University of Exeter
Department of Computer Science

Some Topics on Similarity Metric Learning

Qiong Cao

June 2015

Supervised by Dr. Yiming Ying

Submitted by Qiong Cao, to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Computer Science , June 2015.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

(signature) .....

Abstract

The success of many computer vision problems and machine learning algorithms critically depends on the quality of the chosen distance metrics or similarity functions. Due to the fact that the real-data at hand is inherently task- and data-dependent, learning an appropriate distance metric or similarity function from data for each specific task is usually superior to the default Euclidean distance or cosine similarity. This thesis mainly focuses on developing new metric and similarity learning models for three tasks: unconstrained face verification, person re-identification and kNN classification.

Unconstrained face verification is a binary matching problem, the target of which is to predict whether two images/videos are from the same person or not. Concurrently, person re-identification handles pedestrian matching and ranking across non-overlapping camera views. Both vision problems are very challenging because of the large transformation differences in images or videos caused by pose, expression, occlusion, problematic lighting and viewpoint.

To address the above concerns, two novel methods are proposed. Firstly, we introduce a new dimensionality reduction method called Intra-PCA by considering the robustness to large transformation differences. We show that Intra-PCA significantly outperforms the classic dimensionality reduction methods (e.g. PCA and LDA). Secondly, we propose a novel regularization framework called Sub-SML to learn distance metrics and similarity functions for unconstrained face verification and person re-identification. The main novelty of our formulation is to incorporate both the robustness of Intra-PCA to large transformation variations and the discriminative power of metric and similarity learning, a property that most existing methods do not hold.

Working with the task of kNN classification which relies a distance metric to identify the nearest neighbors, we revisit some popular existing methods for metric learning and develop a general formulation called DML$_p$ for learning a distance metric from data. To obtain the optimal solution, a gradient-based optimization algorithm is proposed which only needs the computation of the largest eigenvector of a matrix per iteration.

Although there is a large number of studies devoted to metric/similarity learning based on different objective functions, few studies address the generalization analysis of such methods. We describe a novel approch for generalization analysis of metric/similarity learning which can deal with general matrix regularization terms including the Frobenius norm, sparse L^1-norm, mixed $(2,1)$-norm and trace-norm.

The novel models developed in this thesis are evaluated on four challenging databases: the Labeled Faces in the Wild dataset for unconstrained face verification in still images; the YouTube Faces database for video-based face verification in the wild; the Viewpoint Invariant Pedestrian Recog-

nition database for person re-identification; the UCI datasets for kNN classification. Experimental results show that the proposed methods yield competitive or state-of-the-art performance.

Acknowledgements

I would very much like to thank the following people without the help and support of whom this work would not have been possible.

First and foremost, I would like to thank my supervisor Yiming Ying who introduced me to the fields of machine learning and computer vision. He has been a constant source of wisdom, encouragement and patience during the course of my PhD. His dedication to producing top research is an inspiration.

I would also like to thank my current supervisor Richard Everson for the great discussions, guidance and comments on my thesis. An important thank as well to Peng Li from Aurora CS Ltd for the collaborations and helpful suggestions.

Many thanks to my colleagues for their kind academic support, including Jacqueline Christmas, Carlos Martinez-Ortiz, Alma Rahat, Philip Sansom and David Walker. I also thank all my friends in the UK and China for their consistent love.

Finally, deepest thanks to my parents Yafeng Cao and Gaishu Pei, my young brother Xu Cao for being there through every step of my life. Their unconditional love and support help me go through the tough time of my PhD study.

Contents

List of Tables

List of Figures

Publications

Cao, Q., Ying, Y., and Li, P. (2012). Distance metric learning revisited. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*.

Cao, Q., Ying, Y., and Li, P. (2013). Similarity metric learning for face recognition. In *Computer Vision (ICCV), IEEE International Conference on.* IEEE.

Cao, Q., Guo, Z., and Ying, Y. (2015). Generalization Bounds for Metric and Similarity Learning. Accepted for *Machine Learning Journal*.

Cao, Q., Pontil, M., Ying, Y and Li P. (2015). Similarity Metric Learning for Unconstrained Face Recognition and Person Re-Identification. For submission to *Image Processing, IEEE Transactions on.* IEEE.

1 Introduction

1.1 Objectives

Distance metrics and similarity functions are fundamental concepts in computer vision and machine learning. For instance, in computer vision, most face recognition methods rely on a similarity function to identify or verify one or more persons from a database of facial images/vidoes. Most of work in person re-identification depends on a similarity function to match observations of individuals across disjoint camera views. In machine learning, k-nearest-neighbour (kNN) classifier depends on a distance metric to identify the nearest neighbors for classification. A common choice of distance metric or similarity function is the Euclidean distance or cosine similarity which gives equal weights to all features. However, it ignores the fact that the real-data at hand is inherently task- and data-dependent. Often, domain experts adjust them manually which is not a robust approach.

Metric and similairty learning aim to learn an appropriate distance metric or similarity function explicitly from the available data for each specific task. Given some side information of constraints, the target of metric learning is to learn a distance metric such that the distances between similar pairs (i.e. from the same class) are small while the distances between dissimilar pairs (i.e. from different classes) are large. Most metric learning methods seek to learn the (squared) Mahalanobis distance defined, for any $x, t \in \mathbb{R}^d$, by $d_M(x,t) = (x-t)^T M(x-t)$, where M is a positive semi-definite (p.s.d.) matrix. Concurrently, similarity learning attempts to learn a similarity function such that it reports large scores for similar pairs and small scores for dissimilar pairs. Most of work in similarity learning focuses on the bilinear similarity function defined by $s_M(x,t) = x^T Mt$ or the cosine similarity $CS_M(x,t) = x^T Mt/\sqrt{x^T Mx}\sqrt{t^T Mt}$, where M is a positive semi-definite matrix. A full description of metric/similarity learning methods will be presented in Chapter 2.

The first objective of this thesis is to develop novel metric/similarity methods for the following three tasks.

Face recognition. Due to various real-life applications such as human-computer interaction, desktop login and biometrics, face recognition has remained an active topic in computer vision in the past decades. It can be divided into two categories: face identification and face verification. Face identification is a one-to-many matching task which attempts to recognize the identity of a query facial image/video from a set of gallery facial images/videos. In contrast, face verification is a one-to-one matching problem, the target of which is to predict whether two facial images/videos represent the same person or not. Face recognition under well-controlled conditions has been extensively studied over decades [Turk and Pentland, 1991a; Belhumeur et al., 1997; Moghaddam et al., 2000]. More recently, a lot of research effort has gone into unconstrained face verification

(e.g. [Wolf et al., 2008; Guillaumin et al., 2009]) where faces are captured under unconstrained conditions. This thesis mainly focuses on this unconstrained setting, see Section 1.2 for more details.

Person re-identification. Person re-identification has attracted a growing interest in the past few years in computer vision. It has important application in video surveillance such as pedestrian tracking, multi-camera event detection and person retrieval. Different from person identification, person re-identification is a process of recognizing if a person has been previously observed over a network of cameras. In particular, it aims to match and rank the images of pedestrians captured from a set of non-overlapping camera views at different locations and times. Many methods have been developed for person re-identification (e.g. [Gray et al., 2007; Zheng et al., 2011; Kostinger et al., 2012]), which advances this field.

Metric and similarity learning aim to learn an appropriate distance metrics or similarity function to compare pairs of examples, which provides a natural solution for the above vision tasks.

KNN classification. KNN classification is a supervised learning algorithm for classification which needs the label information of training data. Given the training samples along with their class labels, the goal of kNN classification is to find the class label for an unlabelled point (a query/test point). To do this, a distance metric is first used to identify the k-nearest neighbors of the test point from training data. The test point is then assigned to the class that is most frequent among its k-nearest neighbors.

For the tasks of unconstrained face verification and person re-identification, the images are usually represented by high dimensional vectors, which can lead to prohibitive computational cost. Therefore, the second objective of this thesis is to introduce a new dimensionality reduction method for the two tasks.

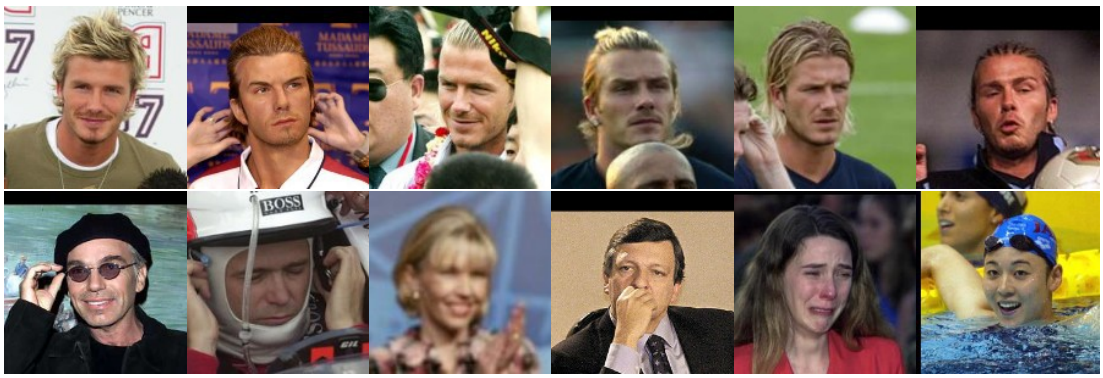
Although there is a large body of work devoted to metric/similarity learning based on different objective functions, less attention has been paid on the generalization ability of such models, i.e. their performance on unseen samples. The third objective of this thesis is thus to describe a novel approach for the generalization analysis of metric/similarity learning methods.

1.2 Research Challenges

In this section, several research challenges arising from unconstrained face verification and person re-identification are described.

Consider the task of unconstrained face verification in still images. Facial images are captured in the wild and often exhibit large transformation differences caused by all kinds of transformations such as pose, lighting, hairstyle and expression. This is the case for the popular the Labeled Faces in the Wild (LFW) dataset [Huang et al., 2007], example images of which are depicted in Figure 1.1a. Section 2.4.1 will provide a detailed description of the LFW data set.

Compared to image-based face verification in the wild, unconstrained face verification in videos is more difficult because video clips are generally recorded by amateurs and are in low quality.



(a) Example images from LFW database.

(b) Example frames from YTF database.

Figure 1.1: Example images/frames from the Labeled Faces in the Wild (LFW) dataset [Huang et al., 2007] and the YouTube Faces (YTF) database [Wolf et al., 2011a] show large transformation differences caused by pose, background, occlusion and problematic lighting: the top rows in (a) and (b) are images and frames from the same person and the bottom rows in (a) and (b) are images and frames from different persons.

Figure 1.2: Example images from the Viewpoint Invariant Pedestrian Recognition (VIPeR) database [Gray et al., 2007]. Images are taken across spatially disjoint cameras under varying illumination conditions and show large transformation differences caused by viewpoint, illumination and background.

Faces in the video clips often suffer from large transformation differences including motion blur, occlusion and problematic lighting, which can negatively influence the verification accuracy. Figure 1.1b shows the example frames from the YouTube Faces (YTF) database [Wolf et al., 2011a], a benchmark for unconstrained face verification in videos. An introduction of the YTF database will be presented in Section 2.4.2.

For the task of person re-identification, despite the best effort from computer vision researchers in the past five years, it remains largely unsolved. This is due to the fact that a person's appearances are captured in different camera views and often undergo significant transformation variations in

view angle, illumination and occlusion. For instance, a person appeared in frontal view under one camera may appear in back view under another camera. Example images from the Viewpoint Invariant Pedestrian Recognition (VIPeR) database [Gray et al., 2007] are illustrated in Figure 1.2. This database is the largest publicly available database for person re-identification, and a brief introduction of this dataset will be given in Section 2.4.3.

Overall, due to the large transformation differences existing in the above vision tasks, variations among images/videos from the same person can vary significantly and variations across images/videos from different persons also varies drastically. Such transformation differences can overwhelm the variations caused by the identity differences, which makes the problem of unconstrained face verification and person re-identification extremely challenging.

In addition, as will be shown in Section 2.4, the evaluation procedure for the above vision tasks typically assumes that the person identities in the training and test sets are exclusive. Therefore, the prediction is required to be of never-seen-before faces/pedestrians, which makes unconstrained face verification and person re-identification more challenging.

1.3 Limitations of Existing Methods

In this section, several limitations of existing dimensionality reduction and metric/similarity learning methods are identified as follows:

- For the tasks of unconstrained face verification and person re-identification, the following two limitations of the classic dimensionality reduction methods are identified. Firstly, dimensionality reduction models such as PCA or Eigenfaces [Turk and Pentland, 1991a], LDA or Fisherfaces [Belhumeur et al., 1997] and Bayesian face recognition [Moghaddam et al., 2000] have demonstrated promising results under well-controlled conditions. Unfortunately, when applied to unconstrained conditions, most of the above methods degenerate seriously (see [Li et al., 2012; Chen et al., 2012]). This is due to the large transformation differences described in Section 1.2. Secondly, the above models such as Fisherfaces and Bayesian face recognition are supervised models which need the label information. However, it is less common and challenging in real world compared to the case that only pairwise information (i.e. similar image-pairs/video-pairs and dissimilar image-pairs/video-pairs) is available while the label information is not provided.

- For the tasks of unconstrained face verification and person re-identification, two limitations of the current metric/similarity learning methods are described. Firstly, most of existing metric learning methods [Xing et al., 2003; JacobGoldberger and GeoffHinton, 2004; Globerson and Roweis, 2005; Weinberger et al., 2006; Davis et al., 2007; Torresani and Lee, 2007] deal with the specific task of improving kNN classification or clustering. However, it was observed in [Guillaumin et al., 2009; Zheng et al., 2011; Ying and Li, 2012] that directly applying such methods only yields a modest performance for unconstrained face verification and person re-identification. Secondly, existing metric/similarity learning methods [Xing et al., 2003; Weinberger et al., 2006; Davis et al., 2007; Guillaumin et al., 2009; Nguyen and Bai, 2011; Ying and Li, 2012] limit in that such methods mainly focus on

the discrimination of distance metrics or similarity functions, while little attention is paid on how to reduce the detrimental effect of the large transformation variations. Thus, the learned distance metrics or similarity functions may not be robust to the transformation variations.

- For the task of kNN classification, metric learning methods proposed by Xing et al. [2003] and Davis et al. [2007] learn distance metrics in a global way, i.e. they use all the pairwise information. However, previous studies [Weinberger et al., 2006; Shen et al., 2009; Ying and Li, 2012] show that metric learning methods using local pairwise information usually outperform methods using global one. In particular, this is reasonable for the task of kNN classification, the performance of which is influenced mostly by the data points that are close to the query examples. On the other hand, in [Xing et al., 2003] the projection gradient descent algorithm was employed to obtain the optimal solution. However, this algorithm is slow since it usually takes a large number of iterations to converge and needs the full eigen-decomposition of a matrix per iteration.

- For the generalization analysis of metric/similarity learning, we identify one limitation of the recent work [Jin et al., 2009]. It has been the first attempt to study the generalization analysis for metric learning using the concept of uniform stability [Bousquet and Elisseeff, 2002]. However, a drawback of this approach is that it only works for the matrix regularization term using Frobenius norm [Jin et al., 2009].

1.4 Thesis Contributions

This thesis proposes four novel approaches to address the above limitations of existing dimensionality reduction and metric/similarity learning methods. The contributions of this thesis are summarised as follows:

- For the task of unconstrained face verification and person re-identification, a new dimensionality reduction model called Intra-PCA is introduced. This new model is formulated by considering the robustness to large transformation differences in the setting that only pairwise information is provided while the label information is not available. It includes two steps. In the first step, WPCA (see Section 2.2.2 for details) is applied on the original images to reduce the noise. In the second step, to reduce the large transformation differences, the projection of the resultant images/videos to the intra-personal subspace by the whitening process is proposed. Furthermore, Intra-PCA is extended to video-based face verification in the wild. Experimental results on the Labeled Faces in the Wild (LFW) [Huang et al., 2007], the YouTube Faces (YTF) [Wolf et al., 2011a] and the Viewpoint Invariant Pedestrian Recognition (VIPeR) [Gray et al., 2007] databases show that Intra-PCA is superior to the classic dimensionality reduction methods such as PCA, LDA and Bayesian face recognition.

- For the tasks of unconstrained face verification and person re-identification, a novel regularization framework called Sub-SML is developed to learn distance metrics and similarity functions using pairwise information. Our learning objective is formulated by combining both the robustness to large transformation differences and the discriminative power of metric/similarity learning, a property that most of existing metric/similarity learning methods

[Xing et al., 2003; Weinberger et al., 2006; Davis et al., 2007; Guillaumin et al., 2009; Chechik et al., 2010; Shalit et al., 2010; Kan et al., 2011; Nguyen and Bai, 2011; Ying and Li, 2012] do not hold. Besides, the proposed formulation is a convex optimization problem, which allows us to apply existing optimization algorithms to efficiently find a global solution. This is, for instance, not the case for the current similarity learning model [Nguyen and Bai, 2011]. This framework is further extended to video-based face verification in the wild. Lastly, it is observed in the experiments that Sub-SML significantly outperforms the state-of-the-art metric/similarity learning methods and is competitive with or even outperforms the domain specific state-of-the-arts on the challenging LFW, YTF and VIPeR datasets.

- For the task of kNN classification, a new metric learning formulation called DML$_p$ is presented by recovering metric learning methods [Xing et al., 2003; Ying and Li, 2012]. The proposed formulation is proved to be convex. By further exploring its special structures, DML$_p$ is shown to be equivalent to a convex optimization over the spectrahedron, which enables us to directly employ the Frank-Wolfe algorithm [Frank and Wolfe, 1956] to obtain the optimal solution. In contrast to the optimization algorithm used by Xing et al. [2003], our proposed algorithm only needs the computation of the largest eigenvector of a matrix per iteration. Finally, it is shown in the experiments that DML$_p$ obtains competitive performance against the state-of-the-art metric learning methods on various UCI datasets for kNN classification. Besides, for the study of unconstrained face verification in still images, DML$_p$ outperforms metric learning methods [Xing et al., 2003; Ying and Li, 2012] and delivers comparable performance with the domain specific state-of-the-arts on the LFW dataset.

- For the generalization analysis of metric/similarity learning, a novel approach is described to deal with the general matrix regularization terms including the Frobenius norm [Jin et al., 2009], sparse L^1-norm [Rosales and Fung, 2006], mixed $(2,1)$-norm [Ying et al., 2009] and trace-norm [Ying et al., 2009; Shen et al., 2009]. It is shown that the generalization analysis for metric/similarity learning can be reduced to the estimation of the Rademacher complexity related to the specific matrix norm. Based on the above observation, the generalization bounds for metric/similarity learning with different matrix-norm regularizers can be derived. From our analysis, it is indicated that sparse metrc/similarity learning with L^1-norm regularization obtains significantly better generalization bounds than that with Frobenius-norm regularization, especially when dealing with the training data with high dimensionality. This novel generalization analysis develops and refines the techniques of Rademacher complexity analysis [Bartlett and Mendelson, 2003; Koltchinskii and Panchenko, 2002] and U-statistics [Clemençon et al., 2008; De la Pena and Giné, 1999].

1.5 Overview of Thesis

This thesis is organized as follows:

- Chapter 2 starts with a large survey of work on unconstrained face recognition and person re-identification by covering the literature on feature extraction techniques and similarity-based

approaches to which metric/similarity learning methods belong. Besides, three challenging databases are presented to evaluate our new models for unconstrained face verification and person re-identification.

- Chapter 3 studies the problems of unconstrained face verification and person re-identification by considering the robustness to large transformation variations. A novel dimensionality reduction model called Intra-PCA is introduced using pairwise information. The formulation of the proposed model and its extension to video-based face verification are described. Experimental study using the Labeled Faces in the Wild (LFW) [Huang et al., 2007], the YouTube Faces (YTF) [Wolf et al., 2011a] and the Viewpoint Invariant Pedestrian Recognition (VIPeR) [Gray et al., 2007] databases is provided.

  Some of the material in this chapter has been published in [Cao et al., 2013].

- Chapter 4 develops a novel regularization framework called Sub-SML to learn distance metrics and similarity functions for unconstrained face verification and person re-identification using pairwise information. Firstly, the formulation of the learning problem is described, followed by the derivation of its dual formulation. Then, an efficient optimization algorithm is designed. Furthermore, this framework is extended to video-based face verification. Lastly, the evaluation of Sub-SML on the LFW, YTF and VIPeR datasets is presented in the experiments.

  Some of the material in this chapter has been published in [Cao et al., 2013].

- Chapter 5 extends metric learning methods [Xing et al., 2003; Ying and Li, 2012] and proposes a general metric learning formulation DML$_p$. Various examples are illustrated. The convexity of the proposed formulation is proved and its equivalent formulation is further established. The Frank-Wolfe algorithm [Frank and Wolfe, 1956] is described to obtain the optimal solution. Experimental results on various UCI datasets for kNN classification are reported. In addition, the evaluation is done on the LFW database for unconstrained face verification in still images.

  This work has been published in [Cao et al., 2012].

- Chapter 6 describes a novel approach for the generalization analysis of metric/similarity learning with general matrix regularization terms including Frobenius norm, sparse L^1-norm, mixed $(2,1)$-norm and trace-norm. Followed by the development and refinement of the techniques of Rademacher complexity analysis [Bartlett and Mendelson, 2003; Koltchinskii and Panchenko, 2002] and U-statistics [Clemençon et al., 2008; De la Pena and Giné, 1999], the generalization bounds for metric/similarity learning with different matrix-norm regularizers are established.

  This work has been accepted for *Machine Learning Journal*.

- Chapter 7 concludes with a summary of the thesis and outlines a few possible directions for future work.

2 Literature Review

2.1 Introduction

Recently, a large amount of work has been devoted to addressing the tasks of face recognition and person re-identification. Face recognition has been extensively studied (e.g. [Turk and Pentland, 1991b; Belhumeur et al., 1997; Moghaddam et al., 2000; Wolf et al., 2008; Guillaumin et al., 2009; Taigman et al., 2009; Wolf et al., 2009b; Cox and Pinto, 2011; Wolf et al., 2011b,a; Li et al., 2012; Wolf and Levy, 2013]). Concurrently, progress has been made on person re-identification (see [Gray et al., 2007; Wang et al., 2007; Farenzena et al., 2010; Zheng et al., 2011; Kostinger et al., 2012; Mignon and Jurie, 2012]). Figure 2.1 illustrates the pipeline of face recognition and person re-identification which is commonly structured as a multi-stage process. Typically, it involves face/pedestrian detection systems, feature extraction techniques, similarity-based approaches and recognition. In this chapter, we review the literature on: 1) feature extraction techniques to extract the relevant features from the raw images/videos; 2) similarity-based approaches to learn similarity measures to compare pairs of images/videos.

Below, Section 2.2 reviews the literature on feature extraction techniques with feature representation methods in Section 2.2.1 and dimensionality reduction methods in Section 2.2.2. Section 2.3 covers the related work on similarity-based approaches, with an emphasis on metric learning approaches in Section 2.3.2 and similarity learning approaches in Section 2.3.3. Section 2.4 describes three benchmarks for the evaluation of the new models for unconstrained face verification and person re-identification. Section 2.5 concludes the chapter.

2.2 Feature Extraction Techniques

Feature extraction techniques aim to learn a transformation from the original image space (i.e. pixel space) to find the most compact and informative set of features for the specific vision task. It consists of two steps: feature representation and dimensionality reduction. Below, the feature representation methods are briefly surveyed in Section 2.2.1. Section 2.2.2 discusses existing methods on dimensionality reduction, which forms the ground of the new model that will be introduced in Chapter 3.

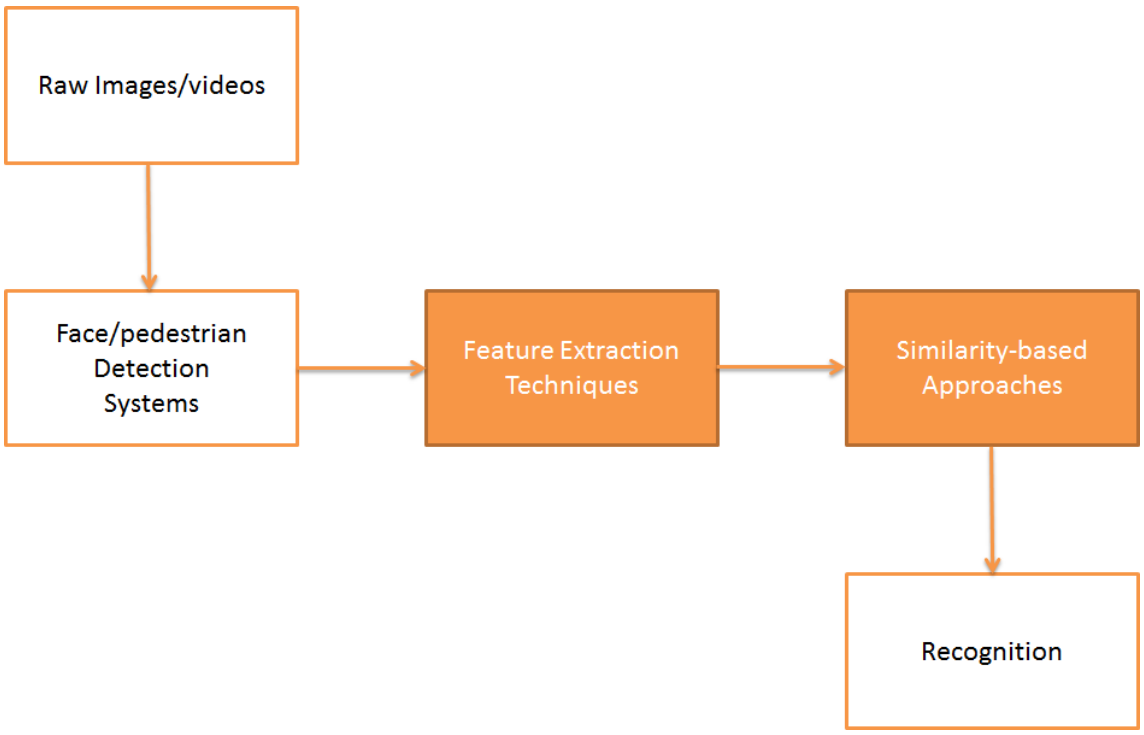


Figure 2.1: The pipeline of face recognition and person re-identification.

2.2.1 Feature Representation

Feature representation approaches for the vision tasks aim to obtain a good vectorial representation to represent a given raw image. This vectorial representation is usually referred to as feature or descriptor.

The simplest type of description for facial images is to directly use the intensity values of pixels in graylevel. Specifically, facial images are often stored as 2D arrays, the element of which is the intensity value of each pixel. This 2D array is then concatenated to form a single vector, which we refer to as Intensity. The dimensionality of the Intensity descriptor is often very high.

A popular texture-based descriptor for facial images is the Local Binary Pattern (LBP) [Ojala et al., 2002]. It detects the properties of the local micro textures (e.g. edges, spots and lines). The simple form of LBP is to extract a binary code for each pixel from a 3×3 neighborhood surrounding the pixel. By thresholding this neighborhood with the intensity value of the central pixel, an 8-bit code is obtained for each pixel. A histogram of each pixel is extracted and then concatenated to form a global description of the face. The invariance to the monotonic transformations of the gray-scale pixel values makes LBP robust to the illumination changes.

Wolf et al. [2008] employed the above binary patterns on the patch-based approaches (e.g. [Shechtman and Irani, 2007]) aiming to capture the properties of the local textures. Two patch-based facial descriptors were proposed: the Three-Patch Local Binary Patterns (TPLBP) and the Four-Patch Local Binary Patterns (FPLBP). Distinct from the LBP descriptor, the TPLBP and FPLBL descriptors produce a binary code for each pixel by comparing the intensity values of three and four patches in the neighbourhood of the pixel respectively.

In [Wolf et al., 2008], the LBP, TPLBP and FPLBP descriptors were evaluated on the challenging

Labeled Faces in the Wild (LFW) database [Huang et al., 2007] and have proven successful for unconstrained face verification.

The Scale-Invariant Feature Transform (SIFT) [Lowe, 2004] descriptor performs well for many computer vision tasks such as object recognition and action recognition in videos. It is a 3D histogram of gradient orientations. The region is quantized to a grid of $4 \times 4 = 16$ locations (histograms) with 8 orientations (bins), resulting in a 128-dimensional descriptor. Then, the magnitude for each orientation of the gradient is computed and weighted by a Gaussian function centered on the keypoint, from which an orientation histogram is formed within each grid. By concatenating the histograms over all the locations, a final descriptor can be formed. To make it more applicable for face recognition, Guillaumin et al. [2009] computed the SIFT descriptor at 3 scales centered on 9 facial points (the corner of mouth, eyes and nose). This multiscale descriptor leads to a $3 \times 9 \times 128 = 3456$ dimensional facial descriptor.

Heikkilä et al. [2006] combined the strengths of the SIFT and LBP descriptors and introduced a Center-Symmetric Local Binary Pattern (CSLBP) operator. The proposed descriptor was constructed similar to the SIFT descriptor. Specifically, it replaced the gradient features used by the SIFT descriptor for each pixel with the center-symmetric local binary pattern features which are captured by comparing 4 center-symmetric pairs among the 8 neighbors of each pixel.

A more recent approach by Kumar et al. [2009] trained 65 attribute classifiers to recognize the presence or absence of describable aspects of visual traits such as gender, race, age, hair color, etc. The outputs of these binary classifiers form the facial descriptor.

Cox and Pinto [2011] generated multiple complimentary representations within a large-scale feature search framework where training set augmentation, alternative face comparison functions, and feature set searches with a varying number of model layers were employed. These individual feature representations were then combined using kernel techniques and obtained state-of-the-art results on the LFW dataset [Huang et al., 2007] for unconstrained face verification.

After the discussion of the feature representation methods for facial images, now we briefly review the methods for representing the pedestrian images.

To represent the pedestrian images, many feature representation methods were proposed, e.g. color histogram [Gray and Tao, 2008], principal axis histogram [Hu et al., 2006], rectangle region histogram [Dollár et al., 2007] and a mixture of color histogram and LBP-based texture histogram [Mignon and Jurie, 2012].

In particular, Kostinger et al. [2012] proposed to extract three types of local features for pedestrian images, i.e. HSV (Hue, Saturation, Value), Lab (L for lightness, and a and b for two color channels) and LBP. Specifically, the images are firstly divided into overlapping blocks and stride of size 8×16 and 8×8, respectively. Then, the color and texture cues are extracted. To describe the color cues, the HSV and Lab histograms were extracted, each with 24 bins per channel. To obtain the texture information, LBP operators are used. Finally, the above three features are concatenated to form a single feature vector, the dimensionality of which is 20480. The proposed descriptor was employed for person re-identification and promising results were obtained on the benchmark Viewpoint Invariant Pedestrian Recognition (VIPeR) database [Gray et al., 2007].

2.2.2 Dimensionality Reduction

After applying feature representation methods, images are usually represented by high dimensional feature vectors, which can lead to prohibitive computational cost. To deal with this problem, it is common to use the dimensionality reduction methods, the task of which is to identify a lower dimensional subspace to represent the large number of the observed dimensions. The dimensionality reduction methods can be linear or nonlinear. Below we start the discussion of the linear dimensionality reduction approaches including Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA), followed by the description of the nonlinear approach Kernel PCA.

Notations: Let $\{x_i\}_{i=1}^n$ denote the training data, where $x_i \in \mathbb{R}^{d_0}$. Each image x_i belongs to one of the G classes $\{C_1, \cdots, C_G\}$. The class label of x_i is denoted as $l(x_i)$. Denote n_i the sample number of each class. The standard Euclidean distance is denoted by $\|\cdot\|$.

Principal Component Analysis

Principal Component Analysis (PCA), also known as Karhunen-Loeve method, is a common technique used for dimensionality reduction in computer vision, particularly in face recognition [Turk and Pentland, 1991b; Swets and Weng, 1996]. It is an unsupervised learning method which aims to learn a linear projection to maximize the variance of the projected data.

More formally, consider the training images $\{x_i\}_{i=1}^n$. The mean image of the samples is given by $m = \frac{1}{n}\sum_{i=1}^n x_i$. Let $X = (x_1 - m, x_2 - m, \ldots, x_n - m) \in \mathbb{R}^{d_0 \times n}$ be the matrix of the centred training data. PCA technique can be achieved by diagonalizing the covariance matrix $C \in \mathbb{R}^{d_0 \times d_0}$ defined as

$$C = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)(x_i - m)^T. \tag{2.1}$$

To diagonalize C, one has to solve the eigenvalue equation

$$Cv_k = \lambda_k v_k, \tag{2.2}$$

where v_k is the eigenvector of C with eigenvalue λ_k satisfying $\lambda_1 \geq \ldots \geq \lambda_k \geq 0$. The eigenvectors v_k are also referred to as Eigenfaces [Turk and Pentland, 1991a]. Denote $V = (v_1, \cdots, v_d)$ the eigenvector matrix with the top leading d eigenvectors of C, then the projection onto the lower-dimensional subspace spanned by d Eigenfaces is given by

$$y = V^T(x - m). \tag{2.3}$$

By selecting the Eigenfaces with large eigenvalues, most noise encoded on the trailing eigenvectors is removed, and therefore PCA is well-suited to object representation. However, it was observed in [Belhumeur et al., 1997; Hariharan et al., 2012] that in the context of face recognition, the variations retained in the leading Eigenfaces often correspond to variations arising from lighting and viewing angles rather than variations caused by identity differences. Consequently, discriminative information may be lost. To deal with this problem, one way is to normalize the

selected Eigenfaces using the whitening process, which is given by

$$z = \Lambda^{-1/2} V^T (x - m), \tag{2.4}$$

where $\Lambda = diag(\lambda_1, \ldots, \lambda_d)$ is the eigenvalue matrix with the top d leading eigenvalues of C. The above whitening process (i.e. equation (2.4)) is referred to as WPCA. From equation (2.4), we can see that the Eigenfaces are weighted by the inverse of the eigenvalues, which penalizes the Eigenfaces with large eigenvalues. Therefore, the negative influences of the leading Eigenfaces are suppressed, and the discriminative information retained in the trailing eigenvectors is magnified. It was empirically shown by Deng et al. [2005] that WPCA outperforms PCA for face recognition.

Linear Discriminant Analysis

As discussed above, PCA does not take into account the discrimination of different classes. Linear Discriminant Analysis (LDA) attempts to preserve the discriminative information as much as possible while performing the dimensionality reduction. To be specific, LDA aims to find a subspace to best discriminate the different face classes by maximizing the ratio of determinant of the between-class scatter matrix S_B to that of the within-class scatter matrix S_W in the projected subspace. S_B and S_W are defined as

$$S_B = \sum_{i=1}^{G} n_i (m_i - m)(m_i - m)^T, \tag{2.5}$$

$$S_W = \sum_{i=1}^{G} \sum_{x_k \in C_i} (x_k - m_i)(x_k - m_i)^T, \tag{2.6}$$

where m_i is the mean image of class C_i, and n_i is the number of samples in class C_i. Note that the ranks of S_B and S_W are at most $G - 1$ and $n - G$ respectively. Comparing with the total covariance matrix C given by equation (2.1), we have $nC = S_W + S_B$. The formulation of LDA is given by

$$W_{opt} = \arg\max_W \frac{|W^T S_B W|}{|W^T S_W W|}. \tag{2.7}$$

The optimal projection matrix W_{opt} can be constructed by the eigenvectors of $S_W^{-1} S_B$ associated with the largest eigenvalues. It was shown in [Fukunaga, 1990] that computing the eigenvectors of $S_W^{-1} S_B$ is equivalent to a two-stage diagonalization of S_W and S_B. Firstly, S_W is whitened by

$$(H\Lambda^{-1/2})^T S_W (H\Lambda^{-1/2}) = I, \tag{2.8}$$

where H and Λ are the eigenvector matrix and the eigenvalue matrix of S_W. Secondly, S_B is transformed to K_B, that is, $K_B = (H\Lambda^{-1/2})^T S_B (H\Lambda^{-1/2})$. After computing the eigenvector matrix U and the eigenvalue matrix Σ such that $K_B = U\Sigma U^T$, the overall projection matrix of LDA is given by

$$W_{opt} = H\Lambda^{-1/2} U, \tag{2.9}$$

where H and U are orthogonal matrices. The projection onto the LDA subspace is then given by

$$y = W_{opt}^T x = U^T \Lambda^{-1/2} H^T x. \tag{2.10}$$

In practice, the within-class scatter matrix $S_W \in \mathbb{R}^{d_0 \times d_0}$ is often singular. This is due to the fact that the rank of S_W is at most $n - G$, which can be much smaller than the dimensionality of the feature vector (i.e. d_0). To avoid the degeneration of S_W, Belhumeur et al. [1997] proposed an approach called Fisherfaces by first using PCA to reduce the dimensionality of the original feature to $n - G$ and then applying LDA on the PCA-reduced subspace for discriminant analysis. Specifically, let W_{pca} be the projection matrix from the original space to the PCA-reduced subspace and W_{lda} be the projection matrix from the PCA-reduced subspace to the LDA subspace, then the projection of Fisherfaces is given by $W_{opt} = W_{pca} W_{lda}$. The performance of Fisherfaces depends on whether or not the within-class scatter captures reliable variation for each face class. When the trailing Eigenfaces with smaller eigenvalues are used in the PCA procedure, the Fisherfaces step has to fit for the noise encoded in the smaller eigenvalues and poor generalization performance can be achieved [Liu and Wechsler, 1998].

Bayesian Face Recognition

Moghaddam et al. [2000] proposed a probabilistic similarity measure based on Bayesian analysis on image differences $\Delta = x_1 - x_2$. It formulates the face recognition task as a binary classification problem. Let Ω_I represent the intra-personal variations between images of the same individual and Ω_E the extra-personal variations between images from different individuals. Then the face recognition problem is cast into classifying the image difference $\Delta = x_1 - x_2$ as the intra-personal variation or extra-personal variation. Based on the maximum a posterior (MAP) rule, the similarity measure is defined as the intra-personal a posterior probability

$$S(x_1, x_2) = P(\Omega_I|\Delta) = \frac{P(\Delta|\Omega_I)P(\Omega_I)}{P(\Delta|\Omega_I)P(\Omega_I) + P(\Delta|\Omega_E)P(\Omega_E)}. \tag{2.11}$$

An alternative similarity measure based on the maximum likelihood (ML) is defined using the intra-personal likelihood alone

$$S^{'}(x_1, x_2) = P(\Delta|\Omega_I). \tag{2.12}$$

The conditional probabilities $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$ in equation (2.11) and (2.12) are assumed as Gaussian-distributed. It was shown in [Moghaddam et al., 2000] that the simplified ML measure was almost as effective as the MAP measure.

To deal with the singularity of the covariance matrix resulting from the high dimensionality of the difference vector Δ and the lack of training samples, the authors estimate the likelihood $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$ using the eigenspace decomposition. Specifically, PCA is applied on the difference set $\{\Delta|\Delta \in \Omega_I\}$ or $\{\Delta|\Delta \in \Omega_E\}$ to divide the difference space into two complimentary subspaces: the intra-personal or extra-personal subspace F spanned by the k largest intra-personal or extra-personal eigenvectors, and its orthogonal complementary $\bar{F}$ containing the residual of the expansion. To estimate $P(\Delta|\Omega_I)$, the intra-personal eigenvectors are computed from the differ-

ence set $\{\Delta_{ij} = x_i - x_j \in \mathbb{R}^{d_0} | l(x_i) = l(x_j)\}$, for which the covariance matrix is given by

$$C_I = \sum_{l(x_i)=l(x_j)} (x_i - x_j)(x_i - x_j)^T. \tag{2.13}$$

Then the projection onto the intra-personal subspace F_I spanned by k_I ($k_I \leq d_0$) largest intra-personal eigenvectors is given by

$$y_i = V_I^T x_i, \tag{2.14}$$

where $V_I = (v_1, \cdots, v_{k_I})$ is the eigenvector matrix with the top leading k_I eigenvectors of C_I. Denote $\Lambda_I = diag(\lambda_1, \ldots, \lambda_{k_I})$ the eigenvalue matrix with the top leading k_I eigenvalues of C_I, the estimate of $P(\Delta|\Omega_I)$ can be written as the product of two independent marginal Gaussian densities in F and $\bar{F}$

$$\hat{P}(\Delta|\Omega_I) = [\frac{\exp(-\frac{1}{2}d_F(\Delta))}{(2\pi)^{k_I/2}\prod_{i=1}^{k_I}\lambda_i^{1/2}}][\frac{\exp(-\frac{1}{2\rho}\epsilon^2(\Delta))}{(2\pi\rho)^{(p-k_I)/2}}]. \tag{2.15}$$

Here,

$$d_F(\Delta) = \|\Lambda_I^{-1/2}(y_i - y_j)\|^2 = \|\Lambda_I^{-1/2}V_I^T(x_i - x_j)\|^2, \tag{2.16}$$

which is a Mahalanobis distance in F, referred to as "distance-in-feature-space" (DIF). We see from equation (2.16) that by further normalizing the projected the images by $\Lambda_I^{-1/2}$, $d_F(\Delta)$ is the Euclidean distance. $\epsilon^2(\Delta) = \|x_i - x_j\|^2 - \|\Lambda_I^{-1/2}V_I^T(x_i - x_j)\|^2$ is the residual error in $\bar{F}_I$, referred to as "distance-from-feature-space" (DFF). ρ is the average of eigenvalues in $\bar{F}_I$, i.e. $\rho = \frac{1}{d_0 - k_I}\sum_{i=k_I+1}^{d_0}\lambda_i$. Similarly, the likelihood $P(\Delta|\Omega_E)$ can be estimated on the extra-personal subspace F_E computed from the difference set $\{\Delta_{ij} = x_i - x_j \in \mathbb{R}^{d_0} | l(x_i) \neq l(x_j)\}$, the covariance matrix of which is given as follows

$$C_I = \sum_{l(x_i)\neq l(x_j)} (x_i - x_j)(x_i - x_j)^T. \tag{2.17}$$

Here, Bayesian face recognition [Moghaddam et al., 2000] is reviewed as a linear dimensionality reduction method, which is from the point of view of implementation. As will be shown in Section 2.3, it can also be regarded as a similarity-based approach from the perspective of its objective.

Kernel Principal Component Analysis

PCA, as a linear dimensionality reduction technique, is not able to detect the nonlinear manifold where the faces usually lie on. This is the case especially when the facial images/videos are taken under unconstrained conditions and show large transformation variations.

A possible approach to alleviate the limitation of linear PCA is to use Kernel PCA which was proposed by Schölkopf et al. [1998]. Kernel PCA is an extension of linear PCA using the techniques of kernel methods. Specifically. let ϕ denote the nonlinear transformation from the original data space to a new feature space $\mathcal{F}$, which is given by

$$\begin{aligned} \phi : \mathbb{R}^{d_0} &\rightarrow \quad \mathcal{F} \\ x \quad &\mapsto \phi(x). \end{aligned} \tag{2.18}$$

Note that the dimensionality of $\mathcal{F}$ could be arbitrarily large, possibly infinite, which could lead to prohibitive computational cost. As will be described in the following, employing the typical tricks of kernel methods simplifies the computation.

Firstly, we assume that the projected data has zero mean, i.e. $\frac{1}{n}\sum_{i=1}^{n}\phi(x_i) = 0$. Denote $\bar{X} = (\phi(x_1), \phi(x_2), \ldots, \phi(x_n))$ the matrix of the training data in the feature space $\mathcal{F}$. Then, the covariance matrix in $\mathcal{F}$ is given by

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\phi(x_i)^T = \frac{1}{n}\bar{X}\bar{X}^T, \tag{2.19}$$

and its eigenvectors with nonnegative eigenvalues are given by $\bar{C}\bar{v}_k = \bar{\lambda}_k\bar{v}_k$, i.e. $\frac{1}{n}\bar{X}\bar{X}^T\bar{v}_k = \bar{\lambda}_k\bar{v}_k$. Multiplying both sides by $\bar{X}^T$ gives

$$\frac{1}{n}\bar{X}^T\bar{X}(\bar{X}^T\bar{v}_k) = \bar{\lambda}_k(\bar{X}^T\bar{v}_k). \tag{2.20}$$

Defining $\bar{\alpha}_k = \bar{X}^T\bar{v}_k \in \mathbb{R}^n$, equation (2.20) can be written as

$$\frac{1}{n}\bar{X}^T\bar{X}\bar{\alpha}_k = \bar{\lambda}_k\bar{\alpha}_k. \tag{2.21}$$

Denote $\bar{K} = \bar{X}^T\bar{X} = (\phi(x_i)^T\phi(x_j))_{ij} \in \mathbb{R}^{n\times n}$, equation (2.21) can then be written as $\frac{1}{n}\bar{K}\bar{\alpha}_k = \bar{\lambda}_k\bar{\alpha}_k$, which is an eigenvector equation for matrix $\bar{K}$. Multiplying both sides of equation (2.21) by $\bar{X}$ gives

$$\frac{1}{n}\bar{X}\bar{X}^T(\bar{X}\bar{\alpha}_k) = \bar{\lambda}_k(\bar{X}\bar{\alpha}_k), \tag{2.22}$$

from which we see $\bar{X}\alpha_k$ is the eigenvector of $\bar{C}$ with eigenvalue $\bar{\lambda}_k$, i.e. $\bar{v}_k \propto \bar{X}\alpha_k$. Assuming that $\bar{v}_k$ has been normalized to unit length (i.e. $\|\bar{v}_k\| = 1$) gives

$$\bar{v}_k = \bar{X}\bar{\alpha}_k, \|\bar{\alpha}_k\|^2 = \frac{1}{n\bar{\lambda}_k}, \bar{\lambda}_k \neq 0. \tag{2.23}$$

Let $\phi(x)$ be a data point in $\mathcal{F}$, the projection onto the principal component $\bar{v}_k$ is then given by

$$\bar{v}_k^T\phi(x) = \bar{\alpha}_k^T\bar{X}^T\phi(x). \tag{2.24}$$

If the projected data $\{\phi(x_i)\}_{i=1}^n$ does not have zero mean, we can center the projected data by $\{\phi(x_i) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\}_{i=1}^n$. Then the corresponding kernel matrix $K = ((\phi(x_i) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i))^T (\phi(x_j) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i))_{ij} \in \mathbb{R}^{n\times n}$ can be obtained by

$$K = \bar{K} - 1_{nn}\bar{K} - \bar{K}1_{nn} + 1_{nn}\bar{K}1_{nn}, \tag{2.25}$$

where 1_{nn} is the $n \times n$ matrix with all elements equal to $\frac{1}{n}$.

