table 3.3 shows the training time and the average time to super-resolve a face image on CAS-PEAL-R1 and CUHK datasets.

Compare to the other methods involving a training process, our method took significantly less time for training. It is due to the small size of matrices involved in the 2D CCA computation. To super-resolve a LR face image, although ICBI [49] and 1D CCA [71] spent less time, our method is able to generate better results while also keeping the computation time to a small value. When comparing to SPR [175] and PP [113], our method requires much less time to super-resolve an image.

3.4.5 Discussion

The high quality of the super-resolved image demands accurate alignment in the preprocessing step. The images for training and testing need to be aligned in the

same manner (with the positions of eyes and the center of the mouth fixed). Without proper alignment, the quality of the output would degrade. However, this constraint also holds for CCA based method [71] and position-patch based method [113].

3.5 Conclusions

In this chapter, a two-step approach for face super-resolution based on 2D canonical correlation analysis is proposed. One major merit of the proposed method is that our method works directly on the original 2D representation of the image data without converting the images into vectors as it is commonly done in the previous work. This important methodology maintains the intrinsic 2D structure of the face images. Experimental results show that compared to the state-of-the-art methods, the super-resolved faces by the proposed approach are visually realistic and very close to the ground-truth. Various image quality metrics also support that the results by our method are superior to the other methods. The super-resolved images are tested in the recognition task and the results suggest that the super-resolved images by the proposed method achieve the highest accuracy. Due to the small matrices involved in our method, the computation in both training and testing processes is very efficient.

Chapter 4

Dynamic Bayesian Network for Unconstrained Face Recognition in Surveillance Camera Networks

4.1 Introduction

With the broad establishment of surveillance video camera systems in recent years in both public and private venues, the recognition/verification of the subjects is often of interest and importance for purposes such as security monitoring, access control, etc. Some biometric traits such as gait can be used to recognize different subjects [58], however, it is preferred to use more distinct biometric clues such as face to identify a subject. Although face recognition has been studied extensively, face recognition in an unconstrained environment such as in surveillance camera videos remains very challenging and the recognition rate could drop dramatically to less than 10% using standard technologies [54]. The challenges to unconstrained face recognition in surveillance cam-

eras are mainly due to the following reasons:

- Low resolution. In the video captured by surveillance cameras, the pixels that account for the faces are very limited. However, previous studies have shown that faces of size 64×64 are required for the existing algorithms to achieve good recognition accuracy [110].

- Arbitrary poses. Usually the subjects are moving freely. Consequently, it is not uncommon that the captured faces have different poses in different cameras.

- Varying lighting conditions. As the lighting is usually not uniform in the coverage area of the surveillance cameras, the illumination on the subject's face could vary significantly as he/she moves (e.g., the subjects walks into the shade from direct sunshine).

- Noise and blurriness. The captured images are often corrupted by noise during transmission and the motion of the subjects usually introduces blurriness.

Figure 4.1 shows an example of a subject's face captured by three surveillance cameras. The cameras are placed above a portal. The cameras have different viewing angles and none of the cameras captures the full frontal face of the subject. The face images exhibit variations in resolution, lighting condition and poses. In addition, noise and blurriness are also observed. Under such circumstance, the standard face recognition algorithms such as Eigenfaces [149] would fail to work effectively. Despite the aforementioned difficulty, a multi-camera system provides different views of subjects which are complementary to each other. This enables the potential to improve the recognition performance with low quality input faces from multiple cameras.

Figure 4.1: The subject's face is captured by 3 cameras from different views in a typical surveillance camera system setup [164]. The pose and resolution of the captured faces vary across different views.

In this chapter, we propose a dynamic Bayesian network (DBN) based approach to tackle the problem of face recognition in multi-camera systems.

4.1.1 Contributions of This Chapter

Previously Bayesian network has been applied to face recognition. Heusch *et al.* [64] combined intensity and color information for face recognition in a Bayesian network where the observation nodes represented different parts of the face and the hidden nodes described the types of the observations. In [123] an embedded Bayesian network was proposed for efficient face recognition. Beyond the image-based recognition, there has been a growing interest to study the temporal dynamics in video sequences to improve the recognition performance in recent years [43, 105]. A head pose estimation framework using Bayesian network was proposed in [78].

We propose a probabilistic approach for video-to-video face recognition using a DBN, utilizing different frames from multiple cameras [6]. DBN has previously been applied to tasks such as speech recognition [122], vehicle classification [77], visual tracking [156] and facial expression recognition [185]. Variant of DBN such as topological DBN has also been proposed to identify human faces across age [24]. In this chapter, the DBN is constructed by repeating a Bayesian network over a certain number of time slices with time-dependent variables. In each time slice the observed nodes are from different cameras. During the training, the temporal information is well encoded and the person-specific dynamics are learned. The identity of the testing subject can be inferred using previously trained network structure and parameters. By using DBN we are able to factor the joint probability distribution considering the temporal relationship of the feature evolution process between consecutive frames. Moreover, the DBN is defined and structured in a way that adding more cameras is easy. In addition, if features from one camera were not extracted due to image capture failure, this information can still be inferred by DBN and, therefore, recognition may not fail.

Compared to the previous work [13] in which the DBN structure is manually defined and only two cameras are used for recognition, in this chapter, the topological structure in each time slice of DBN is learned automatically in an optimal manner with three cameras involved. In addition, the experimental results are examined thoroughly using multiple performance evaluation criteria using much more data with improved evaluation protocol.

In summary, the contributions of this chapter are:

- We propose a probabilistic framework for unconstrained face recognition in a multi-camera surveillance scenario. To the authors' best knowledge, this is the first work

using DBN for video-based face recognition in surveillance cameras with more than two cameras. The framework is flexible and can be easily extended to more complicated multi-camera settings. Besides, any feature descriptor is compatible in this framework.

- We test the proposed method on a publicly available multi-camera surveillance video dataset "ChokePoint" with unconstrained face acquisition [164], in contrast to the other commonly used datasets which were collected in controlled environment. We compare the proposed method with popular benchmark classifiers using different feature descriptors. The superior performance of the proposed DBN approach is verified in different aspects with various evaluation criteria.

- We compare the face recognition performance using all of the three cameras in the ChokePoint dataset against using single camera. Experimental results demonstrate that using multiple cameras improves the recognition performance over any single camera.

The rest of this chapter is organized as follows. Section 4.2 describes the details of the proposed method. In Section 4.3 the experimental results are reported. We conclude this chapter in Section 4.4. Before the detailed algorithms is presented, Table 4.1 gives a summary of the symbols used in the following sections for a better understanding.

$$\begin{aligned} s &= \underset{S}{\operatorname{argmax}}\, p(S|CAM_1, CAM_2, CAM_3) \\ &= \underset{S}{\operatorname{argmax}}\, \frac{p(CAM_1, CAM_2, CAM_3|S)p(S)}{\sum_S p(CAM_1, CAM_2, CAM_3|S)p(S)} \\ &= \underset{S}{\operatorname{argmax}}\, \frac{p(CAM_1|S)p(CAM_2|CAM_1, S)p(CAM_3|CAM_2, S)p(S)}{\sum_S p(CAM_1, CAM_2, CAM_3|S)p(S)} \end{aligned} \tag{4.1}$$

Table 4.1: Definition of the Symbols used in this chapter

Symbol	Definition
K	Number of cameras in the multi-camera setup
k	camera index
T	total number of time slices in the DBN (sequence length)
t	time slice index
CAM_k^t	the random variable representing the feature vector of a face image from the k^{th} camera in time slice t
N	Number of subjects in the gallery
S	the random variable representing the probability distribution over the gallery of subjects

4.2 Technical Details

In the following subsections, we first explain the Bayesian network structure for face recognition from multiple cameras with a single time slice and then the DBN structure with multiple time slices is presented.

4.2.1 Bayesian Network

A Bayesian network (BN) is a graphical model, which is defined using a directed acyclic graph. The nodes in the model represent the random variables and the edges define the dependencies between the random variables. Given the value of its parents, each variable is conditionally independent of its non-descendants. A BN can effectively represent and factor the joint probability distributions and it is suitable for the classification tasks. Mathematically, given a set of ordered random variables X_1, X_2,...,X_n, the full joint distribution is given by:

$$p(x_1, x_2, \ldots, x_n) = p(x_1) \times p(x_2|x_1) \times \ldots$$
$$\times p(x_n|x_1, x_2, \ldots, x_{n-1}) = \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1}). \quad (4.2)$$

In the scope of multi-camera face recognition, when several face images of the same subject are captured by different cameras, we construct the corresponding BN using two different kinds of nodes:

- Root node: This is a discrete node on the top of the BN. The node is represented by a random variable S. S is the probability distribution over all the subjects in the gallery and does not represent the identity of a single subject. The size of the root node indicates the number of the subjects (classes).

- Camera node: This continuous node contains the feature descriptors of the extracted face image from one camera. The number of the camera nodes depends on the number of cameras involved in the surveillance. Different feature descriptors such as local binary patterns (LBP) [3] or local phase quantization (LPQ) [2] can be adopted. The notation CAM is used to represent this random variable.

When a test sequence is provided, the subject's identity s is determined using the *maximum a posterior* (MAP) rule:

$$\begin{aligned} s &= \underset{S}{\operatorname{argmax}}\, p(S|CAM_1, \ldots, CAM_K) \\ &= \underset{S}{\operatorname{argmax}}\, \frac{p(CAM_1, \ldots, CAM_K|S)p(S)}{\sum_S p(CAM_1, \ldots, CAM_K|S)p(S)} \end{aligned} \tag{4.3}$$

where CAM_k is the random variable representing the feature vector from the face image in camera k. $p(S)$ is the prior probability of the presence of each subject and is usually modeled by a uniform distribution. Since the different cameras are capturing the same subject, the camera nodes are not independent. We explain how the BN structure is learned in the next part.

4.2.2 Structure Learning

The structure of the BN would greatly impact the accuracy of the model. However, the number of possible structures is super-exponential in the total number of nodes. Therefore, it is desirable to avoid performing exhaustive search for structure learning. In this chapter, we use the K2 structure learning algorithm [29] to determine the BN's structure. K2 uses a greedy approach to incrementally add parents to a node according to a chosen scoring function. The search space of K2 algorithm is much smaller than the entire space due to the ordering of the nodes and it guarantees no cycle in the generated structure. We use the Completed Likelihood Akaike Information Criterion

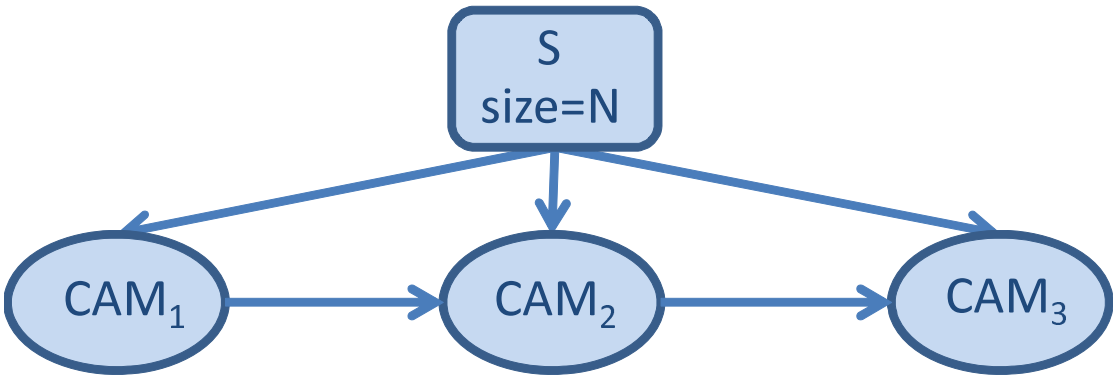

Figure 4.2: K2 learned Bayesian network structure [29]. The training data is from the ChokePoint dataset [164].

(CL-AIC) scoring function for this purpose [166]. Figure 4.2 shows the K2 learned BN structure. In this case, the subject's identity s is determined by Equation 4.1.

4.2.3 Dynamic Bayesian Network for Face Recognition

Compared to the traditional face recognition methods which are typically image based, the video based face recognition is advantageous since the dynamics in different frames for the specific person can be learned to help the recognition of the subject. As suggested in [131], multiple face samples from a video sequence have the potential to boost the performance of the recognition system.

We propose our graphical model as a DBN. DBN differs from HMM in the following aspects: a DBN represents the problem utilizing a set of random variables whereas an HMM uses a single discrete random variable; in a standard first-order HMM modeled as a DBN, the random variables at time slice t depend only on the variables in time slices t and $t-1$ for all $t > 1$; in an HMM all the hidden random variables are combined in a single multi-dimensional node, whereas in a DBN multiple hidden nodes can be present.

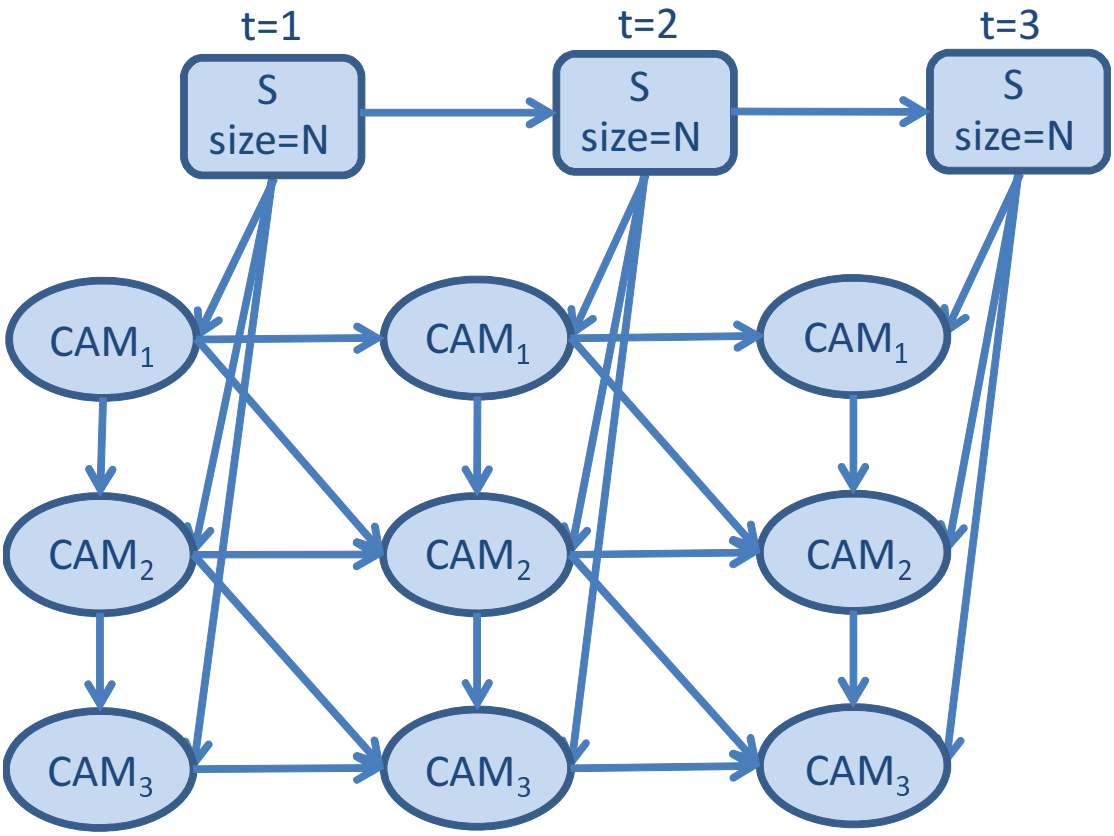

Figure 4.3: The DBN structure for 3 time slices with a 3-camera setup.

$$
\begin{aligned}
p(S^t|CAM^{1:T}_{k=1,2,3}) &= p(S^t, CAM^1_1 = cam^1_1, \ldots, CAM^T_3 = cam^T_3) \times \underbrace{1/p(CAM^1_1 = cam^1_1, \ldots, CAM^T_3 = cam^T_3)}_{=L \text{ (a constant)}} \\
&= \sum_{S^1,\ldots,S^{t-1},S^{t+1},\ldots,S^T} p(S^1, \ldots, S^T, CAM^1_1 = cam^1_1, \ldots, CAM^T_3 = cam^T_3) \times L \\
&= \sum_{S^1,\ldots,S^{t-1},S^{t+1},\ldots,S^T} p(S^1)p(CAM^1_1|S^1)p(CAM^1_2|S^1, CAM^1_1)p(CAM^1_3|S^1, CAM^1_2) \\
&\quad \times \prod_{i=2:T} p(S^i|S^{i-1}) \prod_{i=2:T} p(CAM^i_1|S^i, CAM^{i-1}_1)p(CAM^i_2|S^i, CAM^{i-1}_1, CAM^{i-1}_2, CAM^i_1) \\
&\quad \times p(CAM^i_3|S^i, CAM^{i-1}_2, CAM^{i-1}_3, CAM^i_2) \times L \\
&= \sum_{S^1,\ldots,S^{t-1},S^{t+1},\ldots,S^T} p(S^1)p(CAM^1_1|S^1)p(CAM^1_2|S^1, CAM^1_1)p(CAM^1_3|S^1, CAM^1_2) \\
&\quad \times \prod_{i=2:t} p(S^i|S^{i-1}) \prod_{i=2:t} p(CAM^i_1|S^i, CAM^{i-1}_1)p(CAM^i_2|S^i, CAM^{i-1}_1, CAM^{i-1}_2, CAM^i_1) \\
&\quad \times p(CAM^i_3|S^i, CAM^{i-1}_2, CAM^{i-1}_3, CAM^i_2) \\
&\quad \times \prod_{i=t+1:T} p(S^i|S^{i-1}) \prod_{i=t+1:T} p(CAM^i_1|S^i, CAM^{i-1}_1)p(CAM^i_2|S^i, CAM^{i-1}_1, CAM^{i-1}_2, CAM^i_1) \\
&\quad \times p(CAM^i_3|S^i, CAM^{i-1}_2, CAM^{i-1}_3, CAM^i_2) \times L
\end{aligned}
\tag{4.4}
$$

In terms of complexity, an HMM would require $O(T(N^K)^2)$ for inference, $O(N^{2K})$ parameters to specify $P(S^t|S^{t-1})$, and $O(TN^K)$ space, where T is the sequence length, N is the number of classes, and K is the number of camera observations. For a DBN, $O(TKN^{K+1})$ is required for inference, and $O(KN^2)$ parameters to specify $P(S^t|S^{t-1})$. The DBN has exponentially less parameters and inference is much faster.

Operating a graphic model requires three main steps: defining the structure, learning the parameters, and inference. The structure of the DBN consists of the inter-slice topology and the intra-slice topology. The inter-slice topology is defined as follows. Each time slice $t = 1 \dots T$ has $K + 1$ nodes; one root node S, and K camera nodes $CAM_{k=1\dots K}$. This structure is the same as shown in Figure 4.2 for the 3-camera setting ($K = 3$). The intra-slice topology is illustrated in Figure 4.3 with 3 time slices.

After defining the structure, it is required to learn the parameters of the DBN before recognition is performed. Therefore, the probability distribution for each node given its parents should be determined. In the 3-camera setting, for the first time slice this includes:

$$p(CAM_1^1|S^1),\ p(CAM_2^1|S^1, CAM_1^1),$$
$$p(CAM_3^1|S^1, CAM_2^1),\ p(S^1) \tag{4.5}$$

For time slices $t = 2 \dots T$ it includes:

$$p(CAM_1^t|S^t, CAM_1^{t-1}),$$
$$p(CAM_2^t|S^t, CAM_1^{t-1}, CAM_2^{t-1}, CAM_1^t),$$
$$p(CAM_3^t|S^t, CAM_2^{t-1}, CAM_3^{t-1}, CAM_2^t),\ p(S^t|S^{t-1}) \tag{4.6}$$

With new unseen data (evidence), an inference algorithm is applied to compute the marginal probability from the evidence. Specifically, inference determines the

subject's identity by $p(S^T|CAM_{k=1,2,3}^{(1:T)})$, where $CAM_{k=1,2,3}^{(1:T)}$ refers to features from all of the three cameras for time slices 1 to T. In other words, a probability distribution over the set of all the subjects is defined. The goal is to find the marginal probability of each hidden variable. Equation 4.4 shows how $p(S^t|CAM_{k=1,2,3}^{(1:T)})$ is computed for any $t = 2 \dots T$.

4.3 Experiments

4.3.1 Experimental Settings

4.3.1.1 Dataset

We use the ChokePoint dataset [164] which is designed for evaluating face recognition algorithms under real-world surveillance conditions. This dataset is challenging for face recognition task as the captured faces are unconstrained in terms of pose, lighting, and image quality. Although many face datasets exist, to the authors' best knowledge, the ChokePoint dataset is the only available open surveillance video dataset with multiple cameras. Figure 4.4 shows some sample images from this dataset. The setting for the network involves three cameras mounted above two portals (P1 and P2) that captured the video sequences of the moving subjects while the subjects were either entering (E) or leaving (L) the portals in a natural manner. In total four data subsets are available (P1E, P1L, P2E, and P2L). In each subset, four sequences are provided (S1, S2, S3, and S4) and each sequence contains the recorded videos from three cameras (C1, C2, and C3). In P1 25 subjects were involved and in P2 there were 29 participants. The resolution of the captured frames are 800×600 at a frame rate of 30

Figure 4.4: Sample images from the ChokePoint [164] dataset.

fps and the cropped faces with size 96×96 from the original video frames are provided.

4.3.1.2 DBN Structure

The DBN is constructed with five time slices. The size of the DBN is determined empirically to offset the complexity of the network and to ensure sufficient dynamics to be encoded as a temporal clue. In each time slice we use the learned structure in Figure 4.2. For parameter learning, the EM algorithm is used and the junction tree algorithm is chosen for inference. With non-optimized Matlab implementation, the training takes about 42 seconds on a PC with 3GHz CPU and 8GB RAM. For testing, the inference takes about 60 seconds.

In our experiments, we use faces from all of the four subsets (P1E, P1L, P2E,

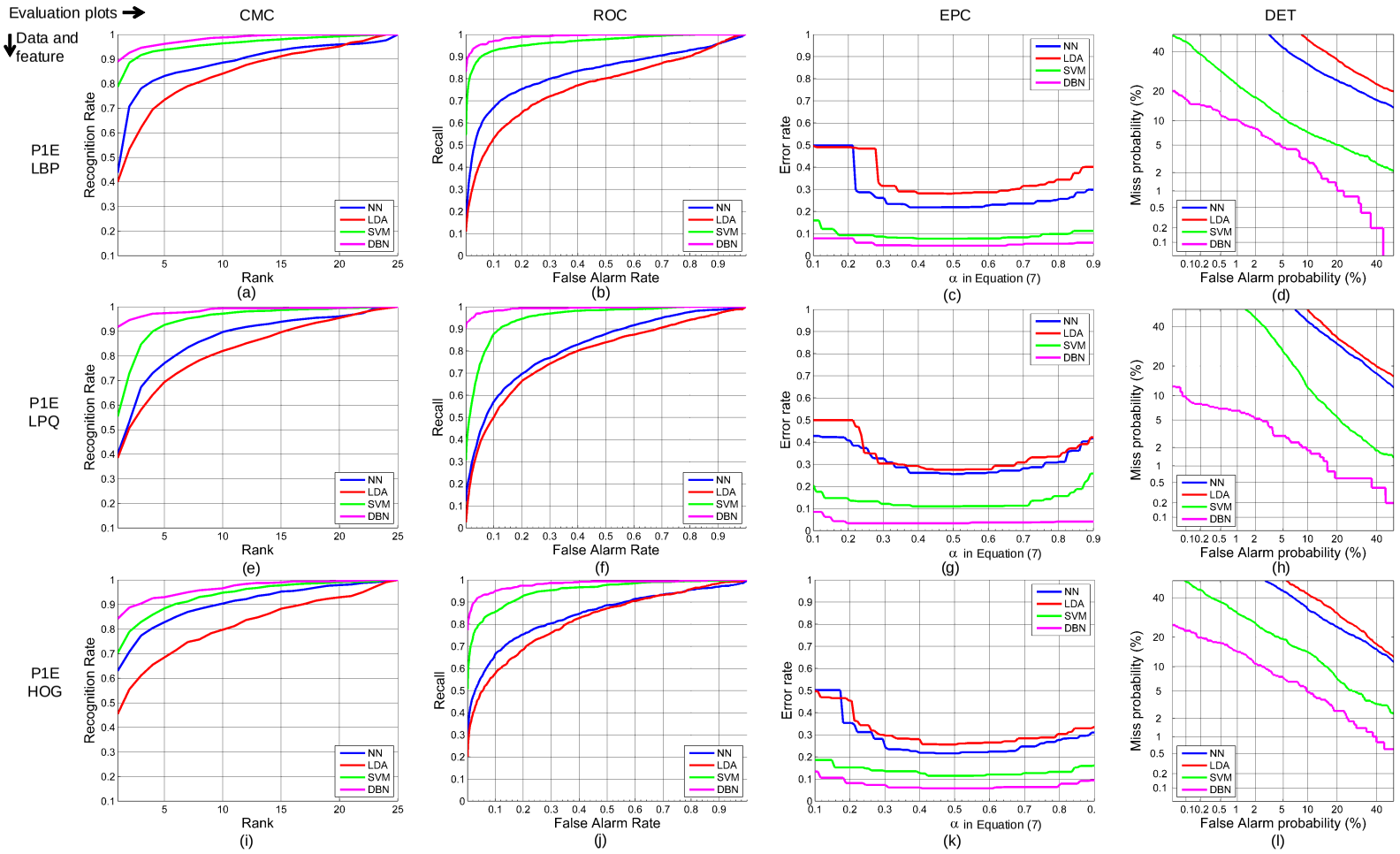

Figure 4.5: The evaluation results for P1E. From left to right: different plots (CMC, ROC, EPC, and DET). From top to bottom: different feature descriptors (LBP, LPQ, and HOG).

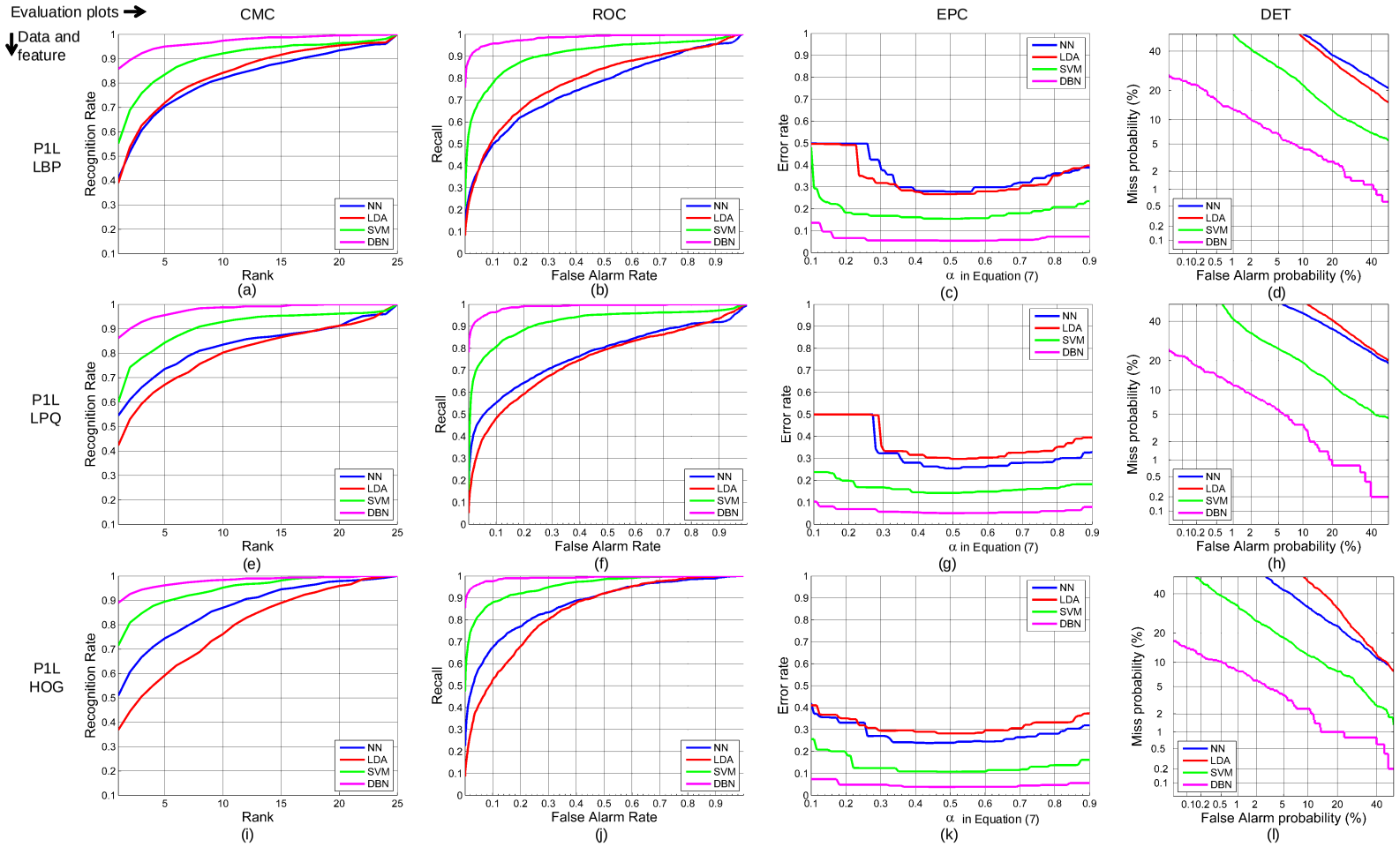

Figure 4.6: The evaluation results for P1L. From left to right: different plots (CMC, ROC, EPC, and DET). From top to bottom: different feature descriptors (LBP, LPQ, and HOG).

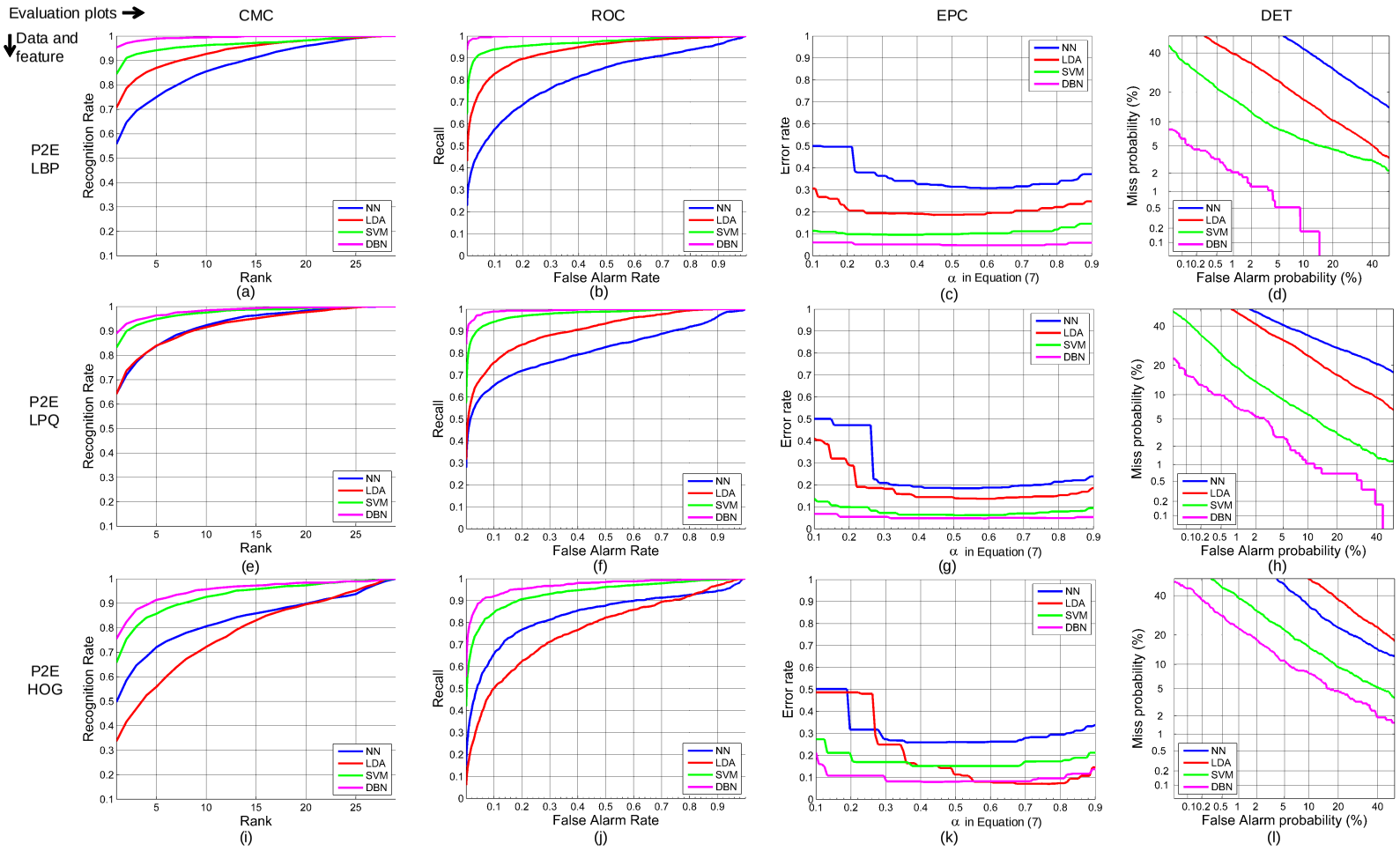

Figure 4.7: The evaluation results for P2E. From left to right: different plots (CMC, ROC, EPC, and DET). From top to bottom: different feature descriptors (LBP, LPQ, and HOG).

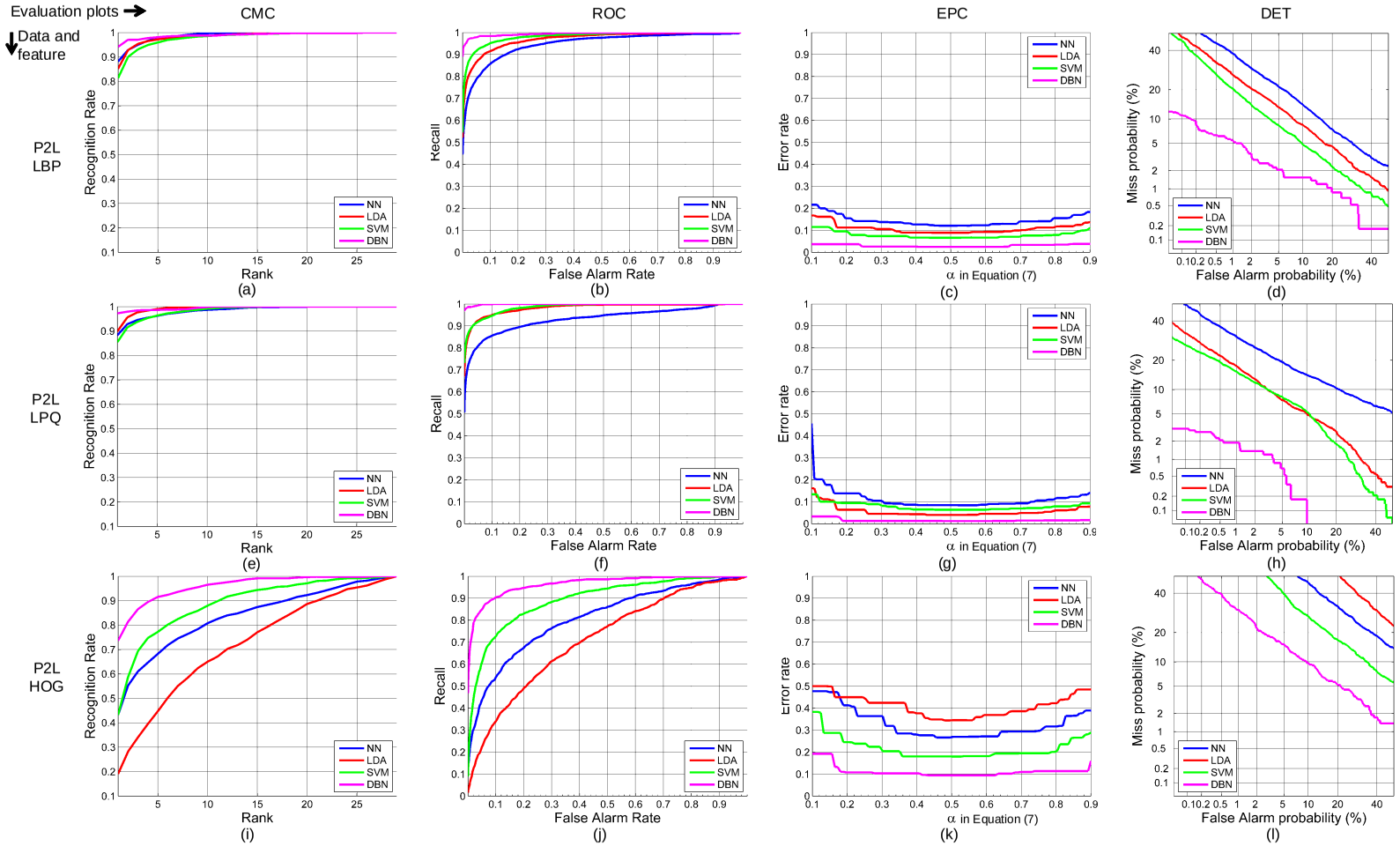

Figure 4.8: The evaluation results for P2L. From left to right: different plots (CMC, ROC, EPC, and DET). From top to bottom: different feature descriptors (LBP, LPQ, and HOG).

and P2L). S1, S2, S3, and S4 are all used except for P2E in which we use only S3 and S4 due to incomplete data in P2E_S1 and P2E_S2. In each sequence 40 instances are used for training or testing. Each instance consists of 15 face images (3 cameras in each time slice, 5 time slices in total). In each run, we perform cross-validation on the same subset (i.e., train on sequence P1E_S1 and test on P1E_S2, P1E_S3, and P1E_S4). The averaged results are reported for each subset separately.

4.3.1.3 Feature Descriptors

For face recognition, various feature descriptors have been proposed and applied. Local binary pattern (LBP) and its derivatives are among the most popular choices [3, 184]. To tackle with blurred face recognition, local phase quantization (LPQ) has been adopted [2]. Recently, inspired by the success in object recognition tasks, histogram of oriented gradients (HOG) has been applied to face recognition [36]. In our experiments, we choose to use these three popular feature descriptors: LPQ, LBP and HOG. For LBP and LPQ operation, the image is divided into the blocks of size 16×16. In LBP, $LBP^{u2}_{8,2}$ is used as suggested in [3]. The parameters for LPQ are set to $M = 7$, $\alpha = 1/7$ and $\rho = 0.9$. For HOG, the image is divided into 9 blocks and the number of orientation bins is set to 15. Note that any feature descriptors can be applied in the proposed framework. The dimensionality of the extracted feature vectors is reduced to 50 using PCA to enforce the efficiency during computation.

4.3.1.4 Classifiers Compared

The DBN is compared with three benchmark classifiers: nearest neighbor (NN), linear discriminant analysis (LDA) and support vector machine (SVM). These classifiers are commonly used in recognition tasks. In the SVM classifier, its linear version is used.

For these classifiers, the same training and testing data are used as for DBN. After multiple testing samples are classified, we adopt the majority voting scheme to decide the final class label (identity) of each subject.

4.3.2 Experimental Results

4.3.2.1 Comparison with Different Classifiers

To compare with other classifiers, the rank-1 recognition rates for the four groups of sequences P1E, P1L, P2E, and P2L are reported in Table 4.2. In most cases, NN and LDA are less able to discriminate the faces from the unconstrained video sequences due to the challenging dataset used. SVM improves the results by seeking for the maximum separation between the features of distinct subjects. Regardless of the choice of the feature descriptor, the proposed DBN classifier, compared to NN, LDA, and SVM, performs best in different sequences as a result of the encoding of the person-specific dynamics in the video and the fusion of multi-camera inputs.

To carefully investigate the performance of the classifiers, for each sequence four evaluation plots are presented: Cumulative Match Characteristic (CMC) curve, Receiver Operating Characteristic (ROC) plot, Expected Performance Curve (EPC) [21], and Detection Error Tradeoff (DET) plot. Figure 4.5 shows the results for the P1E sequence. In Figure 4.5, Figure 4.5(a), Figure 4.5(e), and Figure 4.5(i) present the CMC curves for LBP, LPQ, and HOG, respectively. The recognition rates for the top 25 ranks are reported as the gallery includes 25 subjects in P1. Compared to the other classifiers, the recognition results are more accurate using the proposed DBN at different ranks. The comparison of the results among different feature descriptors confirms the superiority of the proposed method over the other classifiers.

Table 4.2: The rank-1 recognition rates on different testing sequences (in %)

Data→	P1E				P1L				P2E				P2L			
Method→	NN	LDA	SVM	DBN	NN	LDA	SVM	DBN	NN	LDA	SVM	DBN	NN	LDA	SVM	DBN
LBP	43.7	40	78.2	**89.7**	40.9	39	55.2	**85.8**	55.8	70.6	84.5	**95.3**	88.1	85.2	81.3	**94.1**
LPQ	40.1	38.5	55.36	**91.7**	54.5	42.3	60	**86.3**	64.2	64.1	83.2	**89**	88.2	89.9	85.5	**97.2**
HOG	63	45.5	70.3	**84.3**	50.9	36.8	71.5	**89**	49.7	33.7	65.8	**75.5**	43.2	19.1	43.2	**73.6**

Figure 4.5(b), Figure 4.5(f), and Figure 4.5(j) present the ROC plots using the three different feature descriptors. The recognition using LPQ is better than LBP and HOG in terms of ROC performance. The reason is that LPQ is inherently designed as a blur invariant feature descriptor while the captured faces by the surveillance cameras show explicit blurriness due to subject's motion. Note that with different feature descriptors, the performance of the DBN is constantly better than the other classifiers in most cases. This indicates that the performance gain of the proposed method is not entirely feature dependent.

The EPCs in Figure 4.5(c), Figure 4.5(g), and Figure 4.5(k) compare DBN with the other classifiers from the viewpoint of the tradeoff between false alarm and false reject probabilities. The x-axis represents $\alpha \in \mathbb{R}$ where $\alpha \in [0, 1]$ and the y-axis corresponds to the error rate β defined as

$$\beta = \alpha \times \text{FAR} + (1 - \alpha \times \text{FRR}), \tag{4.7}$$

where FAR is the false alarm ratio and FRR represents the false rejection ratio. For all of the three feature descriptors, DBN reports lower error rate compared to the other classifiers. More importantly, the error rate is almost constant for all values of α.

Figure 4.5(d), Figure 4.5(h), and Figure 4.5(l) present the DET plots comparing the decision error rate of DBN vs. the other classifiers. The performance is characterized by the miss and false alarm probabilities. Both x and y axes are scaled non-linearly by their standard normal deviates such that a normal Gaussian distribution will plot as a straight line. The results show that the DBN reports less miss probability with equal false alarm probability compared to NN, LDA, and SVM. It's important to point out that not only the DBN outperforms the other classifiers as shown in the DET plots, but even

Table 4.3: The Rank-1 Recognition rates with different cameras on different testing sequences (in %)

Data→	P1E				P1L				P2E				P2L			
Camera→	C_1	C_2	C_3	ALL	C_1	C_2	C_3	ALL	C_1	C_2	C_3	ALL	C_1	C_2	C_3	ALL
LBP	77.3	81.3	75.3	**89.7**	80	81.3	75.3	**85.8**	74.1	86.2	84.5	**95.3**	55.8	65.2	63.2	**94.1**
LPQ	78.3	84	73.3	**91.7**	75.3	86	60	**86.3**	79.9	84.8	77.9	**89**	65.8	70.4	67.2	**97.2**
HOG	71.3	73.7	73.7	**84.3**	73.7	77	76.7	**89**	67.6	65.9	71	**75.5**	68.4	69.3	69	**73.6**
Average	75.7	79.7	74.1	**88.6**	76.3	81.4	70.7	**87**	73.9	79	77.8	**86.6**	63.3	68.3	66.5	**73.6**

in cases where DBN and SVM seem to have similar performance (e.g., Figure 4.7(e)), the DET plot shows that DBN achieves significantly less miss probability (Figure 4.7(h)).

Figures 4.6 to 4.8 show results for sequences P1E, P1L, P2E, and P2L, respectively. The observations of Figures 4.6 to 4.8 are similar to that of Figure 4.5. Overall, compared to NN, LDA and SVM, DBN is more robust in recognition and less prone to error.

4.3.2.2 Multiple Camera vs. Single Camera

We compare the recognition performance using three cameras against using only one camera in the proposed DBN framework. The DBN structure for a single camera is derived from Figure 4.3 by removing the other two camera nodes. Table 4.3 show the rank-1 recognition rates comparisons using three cameras together (ALL) against using only a single camera on sequences from P1E, P1L, P2E and P2L. As can be seen, regardless of the specific choice of the feature descriptor, the recognition rates with three cameras are higher than using any of a single camera. The reason is that DBN takes into account the relationship of the three cameras through the dependencies, thus the complementary information from each camera is utilized to help improve the recognition performance. Also note that in most cases CAM_2 (C_2) provides higher recognition rates compared to CAM_1 (C_1) and CAM_3 (C_3) due to the near frontal faces it captured with relatively higher video quality. Although the performance using HOG is, in general, inferior to LBP and LPQ, for different camera, HOG gives similar recognition rates. This is due to the tolerance of the HOG descriptor for small pose variations.