As can be seen from the definition of matrix $\bar{K}$ and equations (2.24) and (2.25), we only need to compute the inner product of the mapped data points instead of explicitly dealing with the feature map $\phi(x)$.

2.3 Similarity-based Approaches

By following the pipeline of face recognition and person re-identification (see Figure 2.1), section 2.2 has discussed the feature extraction techniques which attempt to learn a transformation to find the informative set of features. In this section, we review the related work on similarity-based approaches to learn similarity measures to compare pairs of images/videos.

One type of similarity-based approaches is the technique of the probabilistic similarity measures, among which Bayesian face recognition proposed by Moghaddam et al. [2000] is a representative. The authors introduced a probabilistic framework that models the distribution of two classes of facial images variations: intra-personal variation for the same individual and extra-personal variation for different individuals. More recently, Chen et al. [2012] proposed to directly model the joint distribution of two faces in the Bayesian framework by introducing a prior on face representation. For person re-identification, Zheng et al. [2011] proposed a probabilistic relative distance comparison (PRDC) model which aims to maximise the probability of similar pairs having smaller distances than dissimilar pairs.

Another type of similarity-based approaches is a family of the background similarity methods which were recently introduced for face recognition. They aim to learn the similarity scores between image-pairs/video-pairs based on background samples. Wolf et al. [2008, 2009a] developed the One-Shot-Similarity (OSS) for image-based matching problem. Specifically, the one-shot similarity score measures the likelihood of each image sharing the same class as the other image and not belonging to a fixed set of "negative" samples. Recently, Wolf et al. [2011a] extended this One-Shot-Similarity to the Matched Background Similarity (MBGS) for video-based matching problem. More recently, Wolf and Levy [2013] derived a new similarity measure called SVM $\ominus$ based on the additional 3D headpose information for video-based matching problem.

Metric learning or similarity learning also belongs to the group of similarity-based approaches. It aims to learn an appropriate distance metric or similarity function to compare pairs of examples, which provides a natural solution for the matching tasks. For simplicity, later on, metric learning or similarity learning is referred to as *Similarity Metric Learning*. Below we discuss the related work on similarity metric learning, which forms the ground of the novel models that will be developed in Chapters 4, 5 and 6.

2.3.1 Preliminaries

Similarity metric learning aims to learn a distance metric or similarity function from side information which is usually given in two forms: pairwise constraints and relative constraints. For the pairwise constraints, only similar pairs (pairs of samples from the same class) and dissimilar pairs (pairs of samples from different classes) are provided. Let $\mathcal{S}$ and $\mathcal{D}$ denote the index set of similar pairs and dissimilar pairs respectively, then $\mathcal{S}$ and $\mathcal{D}$ are given as follows

$$\mathcal{S} = \{(i,j)|(x_i, x_j) \text{ are pairs of samples from the same class}\}, \tag{2.26}$$

$$\mathcal{D} = \{(i,j)|(x_i, x_j) \text{ are pairs of samples from different classes}\}. \tag{2.27}$$

For instance, the notation $(i, j) \in \mathcal{S}$ means a similar pair (x_i, x_j) and $(i, j) \in \mathcal{D}$ means a dissimilar pair (x_i, x_j). For the relative constraints, a set of triplets is given. Specifically, let $\mathcal{T}$ denote the index set of triplets, then $\mathcal{T}$ is given by

$$\mathcal{T} = \{(i, j, k) | (x_i, x_j) \in \mathcal{S}, (x_i, x_k) \in \mathcal{D}\}. \tag{2.28}$$

In practice, $\mathcal{S}$, $\mathcal{D}$ and $\mathcal{T}$ can be easily collected from the label information.

Metric learning [Xing et al., 2003; Weinberger et al., 2006; Davis et al., 2007; Shen et al., 2009; Ying and Li, 2012] usually focuses on the (squared) Mahalanobis distance defined, for any $x, t \in \mathbb{R}^d$, by

$$d_M(x, t) = (x - t)^T M(x - t), \tag{2.29}$$

where M is a positive semi-definite (p.s.d.) matrix. It was recently used for unconstrained face verification (e.g. [Guillaumin et al., 2009; Mignon and Jurie, 2012; Kostinger et al., 2012]) and person re-identification (e.g. [Dikmen et al., 2011; Zheng et al., 2011]). In contrast, similarity learning aims to learn the bilinear similarity function which is defined by

$$s_M(x, t) = x^T Mt, \tag{2.30}$$

or the cosine similarity defined by

$$CS_M(x, t) = \frac{x^T Mt}{\sqrt{x^T Mx}\sqrt{t^T Mt}}, \tag{2.31}$$

where M is a positive semi-definite (p.s.d.) matrix. Both s_M and CS_M have successful applications on unconstrained face verification (see [Nguyen and Bai, 2011]) and image similarity search (see [Chechik et al., 2010; Shalit et al., 2010]).

Learning a distance metric or similarity function is closely related to learning a linear transformation from the original space. To see this, observe that any positive semi-definite matrix M can be rewritten as $L^T L$, where $L \in \mathbb{R}^{d' \times d}$ and $d' \leq d$. Hence, the Mahalanobis distance can be rewritten as

$$d_M(x, t) = (x - t)^T M(x - t) = \|L(x - t)\|^2, \tag{2.32}$$

the bilinear similarity function can be rewritten as

$$s_M(x, t) = x^T Mt = (Lx)^T (Lt), \tag{2.33}$$

and the cosine similarity is equivalent to

$$CS_M(x, t) = \frac{x^T Mt}{\sqrt{x^T Mx}\sqrt{t^T Mt}} = \frac{(Lx)^T (Lt)}{\|Lx\|\|Lt\|}. \tag{2.34}$$

The above observations imply that learning an appropriate M is equivalent to learning an appropriate projection L. From this perspective, the linear dimensionality reduction methods discussed in Section 2.2.2 can be regarded as similarity metric learning methods.

To well discriminate similar pairs from dissimilar pairs, a common discriminative idea behind similarity metric learning is to learn a distance metric (or similarity function) such that a good

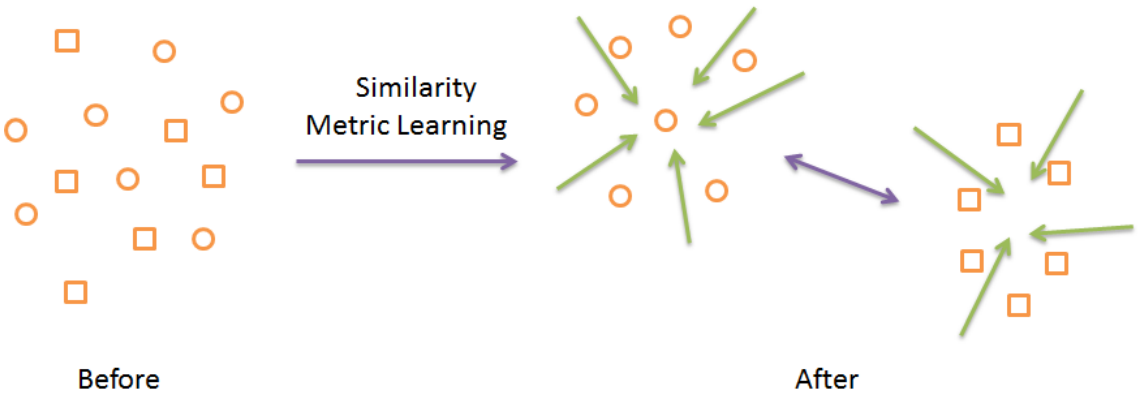


Figure 2.2: Intuition behind similarity metric learning: before learning (left) versus after learning (right). The circles and squares represent samples from two classes. After learning, the distance metric or similarity function is optimized such that the circles and squares are well separated.

distance metric (or similarity function) should report a small distance (or large similarity score) for similar pairs and a large distance (or small similarity score) for dissimilar pairs, see Figure 2.2 for illustration. Similarity metric learning methods used different objective functions to achieve this goal. In the following, we review similarity metric leaning methods: Section 2.3.3 surveys metric learning methods and Section 2.3.3 reviews similarity learning methods.

Notations: Denote $\mathbb{R}^{d\times d}$ the space of $d \times d$ matrices and $\mathbf{Tr}(\cdot)$ the trace of a matrix in $\mathbb{R}^{d\times d}$. The space of symmetric $d \times d$ matrices is denoted by $\mathbb{S}^d$, and the set of positive semi-definite matrices is denoted by $\mathbb{S}^d_+$. The standard Euclidean norm on vectors is denoted by $\|\cdot\|$ and the Frobenius norm on matrices by $\|\cdot\|_F$. Denote $\mathcal{P} = \mathcal{S} \cup \mathcal{D}$ the index set of all pairwise constraints. Denote $y_{ij} = 1$ if $l(x_i) = l(x_j)$, and $y_{ij} = 0$ otherwise.

2.3.2 Metric Learning

This section starts the review of metric learning work using pairwise constraints, followed by the discussion on metric learning methods using relative constraints.

The pioneering work on metric learning was proposed by Xing et al. [2003]. It was developed for k-means clustering. The main idea is to maximize the sum of distances between dissimilar pairs, while maintaining an upper bound on the sum of squared distances between similar pairs. Specifically, the following formulation was presented

$$\begin{array}{ll} \max_{M\in\mathbb{S}^d_+} & \sum_{(i,j)\in\mathcal{D}} \sqrt{d_M(x_i,x_j)} \\ \text{s.t.} & \sum_{(i,j)\in\mathcal{S}} d_M(x_i,x_j) \leq 1. \end{array} \tag{2.35}$$

It is a global metric learning method in the sense that it uses all the similar and dissimilar pairs. The optimal solution is obtained by employing the projected gradient descent algorithm which, however, usually takes a large number of iterations to converge and needs full eigen-decomposition of a matrix per iteration. A detailed review of the proposed formulation and the optimization algorithm will be given in Sections 4.6 and 5.6, respectively. For simplicity, we refer to this method as Xing.

Davis et al. [2007] developed an information theoretic approach called ITML to learn a Mahalanobis matrix M under a set of pairwise constraints. Given the prior information on the Mahalanobis distance M_0, the idea is to regularize M to be as close as possible to M_0. Specifically, the problem is formulated by minimizing the relative entropy between two multivariate Gaussians parameterised by M and M_0, i.e.

$$\begin{array}{ll} \min_{M \in \mathbb{S}_+^d} & \mathrm{KL}(p(x; M_0) \| p(x; M)) \\ \text{s.t.} & d_M(x_i, x_j) \leq u, (i,j) \in \mathcal{S} \\ & d_M(x_i, x_j) \geq l, (i,j) \in \mathcal{D}, \end{array} \tag{2.36}$$

which can be rewritten as the following LogDet optimization problem

$$\begin{array}{ll} \min_{M \in \mathbb{S}_+^d} & \mathbf{Tr}(M {M_0}^{-1}) - \log\det(M {M_0}^{-1}) \\ \text{s.t.} & d_M(x_i, x_j) \leq u, (i,j) \in \mathcal{S} \\ & d_M(x_i, x_j) \geq l, (i,j) \in \mathcal{D}. \end{array} \tag{2.37}$$

To guarantee the existence of a feasible solution, slack variables were incorporated. The proposed optimization algorithm is based on Bregman projection and has the advantage that no eigenvalue computations are needed. The prior M_0, in practice, is chosen to the identity matrix I. Thus, minimising the objective function promotes the closeness between the learned distance and the Euclidean distance. Unfortunately, for specific tasks such as face verification, the Euclidean distance might not be optimal and hand-picking M_0 could have detrimental effect on the performance of the distance metric. Section 4.6 will describe the proposed formulation in details.

Guillaumin et al. [2009] proposed a logistic discriminant approach called LDML to learn a distance metric for face verification from a probabilistic view. Based on the idea that the distances between similar pairs should be smaller than those between dissimilar pairs, the authors modelled the probability p_{ij} that a given pair (x_i, x_j) belongs to the same object as

$$p_{ij} = p(y_{ij} = 1 | x_i, x_j; M, b) = \sigma(b - d_M(x_i, x_j)), \tag{2.38}$$

where $\sigma(z) = (1 + \exp(-z))^{-1}$ is a sigmoid function and b is a bias term. To estimate M, maximum log-likelihood is used

$$\max_{M \in \mathbb{S}^d} \sum_{(i,j) \in \mathcal{P}} y_{ij} \ln p_{ij} + (1 - y_{ij}) \ln (1 - p_{ij}). \tag{2.39}$$

To optimize M, gradient ascent algorithm was employed without constraining M to be positive semi-definite. A further discussion of LDML will be given in Section 4.6.

Motivated by the method Xing [Xing et al., 2003], Ying and Li [2012] recently proposed a metric learning model named DML-eig with eigenvalue optimization. It aims to maximize the minimal squared distances between dissimilar pairs while maintaining the sum of squared distances between similar pairs upper-bounded. Specifically, the following formulation was proposed

$$\begin{array}{ll} \max_{M \in \mathbb{S}_+^d} & \min_{(i,j) \in \mathcal{D}} d_M(x_i, x_j) \\ \text{s.t.} & \sum_{(i,j) \in \mathcal{S}} d_M(x_i, x_j) \leq 1. \end{array} \tag{2.40}$$

It was further shown that formulation (2.40) is equivalent to an eigenvalue optimization problem, based on which efficient first-order algorithms were developed with only the computation of the largest eigenvector of a matrix at each iteration. Section 4.6 will provide an equivalent formulation of DML-eig.

Kan et al. [2011] modified LDA (see Section 2.2.2 for a review of LDA) and developed a Side-Information based Linear Discriminant Analysis (SILD) approach for image-to-image face verification. By exploiting the pairwise information, the authors proposed to replace the within-class scatter matrix S_W (i.e. equation (2.6)) and the between-class scatter matrix S_B (i.e. equation (2.5)) by C_I (i.e. equation (2.13)) and C_E (i.e. equation (2.17)) respectively. The formulation is given by

$$\max_{W} \frac{|W^T C_E W|}{|W^T C_I W|}. \tag{2.41}$$

SILD can be regarded as an variant of LDA, since it was shown that SILD is identical to LDA if the class label information is provided and all the classes have the same number of samples. Let $M = WW^T$, SILD can be rewritten as

$$\max_{M \in \mathbb{S}_+^d} \left[\frac{\sum_{(i,j)\in\mathcal{D}} d_M(x_i, x_j)}{\sum_{(i,j)\in\mathcal{S}} d_M(x_i, x_j)} \right]. \tag{2.42}$$

An equivalent formulation of SILD will be described in Section 4.6.

Metric learning method KISSME [Kostinger et al., 2012] was motivated by the statistical inference based on a likelihood-ratio test in the space of pairwise differences (i.e. $x_i - x_j$). It is similar to the classic bayesian face recognition approach which has been reviewed in Section 2.2.2. Based on the fact that maximizing the likelihood estimate of the Gaussian is equivalent to minimizing the Mahalanobis distance, the distance metric is obtained to reflect the properties of the likelihood-ratio test. In contrast to most metric learning methods, KISSME is not formulated as an iterative optimization procedure. Instead, it only involves the computation of two covariance matrices.

Mignon and Jurie [2012] introduced the Pairwise Constrained Component Analysis (PCCA) to learn a linear transformation $L \in \mathbb{R}^{d' \times d}$ ($d' \ll d$) to project the data points onto a lower dimensional subspace using pairwise constraints. The problem was formulated as

$$\min_{L} \sum_{(i,j)\in\mathcal{P}} l_\beta(y_{ij}(\|L(x_i - x_j)\|^2 - 1)), \tag{2.43}$$

where $l_\beta(x) = \frac{1}{\beta}\log(1 + \exp(\beta x))$ is the generalized logistic loss function. Learning a linear mapping to a lower dimensional space imposes a low rank constraint on the distance matrix $M = L^T L$. Gradient descent algorithm was employed to get the optimal solution. Besides, the authors developed the kernelized version of PCCA and competitive results were reported for person re-identification on the benchmark Viewpoint Invariant Pedestrian Recognition (VIPeR) database [Gray et al., 2007]. However, as will be shown in Section 4.5, the performance of PCCA drops heavily when less training samples are used.

After the review of metric learning methods using pairwise constraints, now we discuss metric learning methods that utilise relative constraints.

Weinberger et al. [2006] proposed a method called LMNN to learn a Mahalanobis distance metric for kNN classification. The metric is optimized with the intuition that k-nearest neighbors always belong to the same class while samples from different classes are separated by a large margin. Specifically, for each input x_i, define the target neighbors of x_i as its k-nearest neighbors with the same label y_i. The index set of similar pairs $\mathcal{S}$ is then constructed using each x_i and its corresponding target neighbors. Let $\eta_{ij} = 1$ if x_j is a target of x_i, and $\eta_{ij} = 0$ otherwise. Given the index set of dissimilar pairs $\mathcal{D}$ (see formulation (2.27) for definition) and the index set of triplets $\mathcal{T}$ (see formulation (2.28) for definition), the cost function is given by

$$\epsilon(M) = \sum_{(i,j)\in\mathcal{S}} \eta_{ij} d_M(x_i, x_j) + \gamma \sum_{\tau=(i,j,k)\in\mathcal{T}} \eta_{ij}(1 - y_{ik})\big[1 + d_M(x_i, x_j) - d_M(x_i, x_k)\big]_+. \tag{2.44}$$

Minimizing the first term of the above formulation penalizes the large distance between each input and its target neighbors. The second term is a hinge loss which incorporates a large margin between each input and all the other samples that share distinct labels from the input. The authors reformulated the model as a semidefinite program, i.e.

$$\begin{aligned} \min_{M,\xi} \quad & \sum_{(i,j)\in\mathcal{S}} \eta_{ij} d_M(x_i, x_j) + \gamma \sum_{\tau=(i,j,k)\in\mathcal{T}} \eta_{ij}(1 - y_{il})\xi_{ijk} \\ \text{s.t.} \quad & d_M(x_j, x_k) - d_M(x_i, x_j) \geq 1 - \xi_{ijk}, \\ & M \in \mathbb{S}_+^d, \ \ \xi_{ijk} \geq 0, \ \ \forall\, (i,j,k) \in \mathcal{T}. \end{aligned} \tag{2.45}$$

The above formulation can be further simplified as

$$\begin{aligned} \min_{M,\xi} \quad & \sum_{(i,j)\in\mathcal{S}} d_M(x_i, x_j) + \gamma \sum_{\tau=(i,j,k)\in\mathcal{T}} \xi_{ijk} \\ \text{s.t.} \quad & d_M(x_j, x_k) - d_M(x_i, x_j) \geq 1 - \xi_{ijk}, \\ & M \in \mathbb{S}_+^d, \ \ \xi_{ijk} \geq 0, \ \ \forall\, (i,j,k) \in \mathcal{T}. \end{aligned} \tag{2.46}$$

In contrast to Xing [Xing et al., 2003], LMNN is a local method in the sense that only triplets from k-nearest neighbors are used. Projected sub-gradient descent algorithm was used to obtain the optimal solution. Due to the lack of the regularization term, LMNN is sometimes prone to over-fitting. Nevertheless, as will be shown in Section 4.5, applying LMNN to the task of person re-identification gives competitive results. A detailed discussion of LMNN and its proposed algorithm will be presented in Sections 4.6 and 5.6 respectively.

Shen et al. [2009] employed the exponential loss to learn a Mahalanobis distance metric using relative constraints in a large margin framework. The following formulation was proposed

$$\begin{aligned} \min_{M\in\mathbb{S}_+^d} \quad & \log\big(\textstyle\sum_{ijk} \exp(-\rho_{ijk})\big) + \gamma \mathbf{Tr}(M) \\ \text{s.t.} \quad & \rho_{ijk} = d_M(x_i, x_k) - d_M(x_i, x_j), (i,j,k) \in \mathcal{T}. \end{aligned} \tag{2.47}$$

Based on the idea that each positive semi-definite matrix can be decomposed into a linear positive combination of trace-one and rank-one matrices, a boosting-based algorithm called BoostMetric was developed. Section 5.6 will provide a detailed discussion of the proposed BoostMetric.

2.3.3 Similarity Learning

In this section, the following similarity learning methods are descibed: cosine similarity metric leaning (CSML [Nguyen and Bai, 2011]) for face verification using pairwise constraints and online algorithm for scalable image similarity learning (OASIS [Chechik et al., 2010]) for image similarity search exploiting the relative constraints.

Nguyen and Bai [2011] proposed a cosine similarity metric leaning named CSML for face verification. Specifically, the authors aim to learn a linear transformation $L \in \mathbb{R}^{d' \times d}$ ($d' \ll d$) such that the cosine similarities between similar pairs are larger than that between dissimilar pairs in the transformed subspace. The following formulation was proposed

$$\min_{L} \sum_{(i,j)\in\mathcal{D}} \frac{x_i^T L^T L x_j}{\sqrt{x_i^T L^T L x_i}\sqrt{x_j^T L^T L x_j}} - \alpha \sum_{(i,j)\in\mathcal{S}} \frac{x_i^T L^T L x_j}{\sqrt{x_i^T L^T L x_i}\sqrt{x_j^T L^T L x_j}} + \beta \|L - L_0\|^2, \quad (2.48)$$

where $\alpha, \beta \geq 0$ and L_0 is a predefined matrix. Conjugate gradient algorithm was used for the optimization and state-of-the-art results were reported on the LFW database [Huang et al., 2007]. However, one limitation of CSML is that its objective function is not convex with respect to L [1] and thus subjects to a local minimum.

Chechik et al. [2010] developed an online algorithm called OASIS for scalable image similarity learning. It aims to minimize a large margin target function based on the hinge loss, that is,

$$\min_{M\in\mathbb{R}^{d\times d}} \sum_{\tau=(i,j,k)\in\mathcal{T}} (1 - s_M(x_i, x_j) + s_M(x_i, x_k))_+. \quad (2.49)$$

For the optimization, the authors do not require M to be positive, or even symmetric. Passive-aggressive algorithm was employed iteratively over triplets to obtain the optimal solution. At each iteration, a triplet (x_i, x_j, x_k) is randomly selected, and the following convex problem with soft margin is solved:

$$\begin{aligned} M_i = \quad & \arg\min_{M\in\mathbb{R}^{d\times d}} \tfrac{1}{2}\|M - M_{i-1}\|_F^2 + C\xi, \\ \text{s.t.} \quad & (1 - s_M(x_i, x_j) + s_M(x_i, x_k))_+ \leq \xi, \xi \geq 0. \end{aligned} \quad (2.50)$$

It was shown in [Chechik et al., 2010] that OASIS is fast and accurate at a wide range of scales from problems with thousands of images to large web-scale problems.

2.4 Benchmark Databases

This section gives a brief introduction of three standard benchmarks that are used in this thesis to demonstrate the effectiveness of new models for unconstrained face verification and person re-identification. The experimental protocol of each dataset are provided, followed by the description of the feature representation that we employ for each database. Section 2.4.1 introduces the Labeled Faces in the Wild (LFW) database [Huang et al., 2007] for unconstrained face verification

[1] The non-convexity of its objective function can be easily proved by contradiction using matlab.

Figure 2.3: Example of image-pairs from the Labeled Faces in the Wild (LFW) database [Huang et al., 2007] exhibit large transformation differences such as pose, hairstyle and background: the first three columns are pairs of images from the same person; the second three columns are pairs of images from different persons.

in still images. Section 2.4.2 describes the YouTube Faces (YTF) database [Wolf et al., 2011a] for video-based face verification. A brief description of the Viewpoint Invariant Pedestrian Recognition (VIPeR) database [Gray et al., 2007] is presented in Section 2.4.3 for person re-identification.

2.4.1 Labeled Faces in the Wild

Dataset. For unconstrained face verification in still images, the challenging Labeled Faces in the Wild (LFW) database [Huang et al., 2007] is used. In this database, there are 13233 face images of 5749 people, and 1680 of them appear in more than two images. It is commonly regarded as a very challenging dataset for face verification, since the faces were detected from images taken from Yahoo! News and show large transformation differences arising from all kinds of transformations such as pose, expression, lighting, age, background etc. Figure 2.3 illustrates the examples of the LFW dataset.

The images are divided into ten folds where the identities are mutually exclusive. In each fold, 300 similar and 300 dissimilar image-pairs are provided. This database has two different training settings: restricted and unrestricted setting. In the restricted setting, only similar and dissimilar pairs are provided while the identity of images is unknown. In the unrestricted setting, the identity information of images is also provided. The performance is measured by the 10-fold cross-validation test. In each repeat, 9 folds containing 2700 similar image-pairs and 2700 dissimilar image-pairs are used for training and the remaining fold containing 600 image-pairs is used for testing. The performance is reported using mean verification rate ($\pm$ standard error) and the ROC curve. An important aspect of the evaluation procedure on LFW is that the individual identities in the training and test set are exclusive and therefore the prediction of never-seen-before faces is required, which makes face verification in the wild extremely challenging.

Feature representation. Faces are harvested from the raw images by the Viola-Jones face detector [Viola and Jones, 2004] and further cropped and rescaled to 250×250 pixels. The cropped and rescaled images are then prepared in two ways: "aligned" using commercial face alignment software by Taigman et al. [2009] and "funneled" available on the LFW website [Huang et al., 2007]. Two facial descriptors are employed on the "aligned" images: Local Binary Patterns (LBP)

Figure 2.4: Example pairs of frames from the YouTube Faces (YTF) database [Wolf et al., 2011a] exhibit large transformation variations arising from occlusion, problematic lighting, and motion blur: the first three columns are pairs of frames within videos from the same person; the second three columns are pairs of frames within videos from different persons.

[Ojala et al., 2002] and Three-Patch Local Binary Patterns (TPLBP) [Wolf et al., 2008]. On the "funneled" images, we use the SIFT-based descriptor provided by Guillaumin et al. [2009]. For simplicity, we refer to this SIFT-based descriptor as SIFT. Section 2.2.1 has provided a brief review of the above three descriptors.

2.4.2 YouTube Faces Database

Dataset. For video-based face verification, we use the challenging benchmark: the YouTube Faces (YTF) database [Wolf et al., 2011a]. It contains 3425 videos of 1595 different subjects and the average length of a video clip is 181.3 frames. Video clips were downloaded from YouTube and images/frames within each video clip show large transformation variations in occlusion, problematic lighting, pose, and motion blur etc (see Figure 2.4 for examples). The protocol of the YTF dataset is similar to that of the LFW dataset. Specifically, the video clips are divided into ten folds where the identities are mutually exclusive. In each fold, 250 similar video-pairs and 250 dissimilar video-pairs are provided. Two training setting are divided: restricted and unrestricted setting. This thesis mainly focuses on the restricted setting where only similar and dissimilar pairs are given while the label information is not available. The 10-fold cross-validation test is used to measure the performance. In each repeat, we use 9 folds for training and the remaining fold for testing, which assumes that the individuals in the test set are not seen during training and thus requires the prediction of never-seen-before faces. Result is reported using mean verification accuracy ($\pm$ standard error), area under curve (AUC) and equal error rate (ERR). Besides, ROC curve is used to demonstrate the performance.

Feature representation. Faces are detected by the Viola-Jones face detector [Viola and Jones, 2004] and cropped to 100×100 pixels and further aligned by fixing the coordinates of automatically detected facial feature points [Everingham et al., 2006]. We use the features provided by Wolf et al. [2011a]: Local Binary Patterns (LBP) [Ojala et al., 2002], Center-Symmetric LBP (CSLBP) [Heikkilä et al., 2006] and Four-Patch LBP [Wolf et al., 2008], which have been described in Section 2.2.1.

Figure 2.5: Example image-pairs from the Viewpoint Invariant Pedestrian Recognition (VIPeR) database [Gray et al., 2007] exhibit large transformation variations in viewpoint, background and illumination: the first five columns are pairs of images from the same person; the second five columns are pairs of images from different persons.

2.4.3 Viewpoint Invariant Pedestrian Recognition database

Dataset. For the problem of person re-identification across spatially disjoint cameras, we use the benchmark Viewpoint Invariant Pedestrian Recognition (VIPeR) database [Gray et al., 2007]. This database is the largest publicly available dataset for person re-identification, consisting of 632 pedestrian image-pairs. Each image-pair contains two images (48×128 pixels) from the same identity. Images are captured from an arbitrary viewpoint under varying illumination conditions and show large transformation differences in viewpoint, background and illumination, see Figure 2.5 for examples.

For the purpose of evaluation, we follow the procedure described in [Gray et al., 2007; Mignon and Jurie, 2012]. Specifically, h persons out of 632 persons are randomly selected to set up the testing set and the rest forms the training set, which leads to the prediction of never-seen-before pedestrians. For training, denote n^- the different ratio of dissimilar/similar pairs for each person. The image-pair from each person forms one similar pair for each person. To obtain n^- dissimilar pairs for each person, we randomly select one image from the person and one image from a different person. For testing, the test set is split into a probe and gallery set by randomly assigning two images of an image-pair to a probe set and a gallery set. The process is repeated 100 times and average result is reported. Performance is evaluated using a Cumulative Matching Characteristic (CMC) curve which represents the expectation of finding the true match within the top r ranks (see [Wang et al., 2007; Gray et al., 2007]). Thus, the matching rate at rank 1 is the recognition rate. Although it is critical computing the rank 1 matching rate, in practice, computing the top r matching rate is also important, since the top retrieved images can be verified by a human operator. The different ratio n^- in the training set is fixed to be 10 as suggested in [Mignon and Jurie, 2012].

Feature representation. For the feature representation, we use the features [2] generated in [Kostinger et al., 2012]. A review of such features has been provided in Section 2.2.1.

[2] Available at: `http://lrs.icg.tugraz.at/research/kissme/`.

2.5 Conclusion

In this chapter, we reviewed a large body of work on face recognition and person re-identification by following their pipeline (see Figure 2.1). In particular, we mainly discussed the work on feature extraction techniques and similarity-based approaches. Section 2.2 described the feature extraction techniques which comprise two steps: feature representation and dimensionality reduction. Related work on feature representation was covered in Section 2.2.1, followed by a detailed review on dimensionality reduction methods in Section 2.2.2. The advantages and disadvantages of different approaches were discussed. We then did a large survey on similarity-based approaches in Section 2.3, with a focus on similarity metric learning methods. A brief introduction of similarity metric learning methods was given in Section 2.3.1. Sections 2.3.2 and 2.3.3 reviewed metric learning and similarity learning methods respectively. Lastly, Section 2.4 provided a brief overview of three publicly available benchmarks which are used to evaluate the new models in this thesis.

From the literature review, several observations are raised. Firstly, dimensionality reduction methods such as PCA or Eigenfaces [Turk and Pentland, 1991b], LDA or Fisherfaces [Belhumeur et al., 1997] and Bayesian face recognition [Moghaddam et al., 2000] perform well under constrained conditions. However, it was shown in [Li et al., 2012; Chen et al., 2012] that most of the above models perform poorly when applied to unconstrained conditions where images exhibit significant transformation differences such as illumination, motion blur, viewpoints etc. Secondly, one drawback of existing similarity metric learning methods is that most of them mainly focused on the discrimination of the distance metrics or similarity functions while ignoring to consider how to reduce the large transformation differences. Thus, the learned distance metrics or similarity functions may not be robust to the transformation differences and their performance can be degenerated.

The following two chapters address the above two limitations of existing methods. Chapter 3 introduces a new dimensionality reduction model to deal with the detrimental effect induced by the large transformation differences. Chapter 4 develops a novel regularization framework to learn distance metrics and similarity functions by incorporating both the robustness to large transformation differences and the discriminative power of similarity metric learning methods.

3 Robustness to Transformation Differences

3.1 Introduction

In the past decades, a lot of research efforts have been devoted to designing efficient algorithms for face recognition. Many face recognition algorithms have demonstrated promising results under well-controlled conditions with cooperative users. It can be traced back to early dimensionality reduction methods such as PCA or Eigenfaces [Turk and Pentland, 1991a], LDA or Fisherfaces [Belhumeur et al., 1997] and Bayesian face recognition [Moghaddam et al., 2000]. PCA, which has been reviewed in Section 2.2.2, learns a linear projection that maximizes the variance of the projected facial images. LDA, which has been reviewed in Section 2.2.2, seeks a subspace that well separates different face classes by maximizing the ratio of determinant of the between-class scatter matrix to that of the within-class scatter matrix in the projected subspace. Recently, unconstrained face verification (e.g. [Guillaumin et al., 2009; Li et al., 2012; Cox and Pinto, 2011; Taigman et al., 2009; Wolf et al., 2009b, 2008, 2011b,a; Wolf and Levy, 2013]) has been extensively studied, in which the task is to verify whether two facial images/videos are from the same person or not. Facial images/videos are captured in the wild and exhibit large transformation differences arising from pose and illumination changes. Concurrently, person re-identification, handling the pedestrian matching and ranking across non-overlapping camera views, has attracted a lot of interest in the past decade (see [Gray et al., 2007; Zheng et al., 2011; Kostinger et al., 2012]). A person's appearances are generally captured in different camera views and often undergo significant transformation variations in view angle, illumination and occlusion. In the above two vision tasks, the transformation differences can overwhelm the variations arising from identity differences, which makes the problem of unconstrained face verification and person re-identification extremely challenging. Due to the large transformation variations, the performance of most of the above methods degrades heavily when applied to the unconstrained environments (see [Li et al., 2012; Chen et al., 2012]).

Among the above models, LDA and Bayesian face recognition are supervised models which need the label information of the training data. However, compared to the supervised setting, it is more common and challenging in real world that only pairwise information (i.e. pairs of images/videos from same person and pairs of images/videos from different persons) is provided while the label information is not available. Kan et al. [2011] developed a Side-Information based Linear Discriminant Analysis (SILD) approach using pairwise information. It has been shown in Section 2.3.2 that SILD can be regarded as an variant of LDA. Unfortunately, experiments on the benchmark the Labeled Faces in the Wild (LFW) database [Huang et al., 2007] showed modest performance

[1] Some of the material in this chapter has been published in [Cao et al., 2013] and the code is available at: `http://www.albany.edu/~yy298919/software.html`.

for unconstrained face verification.

This chapter develops a novel dimensionality reduction model called Intra-PCA for unconstrained face verification and person re-identification using pairwise information. The learning objective is formulated by considering the robustness to large transformation differences. To be specific, Intra-PCA is developed by first applying WPCA (see Section 2.2.2 for details) to reduce the noise and then mapping the resultant images/videos to the intra-personal subspace by the whitening process to reduce the transformation differences. We further extend Intra-PCA to video-based face verification in the wild. Lastly, we conduct experimental study for unconstrained face verification in still images and videos on standard testbeds including the Labeled Faces in the Wild (LFW) [Huang et al., 2007] and the YouTube Faces (YTF) [Wolf et al., 2011a] databases. We also report experimental results for person re-identification on the benchmark Viewpoint Invariant Pedestrian Recognition (VIPeR) database [Gray et al., 2007]. Our proposed method outperforms state-of-the-art dimensionality reduction methods by a large margin.

The rest of the chapter is organized as follows. Section 3.2 presents the proposed model for unconstrained face verification in still images, with a further extension to video-based face verification in Section 3.2.1. Experimental results are reported respectively in Sections 3.3 and 3.4. We conclude in Section 3.5.

3.2 Reducing Transformation Differences

In this section, we first present a new dimensionality reduction approch for image-based face verification and person re-identification. Then we extend the proposed approach to unconstrained face verification in videos.

One challenging issue in unconstrained face verification and person re-identification is to retain the robustness to noise and large transformation variations in facial images or pedestrian.

To remove the redundant noise, we apply WPCA (i.e. equation (2.4)) which has been empirically shown to outperform the standard PCA (see [Deng et al., 2005]). Recall that the training images $\{x_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^{d_0}$. WPCA maps the original images into d-dimensional ($d \leq d_0$) subspace. For simplicity, we denote the WPCA-reduced images by $\{x_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$.

To reduce the effect of the large transformation differences existed in still images, motivated by the idea in [Moghaddam et al., 2000; Wang and Tang, 2004], we further map d-dimensional WPCA-reduced images to the intra-personal subspace and normalize the projected data. Specifically, WPCA (i.e. equation (2.4)) is applied on the difference set $\{x_i - x_j \in \mathbb{R}^d | (i,j) \in \mathcal{S}\}$, for which the covariance matrix is defined by

$$X_{\mathcal{S}} = \sum_{(i,j)\in\mathcal{S}} (x_i - x_j)(x_i - x_j)^T. \tag{3.1}$$

Let $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$ and $V = (v_1, \cdots, v_k)$ be the top leading k eigenvalues and eigenvectors of X_S. Then, the whitening process, i.e. the projection onto the intra-personal subspace spanned by

the k ($k \leq d$) eigenvectors and the normalization, is given by

$$\tilde{x} = \text{diag}(\lambda_1^{-1/2}, \ldots, \lambda_k^{-1/2}) V^T x. \tag{3.2}$$

Note that the features are weighted by the inverse of the eigenvalues, which penalizes the eigenvectors with large eigenvalues and therefore reduces the variance of the features, i.e. the transformation differences. In the special case that the dimensionality of the intra-personal subspace equals the dimensionality of the WPCA-reduced subspace, i.e. $k = d$, if $X_{\mathcal{S}}$ is invertible and denote

$$L_{\mathcal{S}} = V \text{diag}(\lambda_1^{1/2}, \ldots, \lambda_d^{1/2}), \tag{3.3}$$

then $X_{\mathcal{S}} = L_{\mathcal{S}} L_{\mathcal{S}}^T$ and equation (3.2) becomes $\tilde{x} = L_{\mathcal{S}}^{-1} x$. Later on, we refer to this whitening process (i.e. equation (3.2)) as **Intra-PCA**.

For verification in still images, we use Euclidean distance as the similarity function to measure the similarity between image-pairs. Specifically, denote the similarity function by f, then the similarity between an image-pair $(\tilde{x}_i, \tilde{x}_j)$ over the intra-personal subspace is given by

$$f(\tilde{x}_i, \tilde{x}_j) = -\|\tilde{x}_i - \tilde{x}_j\|^2. \tag{3.4}$$

3.2.1 Extension to Unconstrained Face Verification in Videos

In this section, we extend Intra-PCA to unconstrained face verification in videos where, instead of still images, a person is represented by a sequence of facial images/frames.

Denote the video samples by $\{X_i\}_{i=1}^N$, where each video $X_i = \{x_1^i, x_2^i, \ldots, x_{N_i}^i\}$ contains N_i frames and each frame $x_l^i \in \mathbb{R}^{d_0}$ is a d_0-dimensional vector. Given the index set of similar video-pairs $\mathcal{S}$ and that of dissimilar video-pairs $\mathcal{D}$, we generate all the frame-level pairs from $\mathcal{S}$ and $\mathcal{D}$.

Similar to the discussion on image-based face verification, we consider to remain robust to the noise and the large transformation variations in the video sequences. To this end, we perform WPCA and Intra-PCA to all the images within video sequences.

To reduce the noise, we collect all the frames within video sequences and regard them as the input for WPCA. Specifically, the covariance matrix of the video-based data is given by

$$\mathbf{C} = \sum_{i=1}^{N} \sum_{l=1}^{N_i} (x_l^i - m)(x_l^i - m)^T, \tag{3.5}$$

where $m = \frac{1}{\sum_{i=1}^N N_i} \sum_{i=1}^N \sum_{l=1}^{N_i} x_l^i$. WPCA computes d ($d \leq d_0$) eigenvectors with the largest eigenvalues of the covariance matrix $\mathbf{C}$ and maps the original images within video sequences into d-dimensional ($d \leq d_0$) subspace. For simplicity, we denote the WPCA-reduced video sequences by $\{X_i\}_{i=1}^N$, where $X_i = \{x_1^i, x_2^i, \ldots, x_{N_i}^i\}$ and $x_l^i \in \mathbb{R}^d$.

To retain the robustness to transformation differences existed in video sequences, we extend Intra-PCA to video-based data. In particular, the intra-personal covariance matrix for all the frame-level

pairs is defined as

$$\mathbf{X}_{\mathcal{S}} = \sum_{(i,j)\in\mathcal{S}} \frac{1}{N_i N_j} \sum_{l=1}^{N_i}\sum_{k=1}^{N_j} (x_l^i - x_k^j)(x_l^i - x_k^j)^T. \tag{3.6}$$

The mapping of the WPCA-reduced video sequences to k-dimensional intra-personal subspace ($k \leq d$) and the normalization is then given by the whitening process

$$\tilde{X} = \text{diag}(\gamma_1^{-1/2}, \ldots, \gamma_k^{-1/2}) U^T X, \tag{3.7}$$

where $\Gamma = \{\gamma_1, \ldots, \gamma_k\}$ and $U = (u_1, \cdots, u_k)$ are the top leading k eigenvalues and eigenvectors of $\mathbf{X}_{\mathcal{S}}$ respectively. By weighting feature vectors using the inverse of the eigenvalues, the eigenvectors with large eigenvalues are diminished and therefore the variance of the features (i.e. the transformation differences) are reduced. Similar to the previous discussion, in the special case that the dimensionality of the intra-personal subspace equals the dimensionality of WPCA-reduced subspace, i.e. $k = d$, if $\mathbf{X}_{\mathcal{S}}$ is invertible and denote

$$\mathbf{L}_{\mathcal{S}} = U \text{diag}(\gamma_1^{1/2}, \ldots, \gamma_d^{1/2}), \tag{3.8}$$

then $\mathbf{X}_{\mathcal{S}} = \mathbf{L}_{\mathcal{S}}\mathbf{L}_{\mathcal{S}}^T$ and equation (3.7) becomes $\tilde{X} = \mathbf{L}_{\mathcal{S}}^{-1} X$. For simplicity, we also refer to this whitening process (i.e. equation (3.7)) as **Intra-PCA**.

For verification in videos, we extend the similarity function f (i.e. equation (3.4)) to measure the similarity between video-pairs. To be specific, let F denote the similarity function for the comparison between video-pairs, the similarity between a video-pair $(\tilde{X}_i, \tilde{X}_j)$ over the intra-personal subspace is defined as the average of the similarities between all possible frame-level pairs:

$$F(\tilde{X}_i, \tilde{X}_j) = \frac{1}{N_i N_j} \sum_{l=1}^{N_i}\sum_{k=1}^{N_j} f(\tilde{x}_l^i, \tilde{x}_k^j) = -\frac{1}{N_i N_j} \sum_{l=1}^{N_i}\sum_{k=1}^{N_j} \|\tilde{x}_l^i - \tilde{x}_k^j\|^2. \tag{3.9}$$

The summation in equation (3.9) are normalized by the numbers of frames in videos $\tilde{X}_i$ and $\tilde{X}_j$, since the length of video sequences can vary in different videos.

3.3 Experiment One: Unconstrained Face Verification

This section provides an experimental study of Intra-PCA for unconstrained face verification. Specifically, we conduct experiments on the Labeled Faces in the Wild (LFW) database [Huang et al., 2007] for unconstrained face verification in still images and the YouTube Faces (YTF) database [Wolf et al., 2011a] for unconstrained face verification in videos. Section 3.3.1 presents the experimental results on the LFW dataset and the experimental results on the YTF database are reported in Section 3.3.2.

d \ k	50	100	200	300	400
100	60.52 ± 0.51	81.32 ± 0.46	N/A	N/A	N/A
200	55.28 ± 0.38	63.05 ± 0.63	82.32 ± 0.34	N/A	N/A
300	53.85 ± 0.47	59.27 ± 0.40	68.48 ± 0.40	82.18 ± 0.27	N/A
400	52.40 ± 0.27	55.75 ± 0.34	63.27 ± 0.70	70.90 ± 0.41	81.22 ± 0.41

(a)

d \ k	50	100	200	300	400
100	61.83 ± 0.66	82.53 ± 0.33	N/A	N/A	N/A
200	56.25 ± 0.33	63.60 ± 0.46	83.45 ± 0.24	N/A	N/A
300	52.42 ± 0.18	59.13 ± 0.43	68.93 ± 0.53	82.95 ± 0.23	N/A
400	53.05 ± 0.17	56.53 ± 0.34	63.67 ± 0.43	71.63 ± 0.35	82.50 ± 0.21

(b)

Table 3.1: Verification rates (%) of Intra-PCA using the SIFT descriptor in the restricted setting of LFW: (a) the original SIFT descriptor; (b) the square root of the SIFT descriptor. Parameters d and k are the dimensions of the WPCA-reduced subspace and the intra-personal subspace respectively.

3.3.1 Labeled Faces in the Wild

In this section, we evaluate the proposed Intra-PCA (i.e. equation (3.2)) on the Labeled Faces in the Wild (LFW) database [Huang et al., 2007]. Experiments are done in both the restricted and unrestricted setting of the LFW dataset. A detailed description of this database and its experimental protocol have been provided in Section 2.4.1. For feature representation, three facial descriptors are employed: LBP [Ojala et al., 2002], TPLBP [Wolf et al., 2008] and SIFT [Guillaumin et al., 2009]. Both the original values and the square roots of the above descriptors are tested as done in [Guillaumin et al., 2009; Wolf et al., 2008].