4.4 Conclusions

We proposed a multi-camera face recognition system using DBN and this framework is suitable for applications such as surveillance monitoring in camera networks. In the proposed method, videos from multiple cameras are effectively utilized to provide the complementary information for robust recognition results. In addition, the temporal information among different frames are encoded by DBN to establish the person-specific dynamics to help improve the recognition performance. Experiments on a surveillance video dataset with a three-camera setup show that the proposed method performs better than the other benchmark classifiers using different feature descriptors by different evaluation criteria. Regarding the generality of our method, the feature nodes in the DBN can be replaced with any choice of informative feature descriptors and the proposed framework can be extended to the camera systems with different number of cameras.

Chapter 5

Person Re-Identification with Reference Descriptor

5.1 Introduction

Imaging sensors are being deployed widely for many real-world applications such as video surveillance, access control, etc. Particularly in camera networks, there has been an increasing interest and considerable progress has been made for person re-identification recently [151, 50, 39, 20, 182]. Person re-identification is a recognition task that aims to match individuals across non-overlapping cameras at different time and location. Accurate person re-identification enables locating a target subject's whereabouts in video-monitored surroundings. For people tracking in a multi-camera system [69], re-identification results can be used for tracklet association.

Matching people in different cameras is intrinsically difficult due to the imaging disparity among different cameras. The following problems contribute to the complications of person re-identification in a camera network:

- *Low resolution.* Most of the surveillance cameras are not able to capture high-resolution images due to the low resolution of inexpensive cameras and large distance between camera and human subjects.

- *Arbitrary poses.* Since a subject is captured by surveillance cameras with non-overlapping field-of-views, the poses of a subject in different camera are usually quite different.

- *Changing illumination.* The images are captured at different time and/or location. As a consequence, the appearance of a person may change dramatically due to illumination changes.

- *Occlusion.* A subject may carry accessories such as a backpack, briefcase, etc., which may occlude distinctive features of the subject in a certain view.

Figure 5.1 shows some image pairs of the same and different people in two cameras. Due to large variations in pose, illumination and background, the appearance of the same subject may look very different in different cameras while different people may highly resemble in appearance. The significant view and appearance changes across non-overlapping cameras make person re-identification inherently difficult.

The gallery for re-identification usually contains images of known subjects in one camera view and the probes are subjects from another camera view. In order to recognize a given probe from a large gallery, the basic idea is to first extract a robust feature representation for both probe and gallery images, and then perform matching using this representation. This kind of approach is called *appearance-based* and it makes use of visual cues only.

Appearance-based methods can be categorized into two groups. The goal of the

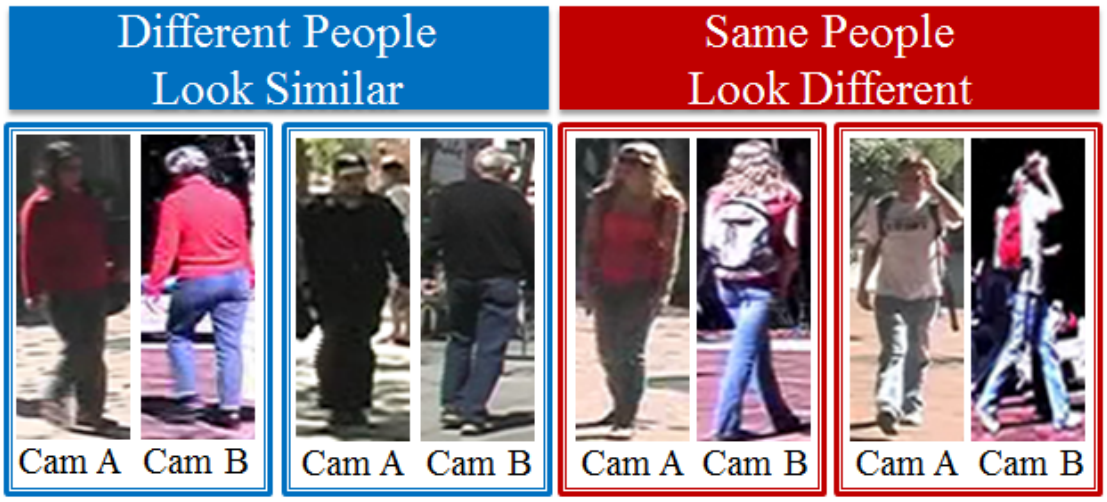

Figure 5.1: In non-overlapping camera views, different people may look very similar (left) while same people's appearance may change dramatically due to variations in pose, illumination (right). Samples are from two cameras (Cam A and Cam B) in the VIPeR dataset [52].

methods in the first group is to extract feature representations that have low intra-class variation for the same subject and high inter-class variation among different subjects (e.g., [52, 44, 157]). However, due to the significant appearance change across different cameras, the intra-class variation is often larger than the inter-class variation. As a result, accurate matching is very difficult.

For the second group of methods, the goal is to learn the optimal distance metric for the image pairs from two different cameras (e.g., [66, 190, 81]). These metric learning approaches learn a transformation for the original feature representation such that the intra-class distances are minimized while the inter-class distances are maximized. The drawback of the metric learning based methods is that the learned model tends to overfit the training data. Also, some popular approaches (e.g., [162, 55, 35]) are computationally expensive due to complex optimization involved.

In this chapter, instead of designing a complex feature representation or learning a specialized distance metric as it has been done in the previous methods, we present

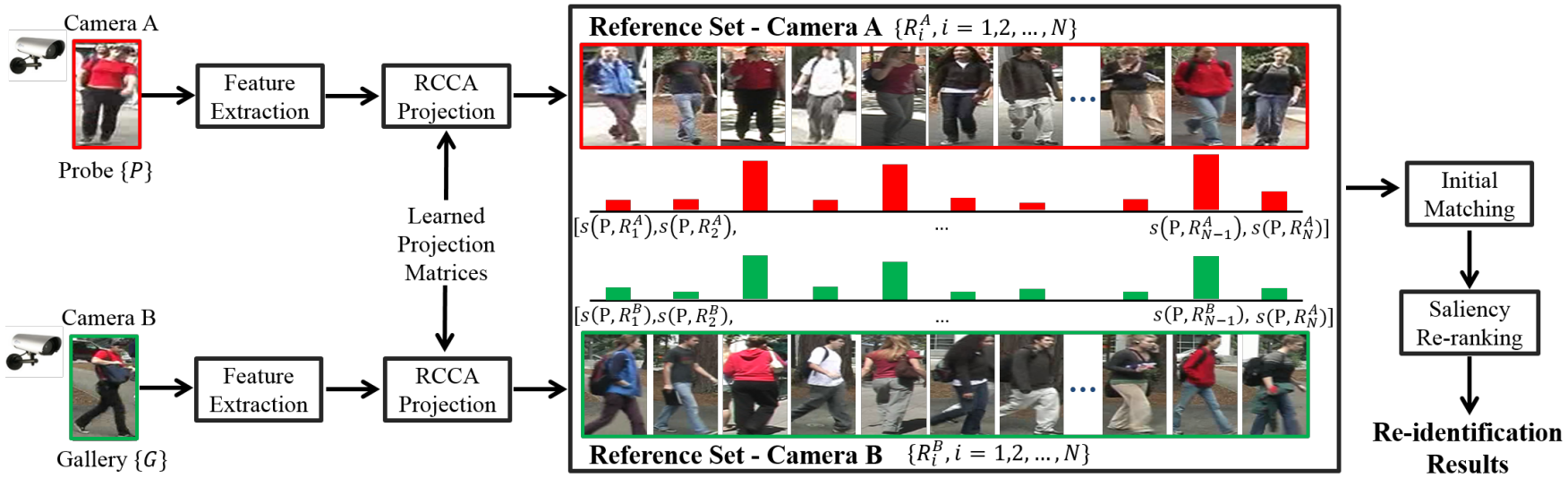

Figure 5.2: Framework of the proposed reference-based re-identification. The appearance features are first extracted from probe and gallery images and are then projected into RCCA subspace with learned projection matrices. A reference descriptor (RD) for a probe or gallery instance is generated by computing and concatenating its similarity scores with respect to a reference set. After the RDs for both probe and gallery are generated, initial matching is performed using RDs. A saliency-based re-ranking scheme is included to further improves the re-identification accuracy.

a new framework for single-shot person re-identification in which the matching is performed using *reference descriptors* (RDs). Figure 5.2 illustrates the framework of the proposed reference-based method. To match a probe and a gallery instance, appearance features are first extracted. Using learned projection matrices, the probe and gallery features are projected into a lower dimensional subspace. We use Regularized Canonical Correlation Analysis (RCCA) to learn the projection matrices since RCCA is able to maximize the correlation between the data from different views. After feature projection, the RDs of the probe and gallery are generated using a *reference set*. The reference set is a set of images of the subjects from different camera views and the identities in the reference set do not overlap with probe or gallery subjects. A RD of a probe or a gallery instance is formed by concatenating the similarity scores between this probe or gallery to the reference set in the RCCA feature space. Thus, the dimension of a RD is determined by the size of the reference set and is irrelevant to the size of the image features. The matching between probe and gallery is performed by computing the similarity between their RDs. In this way, probe and gallery from different views are indirectly compared using a reference set, instead of being matched directly. To improve the initial matching results, a saliency-based re-ranking stage is added to obtain the final re-identification results.

Pattern matching by using a reference set has been explored in different fields. Gyaourova *et al.* [57] generated fixed-length codes for indexing biometric databases. The index codes were constructed by computing match scores between a biometric image and a fixed set of images. Duin *et al.* [40] discussed the dissimilarity space to convert the structural representation of data to a dissimilarity representation using a representation set and some suggestions for prototype selection were provided. Guo *et al.* [56] proposed a prototype embedding of visual appearance by using a representation set of model

prototypes for vehicle matching. Recently, Chen *et al.* [27] developed a reference-based approach for tracking people across non-overlapping cameras using a reference-based appearance model. Li *et al.* [92] proposed a reference-based scheme for scene image classification.

5.1.1 Contributions of This Chapter

As compared to the previous work, the major contributions of this chapter are two-fold. *First*, we tackle the re-identification problem using a reference-based scheme in conjunction with subspace learning. Our framework avoids direct matching of image pairs with significant appearance variation and achieves superior performance compared to the state-of-the-art methods as validated by the experiments. *Second*, we use different methods to pursue optimality for reference set selection and the experiments show that the size of reference set can be reduced without a significant loss of accuracy. Further, the proposed reference-based re-identification framework is compatible with any feature descriptor and can be extended to other applications.

The rest of this chapter is organized as follows. Details of the proposed method for person re-identification are presented in Section 5.2. Section 5.3 provides the experimental results and finally Section 5.4 concludes this chapter and states the future work.

5.2 Person Re-Identification in Reference Space

The proposed method involves an offline process and an on-line re-identification process. In the offline process, the RCCA projection matrices are learned and the RDs of the gallery are generated. During online re-identification process, the RD of a probe is generated and is compared with the RDs of the gallery to obtain the initial matching

result. Re-ranking is then performed to improve the initial results based on image saliency. The details are explained as follows.

5.2.1 Offline Process

5.2.1.1 CCA Subspace Learning

Canonical Correlation Analysis (CCA) is a multivariate statistical analysis technique which was first introduced in [67]. It aims to explore the relationship between two sets of random variables from the different observations on the same data (*e.g.*, images of subjects from different views). CCA finds projections such that the correlation between these two sets of random variables is maximized after projection.

Mathematically, given two sets of data observations, $D^A = \{d_i^A \in \mathbb{R}^m, i = 1, 2, ..., N\}$ and $D^B = \{d_i^B \in \mathbb{R}^n, i = 1, 2, ..., N\}$, CCA aims at obtaining two sets of basis vectors $W_A \in \mathbb{R}^m$ and $W_B \in \mathbb{R}^n$ such that the correlation coefficient ρ of $W_A^T D^A$ and $W_B^T D^B$ is maximized. The objective function to be maximized is

$$\begin{aligned} \rho &= \frac{Cov(W_A^T D^A, W_B^T D^B)}{\sqrt{Var(W_A^T D^A)}\sqrt{Var(W_B^T D^B)}} \\ &= \frac{W_A^T C_{AB} W_B}{\sqrt{W_A^T C_{AA} W_A W_B^T C_{BB} W_B}} \end{aligned} \tag{5.1}$$

where C_{AA} is the covariance matrix of D^A, C_{BB} is the covariance matrix of D^B, and C_{AB} is the cross-covariance matrix between D^A and D^B.

Equivalently, the CCA can be formulated as a constrained optimization problem by

$$\underset{W_A, W_B}{\operatorname{argmax}} W_A^T C_{AB} W_B \tag{5.2}$$

subject to $W_A^T C_{AA} W_A = 1$ and $W_B^T C_{BB} W_B = 1$.

Using Lagrange multiplier, the solution of (5.2) is equivalent to solving the following generalized eigenvalue problems

$$\begin{aligned} C_{AB} W_B &= \lambda C_{AA} W_A \\ C_{BA} W_A &= \lambda C_{BB} W_B \end{aligned} \tag{5.3}$$

where $C_{BA} = C_{AB}^T$. CCA is performed in an unsupervised manner and both correlation maximization and dimensionality reduction can be achieved simultaneously by choosing the number of basis vectors to use.

Often in practice, the feature dimension of the data is significantly larger than the number of data samples. In this case the covariance matrices C_{AA} and C_{BB} may be singular and their inverse would be ill-conditioned. Regularized CCA (RCCA) has been proposed to solve this problem and it prevents overfitting [88]. In the solution of RCCA, the generalized eigenvalue problem becomes

$$\begin{aligned} C_{AB} W_B &= \lambda (C_{AA} + \lambda_1 I_A) W_A \\ C_{BA} W_A &= \lambda (C_{BB} + \lambda_2 I_B) W_B \end{aligned} \tag{5.4}$$

where λ_1 and λ_2 are the two non-negative regularization parameters. I_A and I_B are two identity matrices. Usually λ_1 and λ_2 are determined by cross-validation.

5.2.1.2 Gallery Data in Reference Space

The reference set contains images $\{I_i^A, i = 1, 2, ..., N\}$ and $\{I_i^B, i = 1, 2, ..., N\}$ of N subjects from two different cameras A and B. The features such as color histograms and texture descriptors from each image are extracted and two feature sets $\{F_i^A, i = 1, 2, ..., N\}$ and $\{F_i^B, i = 1, 2, ..., N\}$ are obtained. Since the features are from

images in different views, we first learn a RCCA subspace in which the correlations between the projected feature sets $\{W_A^T F_i^A, i = 1, 2, ..., N\}$ and $\{W_B^T F_i^B, i = 1, 2, ..., N\}$ are maximized. The RCCA projection matrices W_A and W_B are learned as in (5.4). By projecting the original features into the RCCA subspace, we obtain the projected features of the reference set denoted as $\{f_i^A, i = 1, 2, ..., N\}$ and $\{f_i^B, i = 1, 2, ..., N\}$ with reduced dimensionality and enhanced correlation.

Suppose we have a gallery of M subjects from camera A, the features of the gallery subjects are first extracted and then projected into the RCCA subspace using the learned projection matrix W_A. The RCCA feature for the j^{th} subject in the gallery set is denoted by f_j^g. From f_j^g, its RD R_j^g, as a new representation, is generated by

$$R_j^g = [s(f_j^g, f_1^A), s(f_j^g, f_2^A), \ldots, s(f_j^g, f_N^A)]^T \tag{5.5}$$

where $s(a, b)$ denotes the similarity between the features a and b. We use the cosine similarity to compute $s(a, b)$. In this process, the representation of the gallery subject is transformed to a descriptor of length N regardless of the original feature dimension and each element in R_j^g indicates the similarity between this gallery subject and a reference subject. The projected features of the reference set from camera A $\{f_i^A, i = 1, 2, ..., N\}$ are similar to basis functions and in the reference space they jointly describe the appearance of a gallery subject in terms of its similarity to individuals in the reference set. Figure 5.2 shows the basic idea of how the RDs are generated.

The rationale for first projecting the features into the RCCA subspace is to better couple the features $\{f_i^A, i = 1, 2, ..., N\}$ and $\{f_i^B, i = 1, 2, ..., N\}$. In the re-identification, a probe image is described using $\{f_i^B, i = 1, 2, ..., N\}$. Since $\{f_i^A, i = 1, 2, ..., N\}$ and $\{f_i^B, i = 1, 2, ..., N\}$ are maximally correlated after RCCA projection,

the matching between the probe and the gallery becomes meaningful and reliable.

5.2.2 Online Re-Identification

5.2.2.1 Initial Matching

Suppose the probe is from camera B and the detection of a subject (I_p) is given, the appearance features F^p are first extracted. The projected feature f^p of the probe in the RCCA subspace is given by

$$f^p = W_B^T F^p \tag{5.6}$$

The RD of the probe, R^p, is computed in a similar manner as in (5.5) using the projected features of the reference set from camera B $\{f_i^B, i = 1, 2, ..., N\}$ by

$$R^p = [s(f^p, f_1^B), s(f^p, f_2^B), \ldots, s(f^p, f_N^B)]^T \tag{5.7}$$

where f_i^B is the projected features in the RCCA subspace of the reference subject i in camera B.

The identity of the subject is determined by the similarity $sim(R^p, R_i^g)$ between the probe R^p and each gallery R_i^g and then the top match R_k^g is found in the gallery such that

$$k = \underset{i}{\operatorname{argmax}}\, sim(R^p, R_i^g) \tag{5.8}$$

To compute similarity, we use the modified cosine similarity [101] defined as

$$\text{sim}(R^p, R_i^g) = \frac{\left|(R^p)^T \cdot R_i^g\right|}{\|R^p\| \, \|R_i^g\| \, (\|R^p - R_i^g\|_p + \epsilon)} \tag{5.9}$$

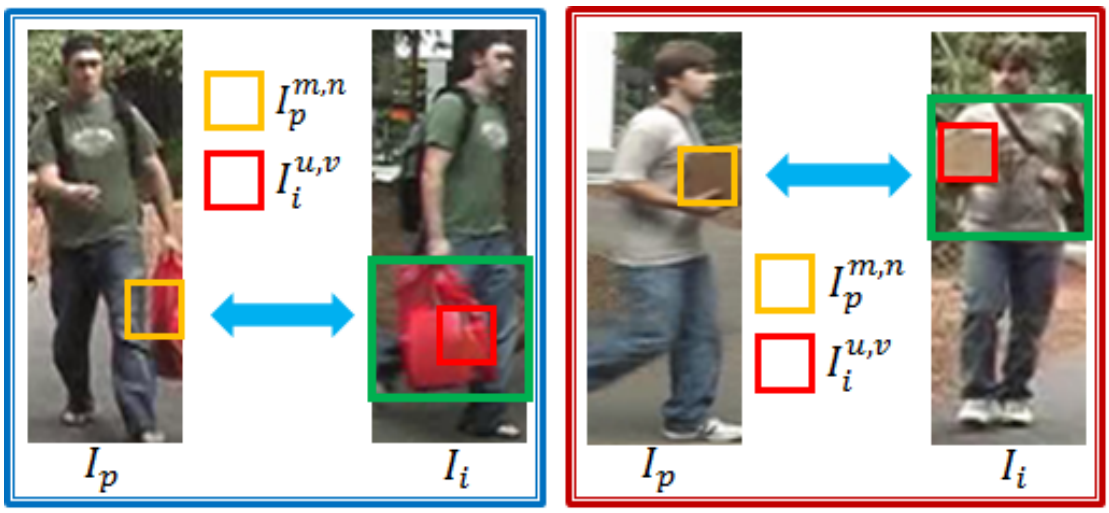

Figure 5.3: Samples of saliency detection in two camera views. To estimate the saliency of a patch $I_p^{m,n}$ (in yellow bounding box) in image I_p in one camera view, a constraint space (in green bounding box) is searched in each image I_i in the reference set in the other camera view. The patch $I_i^{u,v}$ (in red bounding box) is found out as the most similar patch to $I_p^{m,n}$, which will be used to calculate the patch saliency as in (5.10). Best viewed in color.

where $\|\cdot\|_p$ is the l_p norm and ϵ is a small positive number to prevent division by zero. The reason to apply the modified cosine similarity is that the standard cosine similarity does not take into consideration the actual distance between two vectors, while the modified cosine similarity is able to address both the distance measure and angular measure and has improved performance in recognition tasks [101].

5.2.2.2 Saliency Detection

To improve the re-identification accuracy, we opt to high-level image information to re-rank the initially returned results. Specifically, we use image saliency [187, 186] to improve the rank of the correct match. Image saliency, such as carrying item, is a discriminative visual feature to match subjects across different views. Figure 5.3 shows two examples of saliency correspondence across different cameras.

Given the reference set $\mathbf{I} = \{I_i, i = 1, 2, ..., N\}$ in one camera view, to compute

the saliency of an image I_p in another view, image patches are first densely sampled. For each patch $I_p^{m,n}$ in I_p where m and n denote the row and column location of this patch, a constrained search for similar patches in each image in $\mathbf{I}$ is performed in the search space $D(I_p^{m,n}, I_i) = \{I_i^{x,y} | x = m - l, ..., m + l\}$, where l is a small integer that defines the half width of the search space. In other words, the search space for the patch $I_p^{m,n}$ is a strip in I_i located between row $m - l$ and $m + l$. This search space tolerates saliency shift in horizontal direction due to change in camera views and misalignment in vertical direction.

For each image I_i in $\mathbf{I}$, the most similar patch $I_i^{u,v}$ is found from the search space $D(I_p^{m,n}, I_i)$. The distance $d_i(I_p^{m,n}, I_i^{u,v})$ is calculated by Euclidean distance between feature vectors from the two patches $I_p^{m,n}$ and $I_i^{u,v}$. The distances $\{d_i(I_p^{m,n}, I_i^{u,v}) | i = 1, 2, ...N\}$ are then sorted and the saliency score for patch $I_p^{m,n}$ is defined as

$$sal(I_p^{m,n}) = 1 - e^{(-\frac{d_{i_k}(I_p^{m,n}, I_i^{u,v})}{\sigma_1^2})} \tag{5.10}$$

where $d_{i_k}(I_p^{m,n}, I_{i_k}^{u,v})$ is the Euclidean distance of the k-th nearest neighbor (kNN) of $I_p^{m,n}$ from the search space of I_{i_k} and σ_1 controls the bandwidth of Gaussian function. k is set to $\frac{N}{2}$ in the experiments and only the k-th nearest neighbor is involved in saliency computation. In this way, the saliency scores for each patch in the probe and gallery images are calculated. The saliency of a patch $I_p^{m,n}$ is computed from this kNN perspective such that the uniqueness of a patch is approximated by its distance to the samples in the reference set. The interpretation is that the more distinct a patch $I_p^{m,n}$ is, the larger is its distance to the patches in the search space of images in $\mathbf{I}$, thus, the saliency score $sal(I_p^{m,n})$ will be high. In this way, the saliency is calculated without supervision.