In particular, on each of the 10-fold cross-validation test, WPCA is first applied to reduce the dimensionality and remove the noise within facial images. The resultant images are further mapped to the intra-personal subspace by the whitening process given by equation (3.2). The covariance matrix to extract the principal components for WPCA is computed only from 9-fold training set. Also, similar image-pairs from the 9-fold training set are used to compute the intra-personal covariance matrix $X_{\mathcal{S}}$. Image vectors $\tilde{x}$ are L^2-normalized to 1 (i.e. $\|\tilde{x}\| = 1$) before the verification step. For the verification step, a test image-pair is classified to be similar if its similarity score is greater than some threshold, and dissimilar otherwise. In order to learn the threshold, we choose the value that gives the highest verification rate on the 9-fold training set.

Image Restricted Training Paradigm

Here, we demonstrate the effectiveness of Intra-PCA in the restricted setting of the LFW dataset.

Firstly, we conduct experiments to exploit the performance of Intra-PCA. To this end, we investigate Intra-PCA by varying the dimensions of the WPCA-reduced subspace and the intra-personal subspace, i.e. d and k. The maximum value of k is d. Table 3.1 and Table 3.2 report the results

d \ k	50	100	200	300	400
100	64.80 ± 0.49	83.07 ± 0.37	N/A	N/A	N/A
200	59.07 ± 0.67	67.12 ± 0.48	84.55 ± 0.63	N/A	N/A
300	55.50 ± 0.66	62.32 ± 0.43	72.63 ± 0.43	84.23 ± 0.55	N/A
400	54.12 ± 0.31	58.42 ± 0.43	67.53 ± 0.60	75.77 ± 0.51	83.55 ± 0.65

(a)

d \ k	50	100	200	300	400
100	60.52 ± 0.51	83.35 ± 0.45	N/A	N/A	N/A
200	59.40 ± 0.42	67.65 ± 0.36	84.60 ± 0.61	N/A	N/A
300	56.13 ± 0.39	61.70 ± 0.60	72.27 ± 0.64	84.45 ± 0.43	N/A
400	54.00 ± 0.37	58.10 ± 0.61	67.97 ± 0.44	75.22 ± 0.45	83.87 ± 0.36

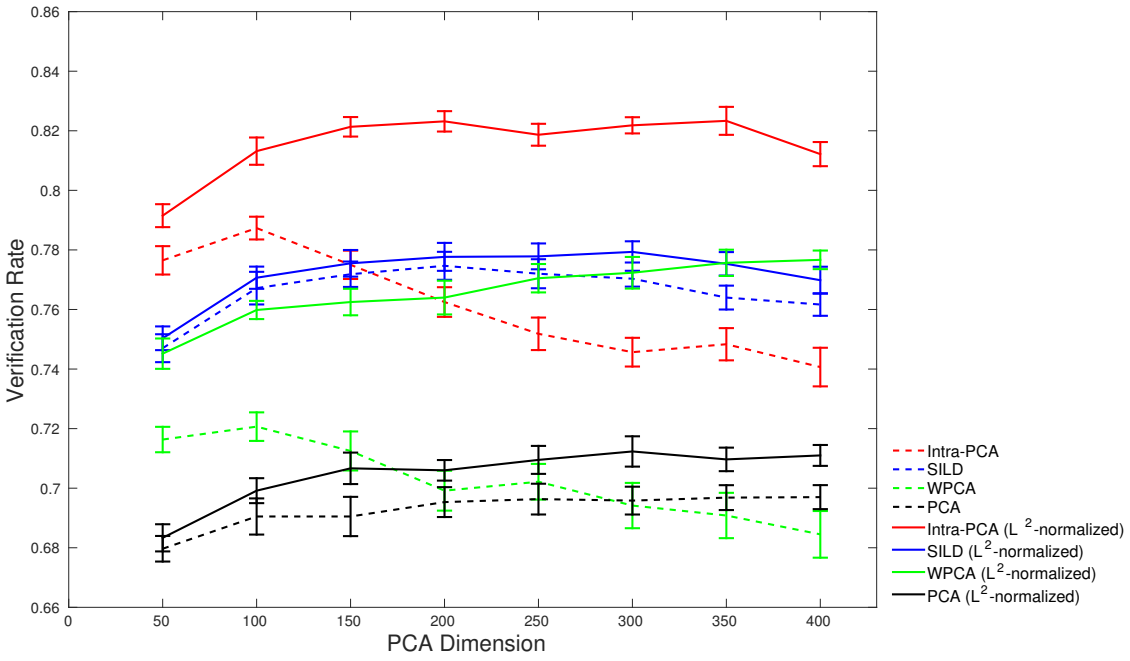
(b)

Table 3.2: Verification rate ($\pm$ standard error) of Intra-PCA using the LBP descriptor in the restricted setting of LFW: (a) the original LBP descriptor; (b) the square root of the LBP descriptor. Parameters d and k are the dimensions of the WPCA-reduced subspace and the intra-personal subspace respectively.

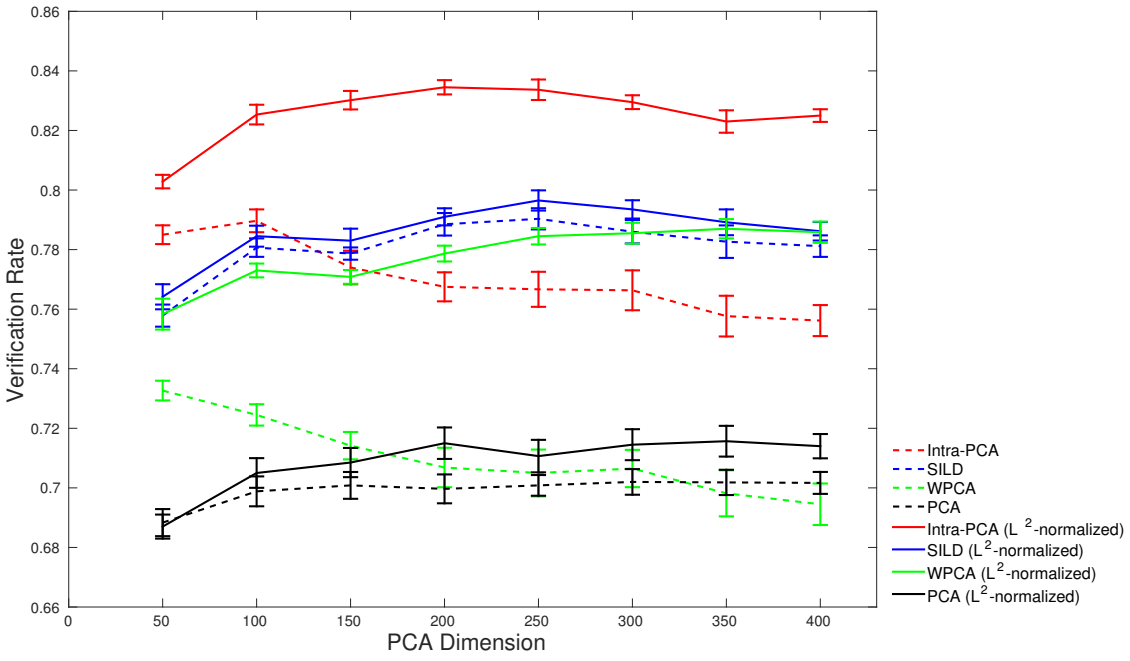
of Intra-PCA on the SIFT and LBP descriptors respectively. As we can see from Table 3.1, across different dimensions of the WPCA-reduced subspace, the verification rate of Intra-PCA on the SIFT descriptor increases with k. This is because as k increases more transformation differences are reduced. In the rest of the experiments, parameter k is tuned via 3-fold cross validation. We also notice that using the square root of the descriptors slightly improves the verification rate in most cases. Similar observations can be made on the LBP descriptor as shown in Table 3.2.

Secondly, we compare Intra-PCA with the dimensionality reduction methods such as PCA, WPCA and SILD [Kan et al., 2011]. We did not compare Intra-PCA with LDA [Belhumeur et al., 1997] and Bayesian face recognition [Moghaddam et al., 2000] since both LDA and Bayesian face recognition need the label information. For fairness of comparison, WPCA is applied before the implementation of SILD. Figure 3.1 and Figure 3.2 depict the comparison results on the SIFT and LBP descriptors respectively. From Figure 3.1 we can see that, on both the original SIFT descriptor and its square root, L^2-normalization improves the performance of most dimensionality reduction methods. This shows the effectiveness of the L^2-normalization as a preprocessing step. Moreover, on the SIFT (L^2- normalized) descriptor we can observe, across different PCA dimensions, Intra-PCA consistently outperforms PCA, WPCA and SILD by a large margin. Similar observations can be made on the LBP descriptor as shown in Figure 3.2. These observations show the effectiveness of Intra-PCA to remove the large transformation variations by mapping WPCA-reduced images into the intra-personal subspace using the whitening process given by equation (3.2).

Thirdly, in Figure 3.3, we present the similarity scores of 600 test image-pairs (300 similar image-pairs and 300 dissimilar image-pairs) obtained by WPCA and Intra-PCA on the SIFT descriptor in 3 folds of the 10-fold cross-validation test. The red and green points represent the similarity scores of similar and dissimilar image-pairs respectively. Figure 3.3 also reports the learned threshold (the black line) for each model. From Figure 3.3 we can observe, the boost in the performance of Intra-PCA in comparison to WPCA is mainly from dissimilar image-pairs. Indeed, the numbers of

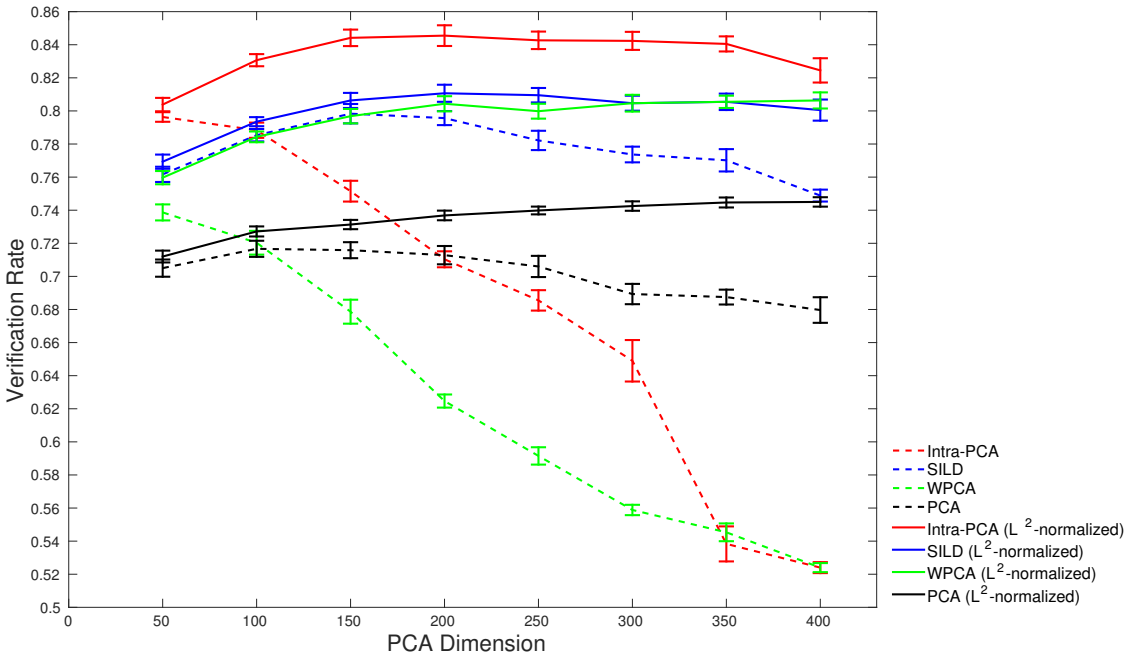


(a)

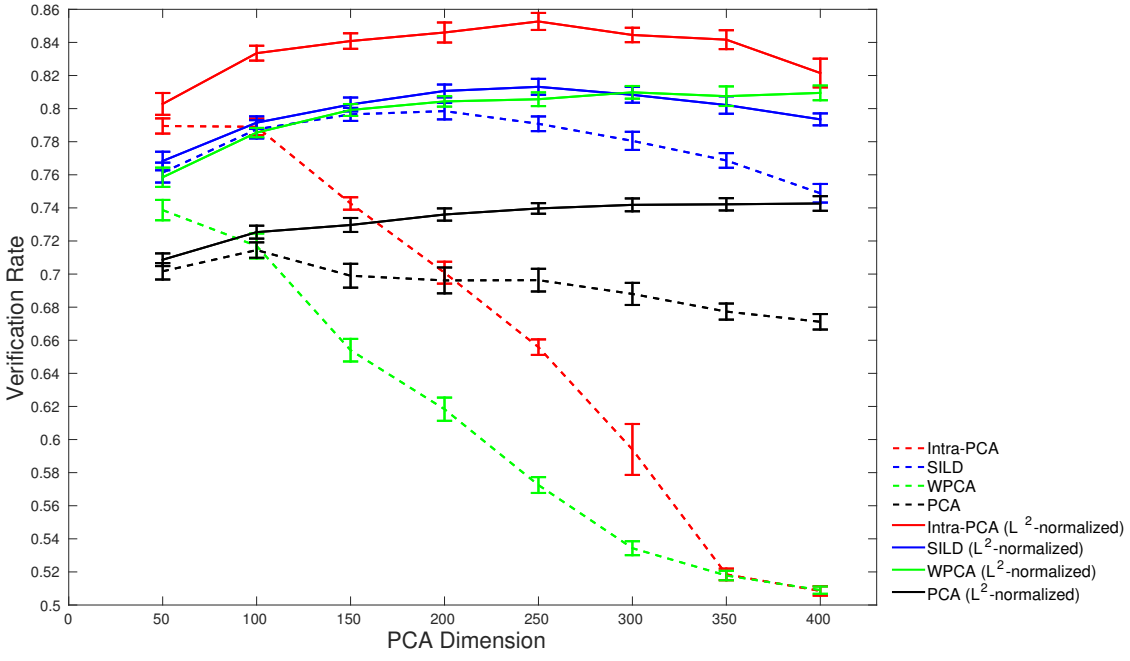


(b)

Figure 3.1: Comparison of PCA, WPCA, SILD and Intra-PCA using the original SIFT descriptor and its square root in the restricted setting of LFW: (a) SIFT descriptor; (b) the square root of the SIFT descriptor. L^2- normalized means the features are L^2-normalized to 1.

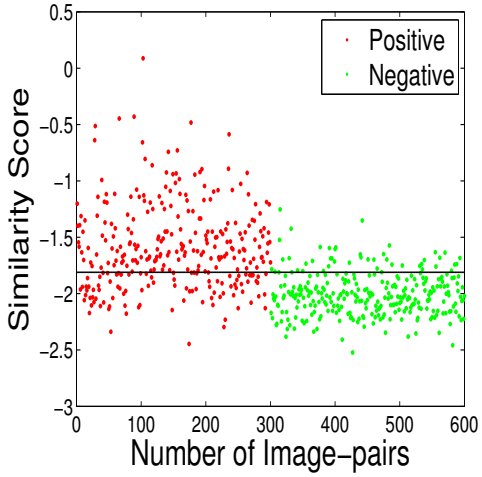


(a)

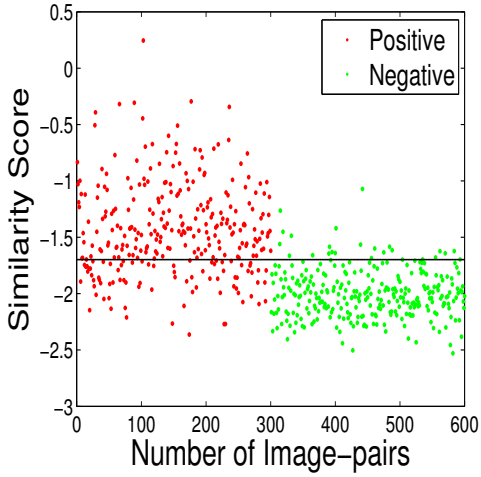


(b)

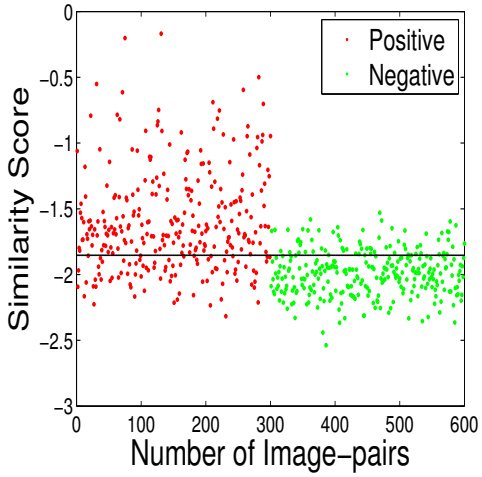
Figure 3.2: Comparison of PCA, WPCA, SILD and Intra-PCA using the original LBP descriptor and its square root in the restricted setting of LFW: (a) LBP descriptor; (b) the square root of the LBP descriptor. L^2- normalized means the features are L^2-normalized to 1.



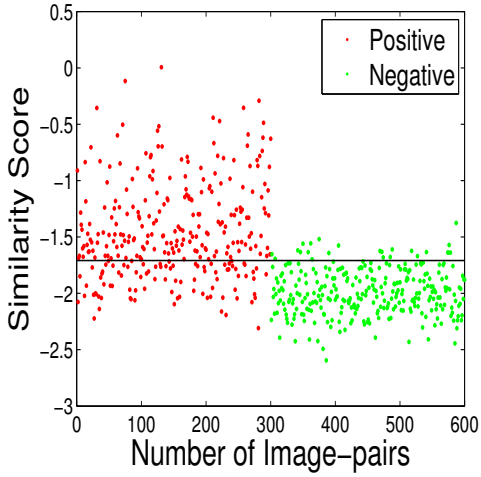
(a) WPCA, Fold 2



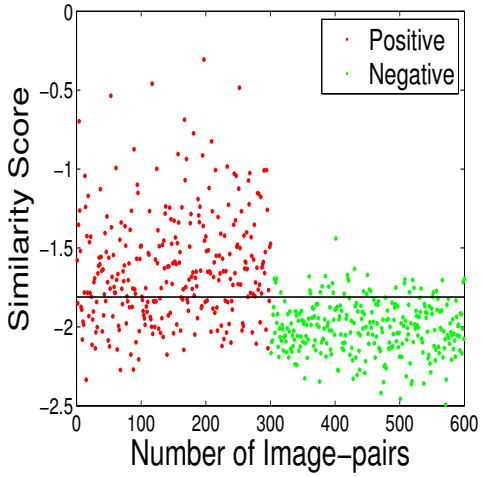
(b) Intra-PCA, Fold 2



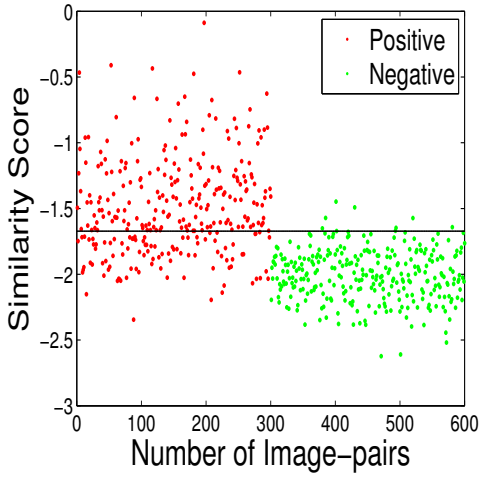
(c) WPCA, Fold 5



(d) Intra-PCA, Fold 5



(e) WPCA, Fold 8



(f) Intra-PCA, Fold 8

Figure 3.3: Similarity scores of 600 test images-pairs (300 similar image-pairs and 300 dissimilar image-pairs) obtained by WPCA and Intra-PCA on the SIFT descriptor in 3 folds of the 10-fold cross-validation test in the restricted setting of LFW: the red and green points represent similar and dissimilar image-pairs respectively; the black line is the learned threshold.

dissimilar pairs that are correctly classified by WPCA (green points that are below the black line in Figure 3.3a, 3.3c and 3.3e) are 253, 232 and 263 respectively, while the numbers of dissimilar pairs that are correctly classified by Intra-PCA (green points that are below the black line in Figure 3.3b, 3.3d and 3.3f) are 275, 276 and 291 respectively. Similar results can be observed in the remaining seven folds. This could be due to the fact that by implementing Intra-PCA, both noise and transformation differences are largely reduced and the main differences remained in image-pairs are identity differences, which makes dissimilar pairs easier to be verified.

Lastly, to examine the efficacy of Intra-PCA in dealing with dissimilar image-pairs that have large transformation differences, in Figure 3.4, we show the qualitative examples for the comparison between WPCA and Intra-PCA. To be specific, Figure 3.4 depicts dissimilar pairs that are correctly classified by Intra-PCA but incorrectly classified by WPCA on the SIFT descriptor. As expected, we see by incorporating the robustness to large transformation variations, Intra-PCA successfully classifies most of the dissimilar pairs that have large transformation variations such as pose, occlusion and background. These observations show the benefit of the proposed Intra-PCA.

Image Unrestricted Training Paradigm

Here, we evaluate the performance of Intra-PCA in the unrestricted setting of the LFW database, where the label information allows us to generate more image-pairs. Following the same procedure as in [Huang et al., 2007], we randomly generate 1000, 1500 and 2000 image-pairs per fold (instead of 600 pairs as provided in the restricted setting), where half are similar image-pairs and half are dissimilar ones.

Methods	d	1000	1500	2000
PCA	100	69.60 ± 0.36	69.87 ± 0.39	69.87 ± 0.34
WPCA	100	76.17 ± 0.46	76.17 ± 0.46	75.88 ± 0.47
LDA	100	70.72 ± 0.36	70.82 ± 0.43	70.95 ± 0.42
SILD [Kan et al., 2011]	100	77.15 ± 0.35	77.12 ± 0.37	76.80 ± 0.34
Intra-PCA	100	81.82 ± 0.34	81.57 ± 0.37	81.65 ± 0.33
PCA	200	70.67 ± 0.46	71.00 ± 0.44	71.03 ± 0.44
WPCA	200	76.65 ± 0.45	76.75 ± 0.46	76.75 ± 0.47
LDA	200	72.12 ± 0.37	71.88 ± 0.49	71.98 ± 0.49
SILD [Kan et al., 2011]	200	78.32 ± 0.48	77.67 ± 0.50	77.45 ± 0.44
Intra-PCA	200	82.20 ± 0.57	82.10 ± 0.51	82.03 ± 0.58
PCA	300	71.03 ± 0.39	70.97 ± 0.55	71.17 ± 0.54
WPCA	300	77.62 ± 0.51	77.42 ± 0.52	77.53 ± 0.52
LDA	300	72.10 ± 0.59	72.03 ± 0.57	72.02 ± 0.33
SILD [Kan et al., 2011]	300	79.25 ± 0.44	78.80 ± 0.38	78.57 ± 0.33
Intra-PCA	300	83.13 ± 0.53	82.95 ± 0.62	82.87 ± 0.62
PCA	400	71.23 ± 0.44	71.37 ± 0.48	71.38 ± 0.43
WPCA	400	77.52 ± 0.26	77.50 ± 0.23	77.50 ± 0.26
LDA	400	71.08 ± 0.56	71.28 ± 0.58	71.45 ± 0.51
SILD [Kan et al., 2011]	400	79.10 ± 0.18	78.77 ± 0.41	78.60 ± 0.28
Intra-PCA	400	82.63 ± 0.50	82.95 ± 0.48	82.95 ± 0.46

Table 3.3: Verification rate ($\pm$ standard error) of PCA, WPCA, LDA, SILD and Intra-PCA versus the number of image-pairs per fold using the SIFT descriptor in the unrestricted setting of LFW.

Figure 3.4: Examples of dissimilar image-pairs that are correctly classified by Intra-PCA while incorrectly classified by WPCA in the restricted setting of LFW.

Methods	d	1000	1500	2000
PCA	100	72.47 ± 0.33	72.63 ± 0.28	72.67 ± 0.30
WPCA	100	78.67 ± 0.34	78.63 ± 0.34	78.58 ± 0.34
LDA	100	74.07 ± 0.33	74.37 ± 0.33	74.47 ± 0.44
SILD [Kan et al., 2011]	100	79.62 ± 0.27	79.13 ± 0.35	78.90 ± 0.30
Intra-PCA	100	83.15 ± 0.34	83.23 ± 0.39	82.83 ± 0.37
PCA	200	73.65 ± 0.28	73.65 ± 0.29	73.63 ± 0.24
WPCA	200	80.32 ± 0.48	80.25 ± 0.45	80.40 ± 0.43
LDA	200	75.33 ± 0.47	75.52 ± 0.38	75.50 ± 0.40
SILD [Kan et al., 2011]	200	81.03 ± 0.35	80.90 ± 0.45	80.73 ± 0.43
Intra-PCA	200	84.65 ± 0.35	84.95 ± 0.30	84.97 ± 0.27
PCA	300	74.12 ± 0.32	73.82 ± 0.32	74.10 ± 0.30
WPCA	300	79.92 ± 0.31	79.88 ± 0.34	79.73 ± 0.26
LDA	300	74.47 ± 0.52	74.57 ± 0.52	74.37 ± 0.39
SILD [Kan et al., 2011]	300	81.37 ± 0.41	81.17 ± 0.35	80.60 ± 0.41
Intra-PCA	300	84.67 ± 0.40	84.70 ± 0.46	84.62 ± 0.37
PCA	400	74.40 ± 0.35	74.43 ± 0.31	74.67 ± 0.33
WPCA	400	80.42 ± 0.49	80.37 ± 0.47	80.37 ± 0.46
LDA	400	75.00 ± 0.50	74.97 ± 0.49	75.03 ± 0.51
SILD [Kan et al., 2011]	400	81.30 ± 0.39	80.98 ± 0.43	80.95 ± 0.54
Intra-PCA	400	84.47 ± 0.47	84.42 ± 0.51	84.83 ± 0.57

Table 3.4: Verification rate ($\pm$ standard error) of PCA, WPCA, LDA, SILD and Intra-PCA versus the number of image-pairs per fold using the LBP descriptor in the unrestricted setting of LFW.

To show the effectiveness of Intra-PCA, we compare Intra-PCA with PCA, WPCA, LDA [Belhumeur et al., 1997] and SILD [Kan et al., 2011]. Table 3.3 and Table 3.4 present the comparison results on the SIFT and LBP descriptors respectively. We can observe from Table 3.3 that for each PCA dimension, across the number of image-pairs per fold, WPCA is better than PCA and LDA, while Intra-PCA is significantly better than PCA, WPCA, LDA and SILD. Similar observation can be made on the LBP descriptor as shown in Table 3.4. These observations verify the effectiveness of Intra-PCA to remove the large transformation variations using the whitening process given by equation (3.2).

3.3.2 YouTube Faces Database

Now we evaluate Intra-PCA (i.e. equation (3.7)) on the YouTube Faces (YTF) database [Wolf et al., 2011a] for unconstrained face verification in videos. Follwing the work in [Wolf et al., 2011a], we mainly focus on the restricted protocol of the YTF database. A brief introduction of the YTF dataset and its experimental setting have been given in Section 2.4.2. For feature representation, we directly use the features provided in [Wolf et al., 2011a], i.e. LBP, CSLBP and FPLBP.

In particular, on each of the 10-fold cross-validation test, Intra-PCA is implemented to reduce the transformation differences. Specifically, WPCA is applied to reduce the dimensionality and remove the noise within facial video sequences, and the resultant video sequences are further mapped to the intra-personal subspace by the whitening process given by equation (3.7). The

parameter k (i.e. the dimensionality of the intra-personal subspace) is tuned via three-fold cross validation. For verification, two video sequences in the test set are ascibed to the same person if their similarity score is greater than some threshold, and different identities otherwise. To learn the threshold, we choose the value that gives the highest verification rate on the 4500 video-pairs of training set.

Method	d	LBP			FPLBP			CSLBP		
		Accuracy ± SE	AUC	EER	Accuracy ± SE	AUC	EER	Accuracy ± SE	AUC	EER
PCA	100	68.4 ± 2.5	74.5	32.2	68.4 ± 2.6	75.3	31.6	65.6 ± 2.2	71.9	34.5
WPCA	100	73.3 ± 2.2	81.1	26.7	71.2 ± 2.4	78.4	28.8	70.7 ± 2.5	77.8	29.4
Intra-PCA	100	76.7 ± 2.0	84.9	23.2	73.7 ± 2.3	81.3	26.3	72.9 ± 2.0	80.3	27.2
PCA	200	68.9 ± 2.5	75.6	31.3	69.5 ± 2.1	76.3	30.4	71.8 ± 2.7	72.4	33.8
WPCA	200	74.8 ± 2.0	82.4	25.8	72.4 ± 1.5	79.5	28.0	72.4 ± 2.5	78.2	28.5
Intra-PCA	200	78.3 ± 1.7	86.3	22.0	74.7 ± 2.0	82.2	25.5	73.5 ± 2.2	80.9	26.6
PCA	300	69.0 ± 2.2	76.0	31.1	70.2 ± 2.1	76.8	30.1	72.2 ± 2.3	72.7	33.9
WPCA	300	75.1 ± 1.6	82.6	25.4	72.6 ± 1.8	79.6	27.5	72.4 ± 2.8	78.6	28.3
Intra-PCA	300	77.8 ± 2.1	86.7	21.1	74.9 ± 1.7	82.2	25.5	74.0 ± 3.1	81.1	26.6
PCA	400	69.5 ± 2.5	76.2	30.8	70.1 ± 2.2	77.1	29.9	67.3 ± 2.0	73.2	33.6
WPCA	400	74.7 ± 1.9	82.8	25.3	72.2 ± 1.5	79.8	27.6	73.3 ± 2.0	79.5	27.3
Intra-PCA	400	78.2 ± 2.0	86.8	21.6	74.7 ± 2.1	82.3	25.3	74.7 ± 2.1	82.0	25.1

Table 3.5: Comparison of PCA, WPCA and Intra-PCA on the LBP, FPLBP and CSLBP descriptors in the restricted setting of YouTube Faces database.

We conduct experiments to demonstrate the effectiveness of Intra-PCA. Firstly, we compare Intra-PCA with PCA and WPCA. We did not compare Intra-PCA with LDA [Belhumeur et al., 1997] and Bayesian face recognition [Moghaddam et al., 2000] and SILD [Kan et al., 2011] since LDA,

Bayesian face recognition and SILD are designed for face verification in still images. Table 3.5 lists the comparison results on the LBP, FPLBP and CSLBP descriptors. We can see from Table 3.5 that on the LBP descriptor, across different PCA dimensions, WPCA is better than PCA and Intra-PCA outperforms WPCA by a large margin in terms of Accuracy, AUC and EER. Similar observations can be made on the FPLBP and CSLBP descriptors. These observations demonstrate the effectiveness of Intra-PCA to remove the transformation differences using the whitening process (i.e. equation (3.7)).

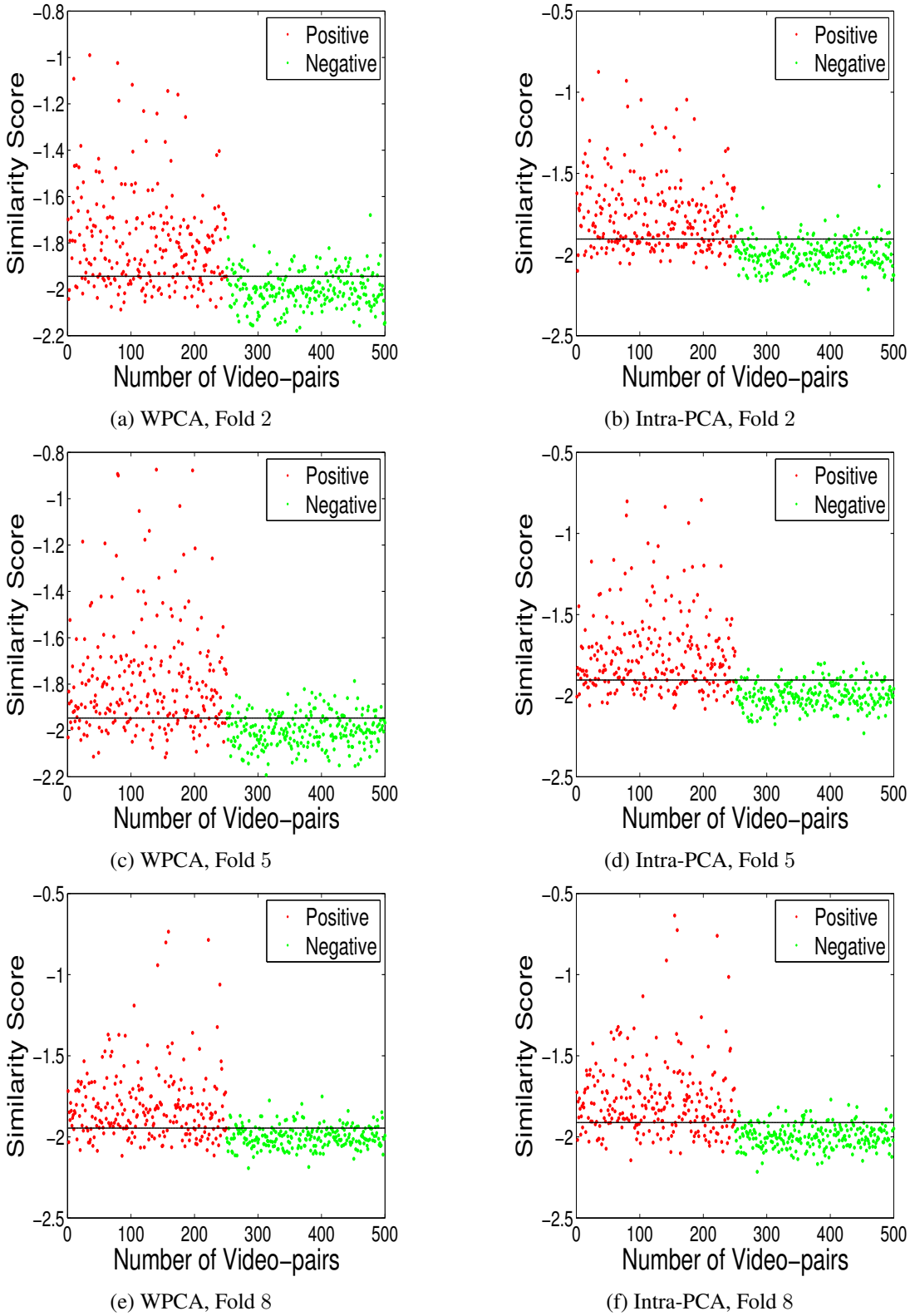
To give an insight of the boost in the performance obtained from Intra-PCA in comparison to WPCA, Figure 3.5 reports the similarity scores of 500 test video-pairs (250 similar video-pairs and 250 dissimilar video-pairs) obtained by WPCA and Intra-PCA using the LBP descriptor in 3 folds of the 10-fold cross-validation test. The red and green points represent the similarity scores of similar and dissimilar video-pairs respectively. In Figure 3.3, we also report the learned threshold (the black line) for each model. From Figure 3.3 we can observe, the improvement of Intra-PCA over WPCA is mainly from the improvement in verifying dissimilar video-pairs, which is consistent with the observation for image-based face verification on the LFW dataset (see Section 3.3.1). Indeed, the numbers of dissimilar pairs that are correctly classified by WPCA (green points that are below the black line in Figure 3.5a, 3.5c and 3.5e) are 193, 192 and 201 respectively, while the numbers of dissimilar pairs that are correctly classified by Intra-PCA (green points that are below the black line in Figure 3.5b, 3.5d and 3.5f) are 215, 219 and 219 respectively. Similar observations can be made in the remaining seven folds. A possible explanation could be that applying Intra-PCA removes most of the noise and the transformation differences existed in video-pairs and the remaining differences are mainly identity differences, from which verifying the dissimilar pairs benefits.

3.4 Experiment Two: Person Re-Identification

In this section, we provide the experimental study of Intra-PCA (i.e. equation (3.2)) for person re-identification across spatially disjoint cameras. We evaluate our method on the benchmark Viewpoint Invariant Pedestrian Recognition (VIPeR) database [Gray et al., 2007]. A brief introduction of the VIPeR dataset and its experimental setting have been given in Section 2.4.3. For feature representation, as described in Section 2.4.3, we use the features [1] generated in [Kostinger et al., 2012]. In each repeat, Intra-PCA is implemented to reduce the transformation differences. In particular, since the features provided by Kostinger et al. [2012] are already PCA-reduced, we apply Intra-PCA by directly mapping the PCA-reduced features to the intra-personal subspace using the whitening process given by equation (3.2).

To demonstrate the effectiveness of Intra-PCA, we compare Intra-PCA with PCA and SILD [Kan et al., 2011]. We did not compare Intra-PCA with WPCA because the features are already PCA-reduced features. We also did not do the comparison with LDA [Belhumeur et al., 1997] and Bayesian face recognition [Moghaddam et al., 2000] since both methods need the label information. The parameters d (i.e. the dimensionality of the PCA-reduced subspace) and k (i.e. the dimensionality of the intra-personal subspace) are tuned via three-fold cross validation.

[1] Available at: `http://lrs.icg.tugraz.at/research/kissme/`.



(a) WPCA, Fold 2

(b) Intra-PCA, Fold 2

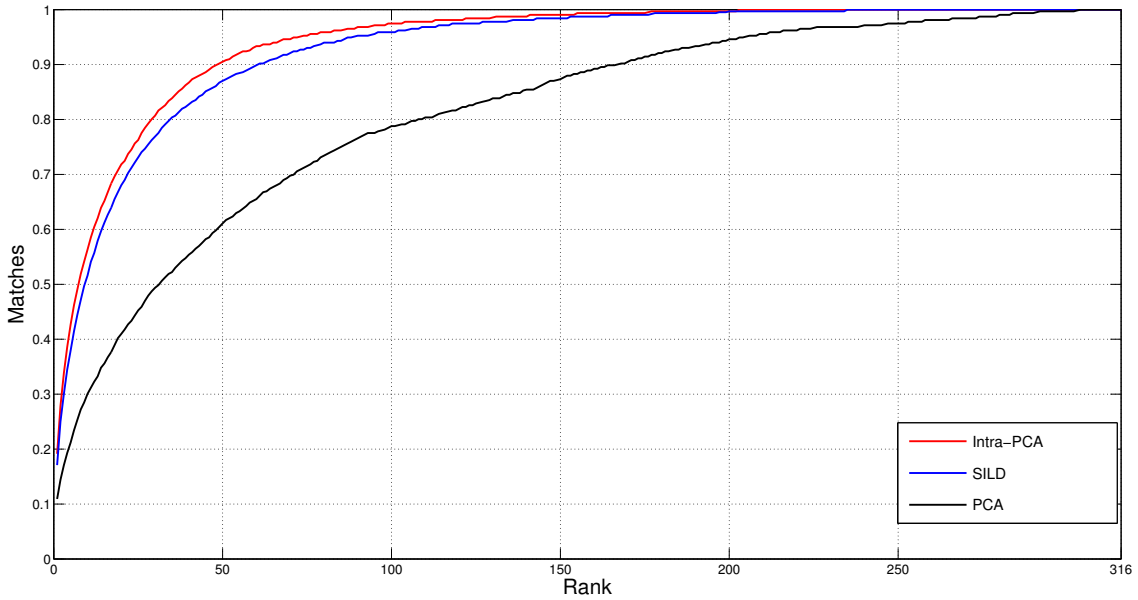
(c) WPCA, Fold 5

(d) Intra-PCA, Fold 5

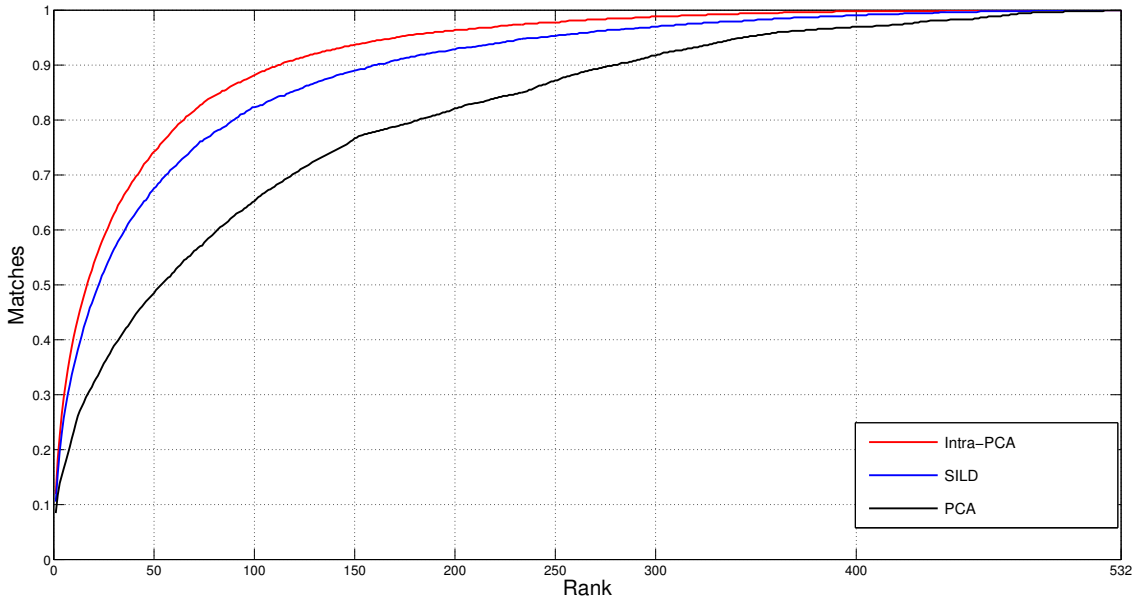
(e) WPCA, Fold 8

(f) Intra-PCA, Fold 8

Figure 3.5: Similarity scores of 500 test video-pairs (250 similar video-pairs and 250 dissimilar video-pairs) obtained by WPCA and Intra-PCA on the LBP descriptor in 3 folds of the 10-fold cross-validation test in the restricted setting of YTF: the red and green points represent similar and dissimilar video-pairs respectively; the black line is the learned threshold.



(a) $h = 316$



(b) $h = 532$

Figure 3.6: CMC curves of PCA, SILD and Intra-PCA on the VIPeR dataset: (a) $h = 316$ and (b) $h = 532$, where h is the number of persons in the test set.

Figure 3.6 depicts the CMC curves of PCA, SILD and Intra-PCA, and Table 3.6 reports the CMC scores in the range of the first 50 ranks. In particular, Figure 3.6a and Table 3.6a are for $h = 316$ (316 persons for testing and 316 persons for training), while Figure 3.6b and Table 3.6b are for $h = 532$ (532 persons for testing and 100 persons for training). Here, CMC curve represents the expectation of finding the true match within the top r ranks, see Section 2.4.3 for details. As we can see from Table 3.6a and Table 3.6b that, for both $h = 316$ and $h = 532$, Intra-PCA is much better than PCA and SILD, which further shows the effectiveness of Intra-PCA to remove the transformation variations by the whitening process given by equation (3.2).

RANK	1	5	10	20	50
PCA	10.92	21.20	30.06	40.98	61.08
SILD	17.09	37.97	51.58	68.04	87.03
Intra-PCA	19.15	42.72	56.33	71.84	90.51

(a) $h = 316$

RANK	1	5	10	20	50
PCA	8.46	16.54	23.31	32.24	48.50
SILD	10.53	25.75	35.15	47.65	67.58
Intra-PCA	11.94	29.51	40.60	53.95	74.25

(b) $h = 532$

Table 3.6: Comparison of matching rates with PCA, SILD and Intra-PCA on the VIPeR dataset: (a) $h = 316$ and (b) $h = 532$, where h is the number of persons in the test set.

3.5 Conclusion

In this chapter, we proposed a novel dimensionality reduction model called Intra-PCA for unconstrained face verification and person re-identification under the scenario that only pairwise information is provided while label information is not available. We formulated our model by incorporating the robustness to large transformation differences. Specifically, WPCA is first applied to reduce the noise and the resultant images are further mapped to the intra-personal subspace by the whitening process given by equation (3.2) to reduce the transformation variations. The proposed model is further extended to unconstrained face verification in videos. We demonstrate Intra-PCA on the benchmark LFW [Huang et al., 2007] and YTF [Wolf et al., 2011a] datasets for unconstrained face verification in still images and videos, and the VIPeR [Gray et al., 2007] database for person re-identification. Experimental results have shown its superior performance to the state-of-the-art dimensionality reduction methods.

Now we discuss some possible future directions. In Section 3.2, WPCA was applied to reduce the redundant noise and dimensionality, and experimental results in Sections 3.3 and 3.4 have shown its large improvement over the standard PCA. However, as a linear dimensionality reduction method, WPCA could not capture the nonlinear manifold where the faces usually lie on, especially when the facial images/videos are taken in the wild and exhibit large transformation differences. Kernel PCA [Schölkopf et al., 1998], as have been reviewed in Section 2.2.2, addresses the above issue by extending linear PCA using techniques of kernel methods. It would be interesting to extend WPCA to the nonlinear case using similar idea as Kernel PCA, so that WPCA could be more applicable to face verification and person re-identification under unconstrained conditions.

In Section 3.2, Euclidean distance (see equations (3.4) and (3.9)) was used as the distance metric to discriminate similar image-pairs/video-pairs from dissimilar image-pairs/video-pairs. However, the straightforward use of the Euclidean distance is often not desirable because it gives equal weights to all features and fails to capture the specific nature of the task at hand. In the next chapter, we will focus on similarity metric learning for discrimination. We propose a novel regularization framework of learning distance metrics and similarity functions for unconstrained face verification and person re-identification, which incorporates both the robustness to large transformation differences and the discriminative power of similarity metric learning.

4 Similarity Metric Learning over the Intra-personal Subspace

4.1 Introduction

In the previous chapter, we proposed a new model called Intra-PCA for unconstrained face verification and person re-identification by incorporating the robustness to large transformation differences caused by pose, illumination and occlusion changes. In this chapter, we explore to combine it with the discriminative power of similarity metric learning (i.e. metric learning or similarity learning, see Section 2.3 for details) methods for discriminating similar image-pair/video-pairs from dissimilar image-pair/video-pairs. A large amount of studies [Xing et al., 2003; JacobGoldberger and GeoffHinton, 2004; Globerson and Roweis, 2005; Weinberger et al., 2006; Davis et al., 2007; Torresani and Lee, 2007] has been devoted to similarity metric learning. Unfortunately, most of the above methods only yield modest performance when applying to unconstrained face verification and person re-identification, as shown in the experiments in [Guillaumin et al., 2009; Zheng et al., 2011; Ying and Li, 2012]. This may be partly because most of such methods are designed for improving kNN classification or clustering, which would be not necessarily suitable for unconstrained face verification and person re-identification.

Another drawback of existing similarity metric learning methods is that such methods mainly focus on the discrimination of distance metrics or similarity functions and do not explicitly take into account on how to reduce the detrimental effect of the transformation differences. Hence, the learned metrics may not be robust to the transformation differences and their recognition performance can be degenerated.

In this chapter, we build on previous studies [Xing et al., 2003; Weinberger et al., 2006; Davis et al., 2007; Guillaumin et al., 2009; Kan et al., 2011; Nguyen and Bai, 2011; Ying and Li, 2012] to show the great potential of similarity metric learning methods to boost the recognition performance for images/videos without hand-crafting advanced feature descriptors. In particular, we develop a novel regularization framework to learn distance metrics and similarity functions for unconstrained face verification and person re-identification in still images, and further extend this framework to video-based face verification. The main novelty of our formulation is to incorporate both the robustness to large transformation variations and the discriminative power of similarity metric learning, a property that most existing similarity metric learning methods do not hold. In addition, our formulation is a convex optimization problem, and hence a global solution can be efficiently found by existing optimization algorithms. This is, for instance, not the case for the

[1] Some of the material in this chapter has been published in [Cao et al., 2013] and the code is available at: `http://www.albany.edu/~yy298919/software.html`.

current similarity metric learning model [Nguyen and Bai, 2011], see Section 2.3.3. We conduct experiments on the Labeled Faces in the Wild (LFW) [Huang et al., 2007] and the YouTube Faces (YTF) [Wolf et al., 2011a] databases, standard testbeds for unconstrained face verification in still images and videos. Further, we provide experimental results on the benchmark Viewpoint Invariant Pedestrian Recognition (VIPeR) [Gray et al., 2007] database, the largest publicly available dataset for person re-identification. Our new approach outperforms state-of-the-art similarity metric learning methods and is competitive with or even outperform the domain specific state-of-the-arts.

The rest of the chapter is organized as follows. Section 4.2 presents the proposed model for face verification in still images and person re-identification. Section 4.3 extends the model to video-based face verification. Experimental results on unconstrained face verification and person re-identification are reported respectively in Sections 4.4 and 4.5. Section 4.6 discusses metric learning models that are closely related to our work. We conclude in Section 4.7.

4.2 Similarity Metric Learning for Recognition in Still Images

In this section, we develop a novel regularization framework to learn distance metrics and similarity functions for image-based face verification in the wild and person re-identification.

4.2.1 Formulation of the Learning Problem

To obtain a good distance metric or similarity function measuring the similarity between images, we formulate the learning objective by considering both the robustness to large transformation differences and the discrimination for separating similar image-pairs from dissimilar image-pairs.

To remain robust to the noise and the large transformation differences in images, we employ the previously proposed Intra-PCA to map the original images onto the intra-personal subspace using the whitening process (see equation (3.2)).

After the images are mapped to the intra-personal subspace, we now consider the discrimination using a distance metric or similarity function, a property that discriminates similar image-pairs from dissimilar image-pairs. To this end, one option is to use the Mahalanobis distance d_M which was observed to significantly improve the results for face identification [Cinbis et al., 2011] and person re-identification [Dikmen et al., 2011]. The other option is to use the cosine similarity function CS_M which was observed to outperform the distance measurement d_M in face verification [Nguyen and Bai, 2011]. However, it is not a convex function with respect to M, see Section 2.3.3 for details. Recent studies [Chechik et al., 2010; Shalit et al., 2010] observed that the similarity function s_M has a promising performance on image similarity search.

Motivated by the above observations, we consider a generalized similarity function $f_{(M,G)}$ to measure the similarity of an image-pair, which can be instantiated by the standard similarity measures d_M and s_M. Specifically, the generalized similarity function $f_{(M,G)}$ is defined, for any image-pair

$(\tilde{x}_i, \tilde{x}_j)$, by

$$f_{(M,G)}(\tilde{x}_i, \tilde{x}_j) = s_G(\tilde{x}_i, \tilde{x}_j) - d_M(\tilde{x}_i, \tilde{x}_j). \tag{4.1}$$

It is easy to see that $f_{(M,G)}$ is linear and convex with respect to variable (M, G). Note that the generalized similarity function $f_{(M,G)} = -d_M$ when $G = 0$ and $f_{(M,G)} = s_G$ when $M = 0$.

Let $\mathcal{P} = \mathcal{S} \cup \mathcal{D}$ denote the index set of all pairwise constraints. The output $y = \{y_{ij} \in \{\pm 1\}, (i, j) \in \mathcal{P}\}$ is defined by, $y_{ij} = 1$ if image $\tilde{x}_i$ is similar to $\tilde{x}_j$ (i.e. images from the same person), and -1 otherwise. To better discriminate similar image-pairs from dissimilar image-pairs, we should learn M and G from the available data such that $f_{(M,G)}(\tilde{x}_i, \tilde{x}_j)$ reports a large score for $y_{ij} = 1$ and a small score otherwise. Based on this rationale, we derive the formulation of the empirical error:

$$\mathcal{E}_{\text{emp}}(M, G) = \sum_{(i,j)\in\mathcal{P}} (1 - y_{ij} f_{(M,G)}(\tilde{x}_i, \tilde{x}_j))_+. \tag{4.2}$$

Minimizing the above empirical error with respect to M and G will encourage the discrimination of similar image-pairs from dissimilar ones. However, directly minimizing the functional $\mathcal{E}_{\text{emp}}$ does not guarantee a robust similar function $f_{(M,G)}$ to the large transformation variations and also will lead to overfitting. Below, we propose a novel regularization framework which learns a robust and discriminative similarity function.

Proposed Framework. Based on the above discussions, our target now is to learn matrices M and G such that $f_{(M,G)}$ not only retains the robustness to large transformation variations but also preserves a good discriminative information. To this end, we propose a new method referred to as *Similarity Metric Learning over the Intra-personal Subspace* which is given by

$$\min_{M,G\in\mathbb{S}^d} \mathcal{E}_{\text{emp}}(M, G) + \frac{\gamma}{2}(\|M - I\|_F^2 + \|G - I\|_F^2). \tag{4.3}$$

By introducing the slacking variables, the above formulation is identical to:

$$\begin{aligned} \min_{M,G\in\mathbb{S}^d} \quad & \sum_{t\in\mathcal{P}} \xi_t + \frac{\gamma}{2}(\|M - I\|_F^2 + \|G - I\|_F^2), \\ \text{s.t.} \quad & y_{ij}[f_{(M,G)}(\tilde{x}_i, \tilde{x}_j)] \geq 1 - \xi_{ij}, \\ & \xi_t \geq 0, \ \ \forall t = (i, j) \in \mathcal{P}. \end{aligned} \tag{4.4}$$

The regularization term $\|M - I\|_F^2$ and $\|G - I\|_F^2$ in the above formulations prevents image vectors $\tilde{x}$ in the intra-personal subspace from being distorted too much, and hence retains the most robustness of the intra-personal subspace. Minimizing the empirical term $\sum_{(i,j)\in\mathcal{P}} \xi_{ij}$ promotes the discriminative power of $f_{M,G}$ for discriminating similar image-pairs from dissimilar ones. The positive parameter γ is trade-offing the effects of the two terms in the objective function of (4.4). We emphasize here that, without loss of generality, we did not constrain M or G to be positive semi-definite in the above formulation. Later on, we refer to formulation (4.4) as ***Sub-SML***.

4.2.2 Optimization Algorithm

We now turn our attention to the optimization algorithm of (4.4). It is easy to see that Sub-SML is a convex optimization problem which guarantees a global solution.

For notational simplicity, for any $t = (i, j) \in \mathcal{P}$, let $\widetilde{X}_t = (\tilde{x}_i - \tilde{x}_j)(\tilde{x}_i - \tilde{x}_j)^T$ and $\widetilde{\widetilde{X}}_t = \tilde{x}_i \tilde{x}_j^T$. We can establish the dual problem of Sub-SML as follows.

Theorem 1. *The dual formulation of Sub-SML (i.e. formulation (4.4)) can be written as*

$$\begin{aligned}\max_{0\le\alpha\le 1} \sum_{t\in\mathcal{P}} \alpha_t + \sum_{t=(i,j)\in\mathcal{P}} \alpha_t y_t(\|\tilde{x}_i - \tilde{x}_j\|^2 - \tilde{x}_i^T \tilde{x}_j) \\ -\tfrac{1}{2\gamma}\big[\big\|\sum_{t\in\mathcal{P}} y_t\alpha_t \widetilde{X}_t\big\|_F^2 + \big\|\sum_{t\in\mathcal{P}} y_t\alpha_t \widetilde{\widetilde{X}}_t\big\|_F^2\big].\end{aligned} \tag{4.5}$$

Moreover, if the optimal solution of (4.5) is denoted by α^* then the optimal solution (M^*, G^*) of (4.4) is given by $M^* = I - \frac{1}{\gamma}\sum_{t\in\mathcal{P}} y_t\alpha_t^* \widetilde{X}_t$ and $G^* = I + \frac{1}{\gamma}\sum_{t\in\mathcal{P}} y_t\alpha_t^* \widetilde{\widetilde{X}}_t$.

Proof. We use the Lagrangian multiplier theorem to prove the desired result. By introducing Lagrangian multipliers $\alpha, \beta \ge 0$, define the Lagrangian function related to (4.4) by

$$\begin{aligned}\mathcal{L}(\alpha, \beta; M, G, \xi) = \sum_{t\in\mathcal{P}} \xi_t + \frac{\gamma}{2}(\|M - I\|_F^2 + \|G - I\|_F^2) \\ - \sum_{t=(i,j)\in\mathcal{P}} \alpha_t\big(y_{ij}[s_G(\tilde{x}_i, \tilde{x}_j) - d_M(\tilde{x}_i, \tilde{x}_j)] - 1 + \xi_t\big) - \sum_{t\in\mathcal{P}} \beta_t\xi_t.\end{aligned}$$

Then, taking the derivatives of $\mathcal{L}$ with respect to the primal variables M, G and ξ implies that $M = I - \frac{1}{\gamma}\sum_{t\in\mathcal{P}} y_t\alpha_t\widetilde{X}_t$, $G = I + \frac{1}{\gamma}\sum_{t\in\mathcal{P}} y_t\alpha_t\widetilde{\widetilde{X}}_t$, and $\alpha_t + \beta_t = 1$. Substituting these equalities back to $\mathcal{L}$, we get the desired result. This completes the proof of the theorem. □

Let $P = |\mathcal{P}|$ denote the number of image-pairs, e denote a P-dimensional vector with its elements all ones and Y = diag(y). The objective function of formulation (4.5) can then be rewritten as

$$f(\alpha) = \alpha^T e + \alpha^T Y\omega - \frac{1}{2\gamma}\alpha^T Y(U + V)Y\alpha, \tag{4.6}$$

and its gradient is given by

$$\nabla f(\alpha) = e + Y\omega - \frac{1}{\gamma}Y(U + V)Y\alpha. \tag{4.7}$$

Here, ω is a P-dimensional single vector (i.e. $\omega \in \mathbb{R}^P$) with the $t = (i, j)$-th element $\omega_t = \|\tilde{x}_i - \tilde{x}_j\|^2 - \tilde{x}_i^T\tilde{x}_j$, where $t = (i, j) \in \mathcal{P}$. In addition, U and V are $P \times P$-dimensional matrices (i.e. $U, V \in \mathbb{R}^{P\times P}$) with the (t, s)-th entries $U_{ts} = ((\tilde{x}_i - \tilde{x}_j)^T(\tilde{x}_l - \tilde{x}_k))^2$ and $V_{ts} = \tilde{x}_i^T\tilde{x}_l\tilde{x}_k^T\tilde{x}_j$, where $t = (i, j) \in \mathcal{P}$ and $s = (k, l) \in \mathcal{P}$.

The following lemma establishes the Lipschitz continuity of $\nabla f(\cdot)$, which is useful to analyse the time complexity of the gradient-based optimization algorithm below.

Lemma 2. *For any α_1, $\alpha_2 \in [0, 1]$, there holds*

$$\|\nabla f(\alpha_1) - \nabla f(\alpha_2)\| \le H\|\alpha_1 - \alpha_2\|, \tag{4.8}$$

where $H = |\lambda|_{\max}(U + V)/\gamma$.

Proof. For any symmetric matrix A, let $|\lambda|_{\max}(A)$ denote the maximum of the absolute values of

FISTA for Sub-SML

1. Initialized $H_0 > 0$, $\alpha_0 = \alpha_1 \in [0, 1]$ and set $c_0 = 0$, $c_1 = 1$.
2. For k = 1,2,...generate $\{\alpha_k\}$ as follows:
3. Set $\beta_k = \alpha_k + \frac{c_{k-1}-1}{c_k}(\alpha_k - \alpha_{k-1})$, $H_k = H_{k-1}$, compute $\nabla f(\beta_k)$
4. For j = 1,2,...

   Set $\alpha_{k+1} = \arg\min_{0\leq\alpha\leq 1} \|\alpha - (\beta_k - \frac{1}{H_k}\nabla f(\beta_k))\|^2$

   If $-f(\alpha_{k+1}) \leq -f(\beta_k) - \langle \nabla f(\beta_k), \alpha_{k+1} - \beta_k\rangle + \frac{H_k}{2}\|\alpha_{k+1} - \beta_k\|^2$

   stop

   else

   $H_k = 2H_k$

   end

   end
5. Set $c_{k+1} = (1 + \sqrt{1 + 4c_k^2})/2$

Table 4.1: Pseudo-code of FISTA for Sub-SML (i.e. formulation (4.5)).

the eigenvalues of matrix A. For any $\alpha_1, \alpha_2 \in [0, 1]$, we have that

$$\begin{aligned}
&\|\nabla f(\alpha_1) - \nabla f(\alpha_2)\| = \tfrac{1}{\gamma}\|Y(U + V)Y(\alpha_1 - \alpha_2)\| \\
&\leq |\lambda|_{\max}(Y(U + V)Y)\, \|\alpha_1 - \alpha_2||/\gamma \\
&= |\lambda|_{\max}(U + V)\, \|\alpha_1 - \alpha_2||/\gamma,
\end{aligned}$$

where the last equality follows from the fact that Y is an orthonormal matrix. This yields the desired result which completes the proof of the lemma. □

Formulation (4.5) is a standard quadratic programming (QP) problem, which can be solved by the standard MATLAB subroutine `quadprog.m`. However, these QP solvers employ the interior-point methods which need the second-order information (Hessian matrix) of the objective function. In the dual problem (4.5), the number of variables equals the number of image-pairs which is usually very large. Hence, the interior methods quickly become infeasible when the number of image-pairs increases. Instead, we use the Nesterov's first-order algorithm [Nesterov, 2004]. This method is guaranteed to converge to the global solution with rate $\mathcal{O}\big(\frac{H}{k^2}\big)$ where k is the iteration number. The original Nesterov's first order algorithm [Nesterov, 2004] needs to estimate the Lipschitz constant H of the gradient of the objective function in advance. However, in practice the Lipschitz constant H is not easily computable and it could be very large, particularly when the regularization parameter γ is small.

Here, we use a variant of Nesterov's first-order algorithm which is usually referred to as the fast iterative shrinkage thresholding algorithm (FISTA) [Nemirovski, 1994; Beck and Teboulle, 2009]. The main idea of this method is to tune the Lipschitz constant at each iteration using a line search scheme. It has the same theoretical convergence $\mathcal{O}\big(\frac{H}{k^2}\big)$ as that of the Nesterov's first-order algorithm. The pseudo-code of the algorithm is given in Table 4.1.

4.3 Extension to Unconstrained Face Verification in Videos

In this section, we extend our proposed framework to unconstrained face verification in videos. We aim to learn distance metrics and similarity functions from video-based data, which is more challenging.

Similar to the discussions in Section 4.2, we formulate the learning objective by considering both the robustness to large transformation differences and the discriminative power of similarity metric learning for discriminating similar video-pairs from dissimilar video-pairs.

To remain robust to the noise and the large transformation differences in videos, we employ the previously proposed Intra-PCA (i.e. equation (3.7)) to project the original video sequences onto the intra-personal subspace using the whitening process.

After the video sequences are mapped to the intra-personal subspace, we need to define a distance metric or similarity function to measure the similarity of a video-pair. Specifically, as an extension of the generalized similarity function $f_{(M,G)}$ (i.e. equation (4.1)), we define the similarity function for a video-pair $(\tilde{X}_i, \tilde{X}_j)$ as the average score of the similarity scores of all possible frame-level pairs, i.e.

$$F_{(M,G)}(\tilde{X}_i, \tilde{X}_j) = \frac{1}{N_i N_j} \sum_{l=1}^{N_i} \sum_{k=1}^{N_j} f_{(M,G)}(\tilde{x}_l^i, \tilde{x}_k^j). \tag{4.9}$$

Obviously, when $G = 0$, $F_{(M,0)}(\tilde{X}_i, \tilde{X}_j)$ is the average of the minus distances between all possible frame-level pairs generated from the video-pair $(\tilde{X}_i, \tilde{X}_j)$, i.e.

$$F_{(M,0)}(\tilde{X}_i, \tilde{X}_j) = \frac{1}{N_i N_j} \sum_{l=1}^{N_i} \sum_{k=1}^{N_j} -d_M(\tilde{x}_l^i, \tilde{x}_k^j), \tag{4.10}$$

When $M = 0$, $F_{(0,G)}(\tilde{X}_i, \tilde{X}_j)$ is the average of the bilinear similarity scores between all possible frame-level pairs generated from the video-pair $(\tilde{X}_i, \tilde{X}_j)$, i.e.

$$F_{(0,G)}(\tilde{X}_i, \tilde{X}_j) = \frac{1}{N_i N_j} \sum_{l=1}^{N_i} \sum_{k=1}^{N_j} s_G(\tilde{x}_l^i, \tilde{x}_k^j). \tag{4.11}$$

Since the length of video sequences can vary in different videos, the summations in equations (4.9), (4.10), and (4.11) are normalized by the numbers of frames in videos $\tilde{X}_i$ and $\tilde{X}_j$.

Let $\mathcal{P} = \mathcal{S} \cup \mathcal{D}$ denote the index set of video-pairs. The binary label is denoted by $y = \{y_{ij} \in \{\pm 1\}, (i,j) \in \mathcal{P}\}$, with $y_{ij} = 1$ indicating videos $\tilde{X}_i$ and $\tilde{X}_j$ are from the same identity and $y_{ij} = -1$ otherwise. With the similarity function $F_{(M,G)}$ defined above, formulation (4.4) is readily adapted to video-to-video matching setting. Specifically, formulation (4.4) is extended to the following problem

$$\begin{aligned} \min_{M,G \in \mathbb{S}^d} \quad & \sum_{t \in \mathcal{P}} \xi_t + \frac{\gamma}{2}(\|M - I\|_F^2 + \|G - I\|_F^2), \\ \text{s.t.} \quad & y_{ij} F_{(M,G)}(\tilde{X}_i, \tilde{X}_j) \geq 1 - \xi_{ij}, \\ & \xi_t \geq 0, \ \ \forall t = (i,j) \in \mathcal{P}. \end{aligned} \tag{4.12}$$

For simplicity, we also refer to formulation (4.12) as ***Sub-SML***.

Following the derivation of the dual formulation of Sub-SML (i.e. formulation (4.4)), the dual formulation of (4.12) can be similarly derived. For any $t = (i, j) \in \mathcal{P}$, denote $\widetilde{X}_t = \frac{1}{N_i N_j} \sum_{lk} (\tilde{x}_l^i - \tilde{x}_k^j)(\tilde{x}_l^i - \tilde{x}_k^j)^T$ and $\widetilde{\widetilde{X}}_t = \frac{1}{N_i N_j} \sum_{lk} \tilde{x}_l^i (\tilde{x}_k^j)^T$. Then, the dual formulation of (4.12) can be established as follows.

Theorem 3. *The dual formulation of Sub-SML (i.e. formulation (4.12)) can be written as*

$$\max_{0 \le \alpha \le 1} \sum_{t \in \mathcal{P}} \alpha_t + \sum_{t=(i,j) \in \mathcal{P}} \alpha_t y_t \frac{1}{N_i N_j} \Big(\sum_{lk} \|\tilde{x}_l^i - \tilde{x}_k^j\|^2 - \sum_{lk} (\tilde{x}_l^i)^T \tilde{x}_k^j \Big) - \frac{1}{2\gamma} \Big[\Big\| \sum_{t \in \mathcal{P}} y_t \alpha_t \widetilde{X}_t \Big\|_F^2 + \Big\| \sum_{t \in \mathcal{P}} y_t \alpha_t \widetilde{\widetilde{X}}_t \Big\|_F^2 \Big].$$

Moreover, if the optimal solution of the above dual formation is denoted by α^* then the optimal solution (M^*, G^*) of (4.12) is given by $M^* = I - \frac{1}{\gamma} \sum_{t \in \mathcal{P}} y_t \alpha_t^* \widetilde{X}_t$ and $G^* = I + \frac{1}{\gamma} \sum_{t \in \mathcal{P}} y_t \alpha_t^* \widetilde{\widetilde{X}}_t$.

Proof. We use the Lagrangian multiplier theorem to prove the desired result. By introducing Lagrangian multipliers $\alpha, \beta \ge 0$, define the Lagrangian function related to (4.12) by

$$\mathcal{L}(\alpha, \beta; M, G, \xi) = \sum_{t \in \mathcal{P}} \xi_t + \frac{\gamma}{2} (\|M - I\|_F^2 + \|G - I\|_F^2) - \sum_{t=(i,j) \in \mathcal{P}} \alpha_t \big(y_{ij} \frac{1}{N_i N_j} \sum_{lk} [s_G(\tilde{x}_l^i, \tilde{x}_k^j) - d_M(\tilde{x}_l^i, \tilde{x}_k^j)] - 1 + \xi_t \big) - \sum_{t \in \mathcal{P}} \beta_t \xi_t.$$

Then, taking the derivatives of $\mathcal{L}$ with respect to the primal variables M, G and ξ implies that $M = I - \frac{1}{\gamma} \sum_{t \in \mathcal{P}} y_t \alpha_t \widetilde{X}_t$, $G = I + \frac{1}{\gamma} \sum_{t \in \mathcal{P}} y_t \alpha_t \widetilde{\widetilde{X}}_t$, and $\alpha_t + \beta_t = 1$. Substituting these equalities back to $\mathcal{L}$, we get the desired result. This completes the proof of the theorem. □

With the dual formulation established above, FISTA [Nemirovski, 1994; Beck and Teboulle, 2009] can be used to obtain the optimal solution in video-based setting.

4.4 Experiment One: Unconstrained Face Verification

In this section, we provide an experimental evaluation of the proposed Sub-SML for unconstrained face verification. Specifically, we carry out experiments on the Labeled Faces in the Wild (LFW) database [Huang et al., 2007] for face verification in still images and the YouTube Faces (YTF) database [Wolf et al., 2011a] for unconstrained face verification in videos. Sections 4.4.1 and 4.4.2 present results on the LFW dataset and the YTF database respectively.

4.4.1 Labeled Faces in the Wild

In this section, we evaluate Sub-SML (i.e. formulation (4.4)) on the Labeled Faces in the Wild (LFW) database [Huang et al., 2007]. We conduct experiments in both the restricted and unre-

stricted setting of the LFW dataset. A brief introduction of this database and its experimental setting can be found in Section 2.4.1. For feature representation, three facial descriptors are employed: SIFT [Guillaumin et al., 2009], LBP [Ojala et al., 2002] and TPLBP [Wolf et al., 2008].

In particular, on each of the 10-fold cross-validation test, Intra-PCA is first implemented to reduce the transformation differences. Image vectors $\tilde{x}$ are then L^2-normalized to 1 (i.e. $\|\tilde{x}\| = 1$) before being fed into Sub-SML. The trade-off parameter γ in Sub-SML are tuned via three-fold cross validation over the remaining 9-fold training sets. For verification, similar to the description in Section 3.3.1, two facial images in the test set are predicted to be from the same person if the similarity score between them is greater than some threshold, and different persons otherwise. To learn the threshold, we choose the value that gives the highest verification rate on the 5400 image-pairs of training set.

Image Restricted Training Paradigm

Here, we evaluate our method in the restricted setting of the LFW dataset.

Method	d	Original Descriptor	Square Root
WPCA	100	75.98 ± 0.31	77.30 ± 0.23
Intra-PCA	100	81.32 ± 0.46	82.53 ± 0.33
Sub-ML	100	81.53 ± 0.37	82.52 ± 0.29
Sub-SL	100	82.47 ± 0.36	83.05 ± 0.58
Sub-SML	100	84.52 ± 0.45	85.27 ± 0.52
WPCA	200	76.40 ± 0.57	77.87 ± 0.27
Intra-PCA	200	82.32 ± 0.34	83.45 ± 0.24
Sub-ML	200	82.20 ± 0.42	83.30 ± 0.26
Sub-SL	200	84.17 ± 0.42	84.60 ± 0.41
Sub-SML	200	85.40 ± 0.42	86.32 ± 0.46
WPCA	300	77.23 ± 0.53	78.55 ± 0.35
Intra-PCA	300	82.18 ± 0.27	82.95 ± 0.23
Sub-ML	300	82.18 ± 0.33	82.65 ± 0.38
Sub-SL	300	83.48 ± 0.47	84.03 ± 0.68
Sub-SML	300	85.55 ± 0.61	86.22 ± 0.27
WPCA	400	77.23 ± 0.53	78.55 ± 0.35
Intra-PCA	400	81.22 ± 0.41	82.50 ± 0.21
Sub-ML	400	81.35 ± 0.39	83.20 ± 0.25
Sub-SL	400	81.30 ± 0.62	80.12 ± 0.83
Sub-SML	400	84.83 ± 0.58	85.57 ± 0.48

Table 4.2: Performance of Sub-ML (i.e. formulation (4.13)), Sub-SL (i.e. formulation (4.14)) and Sub-SML across different WPCA dimension d using the SIFT descriptor in the restricted setting of LFW. Here, the performance is reported using mean verification rate ($\pm$ standard error).

Effectiveness of Sub-SML. We conduct experiments to explore the performance of Sub-SML for face verification in still images. We show the effectiveness of Sub-SML in two main aspects: the generalized similarity function $f_{(M,G)}$ combining d_M and s_G, and Sub-SML as a similarity metric learning method over the intra-personal subspace. To this end, we do the following two comparisons.

Method	d	Original Descriptor	Square Root
WPCA	100	78.43 ± 0.33	78.55 ± 0.28
Intra-PCA	100	83.07 ± 0.37	83.32 ± 0.45
Sub-ML	100	83.35 ± 0.31	83.50 ± 0.41
Sub-SL	100	83.68 ± 0.41	83.40 ± 0.37
Sub-SML	100	84.47 ± 0.56	83.97 ± 0.53
WPCA	200	80.43 ± 0.46	80.43 ± 0.31
Intra-PCA	200	84.55 ± 0.63	84.60 ± 0.61
Sub-ML	200	84.52 ± 0.68	84.57 ± 0.59
Sub-SL	200	85.63 ± 0.55	85.08 ± 0.52
Sub-SML	200	86.08 ± 0.49	86.28 ± 0.55
WPCA	300	80.47 ± 0.51	80.98 ± 0.38
Intra-PCA	300	84.23 ± 0.55	84.45 ± 0.43
Sub-ML	300	84.35 ± 0.56	84.32 ± 0.43
Sub-SL	300	85.00 ± 0.52	85.10 ± 0.58
Sub-SML	300	86.73 ± 0.53	86.88 ± 0.61
WPCA	400	80.47 ± 0.51	80.98 ± 0.38
Intra-PCA	400	83.55 ± 0.65	83.87 ± 0.36
Sub-ML	400	83.48 ± 0.65	83.92 ± 0.38
Sub-SL	400	82.28 ± 0.74	83.52 ± 0.63
Sub-SML	400	86.33 ± 0.47	85.87 ± 0.49

Table 4.3: Performance of Sub-ML (i.e. formulation (4.13)), Sub-SL (i.e. formulation (4.14)) and Sub-SML across different WPCA dimension d using the LBP descriptor in the restricted setting of LFW. Here, the performance is reported using mean verification rate ($\pm$ standard error).

Firstly, we compare Sub-SML with the following formulations, where only the distance metric d_M or the bilinear similarity function s_G is used as the similarity function. More specifically, we compare with the formulation called Sub-ML given by

$$\begin{aligned} \min_{M \in \mathbb{S}^d} \quad & \sum_{t \in \mathcal{P}} \xi_t + \frac{\gamma}{2} \|M - I\|_F^2, \\ \text{s.t.} \quad & y_{ij}[-d_M(\tilde{x}_i, \tilde{x}_j)] \geq 1 - \xi_{ij}, \\ & \xi_t \geq 0, \ \ \forall t = (i,j) \in \mathcal{P}. \end{aligned} \tag{4.13}$$

and the formulation called Sub-SL given by

$$\begin{aligned} \min_{G \in \mathbb{S}^d} \quad & \sum_{t \in \mathcal{P}} \xi_t + \frac{\gamma}{2} \|G - I\|_F^2, \\ \text{s.t.} \quad & y_{ij}[s_G(\tilde{x}_i, \tilde{x}_j)] \geq 1 - \xi_{ij}, \\ & \xi_t \geq 0, \ \ \forall t = (i,j) \in \mathcal{P}. \end{aligned} \tag{4.14}$$

As baselines, we also compare Sub-SML with WPCA and the previously proposed Intra-PCA (i.e. equation (3.2)). It is worth mentioning that, when $\|x_i\| = \|x_j\| = 1$ and M and G are identity matrices, $s_G(x_i, x_j) = (2 - d_M(x_i, x_j))/2 = (f_{(M,G)}(x_i, x_j) - 2)/3$. Hence, in this special case, the verification rate using the Euclidean distance is the same as that using $f_{(M,G)}$.

Table 4.2 reports the comparison results of Sub-SML against Sub-ML and Sub-SL on the SIFT descriptor. We can observe from Table 4.2 that, across different WPCA dimensions, Sub-ML and

Sub-SL are only comparable with or slightly improve Intra-PCA while the performance of Sub-SML is much better than Intra-PCA. Taking WPCA dimension 300 for instance, Sub-SML yields 85.55%, which is better than 82.18% of Sub-ML and 83.48% of Sub-SL. Similar observation can be made on the LBP descriptor as shown in Table 4.3. These observations show the effectiveness of the generalized similarity function $f_{(M,G)}$ by combining the distance metric d_M and the bilinear similarity function s_G.

Secondly, we compare Sub-ML, Sub-SL and Sub-SML with other metric learning methods such as Xing [Xing et al., 2003], ITML [Davis et al., 2007], LDML [Guillaumin et al., 2009], DML-eig [Ying and Li, 2012], SILD [Kan et al., 2011] and KISSME [Kostinger et al., 2012]. For fairness of comparison, we also compare with their variants where image-vectors were processed by Intra-PCA before being fed into metric learning methods. For simplicity, we refer to such a variant of KISSME as Sub-KISSME. For the methods of Xing, ITML, LDML, DML-eig and SILD, as will be shown in Section 4.6, they have implicitly incorporated Intra-PCA. In addition, we conduct the comparison using both the original features and the L^2-normalized features.

From Table 4.4 we can see that, for both the SIFT and LBP descriptors, L^2-normalization improves the performance of most of the metric learning methods, which shows the effectiveness of the L^2-normalization as a preprocessing step. Moreover, on the SIFT (L^2-normalized) descriptor, Sub-ML and Sub-SL are competitive with other metric learning methods, while Sub-SML significantly outperforms these methods by obtaining 85.55% verification rate. Besides, on the LBP (L^2-normalized) descriptor, Sub-ML and Sub-SL yield better performance than existing metric learning methods, while Sub-SML achieves 86.73% verification rate, which outperforms the above metric learning methods by a large margin. The above observations validate the effectiveness of Sub-SML as a similarity metric learning method over the intra-personal subspace. In addition, we can observe that Sub-KISSME improves the performance of KISSME [Kostinger et al., 2012], which again shows the effectiveness of our previously proposed Intra-PCA to reduce the transformation differences.

Comparison with the state-of-the-art results. Now we compare Sub-ML, Sub-SL and Sub-SML with previously published results by combining different descriptors followed the procedure in [Guillaumin et al., 2009; Wolf et al., 2008]. Specifically, we first generate the similarity scores obtained by Sub-ML, Sub-SL and Sub-SML from three descriptors SIFT, LBP and TPLBP and their square roots (six scores). And then we train a Support Vector Machine (SVM) on the vector fused by the above six scores to make prediction. Note that each of these published results uses its own learning technique and different feature extraction approaches. Table 4.5 lists the comparison results and Figure 4.1 depicts the ROC curve comparison. As we can see from Table 4.5, Sub-SML achieves $\mathbf{89.73}\%$ which outperforms Sub-ML, Sub-SL and most of existing methods including the method SFRD+ PMML [Cui et al., 2013] which uses spatial face region descriptors and a multiple metric learning method. Furthermore, Sub-SML is competitive with the state-of-the art methods including DDML [Hu et al., 2014] and VMRS [Barkan et al., 2013]. In particular, DDML builds a deep neural network to learn a nonlinear distance metric and VMRS explores an advanced over-complete LBP feature descriptor (OCLBP). The performance of Sub-SML may be further improved by employing such deep neural network or novel feature descriptors.

Similar to the analysis of Intra-PCA in Section 3.3, we report in Figure 4.2 the similarity scores of

Method	SIFT	SIFT (L^2-normalized)	LBP	LBP (L^2-normalized)
Xing [Xing et al., 2003]	75.93 ± 0.59	75.12 ± 0.55	74.62 ± 0.45	73.92 ± 0.64
DML-eig [Ying and Li, 2012]	80.55 ± 1.71	81.27 ± 2.30	78.92 ± 0.62	81.63 ± 0.43
SILD [Kan et al., 2011]	81.78 ± 0.33	81.77 ± 0.30	81.53 ± 0.66	84.27 ± 0.55
ITML [Davis et al., 2007]	77.38 ± 0.39	80.90 ± 0.41	79.17 ± 0.46	82.45 ± 0.38
LDML [Guillaumin et al., 2009]	77.50 ± 0.50	80.40 ± 0.35	80.65 ± 0.47	81.77 ± 0.50
KISSME [Kostinger et al., 2012]	79.93 ± 0.42	83.08 ± 0.56	79.42 ± 0.61	83.37 ± 0.54
Sub-KISSME	80.87 ± 0.56	83.73 ± 0.43	81.52 ± 0.44	83.75 ± 0.48
Sub-ML	79.93 ± 0.60	82.18 ± 0.33	80.55 ± 0.56	84.35 ± 0.56
Sub-SL	81.47 ± 0.63	83.48 ± 0.47	81.87 ± 0.51	85.00 ± 0.52
Sub-SML	83.37 ± 0.52	$\mathbf{85.55 \pm 0.61}$	84.03 ± 0.60	$\mathbf{86.73 \pm 0.53}$

Table 4.4: Comparison of Sub-ML, Sub-SL and Sub-SML with other metric learning methods on the SIFT and LBP descriptors in the restricted setting of LFW. "L^2-normalized" means the features are L^2-normalized to 1. Sub-KISSME denotes KISSME over the intra-personal subspace. For ITML, $M_0 = X_{\mathcal{S}}^{-1}$.

Method	Accuracy
Combined b/g samples based methods, aligned [Wolf et al., 2009b]	86.83 ± 0.34
LDML combined, funneled [Guillaumin et al., 2009]	79.27 ± 0.60
DML-eig combined, funneled & aligned [Ying and Li, 2012]	85.65 ± 0.56
HTBI Features, aligned [Cox and Pinto, 2011]	88.13 ± 0.58
CSML + SVM, aligned [Nguyen and Bai, 2011]	88.00 ± 0.37
SFRD+ PMML [Cui et al., 2013]	89.35 ± 0.50
VMRS [Barkan et al., 2013]	$\mathbf{91.10 \pm 0.59}$
DDML [Hu et al., 2014]	90.68 ± 1.41
Sub-ML combined, funneled & aligned	86.13 ± 0.39
Sub-SL combined, funneled & aligned	86.68 ± 0.44
Sub-SML combined, funneled & aligned	89.73 ± 0.38

Table 4.5: Comparison of Sub-ML, Sub-SL and Sub-SML with other state-of-the-art methods in the restricted setting of LFW.

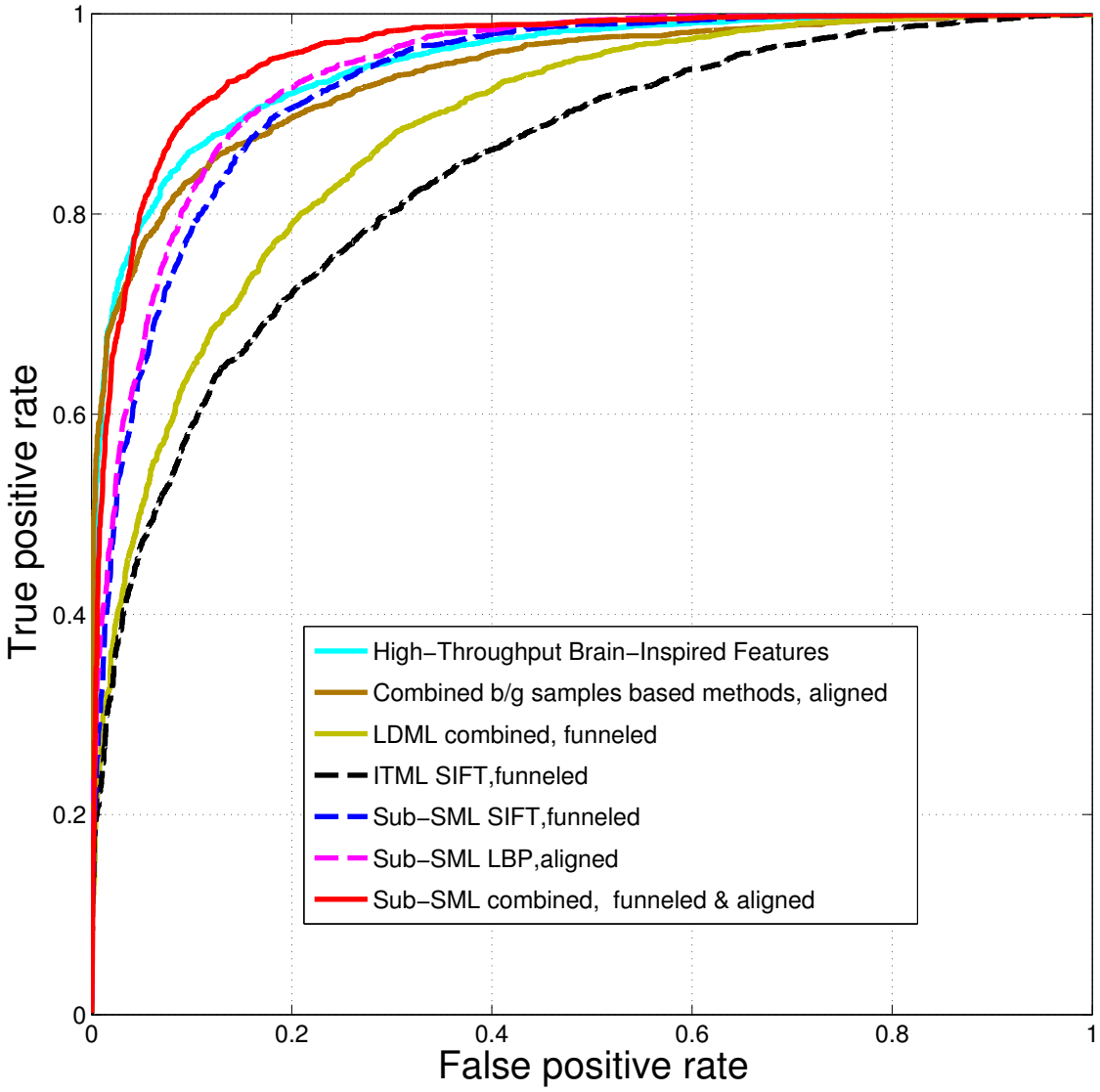


Figure 4.1: ROC curves of Sub-SML and other state-of-the-art methods in the restricted setting of the LFW database.

600 test image-pairs (300 similar image-pairs and 300 dissimilar image-pairs) obtained by Intra-PCA and Sub-SML on the SIFT descriptor in 3 folds of the 10-fold cross-validation test. The red and green points represent the similarity scores of similar and dissimilar image-pairs respectively. The black line for each model represents the threshold which is learned during training. As expected, we can see from Figure 4.2 that the improvement of Sub-SML over Intra-PCA benefits mostly from similar image-pairs. Indeed, the numbers of similar pairs that are correctly classified by Intra-PCA (red points that are above the black line in Figure 4.2a, 4.2c and 4.2e) are 221, 211 and 205 respectively, while the numbers of similar pairs that are correctly classified by Sub-SML (red points that are above the black line in Figure 4.2b, 4.2d and 4.2f) are 261, 228 and 234 respectively. Similar observations can be made in the remaining seven folds. This is because by further incorporating the discriminative power of similarity metric learning, Sub-SML successfully classifies similar image-pairs that are incorrectly classified by Intra-PCA.

Lastly, to understand better the effectiveness of Sub-SML in discriminating images-pairs with large transformation variations, Figure 4.3 depicts the qualitative examples for the comparison between Intra-PCA and Sub-SML. To be specific, Figure 3.4 shows similar pairs that are correctly classified by Sub-SML but incorrectly classified by Intra-PCA on the SIFT descriptor. As expected, we see by further incorporating the discriminative power of similarity metric learning, Sub-SML successfully classifies most of similar pairs that have large transformation variations such as lighting, hairstyle and occlusion. These observations highlight the merit of the proposed Sub-SML.

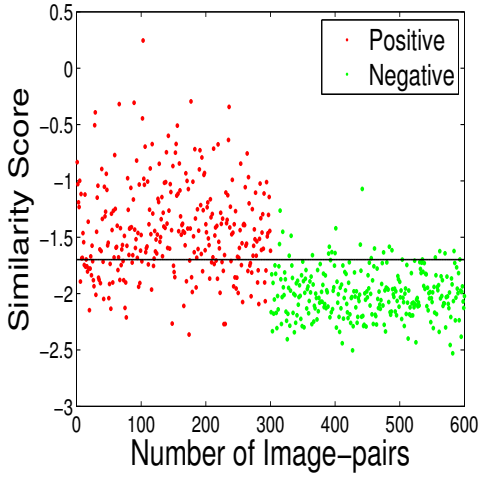
Image Unrestricted Training Paradigm

Here, we evaluate Sub-SML in the unrestricted setting of the LFW database, where the label information allows us to generate more image-pairs during training.

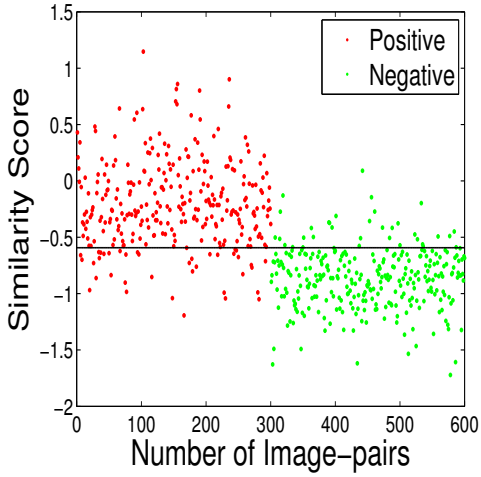
Methods	1000	1500	2000
WPCA	77.62 ± 0.51	77.42 ± 0.52	77.53 ± 0.52
Intra-PCA	83.13 ± 0.53	82.95 ± 0.62	82.87 ± 0.62
Sub-ML	83.07 ± 0.56	82.87 ± 0.60	82.83 ± 0.60
Sub-SL	83.83 ± 0.50	83.72 ± 0.67	83.50 ± 0.58
Sub-SML	$\mathbf{85.62 \pm 0.44}$	$\mathbf{86.42 \pm 0.46}$	$\mathbf{86.13 \pm 0.55}$
Xing [Xing et al., 2003]	75.17 ± 0.56	75.18 ± 0.55	74.95 ± 0.68
ITML [Davis et al., 2007]	79.43 ± 0.47	79.50 ± 0.22	79.88 ± 0.33
LDML [Guillaumin et al., 2009]	78.72 ± 0.38	81.27 ± 0.52	80.93 ± 0.36
DML-eig [Ying and Li, 2012]	81.00 ± 0.65	82.58 ± 0.56	83.67 ± 0.39
SILD [Kan et al., 2011]	79.25 ± 0.44	78.80 ± 0.38	78.57 ± 0.33
KISSME [Kostinger et al., 2012]	83.38 ± 0.42	83.58 ± 0.38	83.10 ± 0.41
Sub-KISSME [Kostinger et al., 2012]	84.18 ± 0.60	84.42 ± 0.51	83.85 ± 0.58

Table 4.6: Verification rate ($\pm$ standard error) of different metric learning methods using the SIFT descriptor versus the number of image-pairs per fold in the unrestricted setting of LFW.