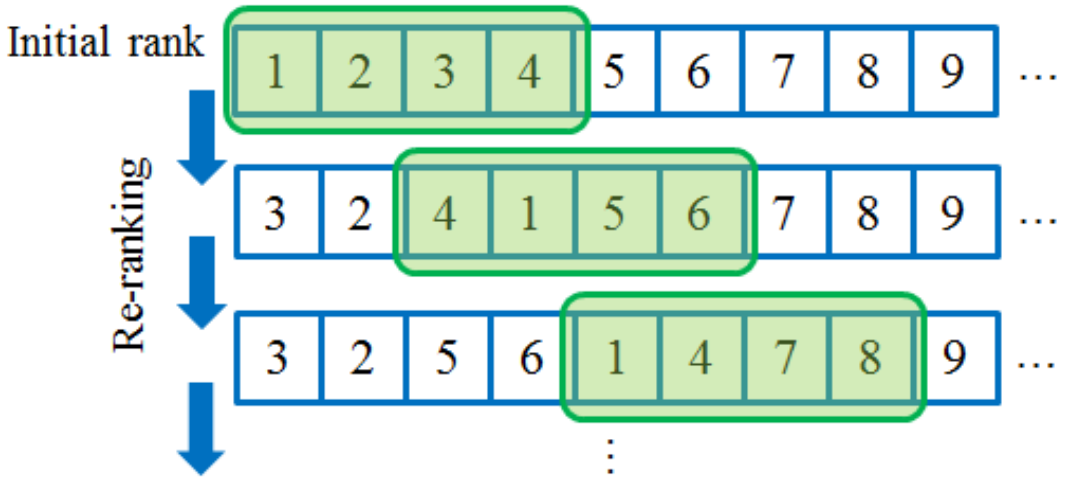

Figure 5.4: Illustration of the re-ranking process. The initial returned ranked list is re-ranked based on saliency similarity of probe and gallery. In this example a local sliding window of size $\alpha = 4$ with a step size of $\beta = 2$ is shown.

5.2.2.3 Re-ranking

Once the saliency of the probe and gallery is detected, the re-ranking of the initial re-identification results is based on the saliency similarity between the probe I_p from camera B and a returned gallery match I_t at rank t from camera A, which is defined as

$$sim_{sal}(I_p, I_t) = \sum_{m,n} sal(I_p^{m,n}) \times sal(I_t^{u,v}) \times e^{-\frac{d(I_p^{m,n}, I_t^{u,v})}{\sigma_2^2}} \tag{5.11}$$

where $I_t^{u,v}$ is the nearest neighbor of $I_p^{m,n}$ found in the search space and σ_2 is a Gaussian parameter. $d(I_p^{m,n}, I_t^{u,v})$ is the Euclidean distance between the features of $I_p^{m,n}$ and $I_t^{u,v}$.

Given a probe image, the reference-based method returns the matching results in descending order based on the similarity between the probe RD and gallery RDs. Based on the saliency similarity sim_{sal} between the probe image and the returned matching candidate, the initial ranked list is re-ranked using a local sliding windows of size α and a step size of β, and the candidate with a higher saliency similarity to the

probe is moved forward in the local window. The re-ranking process is illustrated in Figure 5.4.

5.3 Experimental Results

5.3.1 Datasets

• **VIPeR Dataset.** The VIPeR dataset is one of the most popular benchmark datasets for person re-identification [52]. It contains image pairs of 632 pedestrians. The images are taken by two non-overlapping cameras with a significant view change. For most of the subjects, the view change is more than 90 degrees. In addition, the illumination may also change dramatically. Other aspects such as cluttered background and occlusions further make this dataset more challenging. It is considered as the most challenging dataset currently available for pedestrian re-identification. For each person, a single image is available from each camera view. All of the images in the VIPeR dataset are normalized to 128×48. Some sample images are shown in Figure 5.5(a).

• **CUHK Campus Dataset.** The CUHK Campus dataset contains images of 971 subjects from two non-overlapping camera views [160]. One camera captures the frontal or back view of the subjects and the other camera captures profile views. Each person in each view has two images. The image quality of CUHK Campus dataset is higher compared to the VIPeR dataset. All of the images in the CUHK campus dataset are resized to 128×48 in our experiments. Some sample images are shown in Figure 5.5(b).

5.3.2 Feature Extraction and Parameters

Both color and texture features are extracted as in [66]. Specifically, the HSV and Lab color features are used to describe the color appearance of a subject. For

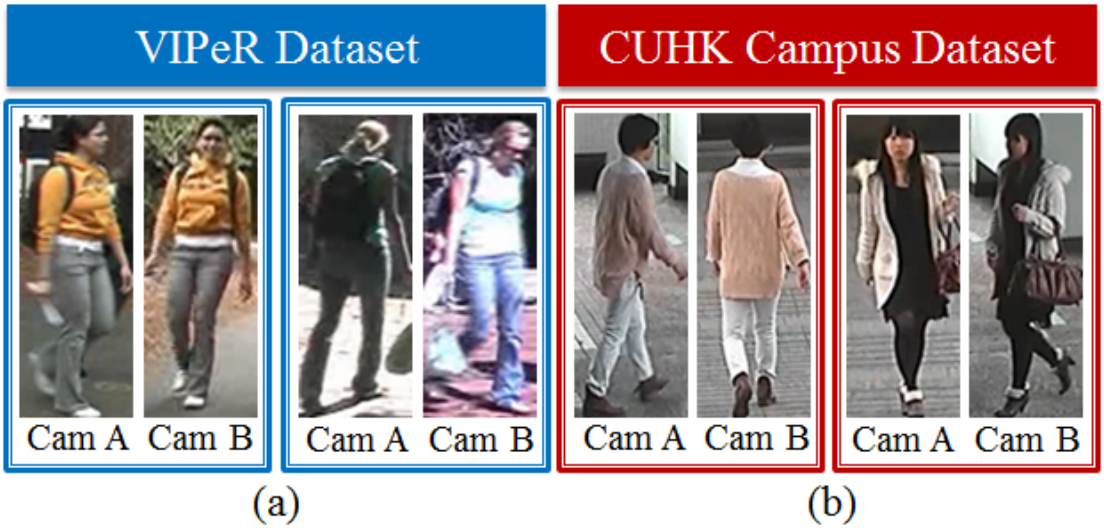

Figure 5.5: Sample images from (a) VIPeR dataset [52] and (b) CUHK Campus dataset [160].

the texture feature we use Local Binary Patterns (LBP) [126]. The image is divided into blocks of size 8×16. The blocks are overlapped by 50% in both horizontal and vertical directions. Thus, the total number of blocks for one image of size 128×48 is $31 \times 5 = 155$. For each block, the quantized mean values of the HSV and Lab color channels are computed and the 8-bit LBP histogram is extracted. The final feature representation for one block is the concatenation of the means of the color channels and the LBP histogram with dimension $3 + 3 + 256 = 262$.

In the RCCA projection the first 50 eigenvectors in the projection matrices W_A and W_B are used (*i.e.*, the RCCA reduces the dimensions of the original features to 50). λ_1 and λ_2 are set to $10^{-1.6}$. For re-ranking a local sliding window of size $\alpha = 4$ is used with a step size of $\beta = 2$. The parameters in our experiments are chosen by cross-validation. For the discovery of saliency, we use the same feature and parameter settings as in [186].

5.3.3 Evaluation Protocol

In our experiments we follow the experimental protocols in the previous work (*e.g.*, [44, 81, 186]). We randomly partition each dataset into two sets of equal size. Half of the data are used for training and constructing the reference set and the other half are used for testing. In the testing, the images from one camera are used as gallery and images from the other camera are used as probes. The recognition rates at major top ranks and the Cumulative Matching Characteristic (CMC) curves are reported. The CMC curve represents the expectation of finding the correct match in the top r matches. In other words, a rank-r recognition rate shows the percentage of the probes that are correctly recognized from the top r matches in the gallery. The experiments are performed 10 times and the average results are reported.

5.3.4 Re-Identification Performance

• **VIPeR Dataset.** The recognition performance on the VIPeR dataset is shown in CMC plots in Figure 5.6 from rank 1 to 10. The results from intermediate steps in the proposed method are also shown. When RCCA is used followed by direct matching only, the rank-1 recognition rate is 24.68%. When the matching is performed in the reference space, the rank-1 recognition rate rises to 31.14%, with an improvement of 26%. The re-ranking step further improves the rank-1 recognition rate to 33.29%. The gain by re-ranking is 7% compared to the results before re-ranking. At each rank in Figure 5.6, the reference-based matching with re-ranking achieves the highest recognition rate.

• **CUHK Campus Dataset.** Figure 5.7 shows the recognition performance as CMC plots for the CUHK Campus dataset. Compared to the rank-1 recognition rates of 23.52% using RCCA only and 29.98% in reference space after RCCA projection, the

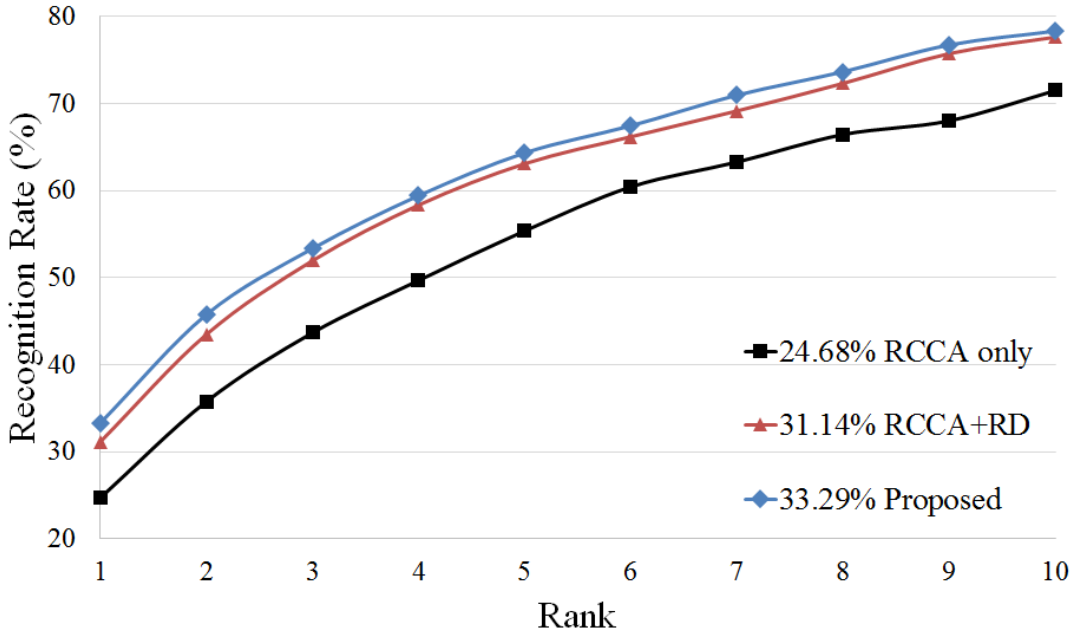

Figure 5.6: CMC curves for the VIPeR dataset. Results using the proposed method, the method using RCCA only, and the method using RCCA and RD without re-ranking are shown.

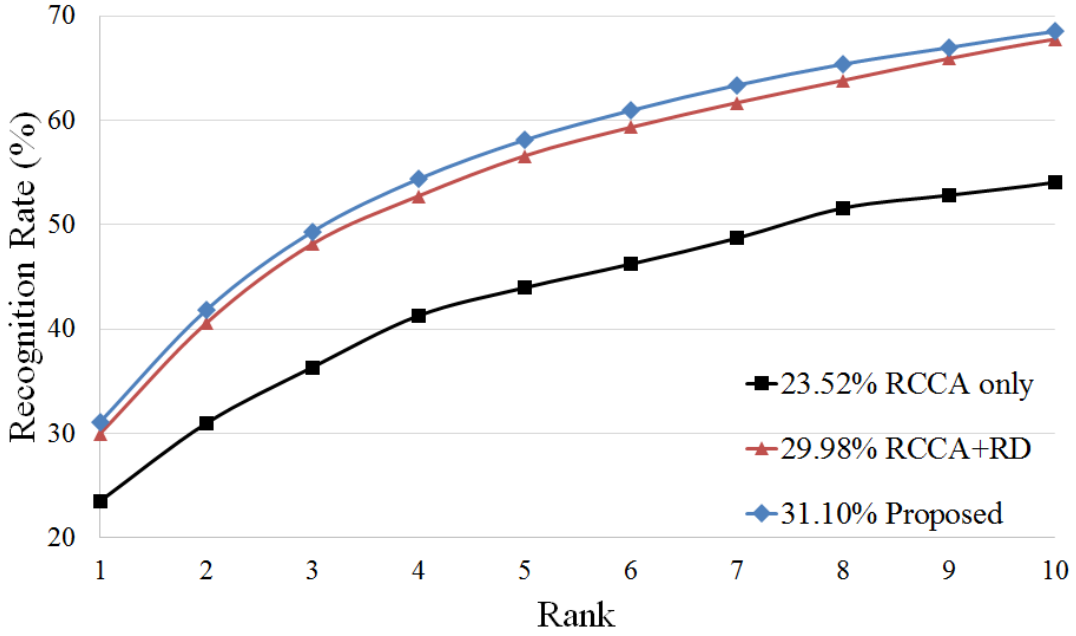

Figure 5.7: CMC curves for the CUHK Campus dataset. Results using the proposed method, the method using RCCA only, and the method using RCCA and RD without re-ranking are shown.

reference-based matching with saliency-based re-ranking as proposed, achieves a rank-1 recognition rate of 31.10%. Figures 5.6 and 5.7 indicate that each step in the proposed method contribute to the recognition performance.

5.3.5 Comparison to Current Methods

• **VIPeR Dataset.** The VIPeR dataset is the most popular benchmark dataset for person re-identification, hence a lot of recent progress in re-identification reports results on this dataset. We compare our approach with recent methods including saliency matching (SalMatch) [186], relaxed pairwise learned metric (RPLM) [66], regularized smoothing KISS metric learning (RS-KISS) [146], custom pictorial structures (CPS) [28], biologically inspired features and covariance descriptors (BiCov) [112], KISS metric (KISSME) [81], large margin nearest neighbor with rejection (LMNN-R) [37], symmetry-driven accumulation of local features (SDALF) [44], manifold ranking (MRank) [109], pairwise constrained component analysis (PCCA) [119], descriptive and discriminative classification (DDC) [65], large margin nearest neighbor (LMNN) [162], attributed-based relative distance comparison (aPRDC) [103], relative distance comparison (PRDC) [190], information-theoretic metric learning (ITML) [35], support vector ranking (RankSVM) [133], and ensemble of localized features (ELF) [53].

The recognition results of the proposed method at rank 1, 10, 20, 50 and 100 are compared to other methods in Table 5.1. As compared to the second best results (SalMatch [186]) with a rank 1 recognition rate of 30.16%, our method achieves a rank 1 recognition rate of 33.29% ,which shows an improvement of over 10%. At other ranks, our method outperforms all of the other listed approaches. Figure 5.8 shows the CMC curves at top 10 ranks for our method and the other six top performers in Table 5.1.

Table 5.1: The comparison of the top ranked recognition rates (in %) on the VIPeR dataset.

Rank→	$r = 1$	10	20	50	100
Proposed	**33.29**	**78.35**	**88.48**	**97.53**	**99.36**
SalMatch [186]	30.16	65.54	79.15	91.49	98.10
RPLM [66]	27.34	69.02	82.69	94.56	98.54
RS-KISS [146]	24.50	66.60	81.70	93.50	98.00
CPS [28]	21.84	57.21	71.00	87.00	91.77
BiCov [112]	20.66	56.18	68.00	81.56	88.66
KISSME [81]	20.03	62.39	77.46	92.81	98.19
LMNN-R [37]	20.00	66.00	79.00	92.50	95.18
SDALF [44]	19.87	49.37	65.73	84.84	90.43
MRank [109]	19.34	55.51	70.44	87.69	96.90
PCCA [119]	19.27	64.91	80.28	95.00	97.01
DDC [65]	19.00	52.00	65.00	80.00	91.00
LMNN [162]	17.41	53.86	67.88	88.13	96.23
aPRDC [103]	16.14	50.98	65.95	88.00	93.00
PRDC [190]	15.66	53.86	70.09	87.79	92.84
ITML [35]	15.54	53.13	69.05	88.54	96.93
RankSVM [133]	14.00	51.00	67.00	85.00	94.00
ELF [53]	12.00	43.00	60.00	81.00	93.00

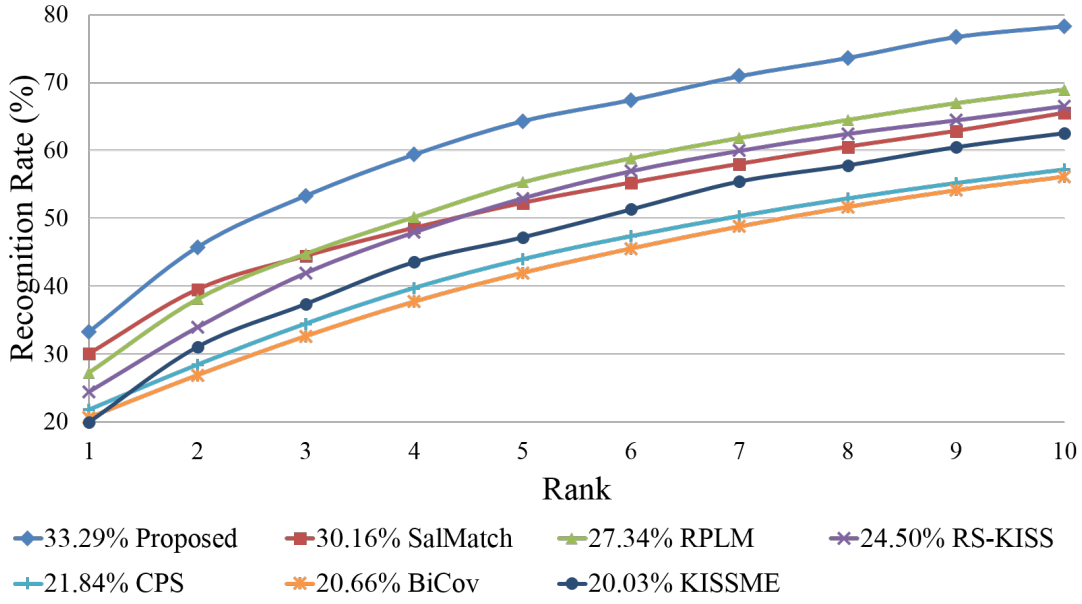

Figure 5.8: The comparison of the CMC curves on the VIPeR dataset for the proposed method and the other methods.

As compared to the improvement by our method at rank 1, at latter ranks our method outperforms the others by a wider margin.

In Table 5.2 we evaluate the performance of our method with reduced training/reference set size. All the data from the VIPeR dataset are used. As the size of the reference set decreases, the number of subjects in the gallery and probe data in-

Table 5.2: The comparison of the recognition rates (in %) with different training (reference) set sizes.

Training size→	N=200			N=100		
Rank→	$r = 1$	10	20	$r = 1$	10	20
PRDC [190]	12.64	44.28	59.95	9.12	34.4	48.55
RPLM [66]	19.51	56.44	71.09	10.88	37.69	51.64
Proposed	**25.93**	**62.73**	**77.31**	**17.86**	**49.44**	**64.29**

Table 5.3: The comparison of the top ranked recognition rates (in %) on the CUHK Campus dataset.

Rank→	$r = 1$	10	20	50	100
Proposed	**31.10**	**68.55**	**79.18**	**90.38**	**95.86**
SalMatch [186]	28.45	55.68	67.95	83.53	92.10
ITML [35]	15.98	45.60	59.81	76.61	88.32
LMNN [162]	13.45	42.25	54.11	73.29	86.65
SDALF [44]	9.90	30.33	41.03	55.99	67.39
L2-norm [186]	9.84	26.42	33.13	46.98	63.48
L1-norm [186]	10.33	26.34	33.52	45.62	61.95

creases, which makes the re-identification more difficult. We compare our results with the reported results by RPLM [66] and PRDC [190]. From the comparison in Table 5.2, it can be observed that with a smaller reference set, the proposed method performs significantly better, with rank 1 recognition rates of 25.93% and 17.86%, when reduced reference sets of size 200 and 100 are used, respectively.

• **CUHK Campus Dataset.** For the CUHK Campus dataset, we compare the proposed approach with the following methods: SalMatch [186], SDALF [44], LMNN [162], ITML [35] as well as baseline methods using L1 norm and L2 norm as reported in [186]. Table 5.3 reports the recognition rates at different ranks. As compared to the most recent method SalMatch [186], our method has over 9% improvement, achieving a rank 1 recognition rate of 31.10%. Figure 5.9 displays the CMC curves of our method and the six competing methods. Similar to the observation in Figure 5.8, at higher ranks our method has a larger improvement in recognition rate over the other methods.

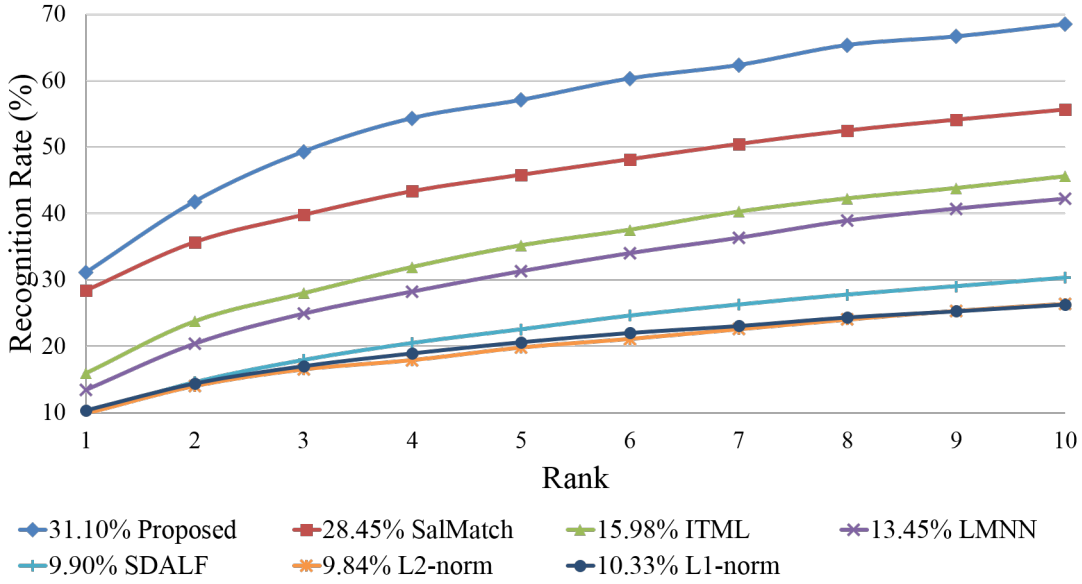

Figure 5.9: The comparison of the CMC curves on the CUHK Campus dataset for the proposed method and other methods.