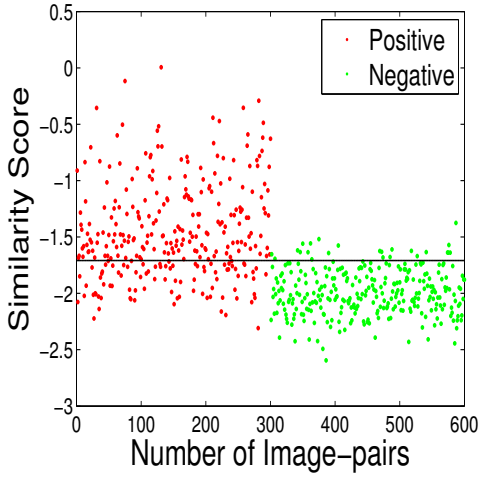
Firstly, we examine the performance of Sub-ML, Sub-SL and Sub-SML when using an increasing number of image-pairs: 1000, 1500 and 2000 per fold. Table 4.6 and Table 4.7 show the comparison results on the SIFT and LBP descriptors against state-of-the-art metric learning methods



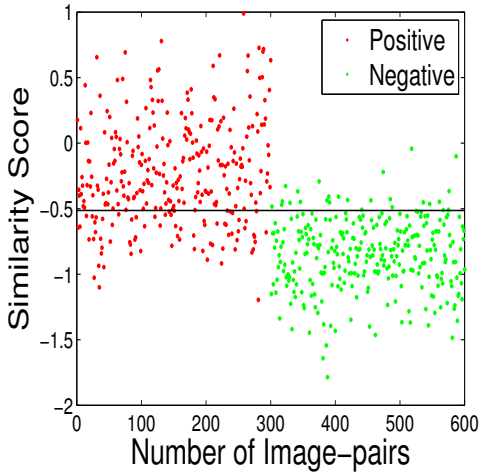
(a) Intra-PCA, Fold = 2



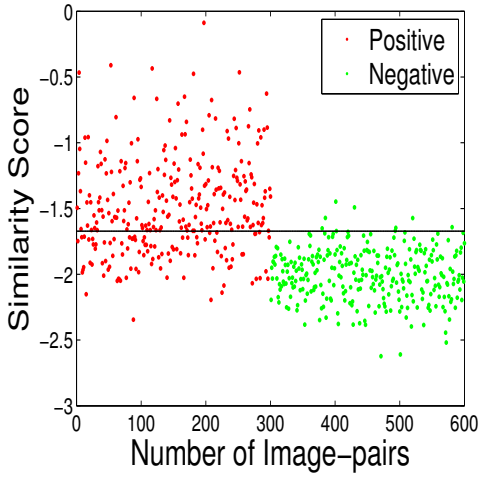
(b) Sub-SML, Fold = 2



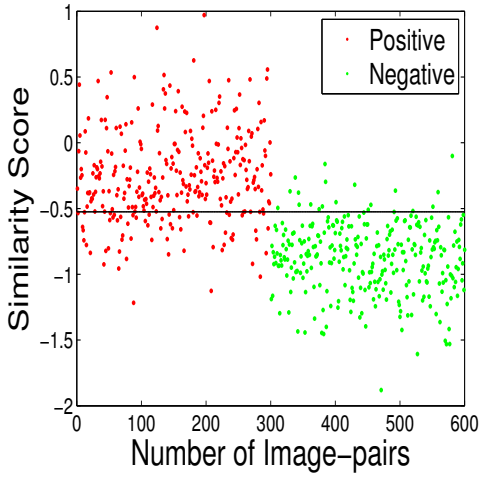
(c) Intra-PCA, Fold = 5



(d) Sub-SML, Fold = 5



(e) Intra-PCA, Fold = 8



(f) Sub-SML, Fold = 8

Figure 4.2: Similarity scores of 600 test images-pairs (300 similar image-pairs and 300 dissimilar image-pairs) obtained by Intra-PCA and Sub-SML on the SIFT descriptor in 3 folds of the 10-fold cross-validation test in the restricted setting of LFW: the red and green points represent similar and dissimilar image-pairs respectively; the black line is the learned threshold.

Figure 4.3: Examples of positive image-pairs that are correctly classified by Sub-SML while incorrectly classified by Intra-PCA in the restricted setting of the LFW dataset.

Methods	1000	1500	2000
WPCA	79.92 ± 0.31	79.88 ± 0.34	79.73 ± 0.26
Intra-PCA	84.67 ± 0.40	84.70 ± 0.46	84.62 ± 0.37
Sub-ML	84.60 ± 0.38	84.82 ± 0.44	84.50 ± 0.32
Sub-SL	85.37 ± 0.51	85.23 ± 0.51	85.08 ± 0.48
Sub-SML	$\mathbf{86.73 \pm 0.59}$	$\mathbf{86.72 \pm 0.51}$	$\mathbf{86.33 \pm 0.54}$
Xing [Xing et al., 2003]	73.92 ± 0.64	73.48 ± 0.79	72.90 ± 0.70
ITML [Davis et al., 2007]	80.27 ± 0.55	80.10 ± 0.39	81.70 ± 0.47
LDML [Guillaumin et al., 2009]	81.70 ± 0.29	82.08 ± 0.37	81.90 ± 0.45
DML-eig [Ying and Li, 2012]	82.42 ± 0.64	82.77 ± 1.2	83.23 ± 0.35
SILD [Kan et al., 2011]	81.37 ± 0.41	81.17 ± 0.35	80.60 ± 0.41
KISSME [Kostinger et al., 2012]	83.13 ± 0.27	83.27 ± 0.30	82.92 ± 0.31
Sub-KISSME	83.77 ± 0.50	83.78 ± 0.49	83.58 ± 0.49

Table 4.7: Verification rate ($\pm$ standard error) of different metric learning methods using the LBP descriptor versus the number of image-pairs per fold in the unrestricted setting of LFW.

Method	Accuracy
SIFT PLDA, funneled [Li et al., 2012]	86.2 ± 1.2
SIFT LMNN, funneled [Weinberger et al., 2006]	80.5 ± 0.5
SIFT LDML, funneled [Guillaumin et al., 2009]	83.2 ± 0.4
SIFT Sub-ML, funneled	83.07 ± 0.56
SIFT Sub-SL, funneled	83.83 ± 0.50
SIFT Sub-SML, funneled	$\mathbf{86.42 \pm 0.46}$
LBP mutishot, aligned [Taigman et al., 2009]	85.17 ± 0.61
LBP PLDA, aligned [Li et al., 2012]	$\mathbf{87.33 \pm 0.55}$
LBP Sub-ML, aligned	84.82 ± 0.55
LBP Sub-SL, aligned	85.37 ± 0.51
LBP Sub-SML, aligned	87.15 ± 0.56
LDML-MkNN, funneled [Guillaumin et al., 2009]	87.50 ± 0.40
Combined multishot, aligned [Taigman et al., 2009]	89.50 ± 0.51
Combined PLDA, funneled & aligned [Li et al., 2012]	90.07 ± 0.51
combined Joint Bayesian [Chen et al., 2012]	90.90 ± 1.48
VMRS [Barkan et al., 2013]	$\mathbf{92.05 \pm 0.49}$
Sub-ML combined, funneled & aligned	0.8775 ± 0.0054
Sub-SL combined, funneled & aligned	0.8768 ± 0.0050
Sub-SML combined, funneled & aligned	90.75 ± 0.64

Table 4.8: Comparison of Sub-ML, Sub-SL and Sub-SML with other state-of-the-art results in the unrestricted setting of LFW: the top 6 rows are based on the SIFT descriptor, the middle 5 rows are based on the LBP descriptor and the bottom 8 rows are based on multiple descriptors.

such as Xing [Xing et al., 2003], ITML [Davis et al., 2007], LDML [Guillaumin et al., 2009], DML-eig [Ying and Li, 2012], SILD [Kan et al., 2011], KISSME [Kostinger et al., 2012] and its variant Sub-KISSME. We observe from Table 4.6 that on the SIFT descriptor, across the number of image-pairs per fold, Intra-PCA is better than WPCA, Sub-ML and Sub-SL are comparable with or slightly improve Intra-PCA, while Sub-SML improves Intra-PCA by a large margin. Similar observations can be made on the LBP descriptor as shown in Table 4.7. These observations verify the effectiveness of the generalized similarity function $f_{(M,G)}$ by combining the distance metric d_M and the bilinear similarity function s_G. Moreover, the performance of Sub-SML is significantly better than the other metric learning methods, which shows its effectiveness as a similarity metric learning method over the intra-personal subspace. We did not directly compare our method with LMNN in Table 4.6, since LMNN needs the information of triplets. However, we notice that the performance of Sub-SML on the SIFT descriptor (see Table 4.6) is much better than the best performance 80.50% of LMNN as reported in [Guillaumin et al., 2009].

Secondly, we compare our method with existing state-of-the-art methods in the unrestricted setting of LFW using single and multiple descriptors. Table 4.8 presents the comparison results and Figure 4.4 depicts the ROC curve comparison. In particular, we see from Table 4.8 that Sub-SML obtains 86.42% on the SIFT descriptor, which outperforms 86.20% of PLDA [Li et al., 2012] and 83.20% of LDML [Guillaumin et al., 2009]. As for the LBP descriptor, Sub-SML is competitive with PLDA. By further combining three descriptors (i.e. SIFT, LBP and TPLBP) and their square roots following the procedure in [Guillaumin et al., 2009; Wolf et al., 2008], Sub-SML using 2000 image-pairs achieves 90.75%, which outperforms 90.07% of PLDA and is competitive with 92.05% of VMRS [Barkan et al., 2013]. The performance of Sub-SML may be further improved by including more image-pairs.

4.4.2 YouTube Faces Database

This section evaluates the efficacy of the proposed Sub-SML (i.e. formulation (4.12)) in the restricted setting of the YouTube Faces Database. A detailed description of this database and its experimental protocol have been provided in Section 2.4.2. For feature extraction, three types of features are employed, i.e. LBP, CSLBP and FPLBP.

In particular, on each of the 10-fold cross-validation test, Intra-PCA is implemented to reduce the transformation differences. The parameters d (i.e. the dimensionality of the WPCA-reduced subspace) and k (i.e. the dimensionality of the intra-personal subspace) are tuned via three-fold cross validation over the remaining 9-fold training sets. The trade-off parameter γ in Sub-SML is also tuned via three-fold cross validation. For the verification step, a test video-pair is classified to be similar if its similarity score is greater than some threshold, and dissimilar otherwise. In order to learn the threshold, we choose the value that gives the highest verification rate on the 9-fold training set.

Firstly, we compare Sub-SML with WPCA and Intra-PCA (i.e. equation (3.7)). Table 4.9 lists the comparison results on the LBP, FPLBP and TPLBP descriptors. As we can see from Table 4.9 that on the three descriptors, Sub-SML outperforms WPCA and Intra-PCA in terms of Accuracy, AUC and EER. This shows the effectiveness of Sub-SML by incorporating both the robustness to

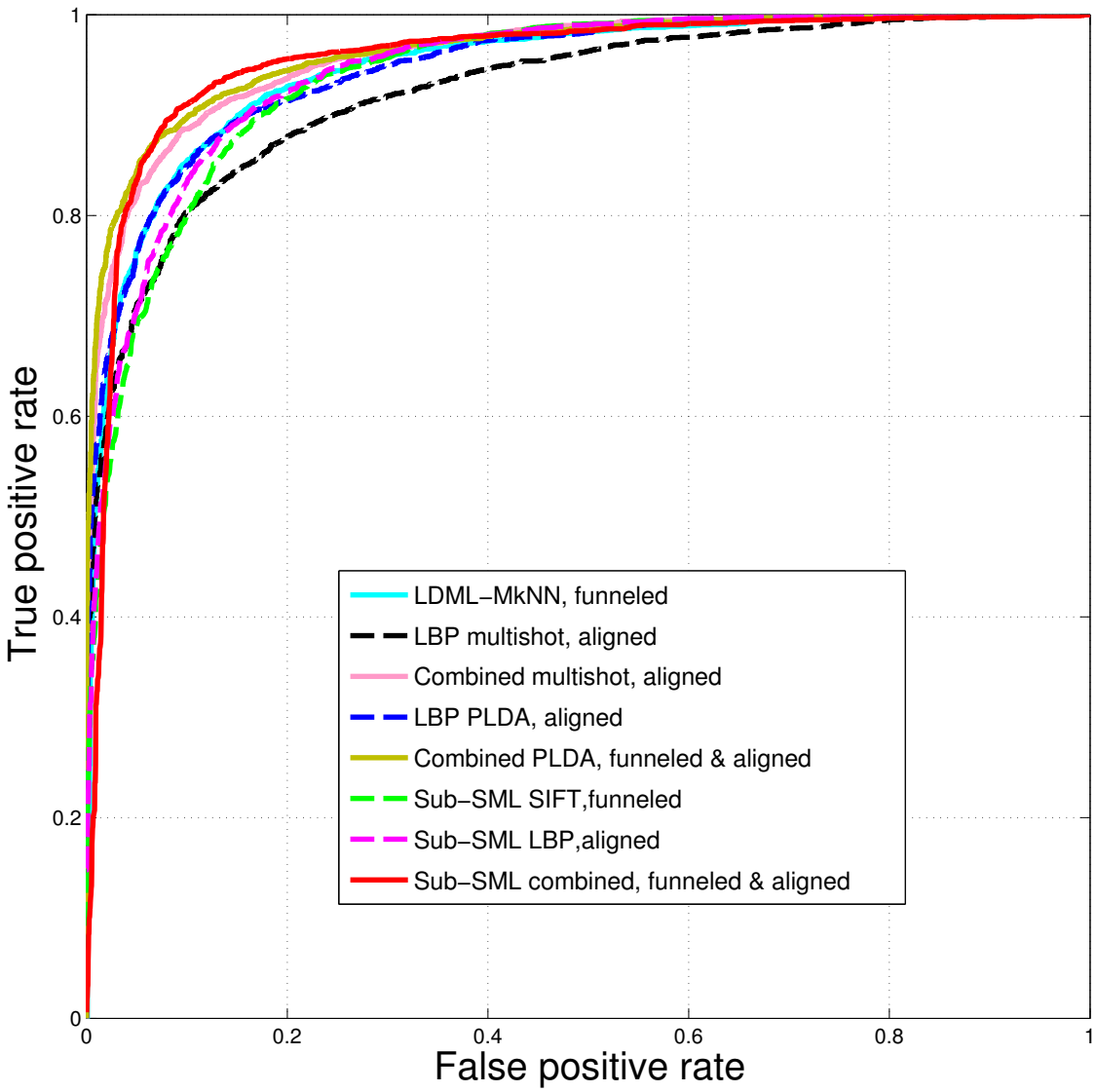


Figure 4.4: ROC curves of Sub-SML and other state-of-the-art methods in the unrestricted setting of LFW.

Descriptor	Method	Accuracy ± SE	AUC	EER
LBP	WPCA	75.0 ± 1.6	82.8	25.4
	Intra-PCA	78.1 ± 1.8	86.8	21.1
	Sub-SML	80.1 ± 1.5	**89.0**	**19.4**
	MBGS [Wolf et al., 2011a]	76.4 ± 1.8	82.6	25.3
	MBGS + SVM ⊖ [Wolf and Levy, 2013]	78.9 ± 1.9	86.9	21.2
	PHL+SILD [Kan et al., 2013]	$\mathbf{80.2 \pm 1.3}$	87.2	20.3
FPLBP	WPCA	72.8 ± 1.3	79.7	27.2
	Intra-PCA	75.1 ± 1.8	82.3	25.0
	Sub-SML	75.5 ± 1.5	**83.9**	**24.1**
	MBGS [Wolf et al., 2011a]	72.6 ± 2.0	80.1	27.7
	MBGS + SVM ⊖ [Wolf and Levy, 2013]	76.0 ± 1.7	83.7	24.9
	PHL+SILD [Kan et al., 2013]	$\mathbf{75.9 \pm 1.5}$	82.5	24.4
CSLBP	WPCA	72.4 ± 2.5	79.0	27.7
	Intra-PCA	74.5 ± 2.6	81.6	25.9
	Sub-SML	74.4 ± 2.5	**82.6**	25.3
	MBGS [Wolf et al., 2011a]	72.4 ± 2.0	78.9	28.7
	MBGS + SVM ⊖ [Wolf and Levy, 2013]	72.6 ± 2.1	81.8	26.1
	PHL+SILD [Kan et al., 2013]	$\mathbf{75.2 \pm 1.0}$	82.3	**24.8**

Table 4.9: Comparison of Sub-SML with the state-of-the-art methods on the LBP, FPLBP and CSLBP descriptors in the restricted setting of the YouTube Faces database.

large transformation differences and the discriminative power using similarity metric learning.

Secondly, we compare our method with the state-of-the-art methods including MBGS [Wolf et al., 2011a], MBGS + SVM ⊖ [Wolf and Levy, 2013] and PHL+SILD [Kan et al., 2013]. From Table 4.9 we can see that on the LBP, FPLBP and TPLBP descriptors, Sub-SML outperforms MBGS by a significant margin in terms of Accuracy, AUC and EER. Taking the LBP descriptor for instance, the improvement of Sub-SML over MBGS are 3.7%, 6.4%, and 5.9% in terms of Accuracy, AUC and EER, respectively. Furthermore, we observe that when considering Accuracy, Sub-SML is competitive to PHL+SILD on the LBP, FPLBP and CSLBP descriptors. In terms of AUC and EER, Sub-SML outperforms PHL+SILD on the LBP and FPLBP descriptors. The above observations validate the competitiveness of Sub-SML for video-based face verification.

Now we compare Sub-SML with the state-of-the-arts for face verification in videos by combining three descriptors following the procedure in [Guillaumin et al., 2009; Wolf et al., 2008]. In our experiment, a SVM classifier is trained on the 3D vector fused by the similarity scores generated using the LBP, FPLBP and CSLBP descriptors. Table 4.10 lists the comparison results and Figure 4.5 depicts the ROC curves. From Table 4.10 we can see that, in terms of Accuracy, AUC and EER, Sub-SML is competitive with or slightly better than the recent state-of-the-arts approaches. In particular, Sub-SML outperforms STFRD + PMML [Cui et al., 2013] which uses spatial-temporal face region descriptor based on the Token-Frequency features and VSOF+OSS (Adaboost) [Mendez-Vazquez et al., 2013] using a local spatio-temporal descriptor based on the volume structured ordinal features. Furthermore, Sub-SML is competitive with DDML (combined) [Hu et al., 2014]. Note that DDML learns a nonlinear distance metric by employing a deep neural network, and the performance of Sub-SML may be further improved by applying such deep neural network.

Method	Accuracy	AUC	EER
STFRD + PMML [Cui et al., 2013]	79.5 ± 2.5	88.6	19.9
APEM FUSION [Li et al., 2013]	79.1 ± 1.5	86.6	21.4
VSOF+OSS (Adaboost) [Mendez-Vazquez et al., 2013]	79.7 ± 1.8	89.4	20.0
DDML (combined) [Hu et al., 2014]	$\mathbf{82.3 \pm 1.5}$	**90.1**	**18.5**
Sub-SML FUSION	80.6 ± 1.4	89.5	19.4

Table 4.10: Comparison of Sub-SML with the state-of-the-art methods in the restricted setting of the YouTube Faces database.

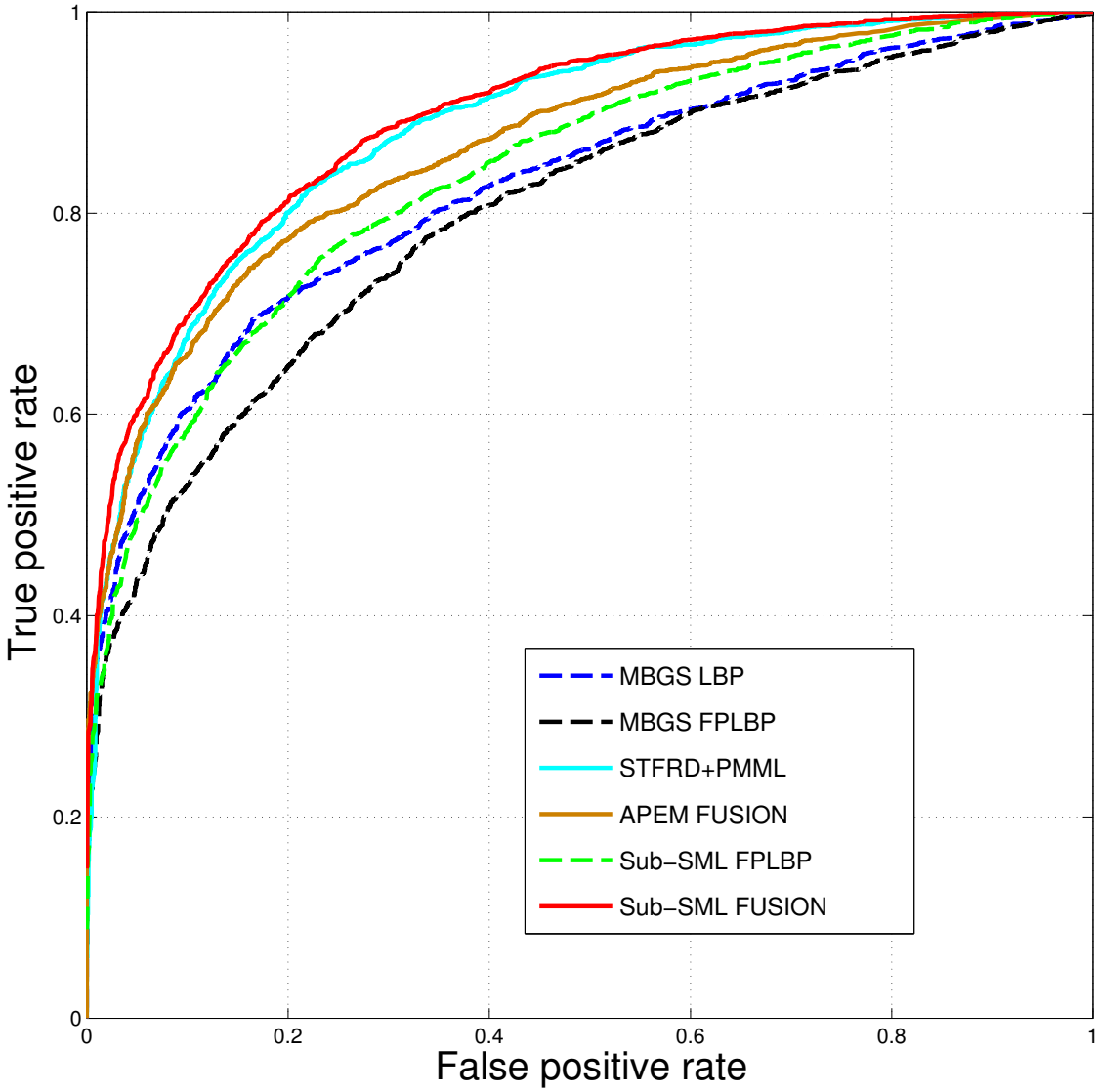


Figure 4.5: ROC curves of Sub-SML and other state-of-the-art methods in the restricted setting of the Youtube Faces database.

Lastly, Figure 4.6 reports the similarity scores of 500 test video-pairs (250 similar video-pairs and 250 dissimilar video-pairs) obtained by Intra-PCA and Sub-SML on the SIFT descriptor in 3 folds of the 10-fold cross-validation test. Similar to the description in Section 3.3.2, the red and green points in Figure 4.6 represent the similarity scores of similar and dissimilar video-pairs respectively and the black line for each model represents the threshold learned during training. From Figure 4.6 we observe that the improvement of Sub-SML over Intra-PCA is mainly from the improvement in classifying similar video-pairs. Indeed, the numbers of similar pairs that are correctly classified by Intra-PCA (red points that are above the black line in Figure 4.6a, 4.6c and 4.6e) are 175, 181 and 188 respectively, while the numbers of similar video-pairs that are correctly classified by Sub-SML (red points that are above the black line in Figure 4.6b, 4.6d and 4.6f) are 199, 216, 207 respectively. Similar observations can be made in the remaining seven folds. An Explanation for the above observations may be that by implementing Sub-SML to further incorporat the discrimination using novel similarity functions, similar video-pairs that are incorrectly verified by Intra-PCA are successfully verified.

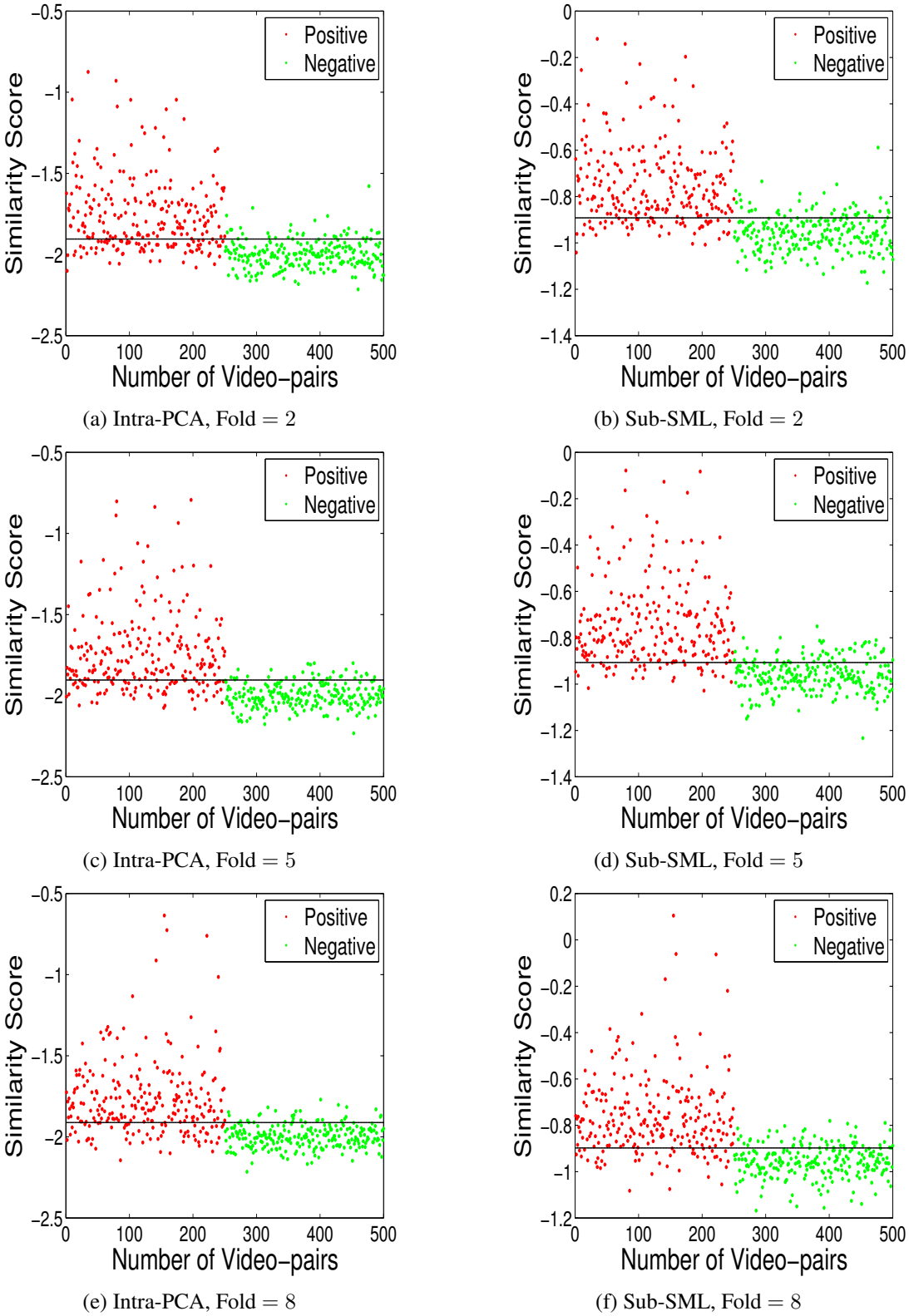
4.5 Experiment Two: Person Re-Identification

In this section, we apply Sub-SML (i.e. formulation (4.4)) to the task of person re-identification on the benchmark Viewpoint Invariant Pedestrian Recognition (VIPeR) database [Gray et al., 2007]. A brief introduction of the VIPeR dataset and its experimental setting have been given in Section 2.4.3. For feature representation, we use the features [1] provided by Kostinger et al. [2012]. The parameter d (i.e. the dimensionality of the PCA-reduced subspace) and the trade-off parameter γ in Sub-SML are tuned via three-fold cross validation.

We first compare Sub-SML with metric learning methods that are closely related to ours, i.e. Xing [Xing et al., 2003], LMNN [Weinberger et al., 2006], ITML [Davis et al., 2007], LDML [Guillaumin et al., 2009], DML-eig [Ying and Li, 2012], SILD [Kan et al., 2011] and KISSME [Kostinger et al., 2012]. As baselines, we also compare Sub-SML with PCA and Intra-PCA. Figure 4.7 depicts the CMC curve comparison and Table 4.11 reports the CMC scores in the range of the first 50 ranks. In particular, Figure 4.7a and Table 4.11a are for $h = 316$ (316 persons for testing and 316 persons for training), while Figure 4.7b and Table 4.11b are for $h = 532$ (532 persons for testing and 100 persons for training). Here, CMC curve represents the expectation of finding the true match within the top r ranks, see Section 2.4.3 for details. We can see from Table 4.11a and Table 4.11b that for both $h = 316$ and $h = 532$, Sub-SML yields better rank 1 matching rate than PCA, Intra-PCA and the above metric learning methods. At higher ranks, Sub-SML outperforms PCA, Intra-PCA, and achieves competitive or even better performance than the above metric learning methods.

Now we compare Sub-SML with four state-of-the-art metric learning approaches on VIPeR, i.e. MCC [Globerson and Roweis, 2005], PRDC [Zheng et al., 2011], KISSME [Kostinger et al., 2012] and PCCA [Mignon and Jurie, 2012]. Note that PCCA learns a linear transformation from the original feature space and is further kernelized by mapping the original feature to a higher dimensional feature space using Φ. Table 4.11 reports the comparison CMC scores in the range

[1] Available at: `http://lrs.icg.tugraz.at/research/kissme/`.



(a) Intra-PCA, Fold = 2

(b) Sub-SML, Fold = 2

(c) Intra-PCA, Fold = 5

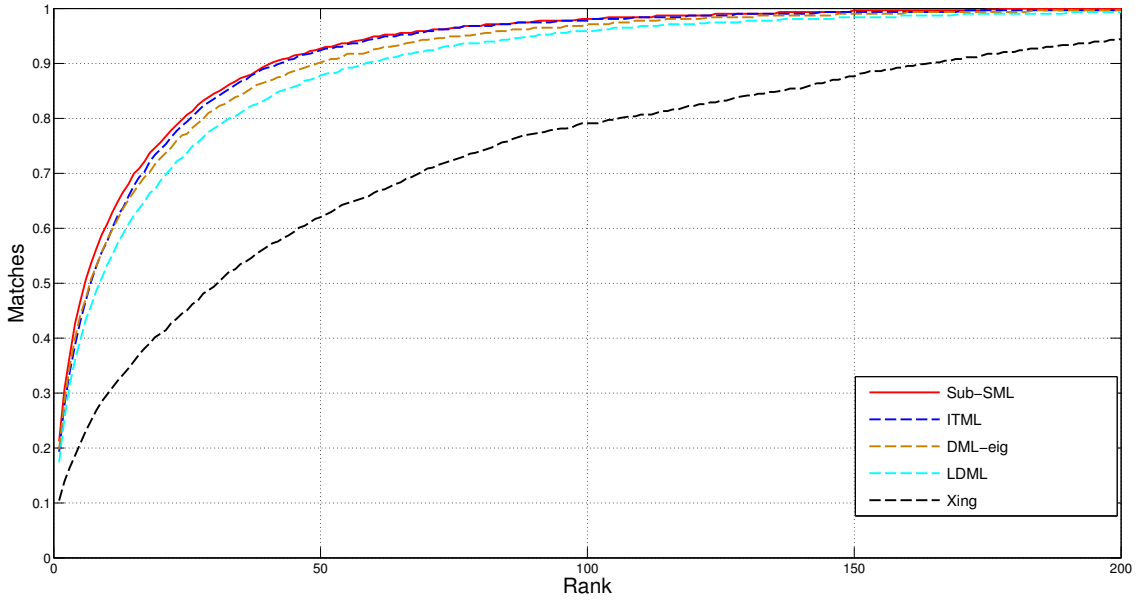
(d) Sub-SML, Fold = 5

(e) Intra-PCA, Fold = 8

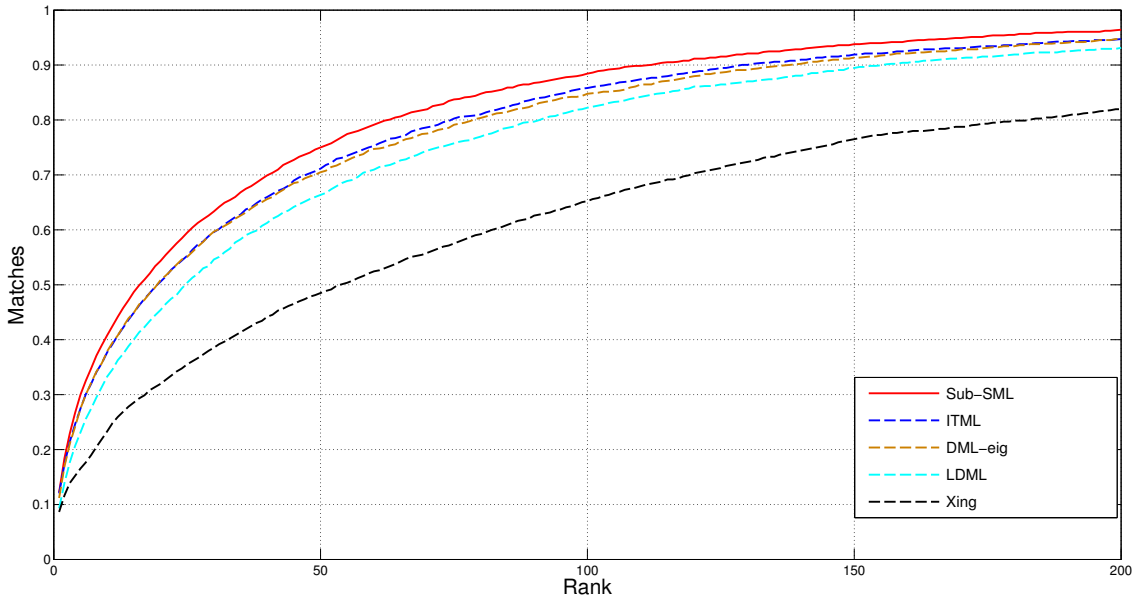
(f) Sub-SML, Fold = 8

Figure 4.6: Similarity scores of 500 test images-pairs (250 similar image-pairs and 250 dissimilar image-pairs) obtained by Intra-PCA and Sub-SML on the LBP descriptor in 3 folds of the 10-fold cross-validation test in the restricted setting of the YTF database: the red and green points represent similar and dissimilar video-pairs respectively; the black line is the learned threshold.

of the first 50 ranks. The results of PRDC and MCC are cited from [Farenzena et al., 2010] and the results of PCCA are cited from [Mignon and Jurie, 2012]. From Table 4.11a and Table 4.11b we can see that Sub-SML obtains the best rank 1 rate for $h = 316$ and $h = 532$. At higher ranks, Sub-SML is competitive with or better than the above four models for $h = 316$. In particular, for $h = 316$, Sub-SML outperforms PCCA using Bhattacharyya kernel (i.e. PCCA-sqrt) by a significant margin, while by employing χ^2_{RBF} kernel PCCA gives the most competitive results. Changing to $h = 532$ where less samples are used for training, Sub-SML is superior to the above four models across the first 50 ranks. The performance of PCCA using χ^2_{RBF} kernel (i.e. PCCA-χ^2_{RBF}) becomes significantly worse when the training samples become less. Our method Sub-SML achieves relatively more stable performance even when the number of training samples become smaller .



(a) $h = 316$



(b) $h = 532$

Figure 4.7: CMC curves of Sub-SML and other metric learning methods on the VIPeR dataset: (a) $h = 316$ and (b) $h = 532$, where h is the number of persons in the test set.

RANK	1	5	10	20	50
PCA	10.92	21.20	30.06	40.98	61.08
Intra-PCA	19.15	42.72	56.33	71.84	90.51
Sub-SML	**21.20**	46.84	60.76	75.63	**92.72**
ITML [Davis et al., 2007]	19.30	43.20	57.59	74.37	92.41
LDML [Guillaumin et al., 2009]	17.41	39.72	53.32	68.67	87.82
LMNN [Weinberger et al., 2006]	18.99	43.67	57.59	73.42	90.51
DML-eig [Ying and Li, 2012]	19.78	44.30	57.91	72.78	90.19
Xing [Xing et al., 2003]	10.44	20.89	29.75	40.82	62.03
SILD [Kan et al., 2011]	17.09	37.97	51.58	68.04	87.03
KISSME [Kostinger et al., 2012]	20.57	47.47	60.44	74.37	90.82
PRDC [Zheng et al., 2011]	15.66	38.42	53.86	70.09	–
MCC [Farenzena et al., 2010]	15.19	41.77	57.59	73.39	–
PCCA-sqrt [Mignon and Jurie, 2012]	17.28	42.41	56.68	74.53	–
PCCA-χ^2_{RBF} [Mignon and Jurie, 2012]	19.27	**48.89**	**64.91**	**80.28**	–

(a) $h = 316$

RANK	1	5	10	20	50
PCA	8.46	16.54	23.31	32.24	48.50
Intra-PCA	11.94	29.51	40.60	53.95	74.25
Sub-SML	**12.22**	**29.98**	**40.79**	54.32	**75.00**
ITML [Davis et al., 2007]	12.03	27.44	37.69	50.56	71.15
LDML [Guillaumin et al., 2009]	9.30	23.12	33.27	45.39	66.35
LMNN [Weinberger et al., 2006]	12.12	29.32	40.51	**54.51**	74.91
DML-eig [Ying and Li, 2012]	11.18	27.44	37.88	50.85	70.49
Xing [Xing et al., 2003]	8.65	16.54	23.31	31.95	48.50
SILD [Kan et al., 2011]	10.53	25.75	35.15	47.65	67.58
KISSME [Kostinger et al., 2012]	12.12	29.32	40.70	53.95	72.93
PRDC [Zheng et al., 2011]	9.12	24.19	34.40	48.55	–
MCC [Farenzena et al., 2010]	5.00	16.32	25.92	39.64	–
PCCA-sqrt [Mignon and Jurie, 2012]	8.44	24.34	35.62	50.07	–
PCCA-χ^2_{RBF} [Mignon and Jurie, 2012]	9.27	24.89	37.43	52.89	–

(b) $h = 532$

Table 4.11: Comparison of the matching rates with various methods on the VIPeR dataset: (a) $h = 316$ and (b) $h = 532$, where h is the number of persons in the test set. The middle 6 rows are metric learning methods closely related to our work, and the bottom 5 are the state-of-the-art metric learning methods for person re-identification. The results of PRDC and MCC are cited from [Farenzena et al., 2010], and the results of PCCA are cited from [Mignon and Jurie, 2012]. The notation '$-$' means that the result was not reported.

4.6 Discussion

This section revisits metric learning methods Xing [Xing et al., 2003], LMNN [Weinberger et al., 2006], ITML [Davis et al., 2007], LDML [Guillaumin et al., 2009], SILD [Kan et al., 2011], DML-eig [Ying and Li, 2012].

We consider the special case that the dimensionality of the intra-personal subspace equals the dimensionality of WPCA-reduced subspace, i.e. $k = d$. In this case, Intra-PCA (i.e. equation (3.2)) becomes $\tilde{x} = L_{\mathcal{S}}^{-1}x$, see equation (3.3) for the definition of $L_{\mathcal{S}}$.

In Section 2.3.2, we have reviewed metric learning methods Xing (i.e. formulation (2.35)), LMNN (i.e. formulation (2.46)), ITML (i.e. formulation (2.37)), SILD (i.e. formulation (2.42)) and DML-eig (i.e. formulation (2.40)). A common term in formulations of the above metric learning methods is the summation of distances between similar image-pairs, i.e. $\sum_{(i,j)\in\mathcal{S}} d_M(x_i, x_j) = \mathbf{Tr}(X_{\mathcal{S}}M) = \mathbf{Tr}(L_{\mathcal{S}}^T M L_{\mathcal{S}})$ (recalling that $X_{\mathcal{S}} = L_{\mathcal{S}}L_{\mathcal{S}}^T$). Let $\widetilde{M} = L_{\mathcal{S}}^T M L_{\mathcal{S}}$, and then Xing is equivalent to the following

$$\begin{array}{ll} \max_{\widetilde{M}\in\mathbb{S}_+^d} & \sum_{(i,j)\in\mathcal{D}} \sqrt{(\tilde{x}_i - \tilde{x}_j)^T \widetilde{M}(\tilde{x}_i - \tilde{x}_j)} \\ \text{s.t.} & \mathbf{Tr}(\widetilde{M}) \leq 1. \end{array} \tag{4.15}$$

LMNN is equivalent to

$$\begin{aligned} \arg\min_{\widetilde{M},\xi} \mathbf{Tr}(\widetilde{M}) + \gamma \sum_{\tau=(i,j,k)\in\mathcal{T}} \xi_{ijk} \\ d_{\widetilde{M}}(\tilde{x}_j, \tilde{x}_k) - d_{\widetilde{M}}(\tilde{x}_i, \tilde{x}_j) \geq 1 - \xi_{ijk} \\ \xi_{ijk} \geq 0, \ \ \forall\, (i,j,k) \in \mathcal{T}, \widetilde{M} \in \mathbb{S}_+^d, \end{aligned} \tag{4.16}$$

SILD is equivalent to

$$\max_{\widetilde{M}\in\mathbb{S}_+^d} \left[\frac{\sum_{(i,j)\in\mathcal{D}} d_{\widetilde{M}}(\tilde{x}_i, \tilde{x}_j)}{\mathbf{Tr}(\widetilde{M})} \right]. \tag{4.17}$$

DML-eig can be rewritten as

$$\begin{array}{ll} \max\limits_{\widetilde{M}\in\mathbb{S}_+^d} & \min\limits_{(i,j)\in\mathcal{D}} d_{\widetilde{M}}(\tilde{x}_i, \tilde{x}_j) \\ \text{s.t.} & \mathbf{Tr}(\widetilde{M}) \leq 1. \end{array} \tag{4.18}$$

LDML can be rewritten as

$$\max_{\widetilde{M}\in\mathbb{S}^d} \sum_{(i,j)\in\mathcal{P}} y_{ij} \ln \widetilde{p}_{ij} + (1 - y_{ij}) \ln (1 - \widetilde{p}_{ij}), \tag{4.19}$$

where $\widetilde{p}_{ij} = (1 + \exp(d_{\widetilde{M}}(\tilde{x}_i, \tilde{x}_j) - b))^{-1}$.
For ITML, let $M_0 = X_{\mathcal{S}}^{-1}$, then we have $\widetilde{M}_0 = L_{\mathcal{S}}^T M_0 L_{\mathcal{S}} = I$. Hence, in this case, ITML is equivalent to

$$\begin{array}{ll} \min_{\widetilde{M}\in\mathbb{S}_+^d} & \mathbf{Tr}(\widetilde{M}) - \log\det(\widetilde{M}) \\ \text{s.t.} & d_{\widetilde{M}}(\tilde{x}_i, \tilde{x}_j) \leq u, (i,j) \in \mathcal{S} \\ & d_{\widetilde{M}}(\tilde{x}_i, \tilde{x}_j) \geq l, (i,j) \in \mathcal{D}. \end{array} \tag{4.20}$$

We should mention that the image-vectors x_i and x_j in formulations (2.35), (2.45), (2.37) , (2.39), (2.42) and (2.40) are WPCA-reduced vectors. We can observe, from their equivalent formulations (4.15), (4.16), (4.20), (4.19), (4.17) and (4.18), that they can also be regarded as metric learning over the intra-personal subspace.

The learned metric on the intra-personal subspace should best reflect the geometry induced by the similarity and dissimilarity of face images/tracks: the distance defined on the intra-personal subspace between similar image-pairs/video-pairs is small while the distance between dissimilar image-pairs/video-pairs is large. Metric learning methods [Xing et al., 2003; Weinberger et al., 2006; Davis et al., 2007; Guillaumin et al., 2009; Kan et al., 2011; Ying and Li, 2012] used different objective functions to achieve this goal.

However, the above methods mainly have two limitations: (**L1**) Although these methods can be regarded as metric learning over the intra-personal subspace, they mainly focused on the discrimination of the distance metric and do not explicitly take into account its robustness. Hence, the learned metrics may not be robust to the transformation variations; (**L2**) Despite the fact that similarity functions s_M and CS_M outperform the distance metric d_M for face verification in still images [Nguyen and Bai, 2011], the above methods only used the distance metric d_M. These limitations degenerate their final verification performance, as has been shown in the experimants. Our proposed method Sub-SML addressed the above limitations by introducing a new generalized similarity function and a novel regularization framework for learning similarity functions.

From the formulation (4.20), we can see that by minimizing the LogDet divergence between matrices M and I, ITML actually retains the robustness to transformation variations. However, its performance on LFW, YTF and VIPeR is inferior to Sub-SML (see Sections 4.4.1 and 4.5 for details). This observation demonstrates the effectiveness of our proposed Sub-SML.

4.7 Conclusion

In this chapter, we first introduced a novel regularization framework of learning a generalized similarity function for unconstrained face verification in still images and person re-identification. The learning objective is formulated by incorporating both the robustness to large transformation differences and the discriminative power of novel distance metrics and similarity functions, a property most of existing similarity metric learning methods do not hold. Our formulation is a convex optimization problem which guarantees the existence of a global solution. We then extend our proposed framework to unconstrained face verification in videos. Lastly, we provide experimental studies on the benchmark LFW [Huang et al., 2007] and YTF [Wolf et al., 2011a] databases for unconstrained face verification in still images and videos, respectively. Besides, we report experimental results on VIPeR database [Gray et al., 2007] for person re-identification. Our proposed methods have achieved the state-of-the-art performances.

Now we discuss some promising future directions. In video-based face verification, a misleading factor in the similarity of facial images is the 3D orientation of head induced by pose, the implication of which should be eliminated. Recently, Wolf and Levy [2013] derived a SVM $\ominus$ classifier to discriminate similar video-pairs from dissimilar video-pairs in a way that the learned similarity

score using feature descriptors such as LBP is uncorrelated with the similarity score induced by the head pose. It would be very interesting to integrate this additional 3D head orientation information into our proposed regularization framework based on the similar idea.

After having dealt with similarity metric learning methods with application to unconstrained face verification and person re-identification, in the next chapter we will focus on metric learning method for improving kNN classification.

5 Metric Learning Revisited

5.1 Introduction

In the previous chapters, we proposed two novel methods called Intra-PCA and Sub-SML for unconstrained face verification and person re-identification and showed their great potential to boost the recognition performance. In this chapter, we focus on metric learning method for improving kNN classification. There is a large amount of studies devoted to metric learning. As described in Section 2.3.2, metric learning methods proposed by Xing et al. [2003] and Davis et al. [2007] are global methods which learn the distance metric satisfying all the pairwise constraints simultaneously. However, it was observed in [Weinberger et al., 2006; Shen et al., 2009; Ying and Li, 2012] that metric learning methods using local pairwise constraints always outperform that using global ones. This is particularly reasonable in the case of learning a distance metric for the kNN classifiers since kNN classifiers are influenced mostly by the data items that are close to the test/query examples.

To obtain optimal solutions, metric learning methods [Xing et al., 2003; Weinberger et al., 2006; Davis et al., 2007; Guillaumin et al., 2009; Shen et al., 2009; Ying and Li, 2012] employed different optimization algorithms. In particular, Xing et al. [2003] used the projection gradient descent algorithm to obtain the optimal distance metric. A drawback of this algorithm is that it usually takes a large number of iterations to become convergent and needs the full eigen-decomposition per iteration.

In this chapter, we revisit the original model [Xing et al., 2003], where the authors proposed to learn a distance metric by maximizing the distances between dissimilar samples whilst keeping the distances between similar points upper-bounded. Ying and Li [2012] proposed to maximize the minimal distance between dissimilar pairs while maintaining an upper bound for the distances between similar pairs. The first contribution of this chapter is to extend the methods in [Xing et al., 2003; Ying and Li, 2012] and propose a general formulation for metric learning. We prove the convexity of this general formulation and illustrate it with various examples. Our second contribution is to show, by exploring its special structures, that the proposed formulation is further equivalent to a convex optimization over the spectrahedron. This equivalent formulation enables us to directly employ the Frank-Wolfe algorithm [Frank and Wolfe, 1956] to obtain the optimal solution. In contrast to the algorithm in [Xing et al., 2003] which needs the full eigen-decomposition of a matrix per iteration, our proposed algorithm only needs to compute the largest eigenvector of a matrix per iteration. We conduct experiments on the UCI datasets for kNN classification and experimental results demonstrate the effectiveness of our proposed method and algorithm. In ad-

[1]This work has been published in [Cao et al., 2012].

dition, the proposed method is shown to compare competitively to those state-of-the-arts on the Labeled Faces in the Wild (LFW) dataset [Huang et al., 2007] for unconstrained face verification.

The rest of the chapter is organized as follows. The next section presents the proposed model and proves its convexity. Section 5.3 establishes its equivalent formulation from which an efficient algorithm is proposed. Section 5.4 reports experimental results for k-NN classification on the UCI datasets and Section 5.5 presents experimental results for unconstrained face verification on the LFW dataset. In Section 5.6, we discuss metric learning models which are closely related to our work. Section 5.7 concludes the chapter.

5.2 Convex Metric Learning Model

For simplicity, we focus on learning a distance metric for kNN classification, although the proposed method below can be easily adapted to metric learning for k-means clustering. We begin by introducing some useful notations. For any $n \in \mathbb{N}$, denote $\mathbb{N}_n = \{1, 2, \ldots, n\}$. Now we rewrite the training data by $\mathbf{z} := \{(x_i, l(x_i)) : i \in \mathbb{N}_n\}$ with input $x_i \in \mathbb{R}^d$ and class label $l(x_i)$ (not necessarily binary). Since we are mainly concerned with metric learning for kNN classifier, the pairwise constraints are generated locally, that is, the similar/dissimilar pairs are k-nearest neighbors. One can follow the mechanism in [Weinberger et al., 2006] to extract local information of similar/dissimilar pairs. The details can be found in the experimental section.

Given a set of similar samples and a set of dissimilar samples, we aim to find a good distance matrix M such that the distances between dissimilar pairs are large while keeping the distances between similar pairs small. There are many formulations to achieve this goal. In particular, recall the formulation proposed by Xing et al. [2003]:

$$\begin{array}{cc} \max_{M \in \mathbb{S}_+^d} & \sum_{(i,j)\in\mathcal{D}} \sqrt{d_M(x_i, x_j)} \\ \text{s.t.} & \sum_{(i,j)\in\mathcal{S}} d_M(x_i, x_j) \leq 1. \end{array} \tag{5.1}$$

An projection gradient descent algorithm was employed to solve the above problem. However, the algorithm generally takes a long time to converge and needs the computation of the full eigendecomposition of a matrix per iteration.

In this chapter, we propose a more general formulation:

$$\begin{array}{cc} \max_{M \in \mathbb{S}_+^d} & \left[\sum_{(i,j)\in\mathcal{D}} [d_M(x_i, x_j)]^p / D\right]^{\frac{1}{p}} \\ \text{s.t.} & \sum_{(i,j)\in\mathcal{S}} d_M(x_i, x_j) \leq 1, \end{array} \tag{5.2}$$

where $p \in (-\infty, \infty)$ and D is the number of dissimilarity pairs. We refer to the above formulation as $\mathbf{DML}_p$. The above formulation is well defined even for the limiting case $p = 0$ as discussed in the examples below.

- $\mathbf{p = 1/2}$: In this case, problem (5.2) can be written as

$$\begin{array}{cc} \max_{M \in \mathbb{S}_+^d} & \left[\sum_{(i,j)\in\mathcal{D}} \sqrt{d_M(x_i, x_j)} / D\right]^2 \\ \text{s.t.} & \sum_{(i,j)\in\mathcal{S}} d_M(x_i, x_j) \leq 1, \end{array} \tag{5.3}$$

which is equivalent to formulation (5.1) proposed in [Xing et al., 2003].

- $\mathbf{p} \to -\infty$: Observe, for any positive sequence $\{\alpha_i > 0 : i \in \mathbb{N}_n\}$, that

$$\lim_{p\to-\infty}\big(\sum_{i\in\mathbb{N}_n} a_i^p/n\big)^{\frac{1}{p}} = \min_{i\in\mathbb{N}_n} a_i.$$

Hence, in the limiting case $p \to -\infty$, problem (5.2) is reduced to the metric learning model DML-eig proposed by Ying and Li [2012]:

$$\begin{array}{cc} \max_{M\in\mathbb{S}_+^d} & \min_{(i,j)\in\mathcal{D}} d_M(x_i,x_j) \\ \text{s.t.} & \sum_{(i,j)\in\mathcal{S}} d_M(x_i,x_j) \le 1. \end{array} \tag{5.4}$$

- $\mathbf{p} \to \mathbf{0}$: Note, for any sequence $\{\alpha_i > 0 : i \in \mathbb{N}_n\}$, that

$$\lim_{p\to 0}\big[\sum_{i\in\mathbb{N}_n} a_i^p/n\big]^{\frac{1}{p}} = \prod_{i=1}^{n}\alpha_i^{\frac{1}{n}}.$$

Hence, in the limiting case $p \to 0$, problem (5.2) becomes

$$\begin{array}{cc} \max_{M\in\mathbb{S}_+^d} & \prod_{(i,j)\in\mathcal{D}}[d_M(x_i,x_j)]^{\frac{1}{D}} \\ \text{s.t.} & \sum_{(i,j)\in\mathcal{S}} d_M(x_i,x_j) \le 1, \end{array}$$

where D is the number of dissimilar pairs in the set $\mathcal{D}$.

The following theorem investigates the convexity/concavity of the objective function in problem (5.2).

Theorem 4. *Let function $\mathcal{L} : \mathbb{S}_+^d \to \mathbb{R}$ be the objective function of DML$_p$, i.e., for any $M \in \mathbb{S}_+^d$, $\mathcal{L}(M) = \big[\sum_{(i,j)\in\mathcal{D}}[d_M(x_i,x_j)]^p/D\big]^{\frac{1}{p}}$ for $p \neq 0$, and $\mathcal{L}(M) = \prod_{(i,j)\in\mathcal{D}}[d_M(x_i,x_j)]^{\frac{1}{D}}$ for $p = 0$. Then, we have that $\mathcal{L}(\cdot)$ is concave for $p < 1$ and otherwise convex.*

Proof. First we prove the concavity of $\mathcal{L}(\cdot)$ when $p < 1$ and $p \neq 0$. It suffices to prove, for any $n \in \mathbb{N}$ and for any $\{\mathbf{a} = (a_1, a_2, \ldots, a_n) : a_i > 0,\ \ i \in \mathbb{N}_n\}$, that function $(\sum_{i\in\mathbb{N}_n} a_i^p)^{1/p}$ is concave w.r.t. variable $\mathbf{a}$. To this end, let f be a function defined, for any $x > 0$ and $y > 0$, by $f(x, y) = -x^{1-p}y^p/p$. We can easily prove that f is jointly convex w.r.t. (x, y), since its Hessian matrix

$$(1-p)\begin{pmatrix} x^{-p-1}y^p & -x^{-p}y^{p-1} \\ -x^{-p}y^{p-1} & x^{1-p}y^{p-2} \end{pmatrix} \in \mathbb{S}_+^d.$$

Consequently, for any $i \in \mathbb{N}_n$, $-x^{1-p}a_i^p/p$ is jointly convex, which implies that its summation $\sum_{i\in\mathbb{N}_n} -x^{1-p}a_i^p/p = -x^{1-p}(\sum_{i\in\mathbb{N}_n} a_i^p)/p$ is jointly convex. Hence, the function defined by $E(x, \mathbf{a}) = (1-p)x/p - x^{1-p}(\sum_{i\in\mathbb{N}_n} a_i^p)/p$ is also jointly convex w.r.t. $(x, \mathbf{a})$. Clearly,

$$-\big(\sum_{i\in\mathbb{N}_n} a_i^p\big)^{1/p} = \min\{E(x,\mathbf{a}) : x \ge 0\}. \tag{5.5}$$

Recalling that the partial minimum of a jointly convex function is convex [Horn and Johnson, Sec.IV.2.4], we obtain the concavity of $(\sum_{i\in\mathbb{N}_n} a_i^p)^{1/p}$ when $p < 1$ and $p \neq 0$. The concavity of $\mathcal{L}$ for $p = 0$ follows from the fact that the limit function of a sequence of concave functions is

concave.

The convexity of $\mathcal{L}$ for $p \geq 1$ can be proved similarly by observing that $E(x, \mathbf{a})$ is jointly concave if $p \geq 1$. Consequently, equation (5.5) should be replaced by $(\sum_{i\in\mathbb{N}_n} a_i^p)^{1/p} = \min\{-E(x, \mathbf{a}) : x \geq 0\}$. This completes the proof of the theorem. □

We conclude this section with three remarks. Firstly, we exclude the extreme case $p = 1$ since, in this case, the optimal solution of DML$_p$ will be always a rank-one matrix (i.e. the data is projected to the line), as argued in [Xing et al., 2003]. Secondly, when $p \in (1, \infty)$, by Theorem 4 we know that formulation (5.2) is indeed a problem of *maximizing a convex function*, which is a challenging task to get a global solution. In this chapter we will only consider the case $p \in (-\infty, 1)$ which guarantees that formulation (5.2) is a convex optimization problem. Lastly, we show that DML$_p$ can be regarded as metric learning over the intra-personal subspace. Let $\widetilde{M} = L_{\mathcal{S}}^T M L_{\mathcal{S}}$, then formulation (5.2) is equivalent to

$$\begin{array}{cc} \max_{\widetilde{M}\in\mathbb{S}_+^d} & \left[\sum_{(i,j)\in\mathcal{D}} [d_{\widetilde{M}}(\tilde{x}_i, \tilde{x}_j)]^p / D\right]^{\frac{1}{p}} \\ \text{s.t.} & \sum_{(i,j)\in\mathcal{S}} d_{\widetilde{M}}(\tilde{x}_i, \tilde{x}_j) \leq 1. \end{array} \tag{5.6}$$

Note that although DML$_p$ can be regarded as metric learning over the intra-personal subspace, we do not explicitly take into account on how to remain robust to large transformation variations because this chapter mainly focuses on metric learning for k-NN classification.

5.3 Equivalent Formulation and Optimization

We turn our attention to an equivalent formulation of problem (5.2), which is key to design its efficient algorithm. For notational simplicity, denote the *spectrahedron* by $\mathcal{Q} = \{M \in \mathbb{S}_+^d : \mathbf{Tr}(M) = 1\}$. For any $X, Y \in \mathbb{R}^{d\times d}$, we denote the inner product in $\mathbb{S}^d$ by $\langle X, Y\rangle = \mathbf{Tr}(X^T Y)$. We use the convention $X_{ij} = (x_i - x_j)(x_i - x_j)^\top$ and then X_S (i.e. equation (3.1)) can be written as $X_{\mathcal{S}} = \sum_{(i,j)\in\mathcal{S}} X_{ij}$. For any $\tau = (i, j) \in \mathcal{D}$, rewrite X_{ij} as X_τ. Without loss of generality, we assume that $X_{\mathcal{S}}$ is invertible throughout the chapter. This can be achieved by adding a small ridge term, i.e. $X_{\mathcal{S}} \longleftarrow X_{\mathcal{S}} + \delta\, \mathbf{I}_d$ where $\mathbf{I}_d$ is the identity matrix and $\delta > 0$ is a small ridge constant. Then, DML$_p$ (i.e. formulation (5.2)) can be rewritten as the following problem:

$$\begin{array}{cc} \max_{M\in\mathbb{S}_+^d} & \left[\sum_{\tau\in\mathcal{D}} \langle X_\tau, M\rangle^p / D\right]^{\frac{1}{p}} \\ \text{s.t.} & \langle X_{\mathcal{S}} + \delta \mathbf{I}_d, M\rangle \leq 1. \end{array} \tag{5.7}$$

Applying the Cholesky decomposition, we get that $X_{\mathcal{S}} = L_C L_C{}^\top$, where L_C is a lower triangular matrix with strictly positive diagonal entries. Equipped with the above preparations, we are now ready to show that problem (5.2) is equivalent to an optimization problem over the spectrahedron $\mathcal{Q} = \{M \in \mathbb{S}_+^d : \mathbf{Tr}(M) = 1\}$. Similar ideas have been used in [Ying and Li, 2012].

Theorem 5. *For any $\tau = (i, j) \in \mathcal{D}$, let $\widetilde{X}_\tau = {L_C}^{-1}(x_i - x_j)({L_C}^{-1}(x_i - x_j))^\top$. Then, problem (5.2) is equivalent to*

$$\max_{S\in\mathcal{Q}} \Big[\sum_{\tau\in\mathcal{D}} \langle \widetilde{X}_\tau, S\rangle^p\Big]^{\frac{1}{p}}, \tag{5.8}$$

Frank-Wolfe algorithm for DML$_p$

Input:

· parameter $p \in (-\infty, 1)$

· tolerance value tol (e.g. 10^{-5})

· step sizes $\{\eta_t = 2/(t+1) : t \in \mathbb{N}\}$

Initialization: $S_1 \in \mathbb{S}_+^d$ with $\mathbf{Tr}(S_1) = 1$

For $t = 1, 2, 3, \ldots$ do

· $Z_t = \arg\max \big\{\langle Z, \nabla f(S_t)\rangle : Z \in \mathbb{S}_+^d,\ \mathbf{Tr}(Z) = 1\big\}$,
i.e. $Z_t = v_t v_t^\top$, where v_t is the maximal eigenvector of matrix $\nabla f(S_t)$

· $S_{t+1} = (1 - \eta_t)S_t + \eta_t Z_t$

· if $|f(S_{t+1}) - f(S_t)| < tol$ then break

Output: $d \times d$ matrix $S_t \in \mathbb{S}_+^d$

Table 5.1: Pseudo-code of the Frank-Wolfe (FW) algorithm for DML$_p$ (i.e. formulation (5.8)).

Proof. Let M^* be an optimal solution of problem (5.2) and $\widetilde{M}^* = \frac{M^*}{\langle X_{\mathcal{S}}, M^*\rangle}$. Then, $\langle X_{\mathcal{S}}, \widetilde{M}^*\rangle = 1$ and $\big[\sum_{\tau\in\mathcal{D}} \frac{\langle X_\tau, \widetilde{M}^*\rangle^p}{D}\big]^{\frac{1}{p}} = \big[\sum_{\tau\in\mathcal{D}} \frac{\langle X_\tau, M^*\rangle^p}{D}\big]^{\frac{1}{p}} / \langle X_{\mathcal{S}}, M^*\rangle \geq \big[\sum_{\tau\in\mathcal{D}} \frac{\langle X_\tau, M^*\rangle^p}{D}\big]^{\frac{1}{p}}$ since $\langle X_{\mathcal{S}}, M^*\rangle \leq 1$. This implies that $\widetilde{M}^*$ is also an optimal solution. Consequently, problem (5.2) is equivalent to, up to a scaling constant,

$$\begin{array}{ll} \max_{M\in\mathbb{S}_+^d} & \big[\sum_{(i,j)\in\mathcal{D}} \langle X_\tau, M\rangle^p / D\big]^{\frac{1}{p}} \\ \text{s.t.} & \langle X_{\mathcal{S}}, M\rangle = 1. \end{array} \tag{5.9}$$

Recall that $X_{\mathcal{S}} = L_C L_C{}^\top$ by Cholesky decomposition. Now the desired equivalence between (5.2) and (5.8) follows from changing variable $S = L_C{}^\top M L_C$ in (5.9). This completes the proof of the theorem. □

By Theorem 5, the original metric learning problem (5.2) is reduced to a maximization problem on the spectrahedron. We rewrite the objective function of the equivalent formulation (5.8)

$$f(S) = \big[\sum_{\tau\in\mathcal{D}} \langle \widetilde{X}_\tau, S\rangle^p\big]^{\frac{1}{p}}. \tag{5.10}$$

The objective function $f(S)$ is not smooth since p can be negative. In order to avoid the numerical instability, we can add a small positive number inside so that it is well defined, i.e. $\big[\sum_{\tau\in\mathcal{D}} (\langle \widetilde{X}_\tau, S\rangle)^p\big]^{\frac{1}{p}}$ is replaced by $\big[\sum_{\tau\in\mathcal{D}} (\langle \widetilde{X}_\tau, S\rangle + \varepsilon)^p\big]^{\frac{1}{p}}$ where ε is a small positive number (e.g. $\varepsilon = 10^{-8}$). Its gradient is then given by

$$\nabla f(S) = \frac{\sum_{\tau\in\mathcal{D}} \langle \widetilde{X}_\tau, S\rangle^{p-1} \widetilde{X}_\tau}{\big[\sum_{\tau\in\mathcal{D}} \langle \widetilde{X}_\tau, S\rangle^p\big]^{1-\frac{1}{p}}}. \tag{5.11}$$

Now we are ready to apply the Frank-Wolfe (FW) algorithm [Frank and Wolfe, 1956; Hazan, 2008] to obtain the optimal solution: the pseudo-code of the algorithm is given in Table 5.1.

Data	No.	n	d	class	T	D
Balance	1	625	4	3	3951	1317
Breast-Cancer	2	569	30	2	3591	1197
Diabetes	3	768	8	2	4842	1614
Image	4	2310	19	2	14553	4851
Iris	5	150	4	3	954	315
Waveform	6	5000	21	3	31509	10503
Wine	7	178	13	3	1134	378

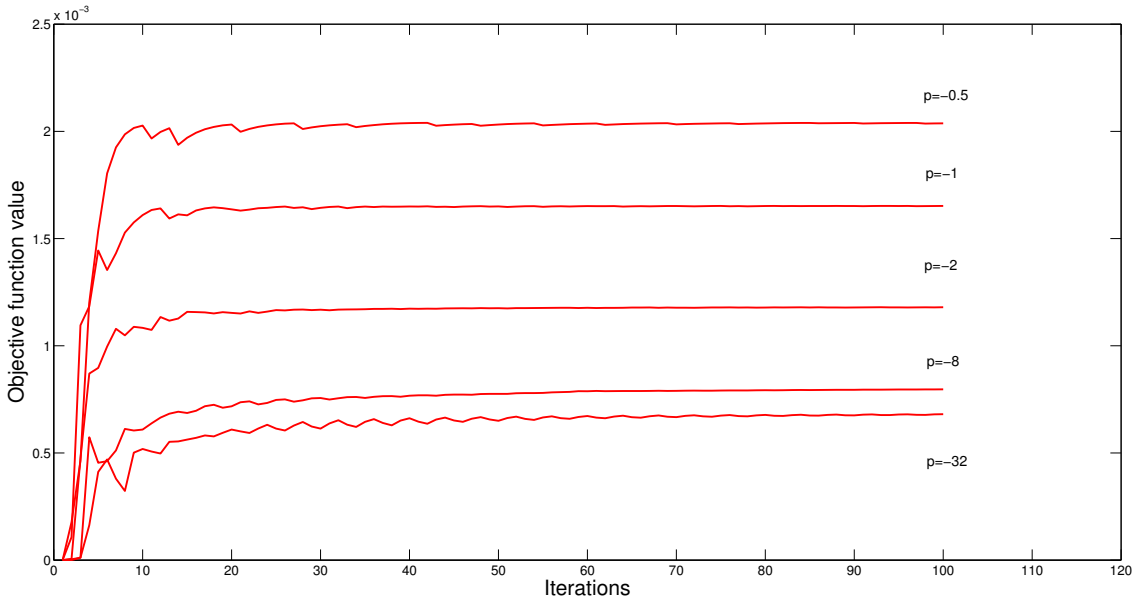
Table 5.2: Description of datasets used in the experiments: n and d respectively denote the number of samples and attributes (feature elements) of the data; T is the number of triplets and D is the number of dissimilar pairs.

5.4 Experiment One: K-NN Classification

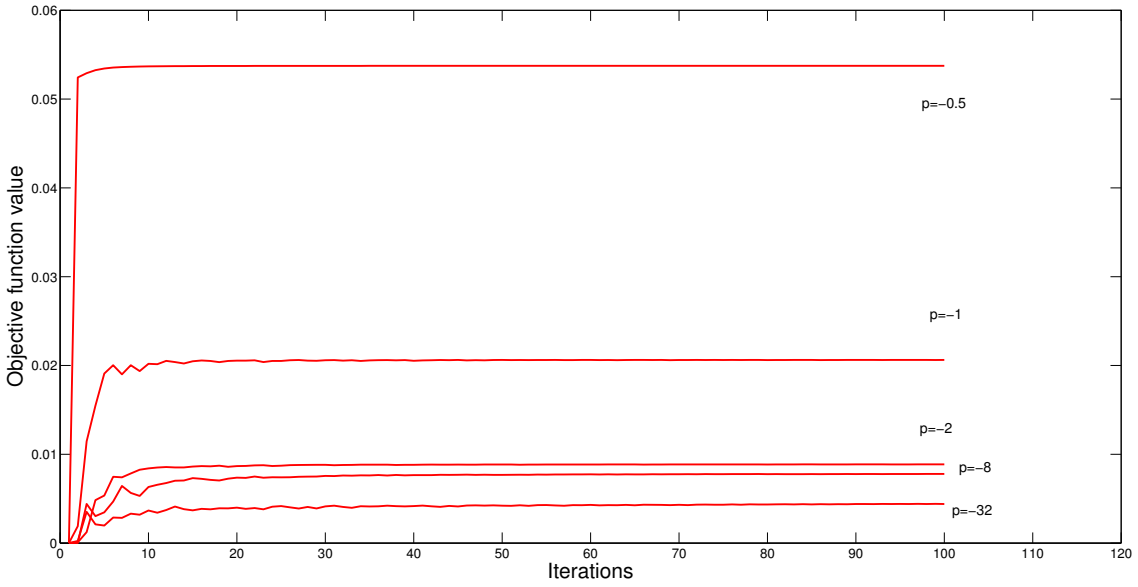
In this section, we aim to assess the convergence and generalization of our proposed DML$_p$ for kNN classification. To this end, we run the experiments on UCI datasets to compare the kNN classification performance ($k = 3$) of different metric learning methods, where the kNN classifier is constructed using the Mahalanobis distance learned by metric learning methods. Specifically, we compare the empirical performance of our proposed method DML$_p$ with six other methods: Xing [Xing et al., 2003], LMNN [Weinberger et al., 2006], ITML [Davis et al., 2007], BoostMetric [Shen et al., 2009], DML-eig [Ying and Li, 2012] and the baseline algorithm using the standard Euclidean distance denoted by Euclidean. The model parameters in ITML, LMNN, BoostMetric and DML$_p$ are tuned via three-fold cross validation. In addition, the maximum iteration number for DML$_p$ is 1000 and the algorithm is terminated when the relative change of the objective function value is less than 10^{-5}.

To investigate the convergence and generalization of DML$_p$, we run experiments on seven UCI datasets: i.e. 1) Balance; 2) Breast-Cancer; 3) Diabetes; 4) Image segmentation; 5) Iris; 6) Waveform; 7) Wine. The statistics of the datasets are summarized in Table 5.2. All the experimental results are obtained by averaging over 10 runs and, for each run, the data is randomly split into 70% for training and 30% for testing. To generate relative constraints and pairwise constraints, we adopt a similar mechanism in [Weinberger et al., 2006]. More specifically, for each training point x_i, k-nearest neighbors that have the same labels as $l(x_i)$ (targets) as well as k-nearest neighbors that have different labels from $l(x_i)$ (imposers) are found. According to x_i and its corresponding targets and imposers, we then construct the set of similar pairs $\mathcal{S}$, the set of dissimilar pairs $\mathcal{D}$ and the set of relative constraints in the form of triplets denoted by $\mathcal{T}$ required by LMNN and BoostMetric. As mentioned above, the original formulation in [Xing et al., 2003] used all pairwise constraints. For fairness of comparison, all methods including Xing used the same set of similar/dissimilar pairs generated locally as above.

Convergence of DML$_p$. We study the convergence of algorithm for DML$_p$ with varying values of p. In Figure 5.1 and Figure 5.2, we plot the objective function value of DML$_p$ versus the number of iteration on Balance (Figure 5.1a); Iris (Figure 5.1b); Diabetes (Figure 5.2a); and Image (Figure 5.2b). We can see from Figure 5.1 and Figure 5.2 that the algorithm converges quickly. The smaller the value of p is and the more iterations algorithm DML$_p$ needs.

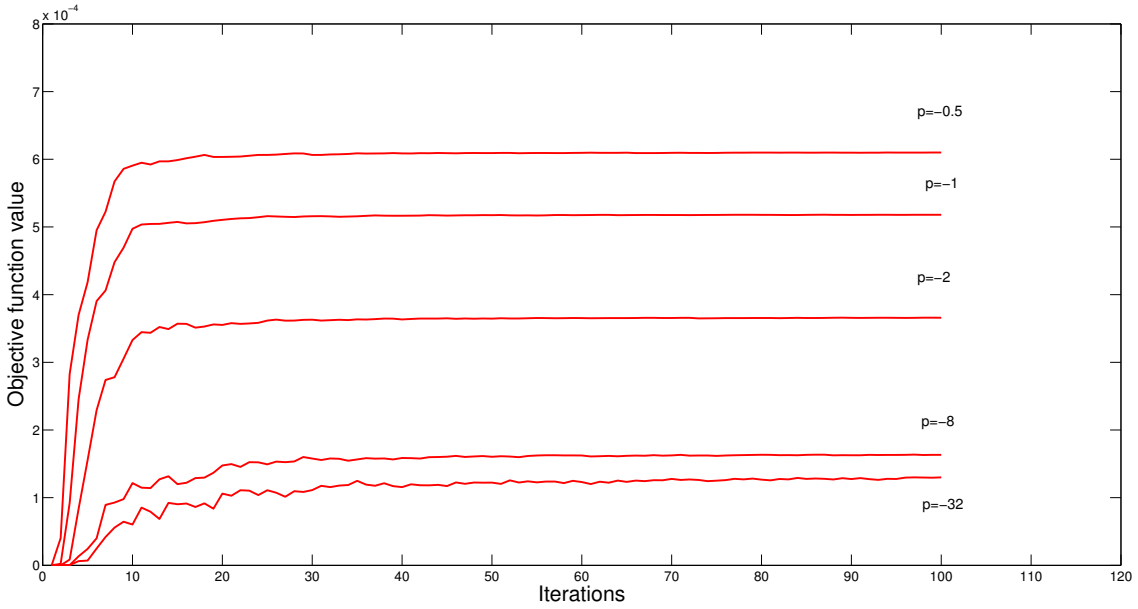


(a) Balance

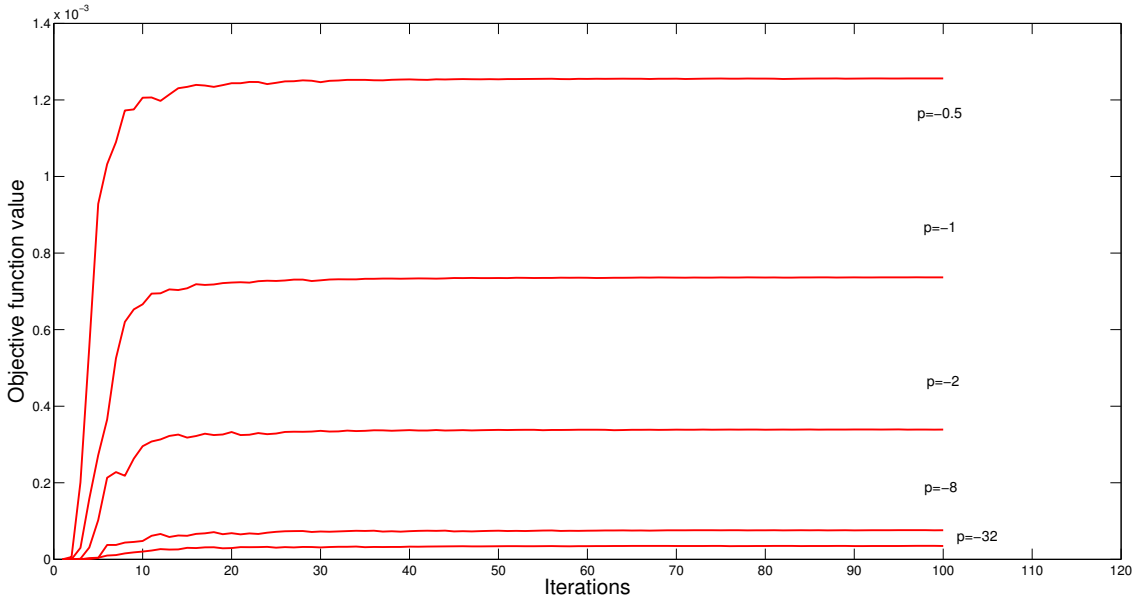


(b) Iris

Figure 5.1: Evolution of the objective function value of DML_p versus the number of iteration with varying p on Balance (a) and Iris (b).

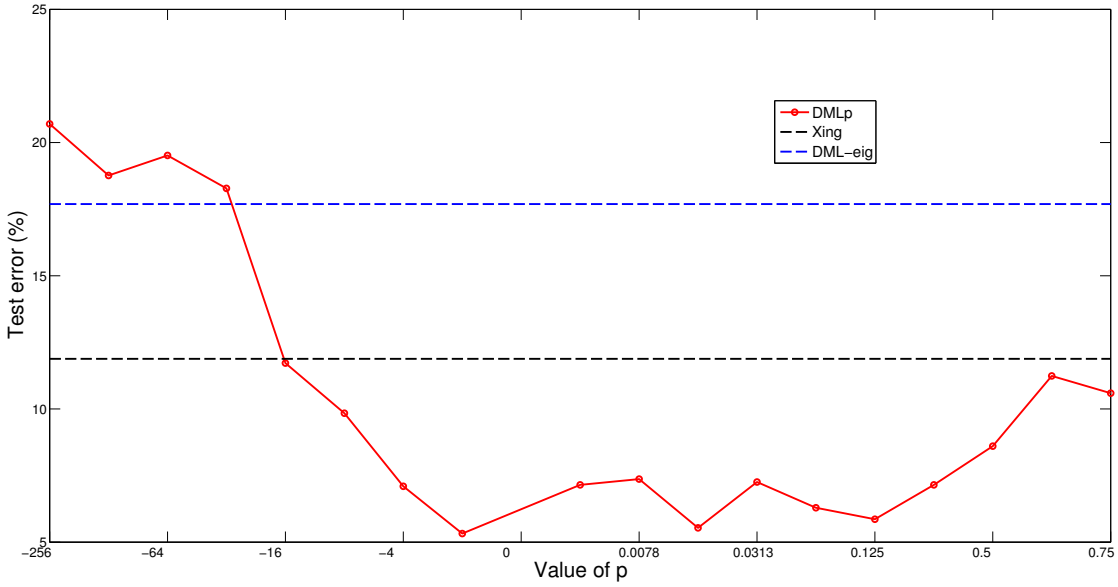


(a) Diabetes

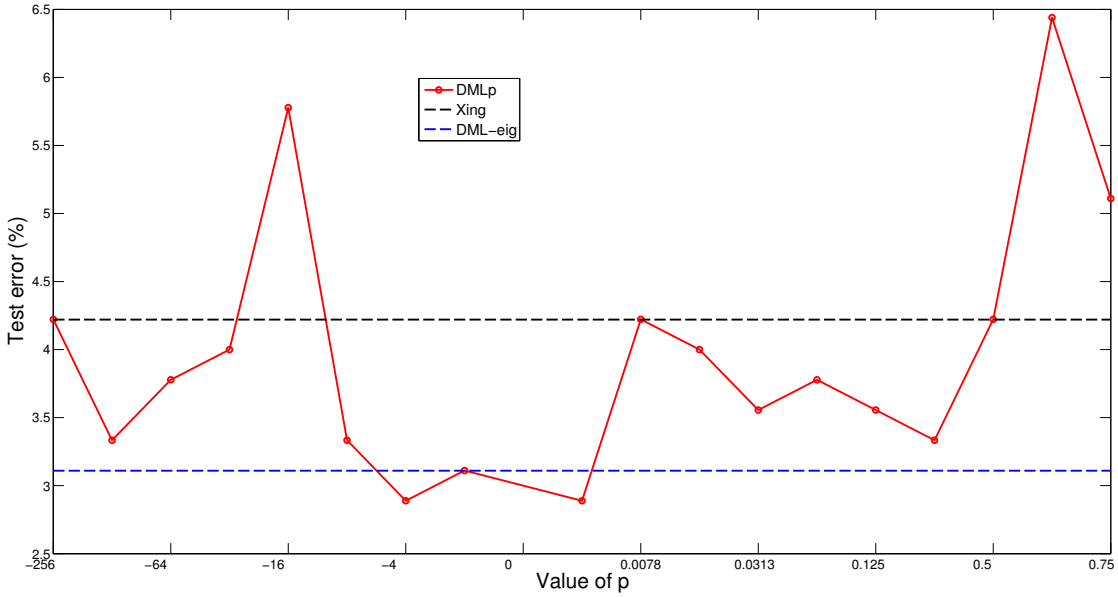


(b) Image

Figure 5.2: Evolution of the objective function value of DML$_p$ versus the number of iteration with varying p on Diabetes (a) and Image (b).

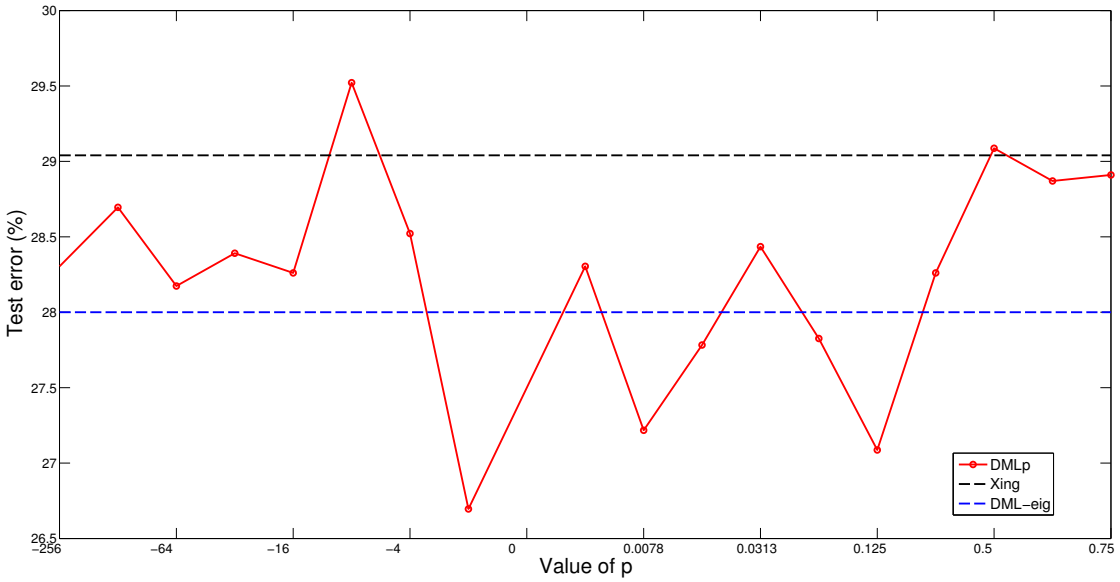


(a) Balance

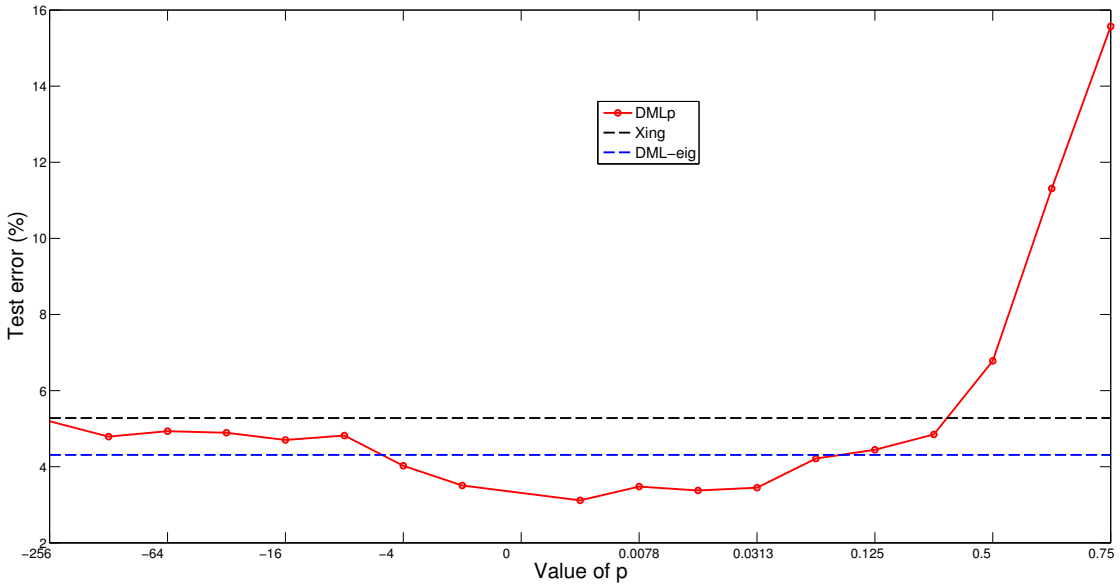


(b) Iris

Figure 5.3: Test error (%) of DML$_p$ versus different values of p on Balance (a) and Iris (b). Red circled line is the result of DML$_p$ across different values of p (log-scaled); blue dashed line is the result of DML-eig [Ying and Li, 2012] and black dashed line represents the result of Xing [Xing et al., 2003].



(a) Diabetes



(b) Image

Figure 5.4: Test error (%) of DML$_p$ versus different values of p on Diabetes (a) and Image (b). Red circled line is the result of DML$_p$ across different values of p (log-scaled); blue dashed line is the result of DML-eig [Ying and Li, 2012] and black dashed line represents the result of Xing [Xing et al., 2003].

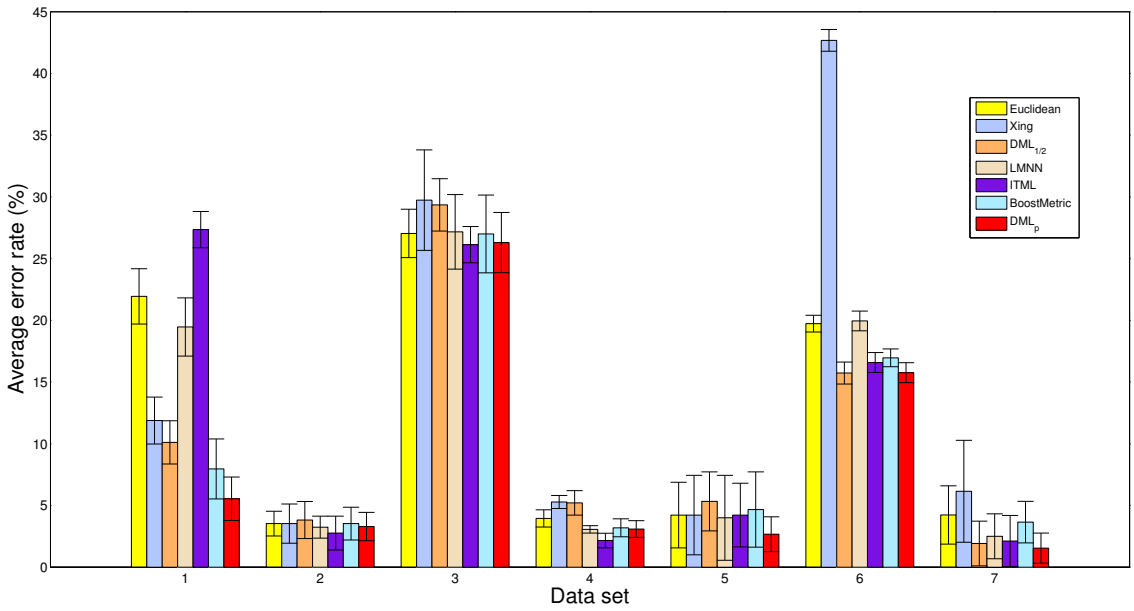


Figure 5.5: Average test error (%) of DML$_p$ against other methods.

Generalization of DML$_p$. We conduct experiments to show the generalization performance of DML$_p$ for kNN classifiers where the distance metric to measure nearest neighbors is learned by metric learning methods.

Firstly, we investigate the performance of DML$_p$ against different values of p. Figure 5.3 and Figure 5.4 depict the test error of DML$_p$ versus the value of p on Balance (Figure 5.3a); Iris (Figure 5.3b); Diabetes (Figure 5.4a); and Image (Figure 5.4b). We can observe from Figure 5.3 and Figure 5.4 that the test error varies on different values of p and the best performance of DML$_p$ is superior to those of DML-eig [Ying and Li, 2012] and Xing [Xing et al., 2003] which are the special cases of DML$_p$ with $p \to -\infty$ and $p = 1/2$ respectively. This observation validates the value of the general formulation DML$_p$ and suggests the importance of choosing an appropriate value of p. In the following experiments, we will tune the value of p by three cross-validation.

Secondly, we study the generalization performance of DML$_p$ for kNN classification. To this end, we compare DML$_p$ with other metric learning methods including Xing [Xing et al., 2003], LMNN [Weinberger et al., 2006], ITML [Davis et al., 2007] and BoostMetric [Shen et al., 2009]. Figure 5.5 depicts the performance of different methods. It shows that almost all metric learning methods improve kNN classification using Euclidean distance on most datasets. Our proposed method DML$_p$ delivers competitive performance with the other state-of-the-art algorithms such as ITML, LMNN and BoostMetric. Indeed, DML$_p$ outperforms the other methods on 4 out of 7 datasets and shows competitive performance against the best one on the rest 3 datasets. From Figure 5.5, it is reasonable to see that the test errors of DML$_{1/2}$ are consistent with those of Xing since they are essentially the same model implemented by different algorithms. The only exception is the performance on Waveform dataset: the test error of Xing is much worse than DML$_{1/2}$. The reason could be that the projection gradient algorithm proposed in [Xing et al., 2003] does not converge in a reasonable time due to the relatively large number of samples on Waveform dataset.