5.3.6 Selection of Reference Set

The reference set can be optimized by selecting the most discriminative reference subjects and removing any redundant data. The goal is to select the "basis" reference subjects that will span the reference space. In other words, dissimilar subjects are preferred to form a more definitive reference set. Different reference selection rules are suggested in [57] and [40]. We use three different methods in order to select n reference subjects out of a total of N available ones.

1. **Random selection.** We randomly sample n reference subjects out of the reference pool of size N.

2. **Max-variation rule.** In this rule, for each image I_i in the reference set $\{I_i, i = 1, 2, ..., N\}$, the similarity $s(f_i, f_j)$ between I_i and $I_{j,j\neq i}$ is computed for all j. The variation score v_i is $Var\{s(f_i, f_j)\}_{j=1,j\neq i}^{N}$. By ranking v_i values in a descending order, top n images are chosen.

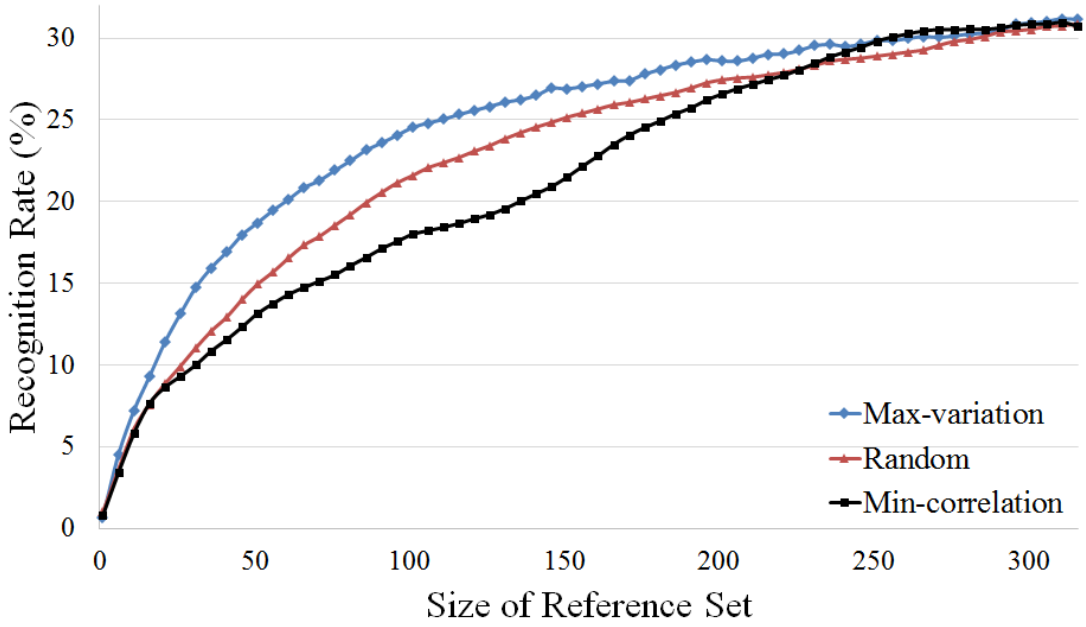

Figure 5.10: Rank-1 re-identification accuracy using reference set with different sizes of the VIPeR dataset.

3. **Min-correlation rule.** This rule is a backward selection process. Starting with the entire reference set $\{I_i, i = 1, 2, ..., N\}$, the sample I_i is removed whose average correlation with other samples $I_{j,j\neq i}$ is the highest. This process is repeated until n samples are left.

• **VIPeR Dataset.** Figure 5.10 shows the rank-1 recognition accuracy on the VIPeR dataset with varying reference set size using the aforementioned selection strategies. For the random selection, as the size of the reference set increases, the recognition rate keeps improving. For the max-variation rule, when the number of selected reference subjects is small, the recognition performance is higher than the results by random selection. As the reference set size reaches above 250, both rules result in a similar performance, with only marginal improvement by adding more reference samples. As compared to random selection and max-variation selection, the min-correlation rule does not select better

reference samples when the size of the reference set is not sufficiently large. Note that the size of the reference set can be reduced without a significant loss in performance. By using the max-variation rule, the size of the reference set is reduced by over 40% from 316 to 180 with a performance drop of less than 10%. While keeping sufficient accuracy with less than 4% degradation, the size of the reference set can be reduced by over 20% from 316 to 250. With constraint such as computational efficiency on reference set, the size of the reference set may be chosen where the improvement in recognition rate starts changing slowly.

• **CUHK Campus Dataset.** Figure 5.11 shows the results with varying reference set size on the CUHK Campus dataset. A trend similar to Figure 5.10 is observed. The max-variation rule is able to select a better subset of the reference samples. As suggested by the experimental results, max-variation is an effective strategy for reference set selection. For the CUHK Campus dataset, by using max-variation rule for selection, the size of the reference set can be reduced by over 40% from 486 to 286 with a performance drop of $\sim$5%.

5.3.7 Computational Cost

The computational cost mainly consists of the following parts: feature extraction, RCCA subspace learning, RD generation, initial matching, re-ranking. The experiments are performed using Matlab implementation without optimization on a laptop with Intel i7 2.4GHz CPU and 8GB RAM. For each image, the feature extraction takes about 0.37 second. On the VIPeR dataset, learning RCCA projection matrices takes about 4.2 seconds. For the CUHK Campus dataset, this procedure takes slightly longer of about 4.4 seconds, due to more data involved. However, the projection learning is

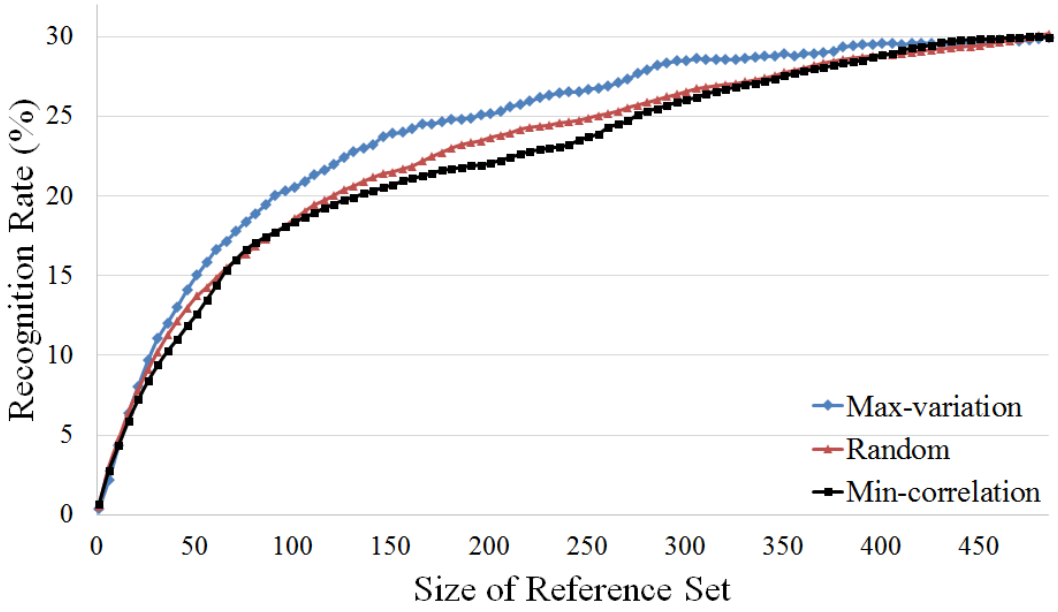

Figure 5.11: Rank-1 re-identification accuracy using reference set with different sizes of the CUHK Campus dataset.

done during the offline process and need to be performed only once. The generation of a RD is very efficient and it takes less than 2×10^{-6} second. Initial matching on the VIPeR dataset for one query takes about 0.8×10^{-4} second, and this goes up to 1.1×10^{-4} second for the CUHK Campus dataset. Saliency-based re-ranking for one probe takes about 0.81 second on the VIPeR dataset and about 0.96 second on the CUHK Campus dataset. The efficiency of re-ranking can be improved using fast patch match technique [18].

5.4 Conclusions

In this chapter, the use of a reference set for person re-identification is proposed. As compared to the previous methods in which either invariant features are extracted or a distance metric is explored, in our approach a reference set is utilized to transfer the matching problem from an appearance space to a reference space. The re-identification

is achieved by matching the reference descriptors (RDs) generated with the reference set and the matching results are improved by a re-ranking step using image saliency information. The experiments on different datasets showed that the proposed method using RCCA in conjunction with reference set outperformed 17 current approaches on VIPeR dataset and six recently published techniques on CUHK Campus dataset.

The proposed method avoided a direct comparison between the gallery and probe using appearance features. Reference-based matching with re-ranking significantly improved upon RCCA-based matching as a baseline method ($\sim$35% improvement on VIPeR Dataset and $\sim$32% improvement on CUHK Campus Dataset). The proposed method can be combined with any advanced feature representation to further improve the re-identification accuracy and the dimension of RDs is determined only by the size of the reference set which can be optimized based on the analysis presented in this chapter. Future work includes study of feature selection and extension of the proposed reference-based framework to multi-shot person re-identification.

Chapter 6

Person Re-Identification Using L2 Regularized Sparse Representation

6.1 Introduction

The vast deployment of video surveillance cameras in public venues drives the need for automated surveillance applications such as people tracking [69], anomaly detection in crowd [97], etc. Many of these applications require the ability to determine the ID of the subject in different camera views, which is a problem referred as person re-identification that is gaining more attentions in the literature recently [52, 44, 157, 94, 66, 190, 146, 182, 12]. Specifically, the goal of person re-identification is to accurately match individuals across non-overlapping cameras at different time and locations. The results of person re-identification can be readily used in further processing tasks such as tracklet association for multi-camera people tracking [27].

(a) (b) (c) (d)

Figure 6.1: Samples of image pairs of the same person in different camera views, showing (a) pose variation, (b) illumination change, (c) occlusion, and (d) low image quality, which make re-identification of people in different cameras a challenging problem.

Despite of the plethora of advanced pattern recognition techniques developed in the past few years, the performance of person re-identification is still not robust enough to warrant high accuracy in practice. The difficulties for person re-identification involve the following aspects: 1) *Pose variation*. In different camera views, a subject may have arbitrary poses (see Fig. 6.1(a)); 2) *Illumination change*. The lighting condition is usually not constant in different camera views. As a result, the appearance of the same subject may vary significantly due to changing illumination (see Fig. 6.1(b)); 3) *Occlusion*. A subject in one camera view may be fully or partially occluded by other subject or carrying items such as a backpack (see Fig. 6.1(c)); 4) *Low image quality*. The captured image of a subject may suffer from low resolution, noise, or blur due to limited imaging quality of surveillance cameras (see Fig. 6.1(d)).

In a person re-identification system usually two steps are involved: 1) extracting feature representations from person detections, and 2) establishing the correspondence between feature representations of probe and gallery. A gallery is a dataset composed of images of people with known ID. A probe is the detection of a person from a different camera. Although other forms of biometrics such as face and gait [59, 192]

can be used to recognize people, however, acquiring such biometrics is difficult in uncontrolled low-resolution video. For person re-identification, most of the existing approaches are appearance-based.

With the availability of tools for person detections, most of the previous work on person re-identification can be categorized into two groups:

1. Extracting feature representations which are robust against pose or illumination change [12, 52, 44, 157].

2. Developing new matching methods using metric learning or ranking classifiers [66, 190, 81, 133].

For the first group, discriminative appearance features are desired. Normally color and texture based features are widely used [84, 66]. However, color or texture feature representations are sensitive to pose and illumination change, which may result in larger intra-person variation (difference between features of same person) than inter-person variation (difference between features of different persons). Besides low level image features, attribute or shape information has been applied in conjunction with color or texture features to improve the recognition accuracy [11, 158].

To pursue more reliable matching, feature transformations or distance metrics are learned such that the distance between feature representations of the same person from different cameras is reduced while the distance between feature representations from different persons is increased [162, 55, 35, 81]. SVM with ranking [133] and transfer learning [189] have also been proposed to obtain better matching correspondence.

In this chapter, we propose a novel feature representation for person re-identification based on sparse coding. Inspired by coherent subspace learning to handle cross-type im-

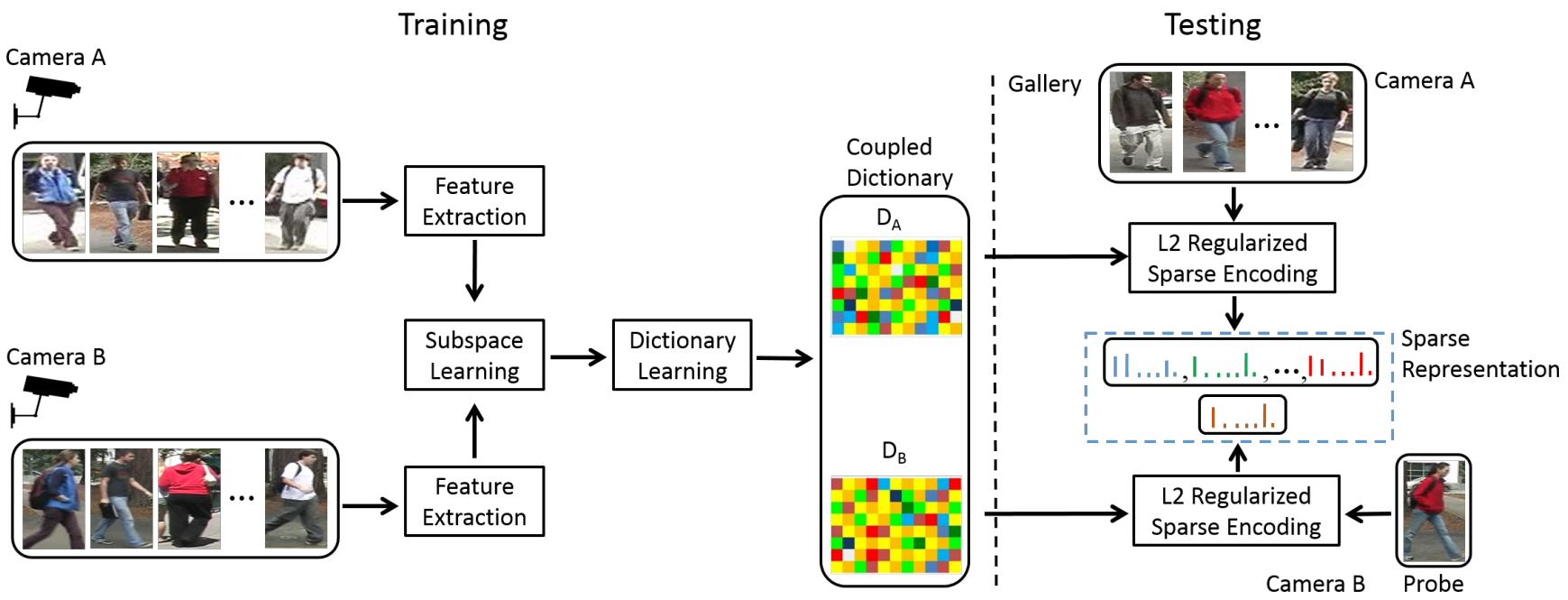

Figure 6.2: Outline of the proposed sparse representation for person re-identification. In the training phase, the appearance features are extracted from images captured by two different cameras. A subspace is learned using CCA to project the features into a coherent subspace with maximized correlation between data in two views. Two dictionaries are learned jointly for each camera. In the testing phase, the features of gallery and probe are first projected into the learned subspace. Then their sparse representations with L2 regularization are obtained using the coupled dictionaries. The sparse representations are then used as a new representation for probe or galley for matching.

age synthesis [154] and face image super-resolution [9], we first learn a transformation to project the original image features into a subspace using canonical correlation analysis (CCA). In this learned subspace, the correlation between the features of the same people from different camera views is maximized. Then, two dictionaries for two camera views are jointly learned using training data in the coherent subspace. Given an image in the gallery, its image features are first projected into the CCA subspace and the sparse coefficients of this gallery subject are obtained using the learned dictionary with an L2 regularization term. These coefficients become the new feature representation for this gallery instance. During re-identification, given a probe, its sparse representation is obtained in the same way using the corresponding dictionary. Fig. 6.2 illustrates the outline of the proposed method for generating the sparse representation. The matching is then performed by computing the similarity between the sparse representations of the probe and gallery. A related work for person re-identification was introduced in which a sparse representation was directly learned using a dictionary [80]. The dictionary was composed of existing data without any learning and the identity of the probe was determined through the non-zero coefficients with a majority voting rule. In contrast to our approach, the sparse representations in [80] were used for determining the identity of the probe, while in our method the sparse representations are used as new feature representations for matching.

6.1.1 Contributions of This Chapter

The main contributions of the method proposed in this chapter are:

1. To mitigate the disparity between image data from different views, we learn a coherent subspace using CCA and project the multiview data (images of the same person in different camera views) into this coherent subspace such that the correla-

tion between two views of the same data are maximized. This subspace projection provides a foundation for robust matching across cameras.

2. We propose a novel framework for generating sparse representations of probe and gallery data in the coherent subspace. The generated sparse representations are used for person re-identification. Compared with matching using features extracted directly from images, using the learned sparse representation achieves the state-of-the-art results on different publicly available datasets.

3. We learn the sparse representation with a coupled dictionary sets. The dictionaries are jointly learned using training data from different camera views. In addition, the sparse representation is regularized with an additional L2 regularization term to ensure the stability of the learned coefficients while preserving sparsity. Experimental results show that L2 regularized sparse representation outperforms standard sparse representation.

The rest of this chapter is organized as follows. Details of the proposed L2 regularized sparse representation based method for person re-identification are presented in Section 6.2. Section 6.3 provides the experimental results and finally Section 6.4 concludes this chapter.

6.2 Sparse Representation for Person Re-Identification

The goal is to re-identify people in non-overlapping cameras. To mitigate significant disparity in appearance feature space for the same subject in different views, the proposed algorithm first finds projection matrices for features from each view such that after projection features of the same person are maximally correlated. To learn this

subspace projection, labeled training image pairs are used in CCA. After the projection matrices are obtained, training data are projected into this coherent subspace. The projected training data are then used to jointly learn coupled dictionaries for each camera view. In the re-identification process given probe and gallery, appearance features are first extracted. These features are then projected into the learned coherent subspace, in which their sparse representations with L2 regularization are obtained using the coupled dictionaries. The calculated sparse coefficients are used as a new feature representation for probe and gallery in the matching process which is based on a modified cosine similarity measure. The pipeline for generating sparse representations is illustrated in Fig. 6.2.

6.2.1 Coherent Subspace Learning

Canonical Correlation Analysis (CCA) was first introduced in [67] and it is a multivariate statistical analysis technique. CCA finds projection matrices for two sets of random variables such that the correlation between the projected random variables is maximized in the correlated or coherent subspace. CCA has been applied to problems involving multi-view or multi-modality data such as image super-resolution [9, 72] and face recognition under pose variation [139].

For person re-identification, given N image pairs from two cameras A and B, appearance features with dimension m are first extracted from the images. These feature vectors are organized into two data matrices $X_A = \{x_A^i \in \mathbb{R}^m, i = 1, 2, ..., N\}$ and $X_B = \{x_B^i \in \mathbb{R}^m, i = 1, 2, ..., N\}$, in which x_A^i and x_B^i correspond to the same person in different views. The goal of CCA is to find a pair of projection vectors $w_A \in \mathbb{R}^m$ and $w_B \in \mathbb{R}^m$ such that the correlation coefficient ρ of $w_A^T X_A$ and $w_B^T X_B$ is maximized. Mathematically, the objective function to be maximized is given by

$$\begin{aligned}\rho &= \frac{cov(w_A^T X_A, w_B^T X_B)}{\sqrt{var(w_A^T X_A) var(w_B^T X_B)}} \\ &= \frac{w_A^T C_{AB} w_B}{\sqrt{w_A^T C_{AA} w_A w_B^T C_{BB} w_B}}\end{aligned} \tag{6.1}$$

where *cov* is short for covariance and *var* computes data variance. $C_{AA} = E[X_A X_A^T]$ and $C_{BB} = E[X_B X_B^T]$ are the covariance matrices of X_A and X_B. $C_{AB} = E[X_A X_B^T]$ is the covariance matrix of X_A and X_B.

Eqn. 6.1 can be reformulated as a constrained optimization problem as follows

$$\begin{aligned}\text{maximize} \quad & w_A^T C_{AB} w_B \\ \text{subject to} \quad & w_A^T C_{AA} w_A = 1 \\ & w_B^T C_{BB} w_B = 1\end{aligned} \tag{6.2}$$

Equivalently, w_A and w_B can be solved through the following generalized eigenvalue problem

$$\begin{bmatrix} 0 & C_{AB} \\ C_{BA} & 0 \end{bmatrix} \begin{bmatrix} w_A \\ w_B \end{bmatrix} = \lambda \begin{bmatrix} C_{AA} & 0 \\ 0 & C_{BB} \end{bmatrix} \begin{bmatrix} w_A \\ w_B \end{bmatrix} \tag{6.3}$$

where $C_{BA} = E[X_B X_A^T] = C_{AB}^T$.

The projection matrices $W_A \in \mathbb{R}^{m \times d}$ and $W_B \in \mathbb{R}^{m \times d}$ are composed of d pairs of projection vectors w_A and w_B corresponding to d largest eigenvalues. In this way, W_A and W_B project the original features from $\mathbb{R}^m$ to a subspace of $\mathbb{R}^d$, where the correlation between the projected features of X_A and X_B is maximized.