5.5 Experiment Two: Unconstrained Face Verification

In this section, we apply our proposed DML$_p$ to the problem of unconstrained face verification in still images. Experiments are carried out in the restricted setting of the Labeled Faces in the Wild (LFW) database [Huang et al., 2007], see Section 2.4.1 for a brief introduction of this database.

For feature representation, we investigate four facial descriptors: Intensity (see Section 2.2.1), SIFT [Guillaumin et al., 2009], LBP [Ojala et al., 2002] and TPLBP [Wolf et al., 2008]. Since the dimension of the original descriptors is quite high (from 3456 to 12000), we reduce the dimension using PCA. These descriptors are tested with both their original values and the square root of them [Wolf et al., 2008; Guillaumin et al., 2009].

The performance is measured by the 10-fold cross-validation test. In each of the 10-fold cross-validation test, the parameter p in DML$_p$ is tuned via three-fold cross validation over the remaining 9-fold training sets. In the restricted protocol, only pairwise constraints are given. Since LMNN [Weinberger et al., 2006] and BoostMetric [Shen et al., 2009] require relative constraints, we only compare our DML$_p$ with ITML [Davis et al., 2007] and LDML [Guillaumin et al., 2009].

Firstly, we investigate the performance of DML$_p$ on the SIFT descriptor by varying the dimension of principal components. Figure 5.6 depicts the verification accuracy versus the dimension of PCA. We can see that, compared to the algorithms ITML and LDML, DML$_p$ using the SIFT descriptor delivers relatively stable performance as PCA dimension varies. In particular, the performance of DML$_p$ becomes stable after the dimension of PCA reaches around 100 and it consistently outperforms ITML across different PCA dimensions. We also observed similar results for other descriptors. Hence, for simplicity we set the PCA dimension to be 100 for the SIFT descriptor and other descriptors. According to [Guillaumin et al., 2009], the best performances of LDML and ITML on the SIFT descriptor are 77.50% and 76.20% respectively. The best performance of DML$_p$ reaches around 80% which outperforms ITML and LDML. We also note that the performance of ITML we get here is consistent with that reported in [Guillaumin et al., 2009].

	DML$_p$	DML$_p$ SQRT
SIFT	80.15 ± 0.55	80.28 ± 0.59
LBP	79.72 ± 0.62	80.05 ± 0.81
TPLBP	77.90 ± 0.58	78.22 ± 0.61
Above combined	85.72 ± 0.55	
Intensity	73.35 ± 0.54	73.48 ± 0.51
All combined	$\mathbf{86.07 \pm 0.58}$	

Table 5.3: Verification rate ($\pm$ standard error) of DML$_p$ on LFW database using different descriptors (mean verification accuracy and standard error) in the restricted setting of LFW. "DML$_p$ SQRT" means DML$_p$ uses the square root of the descriptor. "Intensity" means the raw pixel data by concatenating the intensity value of each pixel in the image. For all feature descriptors, the dimension is reduced to 100 using PCA. See more details in the text.

Secondly, we test the performance of our method using different descriptors and their combinations. Table 5.3 summarizes the results. In Table 5.3, the notation "Above combined" means that we fuse the distance scores from the above listed (six) descriptors in the table and train a linear Support Vector Machine (SVM) on the fused vector to make prediction (see [Guillaumin

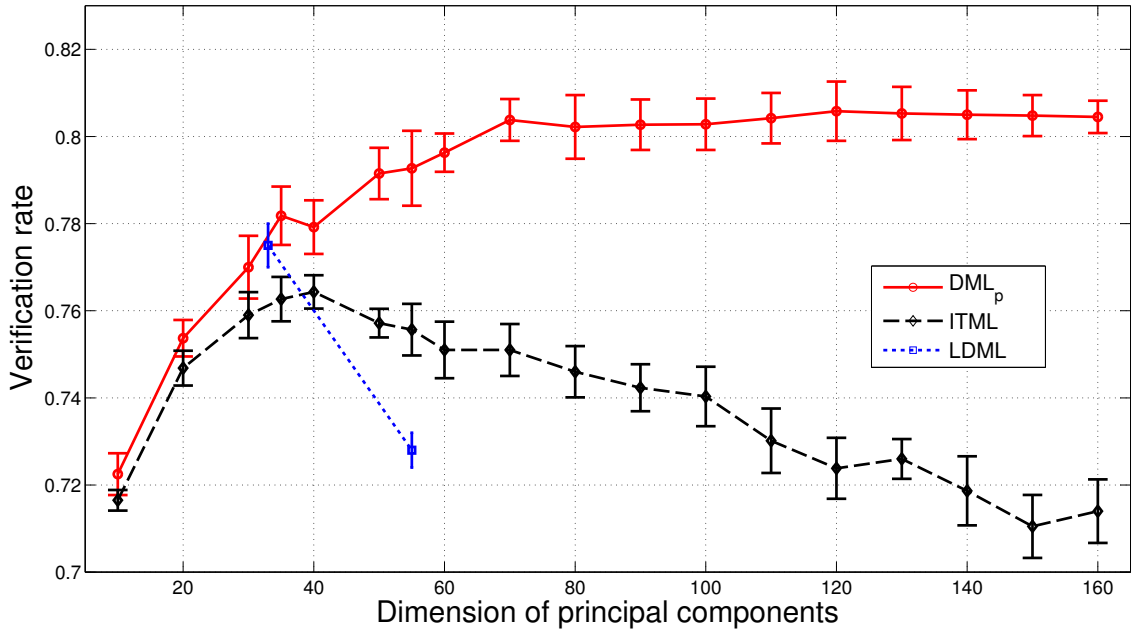


Figure 5.6: Mean verification rate of DML$_p$, ITML, and LDML by varying PCA dimension using the SIFT descriptor in the restricted setting of LFW. The result of LDML is copied from [Guillaumin et al., 2009]: the best performance of LDML and ITML on the SIFT descriptor are respectively 77.50% and 76.20%.

et al., 2009]). "All combined" means that all eight distance scores are combined. We observe that combining four descriptors (Intensity, SIFT, LBP and TFLBP) and their square-root ones yields 86.07% which outperforms 85.65% of DML-eig [Ying and Li, 2012]. As mentioned above, DML-eig can be regarded as a limiting case of DML$_p$ as $p \to -\infty$. This observation also validates the value of the general formulation DML$_p$. From Table 5.3, we can see that, although the individual performance of Intensity is inferior to those of other descriptors, combining it with other descriptors slightly increases the overall performance from 85.72% to 86.07%.

Method	Accuracy $\pm$ SE
High-Throughput Brain-Inspired Features, aligned [Cox and Pinto, 2011]	$\mathbf{88.13 \pm 0.58}$
LDML + Combined, funneled [Guillaumin et al., 2009]	79.27 ± 0.60
DML-eig + Combined [Ying and Li, 2012]	85.65 ± 0.56
DML$_p$ + Combined (this work)	86.07 ± 0.58

Table 5.4: Comparison of DML$_p$ with other state-of-the-art methods in the restricted configuration based on combination of different types of descriptors[1].

Finally, we summarize the performance of DML$_p$ and other state-of-the-art methods in Table 5.4 and plot the ROC curve of our method compared to other published results in Figure 5.7. We observe from Table 5.4 that our method DML$_p$ outperforms metric learning methods LDML [Guillaumin et al., 2009] and DML-eig [Ying and Li, 2012]. Section 4.4.1 has given a detailed comparison of the recent state-of-the-art methods in the restricted setting of LFW and Table 4.5 has listed the comparison results. Note that the results compared in Table 5.4 are system to system where metric learning is only one part of the system. We should point out that the state-of-the art result 88.13% obtained by Cox and Pinto [2011] was not achieved by metric learning method.

[1]Table 4.5 in Section 4.4.1 gives up-to-date results in the restricted setting of LFW.

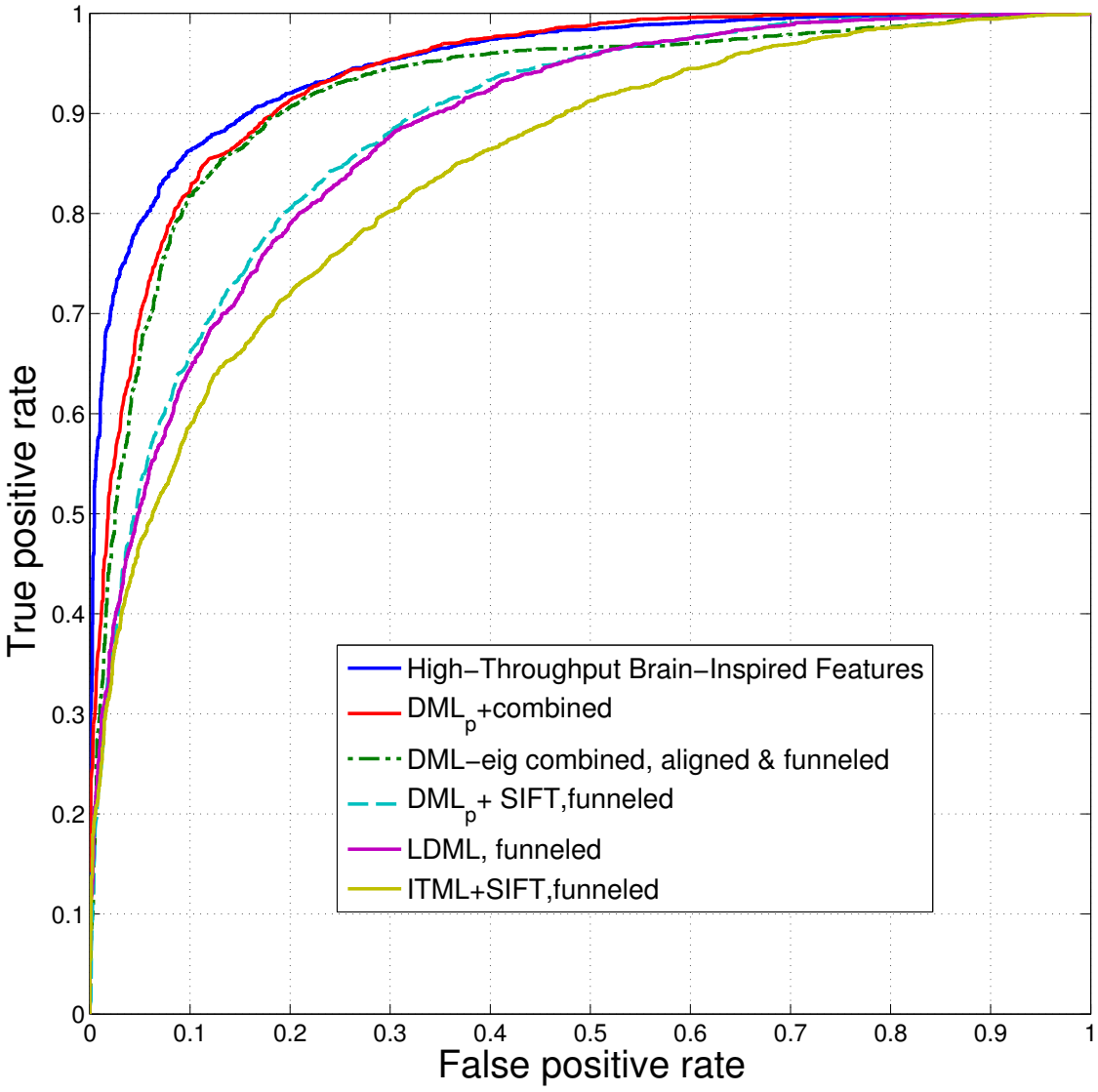


Figure 5.7: ROC curves of DML$_p$ and other state-of-the-art methods on LFW dataset.

Instead, it performs sophisticated large scale feature search which used multiple complimentary representations derived through training set augmentation, alternative face comparison functions, and feature set searches with a varying number of model layers. We believe that the performance of DML$_p$ may be further improved by exploring different types of descriptors such as those used in [Cox and Pinto, 2011].

5.6 Discussion

Below we discuss metric learning models that are closely related to our method.

Xing et al. [2003] presented metric learning formulation (5.1) for k-means clustering and projection gradient algorithm was employed to obtain the optimal solution. Specifically, at each iteration, the algorithm takes a gradient ascent step and then projects it back to the constraints and the cone of the positive semi-definite matrices. One drawback of the above projection gradient method is that a large number of iterations might be taken before its convergence and the computation of full eigen-decomposition per iteration is needed with a time complexity $\mathcal{O}(d^3)$.

Weinberger et al. [2006] developed a method called LMNN to learn a Mahalanobis distance metric in kNN classification setting. Because any positive semi-definite matrix M can be factored as $M = A^T A$, where $A \in \mathbb{R}^{d \times d}$, LMNN is reformulated as an optimization problem with an unconstrained variable A. The sub-gradient descent algorithm was used to obtain the optimal solution. Unfortunately, the reformulated problem is generally not convex with respect to variable A, and thus the proposed sub-gradient method would lead to local minimizers.

Shen et al. [2009] recently proposed a metric learning method called BoostMetric by employing the exponential loss and a boosting-based algorithm was developed for optimization. The rationale behind the proposed algorithm is that each positive semi-definite matrix can be decomposed into a linear positive combination of trace-one and rank-one matrices. This algorithm is very similar to the Frank-Wolfe algorithm [Frank and Wolfe, 1956; Hazan, 2008] we employed for DML$_p$ since both of them iteratively find a linear combination of rank-one matrices to approximate the desired solutions. However, the above boosting-based algorithm is a general column-generation algorithm and its convergence rate is not clear.

In summary, our approach DML$_p$ overcame the above limitations by proposing a general convex formulation for metric learning and applying Frank-Wolfe algorithm [Frank and Wolfe, 1956; Hazan, 2008] to obtain the global solution. Our proposed algorithm only needs the computation of the largest eigenvector of a matrix per iteration and is relatively easy to be implemented by using just a few lines of MATLAB code.

5.7 Conclusion

In this chapter, we extended and developed metric learning models proposed in [Xing et al., 2003; Ying and Li, 2012]. In particular, we proposed a general framework which recovers the models in

[Xing et al., 2003; Ying and Li, 2012] as special cases. This novel framework was shown to be equivalent to a semi-definite program over the spectrahedron. This equivalence is important since it enables us to directly apply the Frank-Wolfe algorithm [Frank and Wolfe, 1956; Hazan, 2008] to obtain the optimal solution. Experiments on the UCI datasets validate the effectiveness of our proposed method and algorithm. In addition, the proposed method performs well on the Labeled Faces in the Wild (LFW) dataset for unconstrained face verification in still images.

We now discuss some possible future work. Firstly, in Section 5.5, PCA was applied to reduce the dimensionality. However, as shown in Chapter 3 that in the context of unconstrained face verification WPCA outperforms the standard PCA (see Sections 3.3 and 3.4). It would be interesting to apply WPCA to reduce the redundant noise and dimensionality for DML_p . Secondly, it would be desirable to investigate the kernelized version of DML_p using similar ideas from [Tsang et al., 2003; Jain et al., 2010]. Thirdly, metric learning can be also regarded as a dimensionality reduction method. However, in its application to face verification, a common approach is to use PCA or WPCA to reduce the dimensionality of the original descriptors. This triggers a natural question for future work on how to design effective metric learning methods to directly deal with the original descriptors of the facial images.

Previous chapters have been mainly concerned with similarity metric learning methods which formulate the problems as tractable optimization procedures. In the next chapter, we look at generalization analysis of similarity metric learning methods that few studies address.

6 Generalization Bounds for Metric and Similarity Learning

6.1 Introduction

In the previous chapters, we mainly focused on developing similarity metric learning models and designing efficient optimization algorithms for the proposed models. Indeed, as described in Sections 2.3.2 and 2.3.3, most of the studies on similarity metric learning have gone into formulating the problems as tractable optimization procedures (e.g. [Xing et al., 2003; Weinberger et al., 2006; Davis et al., 2007; Guillaumin et al., 2009; Shen et al., 2009; Chechik et al., 2010; Shalit et al., 2010; Ying and Li, 2012]). However, few studies address the generalization analysis of such methods. The recent work [Jin et al., 2009] pioneered the generalization analysis for metric learning using the concept of uniform stability [Bousquet and Elisseeff, 2002]. However, this approach only works for the strongly convex norm, e.g. the Frobenius norm, and the bias term is fixed which makes the generalization analysis essentially different.

In this chapter, we develop a novel approach for generalization analysis of metric and similarity learning which can deal with general matrix regularization terms including the Frobenius norm [Jin et al., 2009], sparse L^1-norm [Rosales and Fung, 2006], mixed $(2,1)$-norm [Ying et al., 2009] and trace-norm [Ying et al., 2009; Shen et al., 2009]. In particular, we first show that the generalization analysis for metric/similarity learning reduces to the estimation of the Rademacher average over "sums-of-i.i.d." sample-blocks related to the specific matrix norm, which we refer to as the *Rademacher complexity for metric (similarity) learning*. Then, we show how to estimate the Rademacher complexities with different matrix regularizers. Our analysis indicates that sparse metric/similarity learning with L^1-norm regularization could lead to significantly better generalization bounds than that with Frobenius norm regularization, especially when the dimensionality of the input data is high. This is nicely consistent with the rationale that sparse methods are more effective for high-dimensional data analysis. Our novel generalization analysis develops and extends Rademacher complexity analysis [Bartlett and Mendelson, 2003; Koltchinskii and Panchenko, 2002] to the setting of metric/similarity learning by using techniques of U-statistics [Clemençon et al., 2008; De la Pena and Giné, 1999].

The rest of the chapter is organized as follows. Section 6.2 reviews the models of metric/similarity learning. Section 6.3 establishes the main theorems. We derive and discuss generalization bounds for metric/similarity learning with various matrix-norm regularization terms in Sections 6.4 and 6.5 respectively. Section 6.6 concludes the chapter.

[1]This work has been accepted for *Machine Learning Journal*.

Notation: We equip the cone of positive semi-definite matrices $\mathbb{S}^d_+$ with a general matrix norm $\|\cdot\|$, which can be a Frobenius norm, trace-norm and mixed norm. Its associated dual norm is denoted, for any $M \in \mathbb{S}^d$, by $\|M\|_* = \sup\{\langle X, M\rangle : X \in \mathcal{S}^d, \|X\| \leq 1\}$.

6.2 Metric/Similarity Learning Formulation

In our learning setting, we have an input space $\mathcal{X} \subseteq \mathbb{R}^d$ and output (labels) space $\mathcal{Y}$. Denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and suppose $\mathbf{z} := \{z_i = (x_i, y_i) \in \mathcal{Z} : i \in \mathbb{N}_n\}$ an i.i.d. training set according to an unknown distribution ρ on $\mathcal{Z}$. Denote the $d \times n$ input data matrix by $\mathbf{X} = (x_i : i \in \mathbb{N}_n)$ and the $d \times d$ *distance matrix* by $M = (M_{\ell k})_{\ell,k\in\mathbb{N}_d}$. Then, the (pseudo-) distance between x_i and x_j is measured by

$$d_M(x_i, x_j) = (x_i - x_j)^\top M(x_i - x_j).$$

The bilinear similarity function is defined by

$$s_M(x_i, x_j) = x_i^\top M x_j.$$

It is worth of pointing out that we do not require the positive semi-definiteness of the matrix M throughout this chapter. However, we do assume M to be symmetric, since this will guarantee the distance/similarity between x_i and x_j is equivalent to that between x_j and x_i.

There are two main terms in the metric/similarity learning model: *empirical error* and *matrix regularization term*. The empirical error function is to employ the similarity and dissimilarity information provided by the label information and the appropriate matrix regularization term is to avoid overfitting and improve generalization performance.

For any pair of samples (x_i, x_j), let $r(y_i, y_j) = 1$ if $y_i = y_j$ otherwise $r(y_i, y_j) = -1$. It is expected that there exists a bias term $b \in \mathbb{R}$ such that $d_M(x_i, x_j) \leq b$ for $r(y_i, y_j) = 1$ and $d_M(x_i, x_j) > b$ otherwise. This naturally leads to the empirical error [Jin et al., 2009] defined by

$$\mathcal{E}_\mathbf{z}(M, b) := \frac{1}{n(n-1)} \sum_{i,j\in\mathbb{N}_n, i\neq j} I[r(y_i, y_j)(d_M(x_i, x_j) - b) > 0],$$

where the indicator function $I[x]$ equals 1 if x is true and zero otherwise.

Due to the indicator function, the above empirical error is non-differentiable and non-convex which is difficult to optimize. A usual way to overcome this shortcoming is to upper-bound it with a differentiable and convex loss function. For instance, we can use the hinge loss to upper-bound the indicator function which leads to the following empirical error:

$$\mathcal{E}_\mathbf{z}(M, b) := \frac{1}{n(n-1)} \sum_{i,j\in\mathbb{N}_n, i\neq j} [1 + r(y_i, y_j)(d_M(x_i, x_j) - b)]_+. \tag{6.1}$$

In order to avoid overfitting, we need to enforce a regularization term denoted by $\|M\|$, which will restrict the complexity of the distance matrix. We emphasize here $\|\cdot\|$ denotes a general matrix norm in the linear space $\mathbb{S}^d$. Putting the regularization term and the empirical error term together

yields the following metric learning model:

$$(M_{\mathbf{z}}, b_{\mathbf{z}}) = \arg \min_{M \in \mathcal{S}^d, b \in \mathbb{R}} \big\{ \mathcal{E}_{\mathbf{z}}(M, b) + \lambda \|M\|^2 \big\}, \tag{6.2}$$

where $\lambda > 0$ is a trade-off parameter.

Different regularization terms lead to different metric learning formulations. For instance, the Frobenius norm $\|M\|_F$ was used in [Jin et al., 2009]. To favor the element-sparsity, [Rosales and Fung, 2006] introduced the L^1-norm regularization $\|M\| = \sum_{\ell,k \in \mathbb{N}_d} |M_{\ell k}|$. [Ying et al., 2009] proposed the mixed $(2,1)$-norm $\|M\| = \sum_{\ell \in \mathbb{N}_d} \big(\sum_{k \in \mathbb{N}_d} |M_{\ell k}|^2 \big)^{\frac{1}{2}}$ to encourage the column-wise sparsity of the distance matrix. The trace-norm regularization $\|M\| = \sum_{\ell} \sigma_\ell(M)$ was also considered by [Ying et al., 2009; Shen et al., 2009]. Here, $\{\sigma_\ell : \ell \in \mathbb{N}_d\}$ denote the singular values of a matrix $M \in \mathbb{S}^d$. Since M is symmetric, the singular values of M are identical to the absolute values of its eigenvalues.

In analogy to the formulation of metric learning, we consider the following empirical error for similarity learning [Maurer, 2008; Chechik et al., 2010]:

$$\widetilde{\mathcal{E}}_{\mathbf{z}}(M, b) := \frac{1}{n(n-1)} \sum_{i,j \in \mathbb{N}_n, i \neq j} [1 - r(y_i, y_j)(s_M(x_i, x_j) - b)]_+. \tag{6.3}$$

This leads to the regularized formulation for similarity learning defined as follows:

$$(\widetilde{M}_{\mathbf{z}}, \widetilde{b}_{\mathbf{z}}) = \arg \min_{M \in \mathcal{S}^d, b \in \mathbb{R}} \big\{ \widetilde{\mathcal{E}}_{\mathbf{z}}(M, b) + \lambda \|M\|^2 \big\}. \tag{6.4}$$

The work [Maurer, 2008] used the Frobenius-norm regularization for similarity learning. The trace-norm regularization has been used by [Shalit et al., 2010] to encourage a low-rank similarity matrix M.

6.3 Statistical Generalization Analysis

In this section, we mainly give a detailed proof of generalization bounds for metric and similarity learning. In particular, we develop a novel line of generalization analysis for metric and similarity learning with general matrix regularization terms. The key observation is that the empirical data term $\mathcal{E}_{\mathbf{z}}(M, b)$ for metric learning is a modification of U-statistics and it is expected to converge to its expected form defined by

$$\mathcal{E}(M, b) = \iint (1 + r(y, y')(d_M(x, x') - b))_+ d\rho(x, y) d\rho(x', y'). \tag{6.5}$$

The empirical term $\widetilde{\mathcal{E}}_{\mathbf{z}}(M, b)$ for similarity learning is expected to converge to

$$\widetilde{\mathcal{E}}(M, b) = \iint (1 - r(y, y')(s_M(x, x') - b))_+ d\rho(x, y) d\rho(x', y'). \tag{6.6}$$

The target of generalization analysis [Vapnik, 2000; Cucker and Zhou, 2007] is to bound the true error $\mathcal{E}(\mathcal{M}_{\mathbf{z}}, b_{\mathbf{z}})$ by the empirical error $\mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}})$ for metric learning and $\widetilde{\mathcal{E}}(\widetilde{M}_{\mathbf{z}}, \widetilde{b}_{\mathbf{z}})$ by the

empirical error $\widetilde{\mathcal{E}}_{\mathbf{z}}(\widetilde{M}_{\mathbf{z}}, \widetilde{b}_{\mathbf{z}})$ for similarity learning.

In the sequel, we provide a detailed proof for generalization bounds of metric learning. Since the proof for similarity learning is exactly the same as that for metric learning, we only mention the results followed with some brief comments.

6.3.1 Bounding the Solutions

By the definition of $(M_{\mathbf{z}}, b_{\mathbf{z}})$, we know that

$$\mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}) + \lambda \|M_{\mathbf{z}}\|^2 \leq \mathcal{E}_{\mathbf{z}}(0, 0) + \lambda \|0\| = 1$$

which implies that

$$\|M_{\mathbf{z}}\| \leq \frac{1}{\sqrt{\lambda}}. \tag{6.7}$$

Now we turn our attention to derive the bound of the bias term $b_{\mathbf{z}}$ by modifying the techniques in [Chen et al., 2004] which was originally developed to estimate the offset term of the soft-margin SVM.

Lemma 6. *For any samples* $\mathbf{z}$ *and* $\lambda > 0$*, there exists a minimizer* $(M_{\mathbf{z}}, b_{\mathbf{z}})$ *of formulation (6.2) such that*

$$\min_{i \neq j} [d_{M_{\mathbf{z}}}(x_i, x_j) - b_{\mathbf{z}}] \leq 1, \quad \max_{i \neq j} [d_{M_{\mathbf{z}}}(x_i, x_j) - b_{\mathbf{z}}] \geq -1. \tag{6.8}$$

Proof. We first prove the inequality $\min_{i \neq j} [d_{M_{\mathbf{z}}}(x_i, x_j) - b_{\mathbf{z}}] \leq 1$. To this end, we first consider the special case where the training set $\mathbf{z}$ only contains two examples with distinct labels, i.e. $\mathbf{z} = \{(z_i = (x_i, y_i) : i = 1, 2,\ x_1 \neq x_2,\ y_1 \neq y_2\}$. For any $\lambda > 0$, let $(M_{\mathbf{z}}, b_{\mathbf{z}}) = (\mathbf{0}, -1)$, and observe that $\mathcal{E}_{\mathbf{z}}(\mathbf{0}, -1) + \lambda \|\mathbf{0}\|^2 = 0$. This observation implies that $(M_{\mathbf{z}}, b_{\mathbf{z}})$ is a minimizer of problem (6.2). Consequently, we have the desired result in this extreme case, since $\min_{i \neq j} [d_{M_{\mathbf{z}}}(x_i, x_j) - b_{\mathbf{z}}] = d_{M_{\mathbf{z}}}(x_1, x_2) - b_{\mathbf{z}} = 1$.

Now let us consider the general case where the training set $\mathbf{z}$ has at least two examples with the same label, i.e.

$$\{(z_i = (x_i, y_i) : i = 1, 2,\ x_1 \neq x_2, y_1 = y_2\} \subseteq \mathbf{z}.$$

In this general case, we prove the inequality $\min_{i \neq j} [d_{M_{\mathbf{z}}}(x_i, x_j) - b_{\mathbf{z}}] \leq 1$ by contradiction. Suppose that $s := \min_{i \neq j} [d_{M_{\mathbf{z}}}(x_i, x_j) - b_{\mathbf{z}}] > 1$ which equivalently implies that $d_{M_{\mathbf{z}}}(x_i, x_j) - (b_{\mathbf{z}} + s - 1) \geq 1$ for any $i \neq j$. Hence, for any pair of examples (x_i, x_j) with distinct labels, i.e. $y_i \neq y_j$ (equivalently $r(y_i, y_j) = -1$), there holds

$$\big(1 + r(y_i, y_j)(d_{M_{\mathbf{z}}}(x_i, x_j) - (b_{\mathbf{z}} + s - 1)\big)_+ = \big(1 - (d_{M_{\mathbf{z}}}(x_i, x_j) - (b_{\mathbf{z}} + s - 1)\big)_+ = 0.$$

Consequently,

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}+s-1) &= \tfrac{1}{n(n-1)}\sum_{i\neq j}\Big(1+r(i,j)(d_{M_{\mathbf{z}}}(x_i,x_j)-b_{\mathbf{z}}-s-1)\Big)_+ \\ &= \tfrac{1}{n(n-1)}\sum_{i\neq j, y_i=y_j}(1+d_{M_{\mathbf{z}}}(x_i,x_j)-b_{\mathbf{z}}-(s-1))_+ \\ &< \tfrac{1}{n(n-1)}\sum_{i\neq j, y_i=y_j}(1+d_{M_{\mathbf{z}}}(x_i,x_j)-b_{\mathbf{z}})_+ \leq \mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}).\end{aligned}$$

The above estimation implies that $\mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}+s-1)+\lambda\|M_{\mathbf{z}}\| < \mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}})+\lambda\|M_{\mathbf{z}}\|$ which contradicts the definition of the minimizer $(M_{\mathbf{z}}, b_{\mathbf{z}})$. Hence, $s = \min_{i\neq j}[d_{M_{\mathbf{z}}}(x_i,x_j)-b_{\mathbf{z}}] \leq 1$.

Secondly, we prove the inequality $\max_{i\neq j}[d_{M_{\mathbf{z}}}(x_i,x_j)-b_{\mathbf{z}}] \geq -1$ in analogy to the above argument. Consider the special case where the training set $\mathbf{z}$ contains only two examples with the same label, i.e. $\{(z_i=(x_i,y_i): i=1,2,\ x_1\neq x_2, y_1=y_2\}$. For any given $\lambda>0$, let $(M_{\mathbf{z}}, b_{\mathbf{z}})=(\mathbf{0},1)$. Since $\mathcal{E}_{\mathbf{z}}(\mathbf{0},1)+\lambda\|\mathbf{0}\|^2=0$, $(\mathbf{0},1)$ is a minimizer of problem (6.2). The desired estimation follows from the fact that $\max_{i\neq j} d_{M_{\mathbf{z}}}(x_i,x_j)-b_{\mathbf{z}} = 0-1=-1$.

Now let us consider the general case where the training set $\mathbf{z}$ has at least two examples with distinct labels, i.e.

$$\{(z_i=(x_i,y_i): i=1,2,\ x_1\neq x_2, y_1\neq y_2\}\subseteq \mathbf{z}.$$

We prove the estimation $\max_{i\neq j}[d_{M_{\mathbf{z}}}(x_i,x_j)-b_{\mathbf{z}}]\geq -1$ by contradiction. Assume $s:=\max_{i\neq j}[d_{M_{\mathbf{z}}}(x_i,x_j)-b_{\mathbf{z}}]<-1$, then $d_{M_{\mathbf{z}}}(x_i,x_j)-(b_{\mathbf{z}}+s+1)<-1$ holds for any $i\neq j$. This implies, for any pair of examples (x_i,x_j) with the same label, i.e. $r(i,j)=1$, that $\big(1+r(i,j)(d_{M_{\mathbf{z}}}(x_i,x_j)-b_{\mathbf{z}}-s-1)\big)_+ = 0$. Hence,

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}+s+1) &= \tfrac{1}{n(n-1)}\sum_{i\neq j}\Big(1+r(i,j)(d_{M_{\mathbf{z}}}(x_i,x_j)-b_{\mathbf{z}}-s-1)\Big)_+ \\ &= \tfrac{1}{n(n-1)}\sum_{i\neq j, y_i\neq y_j}\Big(1-d_{M_{\mathbf{z}}}(x_i,x_j)+b_{\mathbf{z}}+(s+1)\Big)_+ \\ &< \tfrac{1}{n(n-1)}\sum_{i\neq j, y_i\neq y_j}(1-d_{M_{\mathbf{z}}}(x_i,x_j)+b_{\mathbf{z}})_+ \leq \mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}).\end{aligned}$$

The above estimation yields that $\mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}}+s+1)+\lambda\|M_{\mathbf{z}}\|^2 < \mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}}, b_{\mathbf{z}})+\lambda\|M_{\mathbf{z}}\|^2$ which contradicts the definition of the minimizer $(M_{\mathbf{z}}, b_{\mathbf{z}})$. Hence, we have the desired inequality $\max_{i\neq j}[d_{M_{\mathbf{z}}}(x_i,x_j)-b_{\mathbf{z}}]\geq -1$ which completes the proof of the lemma. □

Corollary 7. *For any samples* $\mathbf{z}$ *and* $\lambda>0$*, there exists a minimizer* $(M_{\mathbf{z}}, b_{\mathbf{z}})$ *of formulation (6.2) such that*

$$|b_{\mathbf{z}}| \leq 1+\big(\max_{i\neq j}\|X_{ij}\|_*\big)\|M_{\mathbf{z}}\|. \tag{6.9}$$

Proof. From inequality (6.8) in Lemma 6, we see that $-b_{\mathbf{z}}+\min_{i\neq j}[d_{M_{\mathbf{z}}}(x_i,x_j)]\leq 1$ and $\max_{i\neq j}[d_{M_{\mathbf{z}}}(x_i,x_j)]\geq b_{\mathbf{z}}-1$. Equivalently, this implies that $-b_{\mathbf{z}}\leq 1-\min_{i\neq j}[d_{M_{\mathbf{z}}}(x_i,x_j)]$ and $b_{\mathbf{z}}\leq 1+\max_{i\neq j}[d_{M_{\mathbf{z}}}(x_i,x_j)]$. Recall that $X_{ij}=(x_i-x_j)(x_i-x_j)^\top$ and observe, by the definition of the dual norm $\|\cdot\|_*$, that

$$d_M(x_i,x_j)=\langle X_{ij}, M\rangle \leq \|X_{ij}\|_*\|M\|.$$

Combining this observation with the above estimates, we have that $-b_{\mathbf{z}} \leq 1+\big(\max_{i\neq j}\|X_{ij}\|_*\big)\|M_{\mathbf{z}}\|$ and $b_{\mathbf{z}} \leq 1+\big(\max_{i\neq j}\|X_{ij}\|_*\big)\|M_{\mathbf{z}}\|$, which yields the desired result. □

Denote

$$\mathcal{F} = \Big\{(M,b) : \|M\| \leq 1/\sqrt{\lambda},\ \ |b| \leq 1 + X_*\|M\|\Big\}, \tag{6.10}$$

where

$$X_* = \max_{x,x'\in\mathcal{X}} \|(x-x')(x-x')^\top\|_*.$$

From the above corollary, for any samples $\mathbf{z}$ we can easily see that least one optimal solution $(M_{\mathbf{z}}, b_{\mathbf{z}})$ of formulation (6.2) belongs to the bounded set $\mathcal{F} \subseteq \mathcal{S}^d \times \mathbb{R}$.

We end this subsection with two remarks. Firstly, from the proof of Lemma 6 and Corollary 7, we can easily see that, if the set of training samples contains at least two examples with distinct labels and two examples with the same label, all minimizers of formulation (6.2) satisfy inequality (6.8) and inequality (6.9). Hence, in this case all minimizers $(M_{\mathbf{z}}, b_{\mathbf{z}})$ of formulation (6.2) belong to the bounded set $\mathcal{F}$. Consequently, we assume, without loss of generality, that any minimizer $(M_{\mathbf{z}}, b_{\mathbf{z}})$ of formulation (6.2) satisfies inequality (6.9) and belongs to the set $\mathcal{F}$. Secondly, our formulation (6.2) for metric learning focused on the hinge loss which is widely used in the community of metric learning, see e.g [Jin et al., 2009; Weinberger and Saul, 2008; Ying and Li, 2012]. Similar results to those in the above corollary can easily be obtained for q-norm loss given, for any $x \in \mathbb{R}$, by $(1-x)_+^q$ with $q > 1$. However, the question of how to estimate the term b for general loss functions remians open.

6.3.2 Generalization Bounds

Before stating the generalization bounds, we introduce some notations. For any $z = (x, y), z' = (x', y') \in \mathcal{Z}$, let $\Phi_{M,b}(z, z') = (1 + r(y, y')(d_M(x, x') - b))_+$. Hence, for any $(M, b) \in \mathcal{F}$,

$$\sup_{z,z'}\sup_{(M,b)\in\mathcal{F}} \Phi_{M,b}(z,z') \leq B_\lambda := 2\big(1 + X_*/\sqrt{\lambda}\big). \tag{6.11}$$

Let $\lfloor \frac{n}{2} \rfloor$ denote the largest integer less than $\frac{n}{2}$ and recall the definition that $X_{ij} = (x_i - x_j)(x_i - x_j)^\top$. We now define Rademacher average over sums-of-i.i.d. sample-blocks related to the dual matrix norm $\|\cdot\|_*$ by

$$\widehat{R}_n = \frac{1}{\lfloor \frac{n}{2} \rfloor}\mathbb{E}_\sigma\Big\|\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor}\sigma_i X_{i(\lfloor \frac{n}{2} \rfloor + i)}\Big\|_*, \tag{6.12}$$

and its expectation is denoted by $R_n = \mathbb{E}_{\mathbf{z}}\big[\widehat{R}_n\big]$. Our main theorem below shows that the generalization bounds for metric learning critically depend on the quantity of R_n. For this reason, we refer to R_n as the *Radmemcher complexity for metric learning*. It is worth mentioning that metric learning formulation (6.2) depends on the norm $\|\cdot\|$ of the linear space $\mathcal{S}^d$ and the Rademacher complexity R_n is related to its dual norm $\|\cdot\|_*$.

Below, we assemble some facts that are used to establish generalization bounds for metric/similarity learning.

Definition 8. *We say the function* $f : \prod_{k=1}^{n} \Omega_k \to \mathbb{R}$ *with bounded differences* $\{c_k\}_{k=1}^n$ *if, for all* $1 \leq k \leq n$,

$$\max_{z_1,\cdots,z_k,z_k'\cdots,z_n} |f(z_1,\cdots,z_{k-1},z_k,z_{k+1},\cdots,z_n) - f(z_1,\cdots,z_{k-1},z_k',z_{k+1},\cdots,z_n)| \leq c_k$$

Lemma 9. *(McDiarmid's inequality [McDiarmid, 1989]) Suppose* $f : \prod_{k=1}^{n} \Omega_k \to \mathbb{R}$ *with bounded differences* $\{c_k\}_{k=1}^n$ *then , for all* $\epsilon > 0$*, there holds*

$$\mathbf{Pr}_{\mathbf{z}}\Big\{ f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} f(\mathbf{z}) \geq \epsilon \Big\} \leq e^{-\frac{2\epsilon^2}{\sum_{k=1}^n c_k^2}}.$$

Finally we list a useful property for U-statistics. Given the i.i.d. random variables $z_1, z_2, \ldots, z_n \in \mathcal{Z}$, let $q : Z \times Z \to \mathbb{R}$ be a symmetric real-valued function. Denote a U-statistic of order two by $U_n = \frac{1}{n(n-1)} \sum_{i \neq j} q(z_i, z_j)$. Then, the U-statistic U_n can be expressed as

$$U_n = \frac{1}{n!} \sum_{\pi} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} q(z_{\pi(i)}, z_{\pi(\lfloor \frac{n}{2} \rfloor + i)}), \tag{6.13}$$

where the sum is taken over all permutations π of $\{1, 2, \ldots, n\}$ ([Clemençon et al., 2008]). The main idea underlying this representation is to reduce the analysis to the ordinary case of i.i.d. random variable blocks.

Based on the above representation, we can prove the following lemma which plays a critical role in deriving generalization bounds for metric learning. For completeness, we include a proof here. For more details on U-statistics, one is referred to [Clemençon et al., 2008; De la Pena and Giné, 1999].

Lemma 10. *Let* $q_\tau : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ *be real-valued functions indexed by* $\tau \in \mathcal{T}$ *where* $\mathcal{T}$ *is some index set. If* $z_1, \ldots, z_n$ *are i.i.d. then we have that*

$$\mathbb{E}\Big[\sup_{\tau \in \mathcal{T}} \frac{1}{n(n-1)} \sum_{i \neq j} q_\tau(z_i, z_j)\Big] \leq \mathbb{E}\Big[\sup_{\tau \in \mathcal{T}} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} q_\tau(z_i, z_{\lfloor \frac{n}{2} \rfloor + i})\Big].$$

Proof. From the representation of U-statistics (6.13), we observe that

$$\begin{aligned}\mathbb{E}\Big[\sup_{\tau\in\mathcal{T}}\frac{1}{n(n-1)}\sum_{i\neq j}q_\tau(z_i,z_j)\Big] &= \mathbb{E}\sup_\tau\frac{1}{n!}\sum_\pi\frac{1}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}q_\tau\big(z_{\pi(i)},z_{\pi(\lfloor\frac{n}{2}\rfloor+i)}\big)\\ &\leq \tfrac{1}{n!}\mathbb{E}\sum_\pi\sup_\tau\frac{1}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}q_\tau\big(z_{\pi(i)},z_{\pi(\lfloor\frac{n}{2}\rfloor+i)}\big)\\ &= \tfrac{1}{n!}\sum_\pi\mathbb{E}\sup_\tau\frac{1}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}q_\tau\big(z_{\pi(i)},z_{\pi(\lfloor\frac{n}{2}\rfloor+i)}\big)\\ &= \mathbb{E}\Big[\sup_{\tau\in\mathcal{T}}\frac{1}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}q_\tau(z_i,z_{\lfloor\frac{n}{2}\rfloor+i})\Big].\end{aligned}$$

This completes the proof of the lemma. □

We need the following contraction property of the Rademacher averages which is essentially implied by Theorem 4.12 in Ledoux and Talagrand [Ledoux and Talagrand, 1991], see also [Bartlett and Mendelson, 2003; Koltchinskii and Panchenko, 2002].

Lemma 11. *Let F be a class of uniformly bounded real-valued functions on (Ω,μ) and $m\in\mathbb{N}$. If for each $i\in\{1,\ldots,m\}$, $\Psi_i:\mathbb{R}\to\mathbb{R}$ is a function with $\Psi_i(0)=0$ having a Lipschitz constant c_i, then for any $\{x_i\}_{i=1}^m$,*

$$\mathbb{E}_\epsilon\Big(\sup_{f\in F}\big|\sum_{i=1}^m\epsilon_i\Psi_i(f(x_i))\big|\Big)\leq 2\mathbb{E}_\epsilon\Big(\sup_{f\in F}\Big|\sum_{i=1}^m c_i\epsilon_i f(x_i)\Big|\Big). \tag{6.14}$$

Now, we are ready to derive the generalization bounds for metric/similarity learning.

Theorem 12. *Let $(M_{\mathbf{z}},b_{\mathbf{z}})$ be the solution of formulation (6.2). Then, for any $0<\delta<1$, with probability $1-\delta$ we have that*

$$\begin{aligned}\mathcal{E}(M_{\mathbf{z}},b_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}},b_{\mathbf{z}}) &\leq \sup_{(M,b)\in\mathcal{F}}\Big[\mathcal{E}(M,b)-\mathcal{E}_{\mathbf{z}}(M,b)\Big]\\ &\leq \tfrac{4R_n}{\sqrt{\lambda}}+\tfrac{4(3+2X_*/\sqrt{\lambda})}{\sqrt{n}}+2\big(1+X_*/\sqrt{\lambda}\big)\left(\frac{2\ln\left(\frac{1}{\delta}\right)}{n}\right)^{\frac{1}{2}}.\end{aligned} \tag{6.15}$$

Proof. The proof of the theorem can be divided into three steps as follows.