Fig. 6.3 demonstrates the CCA principle that projects the data from different views into a coherent subspace in which the data pair of the same person are maximally correlated. To validate the ability of CCA to find a coherent subspace, we use half of the

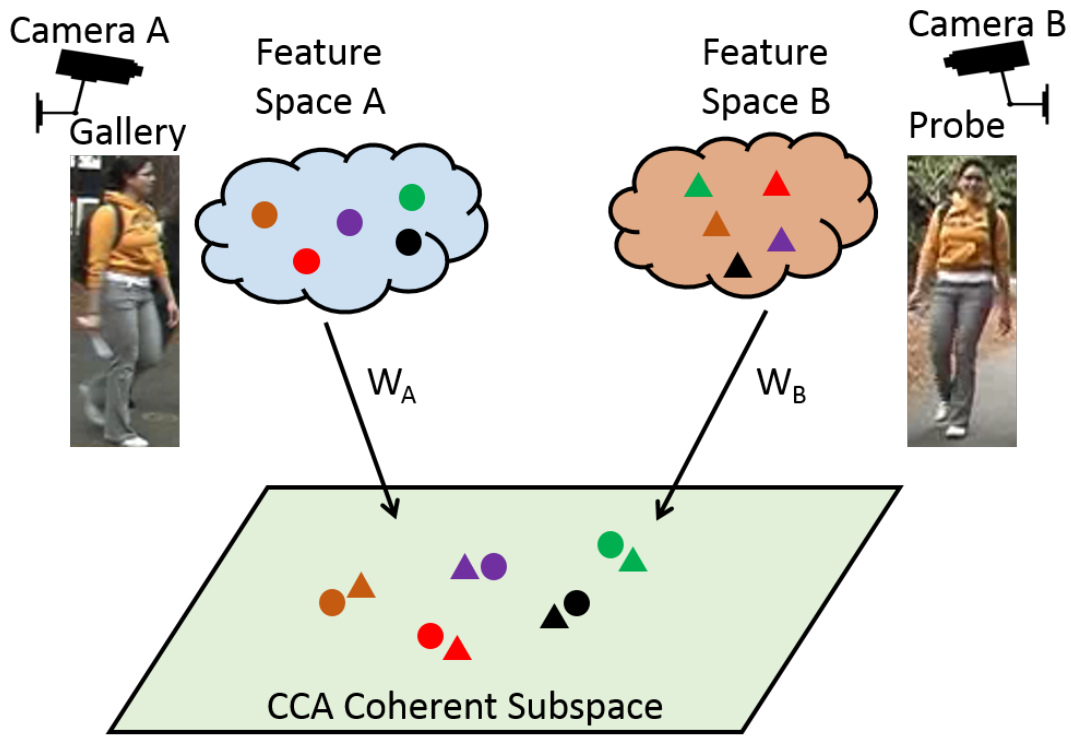

Figure 6.3: Illustration of CCA projection. A pair of symbols with the same color but different shapes indicates features of the same person. The projection matrices W_A and W_B transform the data from the original feature space to a coherent subspace in which the data correlation is maximized.

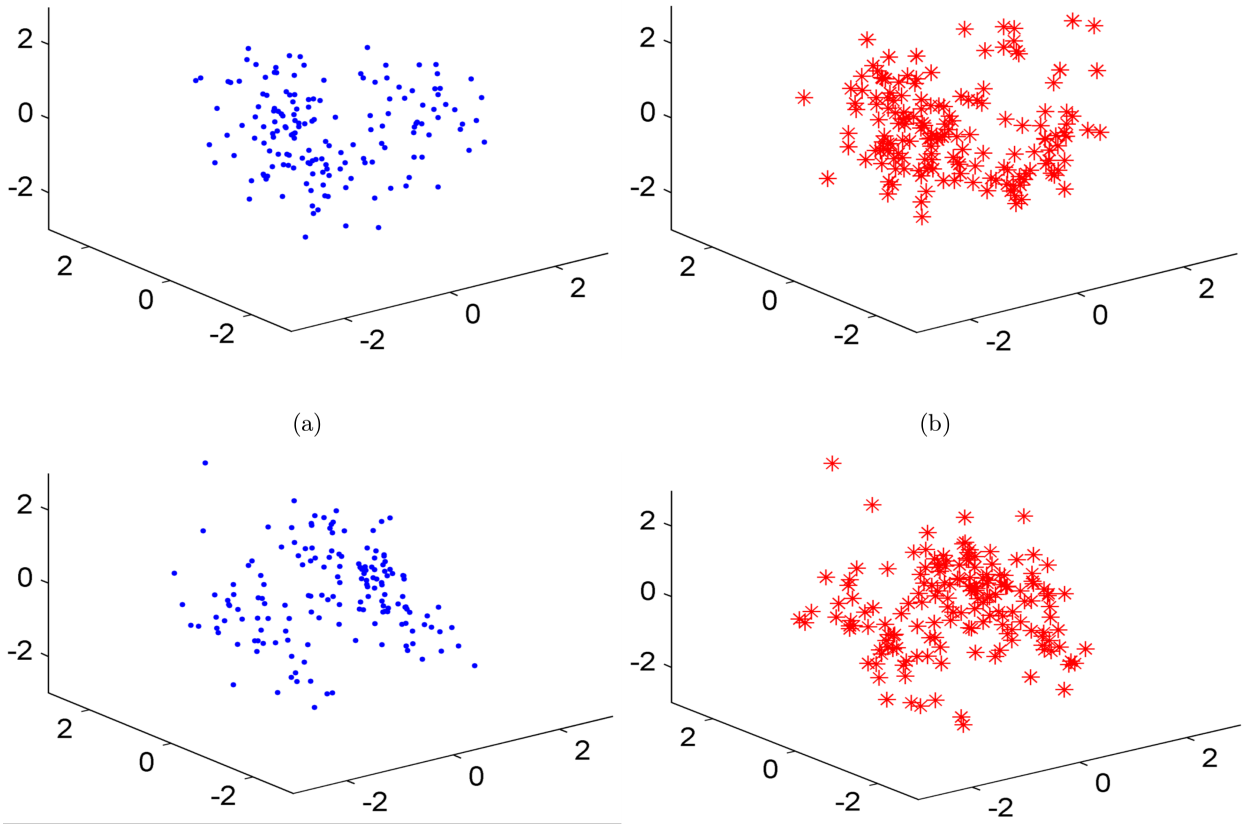

Figure 6.4: Features in subspace using different projections. (a) Embedded manifold for features in PCA subspace of testing data in one camera view (in dots). (b) Embedded manifold for features in PCA subspace of testing data in the other camera view (in asterisks). (c) Embedded manifold for features in CCA subspace of testing data in one camera view (in dots). (d) Embedded manifold for features in CCA subspace of testing data in the other camera view (in asterisks). Different from PCA, CCA projects multi-view data into coherent subspaces and the manifold of data in each view is more similar, i.e., the shapes of data in (c) and (d) by CCA projection are more similar than the shapes of data in (a) and (b) by PCA projection. The training data for PCA and CCA are from the VIPeR dataset [52] and testing data are the rest of the same dataset. The numbers on the axes denote normalized feature values.

data from the VIPeR dataset [52] to learn the PCA and CCA projections, respectively and project the other half of the data into the PCA and CCA subspaces with learned projection matrices. As shown in Fig. 6.4, the CCA projected data from two cameras exhibit more similarity in terms of their manifold structures compared to the structures of the same data in the PCA subspace. This indicates that CCA is able to correlate data from different views through subspace projection.

6.2.2 Coupled Dictionary Learning

Given N training data pairs $x_A^i \in \mathbb{R}^m$ and $x_B^i \in \mathbb{R}^m$ consisting of image pairs from cameras A and B, the projected image features in the CCA subspace are denoted by $p_A^i \in \mathbb{R}^d$ and $p_B^i \in \mathbb{R}^d$, respectively. The goal is to learn the dictionaries $D_A \in \mathbb{R}^{d \times k}$ and $D_B \in \mathbb{R}^{d \times k}$ of size k jointly, such that the sparse representations for p_A^i and p_B^i should be the same. In other words, the idea is that the sparse representations corresponding to the same person but in different camera views should be the same. The energy function to be optimized is

$$\min_{D_A, D_B} \frac{1}{N} \left(\sum_{i=1}^{N} \left\| p_A^i - D_A \alpha_i \right\|_2^2 + \left\| p_B^i - D_B \alpha_i \right\|_2^2 + \gamma_1 \left\| \alpha_i \right\|_1 + \gamma_2 \left\| \alpha_i \right\|_2^2 \right) \tag{6.4}$$

where γ_1 and γ_2 are regularization parameters for L1 and L2 regularization terms respectively.

The L1 regularization term ensures that the coefficients α_i are sparse. Previous study suggested that most images (e.g., faces) can be approximated as a linear combination of the base elements in a dictionary and this representation is naturally sparse [165]. The sparsity resembles human perception system in which the activation of neurons to an image is typically sparse [127]. The L2 regularization term has the

properties as in the Ridge Regression to stabilize the coefficients. Sparse representation with L2 regularization is also referred as Elastic Net in statistics [195].

Similar to the formulation in [174], Eqn. 6.4 can be converted to

$$\min_{D} \frac{1}{N}(\sum_{i=1}^{N} \left\|p^i - D\alpha_i\right\|_2^2 + \gamma_1 \left\|\alpha_i\right\|_1 + \gamma_2 \left\|\alpha_i\right\|_2^2) \tag{6.5}$$

where p^i and D are constructed by

$$p^i = \begin{bmatrix} p_A^i \\ p_B^i \end{bmatrix}, D = \begin{bmatrix} D_A \\ D_B \end{bmatrix} \tag{6.6}$$

To obtain the dictionary in Eqn. 6.5, an online optimization algorithm based on stochastic approximations is used [114].

6.2.3 Sparse Representation with L2 Regularization

With the learned coupled dictionaries D_A and D_B, the sparse representations for probe or gallery data can be generated. Assuming images from camera A serve as the gallery, for each image j in the gallery, the appearance features are first extracted and then projected into the CCA subspace with the projected features denoted as g^j. Its sparse representation α_{g^j} is obtained by

$$\underset{\alpha_{g^j}}{\operatorname{argmin}} \left\|g^j - D_A\alpha_{g^j}\right\|_2^2 + \gamma_1 \left\|\alpha_{g^j}\right\|_1 + \gamma_2 \left\|\alpha_{g^j}\right\|_2^2 \tag{6.7}$$

The L2 regularized sparse coefficients α_{g^j} are used as a new representation for gallery image j. Similarly, given a probe with its appearance feature projected into CCA subspace as p, the sparse representation α_p is learned by solving

$$\underset{\alpha_p}{\operatorname{argmin}} \left\|p - D_B\alpha_p\right\|_2^2 + \gamma_1 \left\|\alpha_p\right\|_1 + \gamma_2 \left\|\alpha_p\right\|_2^2 \tag{6.8}$$

Eqn. 6.7 and Eqn. 6.8 can be solved efficiently with stability using least angle regression (LARS) algorithm [41].

6.2.4 Identity Matching

With the sparse representations of probe and gallery ready, the matching is based on the similarity between the sparse coefficients α_p of a probe and the sparse coefficients α_g^j of a gallery. We adopt a modified cosine similarity measure [101] defined by

$$\mathrm{sim}(\alpha_p, \alpha_{g^j}) = \frac{\left|(\alpha_p)^T \cdot \alpha_{g^j}\right|}{\|\alpha_p\| \left\|\alpha_{g^j}\right\| \left(\left\|\alpha_p - \alpha_{g^j}\right\|_{\mathrm{P}} + \epsilon\right)} \tag{6.9}$$

where $\|\cdot\|_{\mathrm{P}}$ is the L_{P} norm and ϵ is a small positive number to prevent division by zero. The reason to apply the modified cosine similarity is that the standard cosine similarity does not take into account the actual distance between two vectors, while the modified cosine similarity is able to address both the distance measure and angular measure and has shown improved performance in recognition tasks [101].

6.2.5 Summary of the Algorithm

The proposed algorithm for person re-identification consists of training and testing phases. In the training phase, CCA projection matrices are learned as well as the coupled dictionaries for different camera views. Given probe and gallery in the testing phase, the appearance features are extracted and projected into the CCA subspace. Then the L2 regularized sparse representations for probe and gallery are obtained using jointly learned dictionaries. The similarity scores between probe and gallery are computed using their sparse representations. The algorithms for training and testing

Algorithm 1 CCA Subspace and Coupled Dictionary Learning

Input:

Training image pairs from cameras A and B

d: dimension of the CCA subspace

k: size of the dictionaries

1: Extract appearance features from images and obtain data matrices X_A and X_B.

2: Solve the generalized eigenvalue problem in Eqn. 6.3.

3: Project X_A and X_B into the learned CCA subspace.

4: Solve the optimization problem in Eqn. 6.5.

Output:

CCA subspace projection matrices W_A and W_B

Coupled dictionaries D_A and D_B

algorithms are summarized in Algorithm 1 and Algorithm 2, respectively.

6.3 Experiments

6.3.1 Datasets

We evaluate our method on three different publicly available datasets for two typical re-identification scenarios, namely image-based case (single-shot) and video-based case (multi-shot). For single-shot scenario, the VIPeR dataset and the CUHK Campus dataset are used. The VIPeR dataset enrolled 632 persons and is one of the most challenging and widely evaluated benchmark datasets for person re-identification [52]. Some sample image pairs from this dataset are shown in Fig. 6.5. For each person, one detection is available in each of the two non-overlapping cameras with varying illumina-

Algorithm 2 Re-identification with Sparse Representation

Input:

Probe and gallery from cameras A and B

W_A and W_B: CCA projection matrices

D_A and D_B: coupled dictionaries

1: Extract appearance features from probe and gallery.

2: Project appearance features of probe and gallery into the learned CCA subspace using W_A and W_B.

3: Obtain L2 regularized sparse representations for probe and gallery using Eqn. 6.7 and Eqn. 6.8 with D_A and D_B.

4: Calculate similarity scores between probe and gallery using Eqn. 6.9.

5: Rank the similarity scores from high to low.

Output:

Re-identification results

tion and cluttered background. For most of the subjects the view change is more than 90 degrees. In addition, partial occlusion is frequent due to the subjects carrying items such as backpack or handbag.

The CUHK Campus dataset is a recently released dataset which contains images of 971 subjects from two non-overlapping camera views [95]. One camera captures the frontal or rear view of a person and the other camera captures the profile view of a person. Each person has two detections in each camera view. Some sample image pairs from this dataset are shown in Fig. 6.6.

For multi-shot video-based re-identification, we use the Person Re-ID 2011

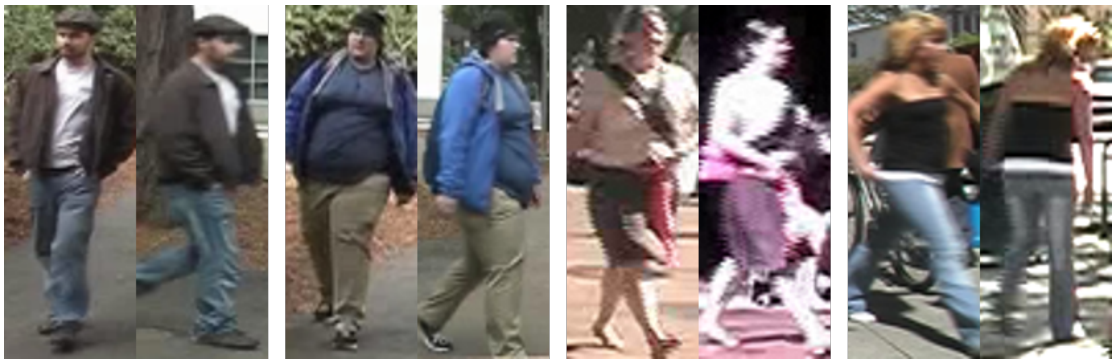

Figure 6.5: Sample image pairs from the VIPeR dataset [52].

Figure 6.6: Sample image pairs from the CUHK Campus dataset [95].

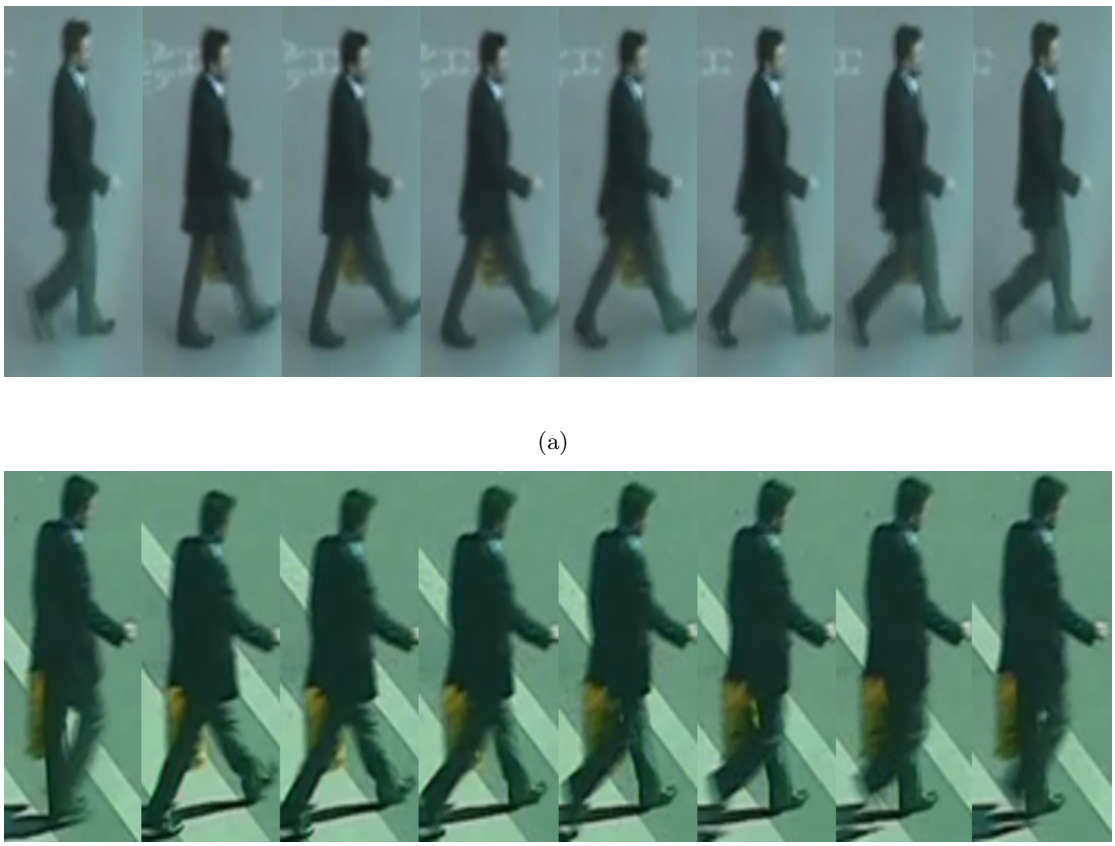

Figure 6.7: Sample image pairs from the PRID dataset [65]. (a) Trajectory of a person in camera A. (b) Corresponding trajectory in camera B.

(PRID) dataset. The PRID dataset consists of multiple person trajectories recorded from two surveillance cameras. Camera A contains 385 persons and camera B shows 749 persons. The first 200 persons appear in both camera views. Each trajectory contains approximately 100 to 150 images depending on the walking speed of a person. Two segments of trajectories of the same person in two cameras are shown in Fig. 6.7.

6.3.2 Feature Extractions

All of the images in these three datasets are normalized to 128×48 in the experiments. Each image is divided into blocks of size 8×16. The blocks are overlapped

Figure 6.8: Sample segmentation results using the method in [111] to separate the foreground subject from the background. The appearance features are extracted from the foreground to mitigate the impact by the cluttered background.

by 50% in both horizontal and vertical directions. The appearance features extracted from the images include both color and texture features as in [66]. For each block, the color features consist of the quantized mean values of the HSV and Lab color channels. In addition, we include semantic color names [83] as an additional color representation. The texture features are represented by the 8-bit Local Binary Patterns (LBP) [126]. The final appearance features are the concatenation of both color and texture features. To minimize the impact of the cluttered background, we use a deep decompositional network based pedestrian parsing method [111] to segment the foreground subject from the background before the appearance features are extracted. Fig. 6.8 shows some segmentation results.

The dimension of the CCA subspace projection matrices W_A and W_B are set to $d = 50$. The L1 and L2 regularization parameters to learn the coupled dictionaries and sparse representations are set to 0.01 and 0.02 respectively. The size of the dictionary is set to $k = 100$. These parameters are determined by cross-validation.

6.3.3 Evaluation Method

For the experiments on the VIPeR and the CUHK Campus datasets, we follow the experimental protocols in the previous work (e.g., [44, 81, 66, 186]) for fair comparison. We randomly partition each dataset into two subsets of equal size. Half of the data are used for training and the other half are used for testing. Gallery consists of images from one camera and images from the other camera are used as probes. For the PRID dataset, the data of the common 200 subjects in two camera views are used. The gallery set is constructed by extracting five evenly sampled images per trajectory as done in [65]. A probe of one subject consists of all the detections in the trajectory and a majority voting rule is applied to determine the identity of each probe.

To evaluate the re-identification performance, recognition rates at selected ranks and the Cumulative Match Characteristic (CMC) curves are reported for comparison. The CMC curve represents the expectation of finding the correct match in the top r matches. In other words, a rank-r recognition rate shows the percentage of the probes that are correctly recognized from the top r matches in the gallery. The experiments are conducted 10 times and the average results are listed.

6.3.4 Experimental Results

6.3.4.1 The VIPeR Dataset

We first conduct experiments on the VIPeR dataset, the results on which have been reported in most of the recent work on person re-identification. We compare our approach with the following 18 state-of-the-art alternatives, which are reference-based approach (RD) [12], saliency matching (SalMatch) [186], relaxed pairwise learned met-

Table 6.1: Person Re-identification recognition rates (in %) on the VIPeR dataset at different ranks.

Rank→	$r = 1$	10	20	50	100
Proposed	**32.91**	**75.93**	**89.24**	**96.84**	**99.73**
RD [12]	30.25	74.68	86.82	95.70	99.24
SalMatch [186]	30.16	65.54	79.15	91.49	98.10
RPLM [66]	27.34	69.02	82.69	94.56	98.54
RS-KISS [146]	24.50	66.60	81.70	93.50	98.00
CPS [28]	21.84	57.21	71.00	87.00	91.77
BiCov [112]	20.66	56.18	68.00	81.56	88.66
KISSME [81]	20.03	62.39	77.46	92.81	98.19
LMNN-R [37]	20.00	66.00	79.00	92.50	95.18
SDALF [44]	19.87	49.37	65.73	84.84	90.43
MRank [109]	19.34	55.51	70.44	87.69	96.90
PCCA [119]	19.27	64.91	80.28	95.00	97.01
DDC [65]	19.00	52.00	65.00	80.00	91.00
LMNN [162]	17.41	53.86	67.88	88.13	96.23
aPRDC [103]	16.14	50.98	65.95	88.00	93.00
PRDC [190]	15.66	53.86	70.09	87.79	92.84
ITML [35]	15.54	53.13	69.05	88.54	96.93
RankSVM [133]	14.00	51.00	67.00	85.00	94.00
ELF [53]	12.00	43.00	60.00	81.00	93.00

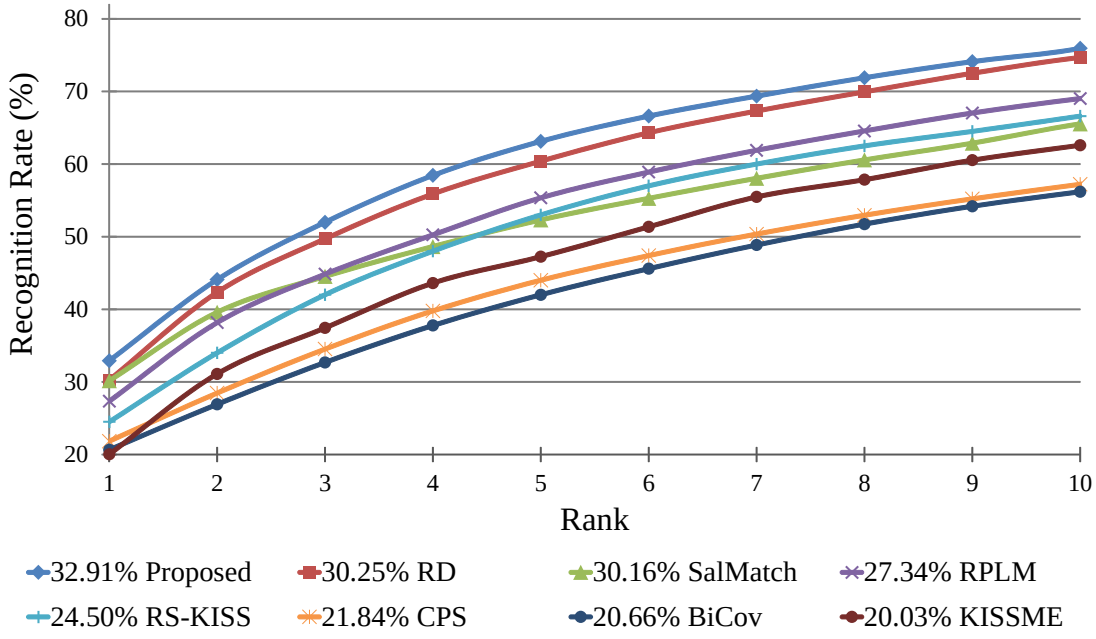

Figure 6.9: CMC curves on the VIPeR dataset for the proposed method and the other methods. The rank-1 rates are shown above the figure caption. Best viewed in color.

ric (RPLM) [66], regularized smoothing KISS metric learning (RS-KISS) [146], custom pictorial structures (CPS) [28], biologically inspired features and covariance descriptors (BiCov) [112], KISS metric (KISSME) [81], large margin nearest neighbor with rejection (LMNN-R) [37], symmetry-driven accumulation of local features (SDALF) [44], manifold ranking (MRank) [109], pairwise constrained component analysis (PCCA) [119], descriptive and discriminative classification (DDC) [65], large margin nearest neighbor (LMNN) [162], attributed-based relative distance comparison (aPRDC) [103], relative distance comparison (PRDC) [190], information-theoretic metric learning (ITML) [35], support vector ranking (RankSVM) [133], and ensemble of localized features (ELF) [53].

The re-identification accuracy of different methods at rank 1, 10, 20, 50 and 100 are reported in Table 6.1. The proposed method achieves a recognition rate of 32.91% at rank 1, which is about 9% improvement over the second best result of 30.25% by RD [12]. Furthermore, at all of the other ranks, the proposed method consistently

Table 6.2: Person Re-identification recognition rates (in %) on the VIPeR dataset at different ranks with reduced training data size.

Training size→	N=200			N=100		
Rank→	$r = 1$	10	20	$r = 1$	10	20
Proposed	**23.34**	**60.07**	**75.26**	**16.82**	**49.14**	**62.97**
RD [12]	21.94	59.26	74.58	15.11	47.14	60.30
RPLM [66]	19.51	56.44	71.09	10.88	37.69	51.64
PRDC [190]	12.64	44.28	59.95	9.12	34.4	48.55

outperforms the competing methods. At rank 100, almost 100% recognition rate is reached. The CMC curves are compared in Fig. 6.9 between our method and the other top performers in Table 6.1. Similar to the observations from Table 6.1, the proposed method achieves higher recognition rates compared to the other methods at different ranks.

To study the impact of reduced training data size and to make comparison with other methods, we report in Table 6.2 the re-identification results with different training data sizes. In this case, all the data from the VIPeR dataset are used. As the size of the training set decreases, the number of subjects in the gallery and probe data increases, which makes the re-identification more difficult. The same experiment protocol was used in RD [12], RPLM [66], and PRDC [190], the results of which are included in Table 6.2 for comparison. The results shown in Table 6.2 suggest that with a smaller training set the proposed method is still able to perform better than the competing methods at different ranks.

Table 6.3: Person Re-identification recognition rates (in %) on the CUHK Campus dataset at different ranks.

Rank→	$r = 1$	10	20	50	100
Proposed	**31.34**	**68.39**	**78.14**	**87.63**	**95.26**
RD [12]	29.97	67.78	77.04	87.24	94.21
SalMatch [186]	28.45	55.68	67.95	83.53	92.10
ITML [35]	15.98	45.60	59.81	76.61	88.32
LMNN [162]	13.45	42.25	54.11	73.29	86.65
SDALF [44]	9.90	30.33	41.03	55.99	67.39

6.3.4.2 The CUHK Campus Dataset

For the CUHK Campus dataset, we compare the proposed approach with the following five methods: RD [12], SalMatch [186], SDALF [44], LMNN [162], ITML [35]. Table 6.3 reports the recognition rates at rank 1, 10, 20, 50, and 100. As compared to the other methods with all the recognition rates below 30%, the proposed method achieves a rank-1 recognition rate of 31.34%. Fig. 6.10 shows the CMC curves of our method and the other methods. The proposed method achieves a higher rate at each rank compared to the second best method (RD), and outperforms the rest of the methods by a large margin.

6.3.4.3 The PRID Dataset

For the PRID dataset, we compare the proposed approach with the following methods: RD [12], KISSME [81], as well as two baseline methods using L1 distance and L2 distance. Table 6.4 reports the recognition rates at rank 1, 5, 10, 20, and

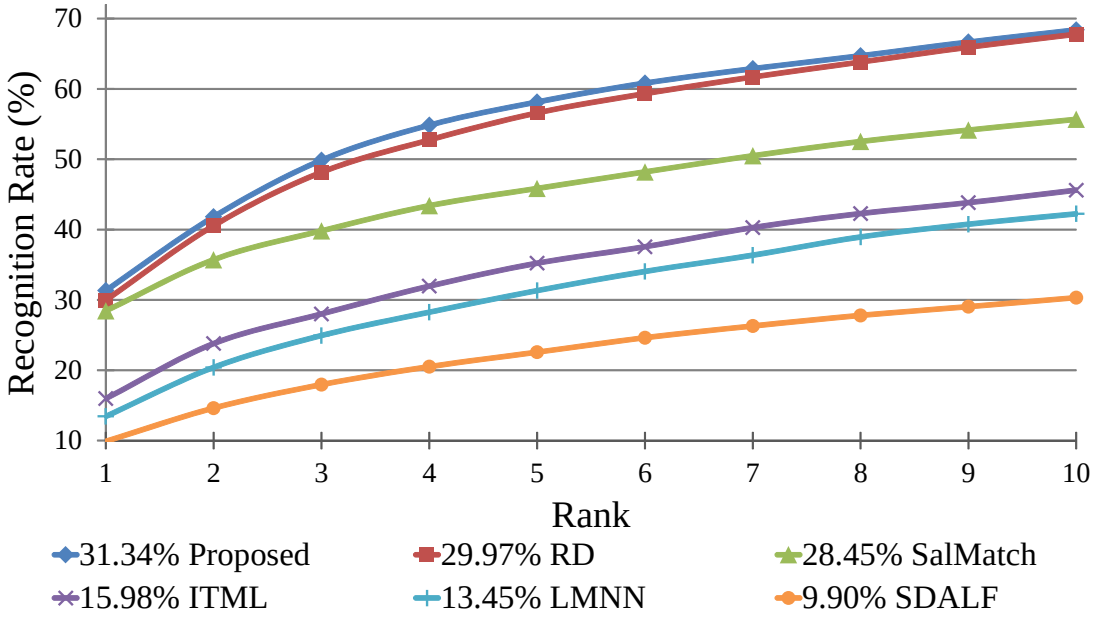

Figure 6.10: CMC curves on the CUHK Campus dataset for the proposed method and the other methods. The rank-1 rates are shown above the figure caption. Best viewed in color.

Table 6.4: Person Re-identification recognition rates (in %) on the PRID dataset at different ranks.

Rank→	$r = 1$	5	10	20	50
Proposed	**27**	**45**	**56**	**69**	**93**
RD [12]	24	41	53	67	92
KISSME [81]	16	38	49	60	92
L1 distance	13	35	47	58	89
L2 distance	11	33	42	57	87

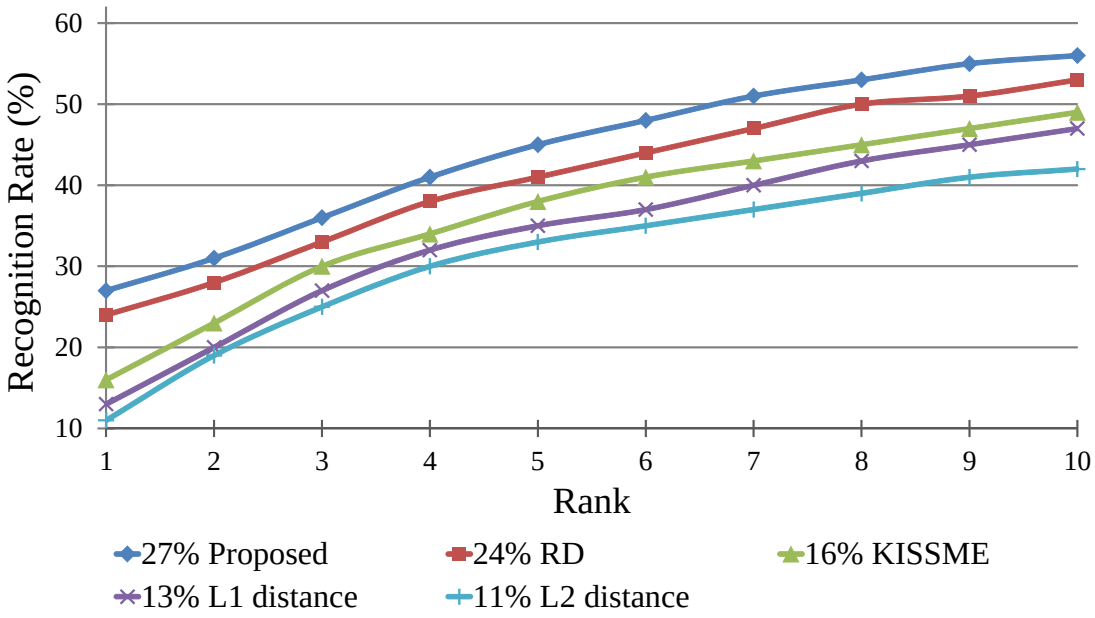

Figure 6.11: CMC curves on the PRID dataset for the proposed method and the other methods. The rank-1 rates are shown above the figure caption. Best viewed in color.

50. Fig. 6.11 compares the CMC curve of our method and the other methods. The performance comparison from Table 6.4 and Fig. 6.11 suggest that compared to the other methods, our method has over 10% improvement in terms of matching accuracy at different ranks. Compared to the baseline methods using L1-norm distance, L2-norm-distance, and KISSME metric [81], in which the low-level appearance features are used for matching, the performance of the proposed method and RD [12] shows significant advantage due to the adoption of new feature representation instead of using low-level appearance features directly for matching.

6.3.4.4 Effects of L2 Regularization

To evaluate the effectiveness of the proposed sparse representation with L2 regularization. Fig. 6.12 shows the re-identification performance in terms of rank-1 recognition rates with different regularization terms. On the VIPeR datraset, when

sparse representation is used with L1 term ($\gamma_2 = 0$ in Eqn. 6.7 and Eqn. 6.8), the rank-1 recognition rate is 25.95%. When the L1 regularization term is dropped while keeping the L2 regularization term ($\gamma_1 = 0$ in Eqn. 6.7 and Eqn. 6.8), a recognition rate of 29.11% is achieved. The combination of L1 and L2, referred as the L2 regularized sparse representation, improves over the results using a single regularization term and brings up the rank-1 recognition rate to 32.91%. This indicates that joint L1 and L2 regularization is effective for the proposed person re-identification approach. On the CUHK Campus dataset, the rank-1 recognition rate is the highest (31.34%) by using both L1 and L2 terms together. The use of a single regularization term (L1 or L2) leads to less accurate recognition rates of 26.39% and 28.87%, respectively. Similar observations hold for the PRID dataset for which L1 + L2 produces a better performance (27%) compared with using each of the regularization terms alone.

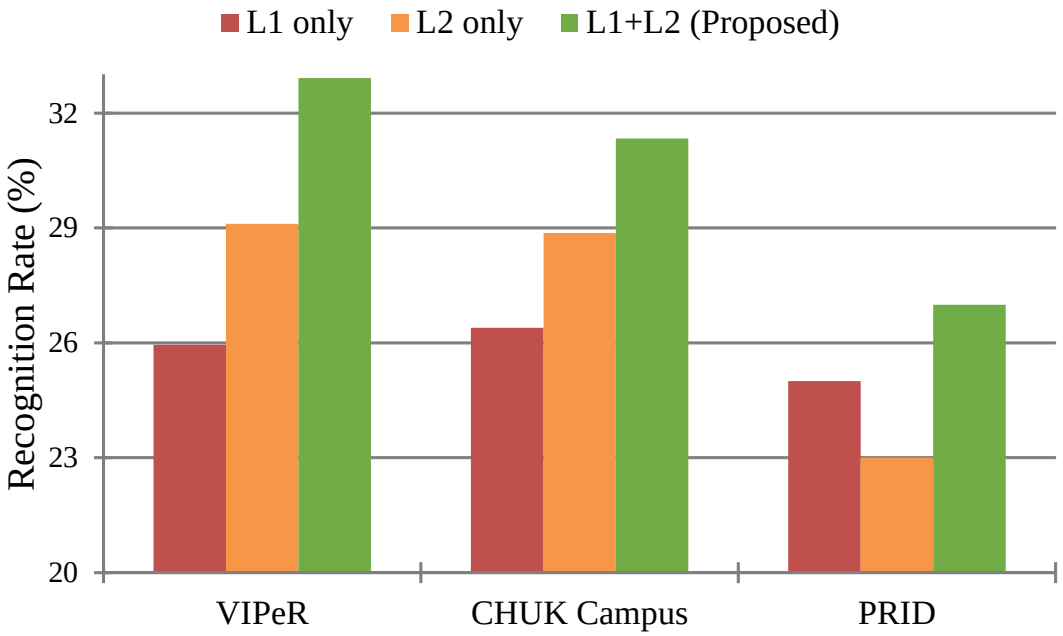

Figure 6.12: Comparisons of the rank-1 recognition rates on different datasets using the proposed method (L1 + L2), sparse representation only (L1), and L2 regularization only. Best viewed in color.

6.4 Conclusions

In this chapter, we proposed a person re-identification method using a sparse representation with L2 regularization. The L2 regularized sparse representations learned in a coherent subspace was used as a new feature representation instead of the appearance features for identity matching. Experiments were conducted on three publicly available datasets to evaluate the performance in single-shot and multi-shot re-identification settings. Compared to the state-of-the-art approaches, the proposed method achieved the highest recognition rates in different scenarios. In addition, the experimental results suggested that sparse representation with L2 regularization has superior performance compared to the baseline methods with a single L1 or L2 regularization term.

Chapter 7

Summary and Future Work

In this dissertation, we proposed several methods for face super-resolution, face recognition, and person re-identification, respectively.

In Chapter 3, we proposed a face super-resolution technique based on 2D canonical correlation analysis. This methods does not need to first convert the face images into vectors and the super-resolution works directly on image. Through thorough experiments with comparisons to the state-of-the-art methods, we demonstrated that the super-resolved faces using our method are visually realistic and very close to the ground-truth. Furthermore, we tested our super-resolved face images in recognition tasks and the results suggested that the super-resolved images by the proposed method achieve the highest accuracy with very low computational cost. Currently our method for super-resolution requires the face images to be aligned. In future work, we are interested in developing alignment-free super-resolution techniques as a pre-processing technique to improve the face recognition accuracy in real-world cases.

In Chapter 4, a multi-camera face recognition system using dynamic Bayesian network was proposed. This application is particulary practical in surveillance applica-

tions where multiple cameras are available that can help identify a person in a collaborative manner. In our video-based approach, information from multiple cameras are fused using a graphical model with temporal links. Specifically, the dynamic Bayesian network encodes the person-specific dynamics to help improve the recognition performance. We performed experiments on a real-world surveillance video dataset with three cameras and experimental results show that the proposed method outperformed standard face recognition modules and the recognition performance using multiple cameras is better than using any single camera. Future work involves study of feature description and selection to further improve the recognition accuracy. In addition, building up an evolving graphical model that improves on itself as more temporal information is acquired is worthwhile to investigate.

In Chapter 5, as compared to the previous methods in which either invariant features are extracted or a distance metric is explored, we used a reference set is utilized to transfer the matching problem from an appearance space to a reference space. The re-identification is achieved by matching the reference descriptors (RDs) generated with the reference set and the matching results are improved by a re-ranking step using image saliency information. Experiments on real-world dataset showed that our method achieves state-of-the-art performance with a simple processing flow. Since the current approach is single image based, we would like to extend it to multi-image video based re-identification. In addition, we would like to perform theoretical analysis on the selection of the reference set.

In Chapter 6, we aimed at solving person re-identification in a different way. In this method, we developed a feature representation based on sparse representation with L2 regularization. Extensive experiments on both image based and video based datasets showed that using this novel representation we achieved state-of-the-art performance.

The choice of L2 regularization term was justified by the experimental results in which we showed that sparse representation with L2 regularization achieved superior performance compared to the baseline methods with a single L1 or L2 regularization term. Future work will extend the current framework to a larger number of cameras. Furthermore, we would like to combine the visual features together with other traits such as gait to improve the re-identification accuracy.

Bibliography

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037 –2041, dec. 2006.

[2] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila. Recognition of blurred faces using Local Phase Quantization. In *19th International Conference on Pattern Recognition (ICPR)*, pages 1 –4, Dec. 2008.

[3] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with Local Binary Patterns. In *European Conference on Computer Vision*, pages 469–481, 2004.

[4] L. An and B. Bhanu. Improved image super-resolution by support vector regression. In *IEEE International Joint Conference on Neural Networks*, pages 696–700, 2011.

[5] L. An, B. Bhanu, and Songfan Yang. Boosting face recognition in real-world surveillance videos. In *IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pages 270–275, 2012.

[6] L. An, M. Kafai, and B. Bhanu. Dynamic bayesian network for unconstrained face recognition in surveillance camera networks. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 3(2):155–164, 2013.

[7] L. An, N. Thakoor, and B. Bhanu. Vehicle logo super-resolution by canonical correlation analysis. In *IEEE International Conference on Image Processing*, 2012.

[8] Le An, B. Bhanu, and Songfan Yang. Face recognition in multi-camera surveillance videos. In *International Conference on Pattern Recognition (ICPR)*, pages 2885–2888, 2012.

[9] Le An and Bir Bhanu. Face image super-resolution using 2D CCA. *Signal Processing*, 2013.

[10] Le An, Bir Bhanu, and Songfan Yang. Unified face representation for individual recognition in surveillance videos. Augmented Vision and Reality, pages 1–14. Springer Berlin Heidelberg, 2013.

[11] Le An, Xiaojing Chen, M. Kafai, Songfan Yang, and B. Bhanu. Improving person re-identification by soft biometrics based reranking. In *ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, pages 1–6, Oct 2013.

[12] Le An, M. Kafai, Songfan Yang, and B. Bhanu. Reference-based person re-identification. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 244–249, Aug 2013.

[13] Le An, Mehran Kafai, and Bir Bhanu. Face recognition in multi-camera surveillance videos using dynamic bayesian network. In *Sixth ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, Oct. 2012.

[14] Le An, Songfan Yang, and Bir Bhanu. Efficient smile detection by extreme learning machine. *Neurocomputing*, 2014.

[15] S.R. Arashloo and J. Kittler. Energy normalization for pose-invariant face recognition based on MRF model image matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1274–1280, June 2011.

[16] Slawomir Bak, Guillaume Charpiat, Etienne Corvee, Francois Bremond, and Monique Thonnat. Learning to match appearances by correlations in a covariance metric space. In *European Conference on Computer Vision (ECCV)*, pages 806–820, October 2012.

[17] Oren Barkan, Jonathan Weill, Lior Wolf, and Hagai Aronowitz. Fast high dimensional vector multiplication face recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1960–1967, Dec 2013.

[18] Connelly Barnes, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. The generalized PatchMatch correspondence algorithm. In *European Conference on Computer Vision (ECCV)*, 2010.

[19] M. Bäuml, K. Bernardin, M. Fischer, H.K. Ekenel, and R. Stiefelhagen. Multi-pose face recognition for person retrieval in camera networks. In *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 441 –447, Sept. 2010.

[20] Apurva Bedagkar-Gala and Shishir K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270 – 286, 2014.

[21] Samy Bengio Bengio, Johnny Marithoz, and Mikaela Keller. The expected performance curve. In *22nd International Conference on Machine Learning*, pages 9–16, 2005.

[22] Thomas Berg and Peter Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. In *Proceedings of the British Machine Vision Conference*, pages 129.1–129.11. BMVA Press, 2012.

[23] S. Biswas, G. Aggarwal, and P.J. Flynn. Face recognition in low-resolution videos using learning-based likelihood measurement model. In *International Joint Conference on Biometrics (IJCB)*, pages 1 –7, Oct. 2011.

[24] D. Bouchaffra. Topological dynamic Bayesian networks: Application to human face identification across ages. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–8, June 2010.

[25] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–275 – I–282 Vol.1, june-2 july 2004.

[26] Xiaojing Chen, Le An, and B. Bhanu. Improving large-scale face image retrieval using multi-level features. In *IEEE International Conference on Image Processing (ICIP)*, pages 4367–4371, Sept 2013.

[27] Xiaojing Chen, Le An, and B. Bhanu. Reference set based appearance model for tracking across non-overlapping cameras. In *ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, 2013.

[28] Dong Seon Cheng, Marco Cristan, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conference (BMVC)*, pages 68.1–68.11, 2011.

[29] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

[30] N.M. Correa, T. Adali, Yi-Ou Li, and V.D. Calhoun. Canonical correlation analysis for data fusion and group inferences. *IEEE Signal Processing Magazine*, 27(4):39 –50, 2010.

[31] Zhen Cui, Wen Li, Dong Xu, Shiguang Shan, and Xilin Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3554–3561, June 2013.

[32] Zhen Cui, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Image sets alignment for video-based face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2626 –2633, June 2012.

[33] Shengyang Dai, Mei Han, Wei Xu, Ying Wu, Yihong Gong, and A.K. Katsaggelos. Softcuts: A soft edge smoothness prior for color image super-resolution. *IEEE Transactions on Image Processing*, 18(5):969 –981, may 2009.

[34] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4):636 –650, apr 2000.

[35] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *International Conference on Machine learning (ICML)*, pages 209–216, 2007.

[36] O. Dèniz, G. Bueno, J. Salido, and F. De la Torre. Face recognition using histograms of oriented gradients. *Pattern Recognition Letters*, 32(12):1598 – 1603, 2011.

[37] Mert Dikmen, Emre Akbas, Thomas S. Huang, and Narendra Ahuja. Pedestrian recognition with a learned metric. In *Asian conference on Computer vision (ACCV)*, pages 501–512, 2011.

[38] Weisheng Dong, D. Zhang, Guangming Shi, and Xiaolin Wu. Nonlocal back-projection for adaptive image enlargement. In *IEEE International Conference on Image Processing*, pages 349–352, 2009.

[39] Gianfranco Doretto, Thomas Sebastian, Peter H. Tu, and Jens Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing*, pages 127–151, 2011.

[40] Robert P. W. Duin and Elbieta Pkalska. The dissimilarity space: Bridging structural and statistical pattern recognition. *Pattern Recognition Letter*, 33(7):826–832, May 2012.

[41] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[42] M. Elad and A. Feuer. Super-resolution reconstruction of image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):817 –834, September 1999.

[43] Wei Fan, Yunhong Wang, and Tieniu Tan. Video-based face recognition using Bayesian inference model. In Takeo Kanade, Anil Jain, and Nalini Ratha, editors, *Audio- and Video-Based Biometric Person Authentication*, volume 3546 of *Lecture Notes in Computer Science*, pages 122–130. Springer Berlin / Heidelberg, 2005.

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

[45] William T. Freeman, Egon C. Pasztor, and Owen T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40:25–47, 2000.

[46] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. The CAS-PEAL large-scale chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 38(1):149 –161, 2008.

[47] Xinbo Gao, Kaibing Zhang, Dacheng Tao, and Xuelong Li. Image super-resolution with sparse neighbor embedding. *IEEE Transactions on Image Processing,*, 21(7):3194–3205, 2012.

[48] Xinbo Gao, Kaibing Zhang, Dacheng Tao, and Xuelong Li. Joint learning for single-image super-resolution via a coupled constraint. *IEEE Transactions on Image Processing,*, 21(2):469–480, 2012.

[49] A. Giachetti and N. Asuni. Real time artifact-free image upscaling. *IEEE Transactions on Image Processing*, PP(99):1, 2011.

[50] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors. *Person Re-Identfication*. Advances in Computer Vision and Pattern Recognition. Springer London, 2014.

[51] Zhenkun Gou and Colin Fyfe. A canonical correlation neural network for multicollinearity and functional data. *Neural Networks*, 17(2):285 – 293, 2004.

[52] Douglas Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, Sept. 2007.

[53] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision (ECCV)*, pages 262–275, 2008.

[54] Mislav Grgic, Kresimir Delac, and Sonja Grgic. SCface — surveillance cameras face database. *Multimedia Tools Appl.*, 51(3):863–879, February 2011.

[55] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 498–505, 2009.

[56] Yanlin Guo, Ying Shan, H. Sawhney, and Rakesh Kumar. PEET: Prototype embedding and embedding transition for matching vehicles over disparate viewpoints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007.

[57] A. Gyaourova and A. Ross. Index codes for multibiometric pattern retrieval. *IEEE Transactions on Information Forensics and Security*, 7(2):518–529, April 2012.

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

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

[60] David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16:2639–2664, December 2004.

[61] J. Harguess, Changbo Hu, and J.K. Aggarwal. Fusing face recognition from multiple cameras. In *IEEE Workshop on Applications of Computer Vision (WACV)*, pages 1 –7, Dec. 2009.

[62] Xiaofei He, Shuicheng Yan, Yuxiao Hu, P. Niyogi, and Hong-Jiang Zhang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328 –340, march 2005.

[63] P.H. Hennings-Yeomans, S. Baker, and B.V.K.V. Kumar. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[64] Guillaume Heusch and Sébastien Marcel. Bayesian networks to combine intensity and color information in face recognition. In *Proceedings of the Third International Conference on Advances in Biometrics*, ICB '09, pages 414–423, Berlin, Heidelberg, 2009. Springer-Verlag.

[65] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian Conference on Image analysis (SCIA)*, pages 91–102, 2011.

[66] Martin Hirzer, Peter M. Roth, Martin Köstinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision (ECCV)*, pages 780–793, 2012.

[67] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):pp. 321–377, 1936.

[68] Chao-Kuei Hsieh, Shang-Hong Lai, and Yung-Chang Chen. Expression-invariant face recognition with constrained optical flow warping. *IEEE Transactions on Multimedia*, 11(4):600–610, 2009.

[69] Weiming Hu, Min Hu, Xue Zhou, Tieniu Tan, Jianguang Lou, and S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):663–671, April 2006.

[70] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07–49, University of Massachusetts, Amherst, Oct. 2007.

[71] Hua Huang, Huiting He, Xin Fan, and Junping Zhang. Super-resolution of human face image using canonical correlation analysis. *Pattern Recognition*, 43(7):2532 – 2543, 2010.

[72] Hua Huang, Huiting He, Xin Fan, and Junping Zhang. Super-resolution of human face image using canonical correlation analysis. *Pattern Recognition*, 43(7):2532 – 2543, 2010.

[73] Hua Huang and Ning Wu. Fast facial image super-resolution via local linear transformations for resource-limited applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(10):1363 –1377, oct. 2011.

[74] Sibt Ul Hussain, Thibault Napolon, and Frederic Jurie. Face recognition using local quantized patterns. In *Proceedings of the British Machine Vision Conference*, pages 99.1–99.11. BMVA Press, 2012.

[75] Omar Javed, Khurram Shafique, Zeeshan Rasheed, and Mubarak Shah. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146 – 162, 2008.

[76] Omar Javed, Khurram Shafique, Zeeshan Rasheed, and Mubarak Shah. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146 – 162, 2008.

[77] M. Kafai and B. Bhanu. Dynamic Bayesian networks for vehicle classification in video. *IEEE Transactions on Industrial Informatics*, 8(1):100 –109, Feb. 2012.

[78] M. Kafai, B. Bhanu, and Le An. Cluster-classification bayesian networks for head pose estimation. In *International Conference on Pattern Recognition (ICPR)*, pages 2869–2872, Nov 2012.

[79] M. Kafai, K. Eshghi, and B. Bhanu. Discrete cosine transform locality-sensitive hashes for face retrieval. *IEEE Transactions on Multimedia*, 16(4):1090–1103, June 2014.

[80] M.I. Khedher, M.A. El Yacoubi, and B. Dorizzi. Multi-shot surf-based person re-identification via sparse representation. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 159–164, Aug 2013.

[81] M. Köstinger, M. Hirzer, P. Wohlhart, P.M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2288–2295, June 2012.

[82] B.G.V. Kumar and R. Aravind. Face hallucination using OLPP and kernel ridge regression. In *IEEE International Conference on Image Processing*, pages 353 –356, 2008.

[83] Cheng-Hao Kuo, S. Khamis, and V. Shet. Person re-identification using semantic color names and rankboost. In *IEEE Workshop on Applications of Computer Vision (WACV)*, pages 281–287, 2013.

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

[85] Ryan Layne, Tim Hospedales, and Shaogang Gong. Person re-identification by attributes. In *British Machine Vision Conference (BMVC)*, pages 24.1–24.11, 2012.

[86] Kuang-Chih Lee, J. Ho, Ming-Hsuan Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 313 – 320, June 2003.

[87] Sun Ho Lee and Seungjin Choi. Two-dimensional canonical correlation analysis. *IEEE Signal Processing Letters,*, 14(10):735 –738, 2007.

[88] S. E. Leurgans, R. A. Moyeed, and B. W. Silverman. Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(3):pp. 725–740, 1993.

[89] Annan Li, Shiguang Shan, Xilin Chen, and Wen Gao. Maximizing intra-individual correlations for face recognition across pose differences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 605 –611, june 2009.

[90] Haoxiang Li, Gang Hua, Zhe Lin, J. Brandt, and Jianchao Yang. Probabilistic elastic matching for pose variant face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3499–3506, June 2013.

[91] Peng Li, Yun Fu, Umar Mohammed, James H. Elder, and Simon J.D. Prince. Probabilistic models for inference about identity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):144–157, 2012.

[92] Qun Li, Honggang Zhang, Jun Guo, B. Bhanu, and Le An. Reference-based scheme combined with K-SVD for scene image categorization. *IEEE Signal Processing Letters*, 20(1):67–70, 2013.

[93] S.Z. Li, RuFeng Chu, Shengcai Liao, and Lun Zhang. Illumination invariant face recognition using near-infrared images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):627–639, 2007.

[94] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3594–3601, June 2013.

[95] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *Asian conference on Computer vision (ACCV)*, pages 31–44, 2012.

[96] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number 152-159, June 2014.

[97] Weixin Li, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32, Jan 2014.

[98] Chia-Te Liao, Shu-Fang Wang, Yun-Jen Lu, and Shang-Hong Lai. Video-based face recognition based on view synthesis from 3d face model reconstructed from a single image. In *IEEE International Conference on Multimedia and Expo*, pages 1589 –1592, Apr. 2008.

[99] Shengcai Liao, Anil K. Jain, and Stan Z. Li. Partial face recognition: Alignment-free approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1193–1205, 2013.

[100] Ce Liu, Heung-Yeung Shum, and Chang-Shui Zhang. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–192 – I–198 vol.1, 2001.

[101] Chengjun Liu. Discriminant analysis and similarity measure. *Pattern Recognition*, 47(1):359 – 367, 2014.

[102] Chunxiao Liu, Shaogang Gong, and Chen Change Loy. On-the-fly feature importance mining for person re-identification. *Pattern Recognition*, 47(4):1602 – 1615, 2014.

[103] Chunxiao Liu, Shaogang Gong, ChenChange Loy, and Xinggang Lin. Person re-identification: What features are important? In *European Conference on Computer Vision Workshops and Demonstrations*, pages 391–401, 2012.

[104] Chunxiao Liu, Chen Change Loy, Shaogang Gong, and Guijin Wang. Pop: Person re-identification post-rank optimisation. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[105] Wei Liu, Zhifeng Li, and Xiaoou Tang. Spatio-temporal embedding for statistical face recognition from video. In *European Conference on Computer Vision*, pages 374–388. 2006.

[106] Wei Liu, Dahua Lin, and Xiaoou Tang. Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 478 – 484 vol. 2, 2005.

[107] Xiaoming Liu and Tsuhan Cheng. Video-based face recognition using adaptive hidden Markov models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 340 – 345, June 2003.

[108] B.C. Lovell, S. Chen, A. Bigdeli, E. Berglund, and C. Sanderson. On intelligent surveillance systems and face recognition for mass transport security. In *International Conference on Control, Automation, Robotics and Vision*, pages 713–718, Dec. 2008.

[109] Chen Change Loy, Chunxiao Liu, and Shaogang Gong. Person re-identification by manifold ranking. In *IEEE International Conference on Image Processing (ICIP)*, pages 3567–3571, Sept 2013.

[110] Yui Man Lui, D. Bolme, B.A. Draper, J.R. Beveridge, G. Givens, and P.J. Phillips. A meta-analysis of face recognition covariates. In *IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, pages 1 –8, Sept. 2009.

[111] Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian parsing via deep decompositional network. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2648–2655, Dec 2013.

[112] Bingpeng Ma, Yu Su, and Frederic Jurie. BiCov: a novel image representation for person re-identification and face verification. In *British Machine Vision Conference (BMVC)*, pages 57.1–57.11, 2012.

[113] Xiang Ma, Junping Zhang, and Chun Qi. Hallucinating face by position-patch. *Pattern Recognition*, 43:2224–2236, June 2010.

[114] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *International Conference on Machine Learning (ICML)*, pages 689–696, 2009.

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

[116] N. Martinel, C. Micheloni, and C. Piciarelli. Learning pairwise feature dissimilarities for person re-identification. In *ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, 2013.

[117] Tao Meng and Mei-Ling Shyu. Leveraging concept association network for multimedia rare concept mining and retrieval. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 860–865, July 2012.

[118] Tao Meng and Mei-Ling Shyu. Model-driven collaboration and information integration for enhancing video semantic concept detection. In *IEEE 13th International Conference on Information Reuse and Integration (IRI)*, pages 144–151, August 2012.

[119] A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2666–2672, June 2012.

[120] Marco K. Müller, Michael Tremer, Christian Bodenstein, and Rolf P. Würtz. Learning invariant face recognition from examples. *Neural Networks*, 41:137 – 146, 2013.

[121] P. Nagesh and Baoxin Li. A compressive sensing approach for expression-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1518–1525, 2009.

[122] Ara V. Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy. Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP J. Appl. Signal Process.*, 2002(1):1274–1288, January 2002.

[123] A.V. Nefian. Embedded Bayesian networks for face recognition. In *IEEE International Conference on Multimedia and Expo*, volume 2, pages 133–136 vol.2, 2002.

[124] K. Nguyen, S. Sridharan, S. Denman, and C. Fookes. Feature-domain super-resolution framework for gabor-based face and iris recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2642 –2649, june 2012.

[125] A.A. Nielsen. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE Transactions on Image Processing*, 11(3):293 –305, mar 2002.

[126] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, Jul 2002.

[127] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311 – 3325, 1997.

[128] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 20(3):21 – 36, may 2003.

[129] Unsang Park, Hong Chen, and A.K. Jain. 3D model-assisted face recognition in video. In *The 2nd Canadian Conference on Computer and Robot Vision*, pages 322 – 329, May 2005.

[130] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3318–3325, June 2013.

[131] N. Poh, Chi Ho Chan, J. Kittler, S. Marcel, C. McCool, E.A. Rúa, J.L.A. Castro, M. Villegas, R. Paredes, V. Štruc, N. Pavešić and, A.A. Salah, Hui Fang, and N. Costen. An evaluation of video-to-video face verification. *IEEE Transactions on Information Forensics and Security*, 5(4):781 –801, Dec. 2010.

[132] U. Prabhu, Jingu Heo, and M. Savvides. Unconstrained pose-invariant face recognition using 3D generic elastic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1952–1961, Oct. 2011.

[133] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by support vector ranking. In *British Machine Vision Conference (BMVC)*, pages 21.1–21.11, 2010.

[134] Lawrence K. Saul and Sam T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, December 2003.

[135] Florian Schroff, Tali Treibitz, David Kriegman, and Serge Belongie. Pose, illumination and expression invariant pairwise face-similarity measure via doppelgänger list comparison. In *IEEE International Conference on Computer Vision (ICCV)*, Barcelona, 2011.

[136] R.R. Schultz and R.L. Stevenson. Extraction of high-resolution frames from video sequences. *IEEE Transactions on Image Processing*, 5(6):996 –1011, June 1996.

[137] Caifeng Shan. Face recognition and retrieval in video. In Dan Schonfeld, Caifeng Shan, Dacheng Tao, and Liang Wang, editors, *Video Search and Mining*. Springer Berlin / Heidelberg, 2010.

[138] Shiguang Shan, Yizheng Chang, Wen Gao, Bo Cao, and Peng Yang. Curse of mis-alignment in face recognition: problem and a novel mis-alignment learning solution. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 314–320, 2004.

[139] Abhishek Sharma, Murad Al Haj, Jonghyun Choi, Larry S. Davis, and David W. Jacobs. Robust pose invariant face recognition using coupled latent space discriminant analysis. *Computer Vision and Image Understanding*, 116(11):1095 – 1110, 2012.

[140] A. Shnayderman, A. Gusev, and A.M. Eskicioglu. An SVD-based grayscale image quality measure for local and global assessment. *IEEE Transactions on Image Processing*, 15(2):422 –429, feb. 2006.

[141] J. Stallkamp, H.K. Ekenel, and R. Stiefelhagen. Video-based face recognition on real-world data. In *IEEE 11th International Conference on Computer Vision (ICCV)*, pages 1 –8, Oct. 2007.

[142] Quan-Sen Sun, Sheng-Gen Zeng, Pheng-Ann Heng, and De-Sen Xia. Feature fusion method based on canonical correlation analysis and handwritten character

recognition. In *Control, Automation, Robotics and Vision Conference*, volume 2, pages 1547 – 1552 Vol. 2, dec. 2004.

[143] Tingkai Sun and Songcan Chen. Locality preserving CCA with applications to data visualization and pose estimation. *Image and Vision Computing*, 25(5):531 – 543, 2007.

[144] H. Takeda, S. Farsiu, and P. Milanfar. Kernel regression for image processing and reconstruction. *IEEE Transactions on Image Processing*, 16(2):349 –366, 2007.

[145] Xiaoyang Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing*, 19(6):1635–1650, 2010.

[146] Dapeng Tao, Lianwen Jin, Yongfei Wang, Yuan Yuan, and Xuelong Li. Person re-identification by regularized smoothing kiss metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(10):1675–1685, 2013.

[147] MarshallF. Tappen and Ce Liu. A Bayesian approach to alignment-based image hallucination. In *European Conference on Computer Vision*, pages 236–249, 2012.

[148] Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra*. SIAM: Society for Industrial and Applied Mathematics, 1997.

[149] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, January 1991.

[150] M. Vernier, N. Martinel, C. Micheloni, and G. L. Foresti. Remote feature learning for mobile re-identification. In *ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, 2013.

[151] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People re-identification in surveillance and forensics: a survey. *ACM Computing Surveys*, 46(2):29:1–29:3, November 2013.

[152] Paul Viola and Michael J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, May 2004.

[153] Haixian Wang. Local two-dimensional canonical correlation analysis. *IEEE Signal Processing Letters*, 17(11):921 –924, nov. 2010.

[154] Shenlong Wang, D. Zhang, Yan Liang, and Quan Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2216–2223, June 2012.

[155] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2216 –2223, june 2012.

[156] Tao Wang, Qian Diao, Yimin Zhang, Gang Song, Chunrong Lai, and G. Bradski. A dynamic Bayesian network approach to multi-cue based visual tracking. In *International Conference on Pattern Recognition*, volume 2, pages 167–170 Vol.2, Aug. 2004.

[157] Xiaogang Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, Oct. 2007.

[158] Xiaogang Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *IEEE 11th International Conference on Computer Vision (ICCV)*, pages 1–8, Oct 2007.

[159] Xiaogang Wang and Xiaoou Tang. Hallucinating face by eigentransformation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 35(3):425 –434, 2005.

[160] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955 –1967, nov. 2009.

[161] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600 –612, april 2004.

[162] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, June 2009.

[163] Lior Wolf, Tal Hassner, and Yaniv Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1978–1990, Oct. 2011.

[164] Yongkang Wong, Shaokang Chen, S. Mau, C. Sanderson, and B.C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 74 –81, June 2011.

[165] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, Feb 2009.

[166] Tao Xiang and Shaogang Gong. Visual learning given sparse data of unknown complexity. In *Tenth IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 701 – 708, Oct. 2005.

[167] Binglong Xie, V. Ramesh, Ying Zhu, and T. Boult. On channel reliability measure training for multi-camera face recognition. In *IEEE Workshop on Applications of Computer Vision (WACV)*, page 41, Feb. 2007.

[168] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *European Conference on Computer Vision (ECCV)*, pages 1–16, 2014.

[169] Xiaona Xu and Zhichun Mu. Feature fusion method based on KCCA for ear and profile face based multimodal recognition. In *IEEE International Conference on Automation and Logistics*, pages 620 –623, aug. 2007.

[170] Yilei Xu, A. Roy-Chowdhury, and K. Patel. Pose and illumination invariant face recognition in video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop on Biometrics (CVPRW)*, pages 1 –7, June 2007.

[171] Shuicheng Yan, Huan Wang, Jianzhuang Liu, Xiaoou Tang, and T.S. Huang. Misalignment robust face recognition. *IEEE Transactions on Image Processing*, 19:1087–1096, 2010.

[172] Chih-Yuan Yang, Jia-Bin Huang, and Ming-Hsuan Yang. Exploiting self-similarities for single frame super-resolution. In *Asian Conference on Computer Vision*, pages 1807–1818, 2010.

[173] Jian Yang, D. Zhang, A.F. Frangi, and Jing yu Yang. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):131 –137, 2004.

[174] Jianchao Yang, Zhaowen Wang, Zhe Lin, S. Cohen, and T. Huang. Coupled dictionary training for image super-resolution. *IEEE Transactions on Image Processing*, 21(8):3467–3478, Aug 2012.

[175] Jianchao Yang, J. Wright, T. Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1 –8, 2008.

[176] Songfan Yang, Le An, B. Bhanu, and N. Thakoor. Improving action units recognition using dense flow-based face registration in video. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, April 2013.

[177] Songfan Yang and Bir Bhanu. Understanding discrete facial expressions in video using an emotion avatar image. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4):980–992, 2012.

[178] Weilong Yang, Dong Yi, Zhen Lei, Jitao Sang, and S.Z. Li. 2D-3D face matching using CCA. In *IEEE International Conference on Automatic Face Gesture Recognition*, pages 1 –6, 2008.

[179] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and StanZ. Li. Salient color names for person re-identification. In *European Conference on Computer Vision (ECCV)*, pages 536–551, 2014.

[180] Qi Yin, Xiaoou Tang, and Jian Sun. An associate-predict model for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '11, pages 497–504, Washington, DC, USA, 2011. IEEE Computer Society.

[181] Liyan Zhang, Dmitri V. Kalashnikov, and Sharad Mehrotra. A unified framework for context assisted face clustering. In *3rd ACM Conference on International Conference on Multimedia Retrieval*, pages 9–16, 2013.

[182] Liyan Zhang, Dmitri V. Kalashnikov, Sharad Mehrotra, and Ronen Vaisenberg. Context-based person identification framework for smart video surveillance. *Machine Vision and Applications*, pages 1–15, 2013.

[183] Liyan Zhang, DmitriV. Kalashnikov, and Sharad Mehrotra. Context-assisted face clustering framework with human-in-the-loop. *International Journal of Multimedia Information Retrieval*, 3(2):69–88, 2014.

[184] Wenchao Zhang, Shiguang Shan, Wen Gao, Xilin Chen, and Hongming Zhang. Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition. In *Tenth IEEE International Conference onComputer Vision (ICCV)*, volume 1, pages 786 – 791, Oct. 2005.

[185] Yongmian Zhang and Qiang Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):699 –714, May 2005.

[186] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person re-identification by salience matching. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[187] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3586–3593, June 2013.

[188] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number 144-151, June 2014.

[189] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Transfer re-identification: From person to set-based verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2650–2657, June 2012.

[190] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2013.

[191] Wenming Zheng, Xiaoyan Zhou, Cairong Zou, and Li Zhao. Facial expression recognition using kernel canonical correlation analysis (KCCA). *IEEE Transactions on Neural Networks*, 17(1):233 –238, jan. 2006.

[192] Xiaoli Zhou and B. Bhanu. Integrating face and gait for human recognition at a distance in video. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(5):1119–1137, Oct 2007.

[193] Yueting Zhuang, Jian Zhang, and Fei Wu. Hallucinating faces: LPH super-resolution and neighbor reconstruction for residue compensation. *Pattern Recognition*, 40(11):3178 – 3194, 2007.

[194] A. Zomet, A. Rav-Acha, and S. Peleg. Robust super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–645 – I–650 vol.1, 2001.

[195] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.