Step 1: Let $\mathbb{E}_{\mathbf{z}}$ denote the expectation with respect to samples $\mathbf{z}$. Observe that $\mathcal{E}(M_{\mathbf{z}},b_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}},b_{\mathbf{z}})\leq\sup_{(M,b)\in\mathcal{F}}\Big[\mathcal{E}(M,b)-\mathcal{E}_{\mathbf{z}}(M,b)\Big]$. For any $z=(z_1,\ldots,z_{k-1},z_k,z_{k+1},\ldots,z_n)$ and $z'=(z_1,\ldots,z_{k-1},z'_k,z_{k+1},\ldots,z_n)$ we know from inequality (6.11) that

$$\begin{aligned}&\Big|\sup_{(M,b)\in\mathcal{F}}\Big[\mathcal{E}(M,b)-\mathcal{E}_{\mathbf{z}}(M,b)\Big]-\sup_{(M,b)\in\mathcal{F}}\Big[\mathcal{E}(M,b)-\mathcal{E}_{\mathbf{z}'}(M,b)\Big]\Big|\\ &\leq\sup_{(M,b)\in\mathcal{F}}|\mathcal{E}_{\mathbf{z}}(M,b)-\mathcal{E}_{\mathbf{z}'}(M,b)|\\ &=\tfrac{1}{n(n-1)}\sup_{(M,b)\in\mathcal{F}}\sum_{j\in\mathbb{N}_n,j\neq k}|\Phi_{M,b}(z_k,z_j)-\Phi_{M,b}(z'_k,z_j)|\\ &\leq\tfrac{1}{n(n-1)}\sup_{(M,b)\in\mathcal{F}}\sum_{j\in\mathbb{N}_n,j\neq k}|\Phi_{M,b}(z_k,z_j)|+|\Phi_{M,b}(z'_k,z_j)|\\ &\leq 4\big(1+X_*/\sqrt{\lambda}\big)/n.\end{aligned}$$

Applying McDiarmid's inequality [McDiarmid, 1989] (Lemma 9) to the term $\sup_{(M,b)\in\mathcal{F}}\Big[\mathcal{E}(M,b) - \mathcal{E}_{\mathbf{z}}(M,b)\Big]$, with probability $1-\delta$ there holds

$$\begin{aligned}\sup_{(M,b)\in\mathcal{F}}\Big[\mathcal{E}(M,b) - \mathcal{E}_{\mathbf{z}}(M,b)\Big] \quad &\leq \mathbb{E}_{\mathbf{z}}\sup_{(M,b)\in\mathcal{F}}\Big[\mathcal{E}(M,b) - \mathcal{E}_{\mathbf{z}}(M,b)\Big] \\ &\quad + 2\big(1+X_*/\sqrt{\lambda}\big)\left(\frac{2\ln\left(\frac{1}{\delta}\right)}{n}\right)^{\frac{1}{2}}.\end{aligned} \tag{6.16}$$

Now we only need to estimate the first term in the expectation form on the right-hand side of the above equation by symmetrization techniques.

Step 2: To estimate $\mathbb{E}_{\mathbf{z}}\sup_{(M,b)\in\mathcal{F}}\Big[\mathcal{E}(M,b) - \mathcal{E}_{\mathbf{z}}(M,b)\Big]$, applying Lemma 10 with $q_{(M,b)}(z_i,z_j) = \mathcal{E}(M,b) - (1+r(y_i,y_j)(d_M(x_i,x_j)-b))_+$ implies that

$$\mathbb{E}_{\mathbf{z}}\sup_{(M,b)\in\mathcal{F}}\Big[\mathcal{E}(M,b) - \mathcal{E}_{\mathbf{z}}(M,b)\Big] \leq \mathbb{E}_{\mathbf{z}}\sup_{(M,b)\in\mathcal{F}}\Big[\mathcal{E}(M,b) - \overline{\mathcal{E}}_{\mathbf{z}}(M,b)\Big], \tag{6.17}$$

where $\overline{\mathcal{E}}_{\mathbf{z}}(M,b) = \frac{1}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\Phi_{M,b}(z_i,z_{\lfloor\frac{n}{2}\rfloor+i})$. Now let $\bar{\mathbf{z}} = \{\bar{z}_1,\bar{z}_2,\ldots,\bar{z}_n\}$ be i.i.d. samples which are independent of $\mathbf{z}$, then

$$\begin{aligned}\mathbb{E}_{\mathbf{z}}\sup_{(M,b)\in\mathcal{F}}\Big[\mathcal{E}(M,b) - \overline{\mathcal{E}}_{\mathbf{z}}(M,b)\Big] \quad &= \mathbb{E}_{\mathbf{z}}\sup_{(M,b)\in\mathcal{F}}\Big[\mathbb{E}_{\bar{\mathbf{z}}}\big[\,\overline{\mathcal{E}}_{\bar{\mathbf{z}}}(M,b)\big] - \overline{\mathcal{E}}_{\mathbf{z}}(M,b)\Big] \\ &\leq \mathbb{E}_{\mathbf{z},\bar{\mathbf{z}}}\sup_{(M,b)\in\mathcal{F}}\Big[\,\overline{\mathcal{E}}_{\bar{\mathbf{z}}}(M,b) - \overline{\mathcal{E}}_{\mathbf{z}}(M,b)\Big]\end{aligned} \tag{6.18}$$

By standard symmetrization techniques (see e.g. [Bartlett and Mendelson, 2003]), for i.i.d. Rademacher variables $\{\sigma_i\in\{\pm1\}: i\in\mathbb{N}_{\lfloor\frac{n}{2}\rfloor}\}$, we have that

$$\begin{aligned}&\mathbb{E}_{\mathbf{z},\bar{\mathbf{z}}}\sup_{(M,b)\in\mathcal{F}}\Big[\overline{\mathcal{E}}_{\bar{\mathbf{z}}}(M,b) - \overline{\mathcal{E}}_{\mathbf{z}}(M,b)\Big] \\ &= \mathbb{E}_{\mathbf{z},\bar{\mathbf{z}}}\frac{1}{\lfloor\frac{n}{2}\rfloor}\sup_{(M,b)\in\mathcal{F}}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i\Big[\Phi_{M,b}(\bar{z}_i,\bar{z}_{\lfloor\frac{n}{2}\rfloor+i}) - \Phi_{M,b}(z_i,z_{\lfloor\frac{n}{2}\rfloor+i})\Big] \\ &= 2\mathbb{E}_{\mathbf{z},\sigma}\frac{1}{\lfloor\frac{n}{2}\rfloor}\sup_{(M,b)\in\mathcal{F}}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i\Phi_{M,b}(\bar{z}_i,\bar{z}_{\lfloor\frac{n}{2}\rfloor+i}) \\ &\leq 2\mathbb{E}_{\mathbf{z},\sigma}\frac{1}{\lfloor\frac{n}{2}\rfloor}\sup_{(M,b)\in\mathcal{F}}\Big|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i\Phi_{M,b}(\bar{z}_i,\bar{z}_{\lfloor\frac{n}{2}\rfloor+i})\Big|.\end{aligned} \tag{6.19}$$

Applying the contraction property of Rademacher averages (Lemma 11) with $\Psi_i(t) = \big(1 + r(y_i,y_{\lfloor\frac{n}{2}\rfloor+i})t\big)_+ - 1$, we have the following estimation for the last term on the righthand side

of the above inequality:

$$\begin{aligned}
&\mathbb{E}_\sigma \tfrac{1}{\lfloor \frac{n}{2} \rfloor} \sup_{(M,b)\in\mathcal{F}} \Big| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \Phi_{M,b}(\bar{z}_i, \bar{z}_{\lfloor \frac{n}{2} \rfloor + i}) \Big| \\
&\leq \mathbb{E}_\sigma \tfrac{1}{\lfloor \frac{n}{2} \rfloor} \sup_{(M,b)\in\mathcal{F}} \Big| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (\Phi_{M,b}(\bar{z}_i, \bar{z}_{\lfloor \frac{n}{2} \rfloor + i}) - 1) \Big| + \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_\sigma \Big| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \Big| \\
&\leq \tfrac{2}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_\sigma \sup_{(M,b)\in\mathcal{F}} \Big| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \big(d_M(x_i, x_{\lfloor \frac{n}{2} \rfloor + i}) - b \big) \Big| + \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_\sigma \Big| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \Big| \\
&\leq \tfrac{2}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_\sigma \sup_{\|M\| \leq \frac{1}{\sqrt{\lambda}}} \Big| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i d_M(x_i, x_{\lfloor \frac{n}{2} \rfloor + i}) \Big| + \frac{(3 + 2X_*/\sqrt{\lambda})}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_\sigma \Big| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \Big|
\end{aligned} \tag{6.20}$$

Step 3 : It remains to estimate the terms on the righthand side of inequality (6.20). To this end, observe that

$$\mathbb{E}_\sigma \Big| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \Big| \leq \Big(\mathbb{E}_\sigma \Big| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \Big|^2 \Big)^{\frac{1}{2}} \leq \sqrt{\lfloor \frac{n}{2} \rfloor}.$$

Moreover,

$$\begin{aligned}
\mathbb{E}_\sigma \sup_{\|M\| \leq \frac{1}{\sqrt{\lambda}}} \Big| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i d_M(x_i, x_{\lfloor \frac{n}{2} \rfloor + i}) \Big| &= \mathbb{E}_\sigma \sup_{\|M\| \leq \frac{1}{\sqrt{\lambda}}} \Big| \langle \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (x_i - x_{\lfloor \frac{n}{2} \rfloor + i})(x_i - x_{\lfloor \frac{n}{2} \rfloor + i})^\top, M \rangle \Big| \\
&\leq \tfrac{1}{\sqrt{\lambda}} \mathbb{E}_\sigma \Big\| \textstyle\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i X_{i(\lfloor \frac{n}{2} \rfloor + i)} \Big\|_*.
\end{aligned}$$

Putting the above estimations and inequalities (6.19), (6.20) together yields that

$$\mathbb{E}_{\mathbf{z},\bar{\mathbf{z}}} \sup_{(M,b)\in\mathcal{F}} \Big[\overline{\mathcal{E}}_{\bar{\mathbf{z}}}(M,b) - \overline{\mathcal{E}}_{\mathbf{z}}(M,b) \Big] \leq \frac{2(3 + 2X_*/\sqrt{\lambda})}{\sqrt{\lfloor \frac{n}{2} \rfloor}} + \frac{4R_n}{\sqrt{\lambda}} \leq \frac{4(3 + X_*/\sqrt{\lambda})}{\sqrt{n}} + \frac{2R_n}{\sqrt{\lambda}}.$$

Consequently, combining this with inequalities (6.17), (6.18) implies that

$$\mathbb{E}_{\mathbf{z}} \sup_{(M,b)\in\mathcal{F}} \Big[\mathcal{E}(M,b) - \mathcal{E}_{\mathbf{z}}(M,b) \Big] \leq \frac{4(3 + 2X_*/\sqrt{\lambda})}{\sqrt{n}} + \frac{4R_n}{\sqrt{\lambda}}.$$

Putting this estimation with (6.16) completes the proof the theorem. □

In the setting of similarity learning, X_* and R_n are replaced by

$$\widetilde{X}_* = \sup_{x,t\in\mathcal{X}} \|x t^\top\|_* \quad \text{and} \quad \widetilde{R}_n = \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\mathbf{z}} \mathbb{E}_\sigma \Big\| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \widetilde{X}_{i(\lfloor \frac{n}{2} \rfloor + i)} \Big\|_*, \tag{6.21}$$

where $\widetilde{X}_{i(\lfloor \frac{n}{2} \rfloor + i)} = x_i x_{\lfloor \frac{n}{2} \rfloor + i}^\top$. Let $\widetilde{\mathcal{F}} = \Big\{ (M,b) : \|M\| \leq 1/\sqrt{\lambda}, \;\; |b| \leq 1 + \widetilde{X}_* \|M\| \Big\}$. Using the exactly same argument as above, we can prove the following bound for similarity learning formulation (6.4).

Theorem 13. *Let $(\widetilde{M}_{\mathbf{z}}, \widetilde{b}_{\mathbf{z}})$ be the solution of formulation (6.4). Then, for any $0 < \delta < 1$, with*

probability $1-\delta$ *we have that*

$$\begin{aligned}\widetilde{\mathcal{E}}(\widetilde{M}_{\mathbf{z}},\widetilde{b}_{\mathbf{z}})-\widetilde{\mathcal{E}}_{\mathbf{z}}(\widetilde{M}_{\mathbf{z}},\widetilde{b}_{\mathbf{z}}) &\leq \sup_{(M,b)\in\widetilde{\mathcal{F}}}\Big[\widetilde{\mathcal{E}}(M,b)-\widetilde{\mathcal{E}}_{\mathbf{z}}(M,b)\Big]\\ &\leq \tfrac{4\widetilde{R}_n}{\sqrt{\lambda}}+\tfrac{4(3+2\widetilde{X}_*/\sqrt{\lambda})}{\sqrt{n}}+2\big(1+\widetilde{X}_*/\sqrt{\lambda}\big)\left(\tfrac{2\ln\left(\frac{1}{\delta}\right)}{n}\right)^{\frac{1}{2}}.\end{aligned} \tag{6.22}$$

6.4 Estimation of R_n

From Theorem 12, we need to estimate the Rademacher average for metric learning, i.e. R_n, and the quantity X_* for different matrix regularization terms. We first recall another property of Rademacher averages, which is the Khinchin-Kahne inequality (see e.g. [De la Pena and Giné, 1999, Theorem 1.3.1]).

Lemma 14. *For* $n\in\mathbb{N}$*, let* $\{f_i\in\mathbb{R}: i\in\mathbb{N}_n\}$*, and* $\{\sigma_i: i\in\mathbb{N}_n\}$ *be a family of i.i.d. Rademacher variables. Then, for any* $1<p<q<\infty$ *we have*

$$\left(\mathbb{E}_\sigma\big|\sum_{i\in\mathbb{N}_n}\sigma_i f_i\big|^q\right)^{\frac{1}{q}}\leq\left(\frac{q-1}{p-1}\right)^{\frac{1}{2}}\left(\mathbb{E}_\sigma\big|\sum_{i\in\mathbb{N}_n}\sigma_i f_i\big|^p\right)^{\frac{1}{p}}.$$

Now we can estimate R_n. Without loss of generality, we only focus on popular matrix norms such as the Frobenius norm [Jin et al., 2009], L^1-norm [Rosales and Fung, 2006], trace-norm [Ying et al., 2009; Shen et al., 2009] and mixed $(2,1)$-norm [Ying et al., 2009].

Example 1 (Frobenius norm)**.** *Let the matrix norm be the Frobenius norm i.e.* $\|M\|=\|M\|_F$*, then the quantity* $X_*=\sup_{x,x\in\mathcal{X}}\|x-x'\|_F^2$ *and the Rademacher complexity is estimated as follows:*

$$R_n\leq\frac{2X_*}{\sqrt{n}}=\frac{2\sup_{x,x'\in\mathcal{X}}\|x-x'\|_F^2}{\sqrt{n}}.$$

Let $(M_{\mathbf{z}},b_{\mathbf{z}})$ *be a solution of formulation (6.2) with Frobenius norm regularization. For any* $0<\delta<1$*, with probability* $1-\delta$ *there holds*

$$\begin{aligned}\mathcal{E}(M_{\mathbf{z}},b_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}},b_{\mathbf{z}}) &\leq 2\Big(1+\tfrac{\sup_{x,x\in\mathcal{X}}\|x-x'\|_F^2}{\sqrt{\lambda}}\Big)\sqrt{\tfrac{2\ln\left(\frac{1}{\delta}\right)}{n}}\\ &\quad+\tfrac{16\sup_{x,x'\in\mathcal{X}}\|x-x'\|_F^2}{\sqrt{n\lambda}}+\tfrac{12}{\sqrt{n}}.\end{aligned} \tag{6.23}$$

Proof. Note that the dual norm of the Frobenius norm is itself. The estimation of X_* is straightforward. The Rademacher complexity R_n is estimated as follows:

$$\begin{aligned}R_n &= \tfrac{1}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}\left(\sum_{i,j=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i\sigma_j\langle x_i-x_{\lfloor\frac{n}{2}\rfloor+i},x_j-x_{\lfloor\frac{n}{2}\rfloor+j}\rangle^2\right)^{\frac{1}{2}}\\ &\leq \tfrac{1}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}_{\mathbf{z}}\left(\mathbb{E}_\sigma\sum_{i,j=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i\sigma_j\langle x_i-x_{\lfloor\frac{n}{2}\rfloor+i},x_j-x_{\lfloor\frac{n}{2}\rfloor+j}\rangle^2\right)^{\frac{1}{2}}\\ &= \tfrac{1}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}_{\mathbf{z}}\left(\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\|x_i-x_{\lfloor\frac{n}{2}\rfloor+i}\|_F^4\right)^{\frac{1}{2}}\\ &\leq X_*/\sqrt{\lfloor\tfrac{n}{2}\rfloor}\leq\tfrac{2X_*}{\sqrt{n}}.\end{aligned}$$

Putting the above estimation back into equation (6.15) completes the proof of Example 1. □

Other popular matrix norms for metric learning are the L^1-norm, trace-norm and mixed $(2,1)$-norm. The dual norms are respectively L^∞-norm, spectral norm (i.e. the maximum of singular values) and mixed $(2,\infty)$-norm. All these dual norms mentioned above are less than the Frobenius norm. Hence, the following estimation always holds true for all the norms mentioned above:

$$X_* \leq \sup_{x,x\in\mathcal{X}} \|x-x'\|_F^2, \quad \text{and} \quad R_n \leq \frac{2\sup_{x,x'\in\mathcal{X}}\|x-x'\|_F^2}{\sqrt{n}}.$$

Consequently, the generalization bound (6.23) holds true for metric learning formulation (6.2) with L^1-norm, or trace-norm or mixed $(2,1)$-norm regularization. However, in some cases, the above upper-bounds are too conservative. For instance, in the following examples we can show that more refined estimation of R_n can be obtained by applying the Khinchin inequalities for Rademacher averages [De la Pena and Giné, 1999].

Example 2 (Sparse L^1-norm)**.** *Let the matrix norm be the L^1-norm (i.e. $\|M\| = \sum_{\ell,k\in\mathbb{N}_d}|M_{\ell k}|$). Then, $X_* = \sup_{x,x'\in\mathcal{X}}\|x-x'\|_\infty^2$ and*

$$R_n \leq 4\sup_{x,x'\in\mathcal{X}}\|x-x'\|_\infty^2\sqrt{\frac{e\log d}{n}}.$$

Let $(M_{\mathbf{z}}, b_{\mathbf{z}})$ be a solution of formulation (6.2) with L^1-norm regularization. For any $0<\delta<1$, with probability $1-\delta$ there holds

$$\begin{aligned}\mathcal{E}(M_{\mathbf{z}},b_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}},b_{\mathbf{z}}) \quad &\leq 2\Big(1+\frac{\sup_{x,x\in\mathcal{X}}\|x-x'\|_\infty^2}{\sqrt{\lambda}}\Big)\sqrt{\frac{2\ln\left(\frac{1}{\delta}\right)}{n}} \\ &\quad + \frac{8\sup_{x,x'\in\mathcal{X}}\|x-x'\|_\infty^2(1+2\sqrt{e\log d})}{\sqrt{n\lambda}} + \frac{12}{\sqrt{n}}.\end{aligned} \tag{6.24}$$

Proof. The dual norm of the L^1-norm is the L^∞-norm. Hence, $X_* = \sup_{x,x'\in\mathcal{X}}\|x-x'\|_\infty^2$. To estimate R_n, we observe, for any $1<q<\infty$, that

$$\begin{aligned} R_n &= \tfrac{1}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}_{\mathbf{z}}\mathbb{E}_\sigma\Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i X_{i(\lfloor\frac{n}{2}\rfloor+i)}\Big\|_\infty \leq \tfrac{1}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}_{\mathbf{z}}\mathbb{E}_\sigma\Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i X_{i(\lfloor\frac{n}{2}\rfloor+i)}\Big\|_q \\ &:= \tfrac{1}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}_{\mathbf{z}}\mathbb{E}_\sigma\left(\sum_{\ell,k\in\mathbb{N}_d}\Big|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i(x_i^k - x^k_{\lfloor\frac{n}{2}\rfloor+i})(x_i^\ell - x^\ell_{\lfloor\frac{n}{2}\rfloor+i})\Big|^q\right)^{\frac{1}{q}} \\ &\leq \tfrac{1}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}_{\mathbf{z}}\left(\sum_{\ell,k\in\mathbb{N}_d}\mathbb{E}_\sigma\Big|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i(x_i^k - x^k_{\lfloor\frac{n}{2}\rfloor+i})(x_i^\ell - x^\ell_{\lfloor\frac{n}{2}\rfloor+i})\Big|^q\right)^{\frac{1}{q}} \end{aligned} \tag{6.25}$$

where x_i^k represents the k-th coordinate element of vector $x_i\in\mathbb{R}^d$. To estimate the term on the right-hand side of inequality (6.25), we apply the Khinchin-Kahane inequality (Lemma 14) with $p=2<q<\infty$ yields that

$$\begin{aligned} &\mathbb{E}_\sigma\Big|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i(x_i^k - x^k_{\lfloor\frac{n}{2}\rfloor+i})(x_i^\ell - x^\ell_{\lfloor\frac{n}{2}\rfloor+i})\Big|^q \\ &\leq q^{\frac{q}{2}}\Big(\mathbb{E}_\sigma\Big|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i(x_i^k - x^k_{\lfloor\frac{n}{2}\rfloor+i})(x_i^\ell - x^\ell_{\lfloor\frac{n}{2}\rfloor+i})\Big|^2\Big)^{\frac{q}{2}} \\ &= q^{\frac{q}{2}}\Big(\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}(x_i^k - x^k_{\lfloor\frac{n}{2}\rfloor+i})^2(x_i^\ell - x^\ell_{\lfloor\frac{n}{2}\rfloor+i})^2\Big)^{\frac{q}{2}} \\ &\leq \max_{x,x'\in\mathcal{X}}\|x-x'\|_\infty^{2q}(\lfloor\tfrac{n}{2}\rfloor)^{\frac{q}{2}}q^{\frac{q}{2}}. \end{aligned} \tag{6.26}$$

Putting the above estimation back into (6.25) and letting $q = 4\log d$ implies that

$$\begin{aligned} R_n &\leq \max_{x,x'\in\mathcal{X}} \|x-x'\|_\infty^2 d^{\frac{2}{q}}\sqrt{q}/\sqrt{\lfloor\frac{n}{2}\rfloor} = 2\sup_{x,x'\in\mathcal{X}} \|x-x'\|_\infty^2\sqrt{e\log d/\lfloor\frac{n}{2}\rfloor} \\ &\leq 4\sup_{x,x'\in\mathcal{X}} \|x-x'\|_\infty^2\sqrt{e\log d/n}. \end{aligned}$$

Putting the estimation for X_* and R_n into Theorem 6.15 yields inequality (6.24). This completes the proof of Example 2. □

Example 3 (Mixed $(2,1)$-norm)**.** *Consider* $\|M\| = \sum_{\ell\in\mathbb{N}_d}\sqrt{\sum_{k\in\mathbb{N}_d}|M_{\ell k}|^2}$. *Then, we have* $X_* = \big[\sup_{x,x'\in\mathcal{X}}\|x-x'\|_F\big]\big[\sup_{x,x'\in\mathcal{X}}\|x-x'\|_\infty\big]$, *and*

$$R_n \leq 4\big[\sup_{x,x'\in\mathcal{X}}\|x-x'\|_\infty\big]\big[\sup_{x,x'\in\mathcal{X}}\|x-x'\|_F\big]\sqrt{\frac{e\log d}{n}}.$$

Let $(M_{\mathbf{z}}, b_{\mathbf{z}})$ *be a solution of formulation (6.2) with mixed* $(2,1)$*-norm. For any* $0<\delta<1$*, with probability* $1-\delta$ *there holds*

$$\begin{aligned} \mathcal{E}(M_{\mathbf{z}},b_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(M_{\mathbf{z}},b_{\mathbf{z}}) &\leq 2\Big(1+\frac{\big[\sup_{x,x'\in\mathcal{X}}\|x-x'\|_\infty\big]\big[\sup_{x,x'\in\mathcal{X}}\|x-x'\|_F\big]}{\sqrt{\lambda}}\Big)\sqrt{\frac{2\ln\left(\frac{1}{\delta}\right)}{n}} \\ &\quad + \frac{8\big[\sup_{x,x'\in\mathcal{X}}\|x-x'\|_\infty\big]\big[\sup_{x,x'\in\mathcal{X}}\|x-x'\|_F\big](1+2\sqrt{e\log d})}{\sqrt{n\lambda}} + \frac{12}{\sqrt{n}}. \end{aligned} \tag{6.27}$$

Proof. The estimation of X_* is straightforward and we estimate R_n as follows. For any $q>1$, there holds

$$\begin{aligned} R_n &= \tfrac{1}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}_{\mathbf{z}}\mathbb{E}_\sigma\Big\|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i X_{i(\lfloor\frac{n}{2}\rfloor+i)}\Big\|_{(2,\infty)} \\ &= \tfrac{1}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}_{\mathbf{z}}\mathbb{E}_\sigma\sup_{\ell\in\mathbb{N}_d}\Big(\sum_{k\in\mathbb{N}_d}\big|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i(x_i^k - x^k_{\lfloor\frac{n}{2}\rfloor+i})(x_i^\ell - x^\ell_{\lfloor\frac{n}{2}\rfloor+i})\big|^2\Big)^{\frac{1}{2}} \\ &\leq \tfrac{1}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}_{\mathbf{z}}\Big(\sum_{k\in\mathbb{N}_d}\mathbb{E}_\sigma\sup_{\ell\in\mathbb{N}_d}\big|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i(x_i^k - x^k_{\lfloor\frac{n}{2}\rfloor+i})(x_i^\ell - x^\ell_{\lfloor\frac{n}{2}\rfloor+i})\big|^2\Big)^{\frac{1}{2}}. \end{aligned} \tag{6.28}$$

It remains to estimate the terms inside the parenthesis on the right-hand side of the above inequality. To this end, we observe, for any $q'>1$, that

$$\begin{aligned} &\mathbb{E}_\sigma\sup_{\ell\in\mathbb{N}_d}\big|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i(x_i^k - x^k_{\lfloor\frac{n}{2}\rfloor+i})(x_i^\ell - x^\ell_{\lfloor\frac{n}{2}\rfloor+i})\big|^2 \\ &\leq \mathbb{E}_\sigma\Big(\sum_{\ell\in\mathbb{N}_d}\big|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i(x_i^k - x^k_{\lfloor\frac{n}{2}\rfloor+i})(x_i^\ell - x^\ell_{\lfloor\frac{n}{2}\rfloor+i})\big|^{2q'}\Big)^{\frac{1}{q'}} \\ &\leq \Big(\sum_{\ell\in\mathbb{N}_d}\mathbb{E}_\sigma\big|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i(x_i^k - x^k_{\lfloor\frac{n}{2}\rfloor+i})(x_i^\ell - x^\ell_{\lfloor\frac{n}{2}\rfloor+i})\big|^{2q'}\Big)^{\frac{1}{q'}}. \end{aligned}$$

Applying the Khinchin-Kahane inequality (Lemma 14) with $q = 2q' = 4\log d$ and $p=2$ to the

above inequality yields that

$$\begin{aligned}
&\mathbb{E}_\sigma \sup_{\ell\in\mathbb{N}_d} \big|\textstyle\sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} \sigma_i (x_i^k - x_{\lfloor \frac{n}{2}\rfloor+i}^k)(x_i^\ell - x_{\lfloor \frac{n}{2}\rfloor+i}^\ell)\big|^2 \\
&\leq \Big(\textstyle\sum_{\ell\in\mathbb{N}_d} (2q')^{q'} \big[\mathbb{E}_\sigma \big|\sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} \sigma_i (x_i^k - x_{\lfloor \frac{n}{2}\rfloor+i}^k)(x_i^\ell - x_{\lfloor \frac{n}{2}\rfloor+i}^\ell)\big|^2\big]^{q'}\Big)^{\frac{1}{q'}} \\
&= \Big(\textstyle\sum_{\ell\in\mathbb{N}_d} (2q')^{q'} \big[\sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} (x_i^k - x_{\lfloor \frac{n}{2}\rfloor+i}^k)^2 (x_i^\ell - x_{\lfloor \frac{n}{2}\rfloor+i}^\ell)^2\big]^{q'}\Big)^{\frac{1}{q'}} \\
&\leq 2q' \sup_{x,x'\in\mathcal{X}} \|x-x'\|_\infty^2 d^{\frac{1}{q'}} \big[\textstyle\sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} (x_i^k - x_{\lfloor \frac{n}{2}\rfloor+i}^k)^2\big] \\
&\leq 4e(\log d) \sup_{x,x'\in\mathcal{X}} \|x-x'\|_\infty^2 \big[\textstyle\sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} (x_i^k - x_{\lfloor \frac{n}{2}\rfloor+i}^k)^2\big]
\end{aligned}$$

Putting the above estimation back into (6.28) implies that

$$\begin{aligned}
R_n &\leq \sqrt{4e\log d}\big[\sup_{x,x'\in\mathcal{X}} \|x-x'\|_\infty\big] \mathbb{E}_{\mathbf{z}} \Big(\textstyle\sum_{i=1}^{\lfloor \frac{n}{2}\rfloor} \|x_i - x_{\lfloor \frac{n}{2}\rfloor+i}\|_F^2\Big)^{\frac{1}{2}} \big/ \lfloor \tfrac{n}{2}\rfloor \\
&\leq \sqrt{4e\log d}\big[\sup_{x,x'\in\mathcal{X}} \|x-x'\|_\infty\big] \big[\sup_{x,x'\in\mathcal{X}} \|x-x'\|_F\big] \big/ \sqrt{\lfloor \tfrac{n}{2}\rfloor} \\
&\leq 4\sqrt{e\log d}\big[\sup_{x,x'\in\mathcal{X}} \|x-x'\|_\infty\big] \big[\sup_{x,x'\in\mathcal{X}} \|x-x'\|_F\big] \big/ \sqrt{n}.
\end{aligned}$$

Combining this with Theorem 12 implies the inequality (6.27). This completes the proof of the example. □

We end this section with two remarks. Firstly, in the setting of trace-norm regularization, it remains a question to us on how to establish more accurate estimation of R_n by using the Khinchin-Kahane inequality. Secondly, the bounds in the above examples are true for similarity learning with different matrix-norm regularization. Indeed, the generalization bound for similarity learning in Theorem 13 tells us that it suffices to estimate $\widetilde{X}_*$ and $\widetilde{R}_n$. In analogy to the arguments in the above examples, we can get the following results. For similarity learning formulation (6.4) with Frobenius-norm regularization, there holds

$$\widetilde{X}_* = \sup_{x\in\mathcal{X}} \|x\|_F^2, \qquad \widetilde{R}_n \leq \frac{2\sup_x \|x\|_F^2}{\sqrt{n}}.$$

For L^1-norm regularization, we have

$$\widetilde{X}_* = \sup_{x\in\mathcal{X}} \|x\|_\infty^2, \qquad \widetilde{R}_n \leq 4\sup_{x\in\mathcal{X}} \|x\|_\infty^2 \sqrt{e\log d}/\sqrt{n}.$$

In the setting of $(2,1)$-norm, we obtain

$$\widetilde{X}_* = \sup_{x\in\mathcal{X}} \|x\|_\infty \sup_{x\in\mathcal{X}} \|x\|_F, \qquad \widetilde{R}_n \leq 4\big[\sup_{x\in\mathcal{X}} \|x\|_F \sup_{x\in\mathcal{X}} \|x\|_\infty\big] \sqrt{e\log d}/\sqrt{n}.$$

Putting these estimations back into Theorem 13 yields generalization bounds for similarity learning with different matrix norms. For simplicity, we omit the details here.

6.5 Discussion

In this section, we discuss the derived generalization bounds for metric/similarity learning with different matrix-norm regularization terms. In the Frobenius-norm case, the main term of the bound (6.23) is $\mathcal{O}\big(\frac{\sup_{x,x'\in\mathcal{X}} \|x-x'\|_F^2}{\sqrt{n\lambda}}\big)$. This bound is consistent with that given by [Jin et al.,

2009] where $\sup_{x\in\mathcal{X}} \|x\|_F$ is assumed to be bounded by some constant B. Comparing the generalization bounds in the above examples in Section 6.4, we see that the key terms X_* and R_n mainly differ in two quantities, i.e. $\sup_{x,x'\in\mathcal{X}} \|x-x'\|_F$ and $\sup_{x,x'\in\mathcal{X}} \|x-x'\|_\infty$. We argue that $\sup_{x,x'\in\mathcal{X}} \|x-x'\|_\infty$ can be much less than $\sup_{x,x'\in\mathcal{X}} \|x-x'\|_F$. For instance, consider the input space $\mathcal{X}=[0,1]^d$. It is easy to see that $\sup_{x,x'\in\mathcal{X}} \|x-x'\|_F = \sqrt{d}$ while $\sup_{x,x'\in\mathcal{X}} \|x-x'\|_\infty \equiv 1$. Consequently, we can summarise the estimations as follows:

- **Frobenius-norm**: $X_* = d, \quad \text{and} \quad R_n \leq \frac{2d}{\sqrt{n}}$.
- **Sparse L^1-norm**: $X_* = 1, \quad \text{and} \quad R_n \leq \frac{4\sqrt{e\log d}}{\sqrt{n}}$.
- **Mixed $(2,1)$-norm**: $X_* = \sqrt{d}, \quad \text{and} \quad R_n \leq \frac{4\sqrt{ed\log d}}{\sqrt{n}}$.

Therefore, when d is large, the generalization bound with sparse L^1-norm regularization is much better than that with Frobenius-norm regularization while the bound with mixed $(2,1)$-norm are between the above two. These theoretical results are nicely consistent with the rationale that sparse methods are more effective in dealing with high-dimensional data.

6.6 Conclusion

In this chapter we were mainly concerned with theoretical generalization analysis of the regularized metric and similarity learning. In particular, we first showed that the generalization analysis for metric/similarity learning reduces to the estimation of the Rademacher average over "sums-of-i.i.d." sample-blocks. Then, we derived their generalization bounds with different matrix regularization terms. Our analysis indicates that sparse metric/similarity learning with L^1-norm regularization could lead to significantly better bounds than that with the Frobenius norm regularization, especially when the dimensionality of the input data is high. Our novel generalization analysis develops the techniques of U-statistics [De la Pena and Giné, 1999; Clemençon et al., 2008] and Rademacher complexity analysis [Bartlett and Mendelson, 2003; Koltchinskii and Panchenko, 2002]. Below we mention several questions remaining to be studied in the future.

Firstly, this study only investigated the generalization bounds for metric and similarity learning. We can further get the consistency estimation for $\|M-M_*\|_F^2$ under strong assumptions on the loss function and the underlying distribution. In particular, assume that the loss function is the least square loss, the bias term b is fixed (e.g. $b\equiv 0$) and let $M_* = \arg\min_{M\in\mathbb{S}^d} \mathcal{E}(M,0)$, we can get the estimation:

$$\begin{aligned}\mathcal{E}(M_{\mathbf{z}},0) - \mathcal{E}(M_*,0) &= \iint \langle M-M_*, x(x')^T\rangle^2 d\rho(x)\rho(x') \\ &= \langle \mathcal{C}(M-M_*), M-M_*\rangle.\end{aligned} \tag{6.29}$$

Here, $\mathcal{C}$ is $d^2\times d^2$ matrix representing a linear mapping from $\mathbb{S}^d$ to $\mathbb{S}^d$:

$$\mathcal{C} = \iint (x(x')^T)\otimes(x(x')^T) d\rho(x)\rho(x').$$

Here, the notation $\otimes$ represents the tensor product of matrices. Equation (6.29) implies that

$\mathcal{E}(M_{\mathbf{z}},0)-\mathcal{E}(M_*,0) = \iint \langle M-M_*, x(x')^T\rangle^2 d\rho(x)\rho(x') \geq \lambda_{\min}(\mathcal{C})\|M-M_*\|_F^2$, where $\lambda_{\min}(\mathcal{C})$ is the minimum eigenvalue of the $d^2 \times d^2$ matrix $\mathcal{C}$. Consequently, under the assumption that C is non-singular, we can get the consistency estimation for $\|M - M_*\|_F^2$ for the least square loss. For the hinge loss, the equality (6.29) does not hold true any more. Hence, it remains a question on how to get the consistency estimation for metric and similarity learning under general loss functions.

We can get the consistency estimation for $\|M - M_*\|_F^2$ under very strong assumption on the loss function and the underlying distribution.

Secondly, the target of supervised metric learning for kNN classifications is to improve the generalization performance of kNN classifiers. It remains a challenging question to investigate how the generalization performance of kNN classifiers relates to the generalization bounds of metric learning given here.

7 Conclusion and Perspectives

This thesis has mainly focused on developing similarity metric learning models for the tasks of unconstrained face verification, person re-identification and kNN classification. In particular, four new models have been proposed. To address the issue of large transformation differences existing in unconstrained face verification and person re-identification, Chapter 3 has introduced a new dimensionality reduction model, Intra-PCA. Its objective function is formulated by remaining robust to large transformation differences. The limitation of most existing similarity metric learning methods [Xing et al., 2003; Weinberger et al., 2006; Davis et al., 2007] is addressed by presenting a novel regularization framework Sub-SML in Chapter 4. It learns novel distance metrics and similarity functions for unconstrained face verification and person re-identification by incorporating both the robustness of Intra-PCA to large transformation variations and the discriminative power of similarity metric learning. Chapter 5 has proposed a general metric learning model DML_p for kNN classification by recovering the methods in [Xing et al., 2003; Ying and Li, 2012]. In Chapter 6, a novel generalization analysis for metric and similarity learning has been described. Our analysis can deal with general matrix regularization terms including the Frobenius norm, sparse L^1-norm, mixed $(2,1)$-norm and trace-norm, which overcomes the limitation of the approach [Jin et al., 2009] that only works for the strongly convex norm (e.g. the Frobenius norm).

Four benchmark databases were used for the evaluation of the proposed approaches. For unconstrained face verification in still images, experiments were conducted on the Labeled Faces in the Wild (LFW) dataset [Huang et al., 2007], a current state-of-the-art dataset for face verification. For video-based face verification in the wild, the comprehensive YouTube Faces (YTF) database [Wolf et al., 2011a] was used in the experiments. Experiments for person re-identification were done on the largest publicly available dataset, i.e. the Viewpoint Invariant Pedestrian Recognition (VIPeR) database [Gray et al., 2007]. Experiments were done on the popular UCI datasets for kNN classification.

Below, Section 7.1 summarizes the contributions of this thesis and Section 7.2 presents several promising directions for future work.

7.1 Contributions

In this section, we summarize the contributions of this thesis work as follows:

- For the tasks of unconstrained face verification and person re-identification, to overcome the detrimental effect of the large transformation differences, Chapter 3 introduced a novel dimensionality reduction model called Intra-PCA under the scenario that only pairwise in-

formation is given while the label information is not provided. Our learning objective is to remain robust to large transformation variations. Specifically, we formulate Intra-PCA by first applying WPCA to reduce the noise and then mapping the resultant images to the intra-personal subspace by the whitening process (see equation (3.2)). The proposed Intra-PCA was further extended to unconstrained face verification in videos. Experiments were conducted on three benchmarks: the LFW dataset [Huang et al., 2007] for unconstrained face verification in still images; the YTF database [Wolf et al., 2011a] for video-based face verification in the wild; the VIPeR database [Gray et al., 2007] for person re-identification. In the experiments, we compared Intra-PCA with the classic dimensionality reduction models such as PCA, WPCA and LDA. It was shown in Sections 3.3 and 3.4 that Intra-PCA outperforms the other dimensionality reduction methods, which demonstrates its effectiveness.

- For the tasks of unconstrained face verification and person re-identification, Chapter 4 explored to combine the robustness to large transformation differences with the discriminative power of similarity metric learning methods. A novel regularized framework called Sub-SML was developed using pairwise information. Unlike most of existing metric learning methods [Xing et al., 2003; Weinberger et al., 2006; Davis et al., 2007; Guillaumin et al., 2009; Chechik et al., 2010; Kan et al., 2011; Ying and Li, 2012] which do not take into account to reduce the transformation variations, we formulated the learning objective by incorporating both the robustness to large transformation variations and the discriminative power using distance metrics and similarity functions. Additionally, the proposed Sub-SML (i.e. formulation (4.4)) is a convex optimization problem, and thus by employing existing optimization algorithms a global solution can be efficiently found. This is, for instance, not the case for the current similarity metric learning model [Nguyen and Bai, 2011]. Furthermore, Sub-SML was extended to video-based face verification. Similar to the experimental study for Intra-PCA, Sub-SML was evaluated on the LFW [Huang et al., 2007] and YTF [Wolf et al., 2011a] databases for unconstrained face verification in still images and videos, respectively. Besides, we conducted experiment on the VIPeR dataset [Gray et al., 2007] for person re-identification. We compared Sub-SML with metric learning models such as Xing [Xing et al., 2003], ITML [Davis et al., 2007], LDML [Guillaumin et al., 2009], DML-eig [Ying and Li, 2012], SILD [Kan et al., 2011] and KISSME [Kostinger et al., 2012]. It was observed that Sub-SML yields significantly better performance than the other metric learning methods, which shows its effectiveness as a similarity metric learning method over the intra-personal subspace. In addition to the above comparison, we also compared Sub-SML with the domain specific state-of-the-arts and experimental results showed that Sub-SML is competitive with or even better than these methods.

- For the task of kNN classification, Chapter 5 revisited the original model in [Xing et al., 2003] and proposed a general formulation of learning a Mahalanobis distance from data. It was shown that the proposed DML$_p$ recovers the models in [Xing et al., 2003; Ying and Li, 2012] as special cases.The convexity of this formulation was also proved. Furthermore, by looking at the special structure of DML$_p$, we showed that DML$_p$ can be rewritten as a convex optimization problem over the spectrahedron and thus Frank-Wolfe algorithm [Frank and Wolfe, 1956] can be used to obtain the optimal solution. Compared to the optimiza-

tion algorithm in [Xing et al., 2003] which needs the full eigen-decomposition per iteration, our proposed algorithm only involves the computation of the largest eigenvector of a matrix per iteration. The evaluation of DML$_p$ for kNN classification was done on various UCI datasets, with the comparison to the state-of-the-art metric learning methods including Xing [Xing et al., 2003], LMNN [Weinberger et al., 2006], ITML [Davis et al., 2007], BoostMetric [Shen et al., 2009] and DML-eig [Ying and Li, 2012]. Experimental results showed that DML$_p$ compares competitively to those state-of-the-art metric learning methods for kNN classification. Additionally, experiments were conducted on the LFW database [Huang et al., 2007] for unconstrained face verification in still images. It was shown that DML$_p$ outperforms metric learning methods in [Xing et al., 2003; Ying and Li, 2012] and obtains comparable performance with the domain specific state-of-the-arts, which showed its applicability.

- For the general analysis of metric and similarity learning methods, Chapter 6 proposed a novel approach for establishing generalization bounds for metric/similarity learning with general matrix regularization terms. The regularization terms discussed in this work include the Frobenius norm [Jin et al., 2009], sparse L^1-norm [Rosales and Fung, 2006], mixed $(2,1)$-norm [Ying et al., 2009] and trace-norm [Ying et al., 2009; Shen et al., 2009]. It was shown that this novel generalization analysis firstly reduces to the estimation of the Rademacher average over "sums-of-i.i.d." sample-blocks related to the specific matrix norm, i.e. Rademacher complexities for metric/simialrtiy learning. Then, by developing and refining the techniques of U-statistics [Clemençon et al., 2008; De la Pena and Giné, 1999] and Rademacher complexity analysis [Bartlett and Mendelson, 2003; Koltchinskii and Panchenko, 2002], the Rademacher complexities with different matrix-norm regularizers were estimated. Lastly, with the estimated Rademacher complexities, generalization bounds for metric/similarity learning with different matrix-norm regularizers were derived. It was indicated from our analysis that sparse metric or similarity learning with L^1-norm regularization could lead to significantly better bounds than those with Frobenius-norm regularisation.

7.2 Future Work

This section outlines several promising directions for future work.

Looking first at the proposed Sub-SML, the improvement of its scalability would be a possible future direction. In Sections 4.2 and 4.3, FISTA [Nemirovski, 1994; Beck and Teboulle, 2009] was used as the optimization algorithm. However, when dealing with large scale problems which often involve millions or even billions of training samples, FISTA may become infeasible because it has to go through all the data points many times in order to find the optimal solution. It would be very interesting to develop online learning algorithms such as the averaged stochastic gradient descent (ASGD) algorithm [Xu, 2011]. ASGD goes through the data in only several passes, which allows Sub-SML to be more suitable for the large scale problems.

In terms of unconstrained face verification and person re-identification, the following promising

work is identified. As seen in Sections 4.4.1 and 4.4.2, DDML [Hu et al., 2014] obtains competitive results on both the LFW and YTF databases. Indeed, DDML trains a deep neural network to learn a set of hierarchical nonlinear transformations for face verification in the wild. It would be very interesting to adapt this deep learning model to our proposed Sub-SML. A possible starting point would be to develop a deep neural network to learn a set of hierarchical nonlinear transformations that map the images/frames onto a new subspace, under which the discriminative power of Sub-SML is retained. It remains a question on how to incorporate the robustness to large transformation differences under this neural network framework.

Consider the generalization analysis of metric/similarity learning. The following future work is identified. In Section 6.3, the derived bounds for metric and similarity learning with trace-norm regularization were the same as those with Frobenius-norm regularization. One interesting direction would be to derive the bounds for metric/similarity learning with trace-norm regularization similar to those with sparse ℓ^1-norm regularization. The key issue is to estimate the Rademacher complexity term (6.12) related to the spectral norm using the Khinchin-Kahne inequality. However, we are not aware of such Khinchin-Kahne inequalities for general matrix spectral norms. Another alternative is to apply the advanced oracle inequalities in [Koltchinskii, 2011].

In many applications involving multi-media data, different aspects of the data may lead to several different, and obviously equally valid notions of similarity. This leads to a natural question to combining multiple similarities and metrics for a unified data representation. An extension of multiple kernel learning approach was proposed in [McFee and Lanckriet, 2011] to address this issue. Another promising avenue would be to investigate the theoretical generalization analysis for this multi-modal similarity learning framework using techniques established for learning the kernel problem [Ying and Campbell, 2009, 2010].

Bibliography

Barkan, O., Weill, J., Wolf, L., and Aronowitz, H. (2013). Fast high dimensional vector multiplication face recognition. In *Proc. IEEE Intl Conf. Computer vision*.

Bartlett, P. L. and Mendelson, S. (2003). Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526.

Cao, Q., Ying, Y., and Li, P. (2012). Distance metric learning revisited. In *Machine Learning and Knowledge Discovery in Databases*, pages 283–298. Springer.

Cao, Q., Ying, Y., and Li, P. (2013). Similarity metric learning for face recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE.

Chechik, G., Sharma, V., Shalit, U., and Bengio, S. (2010). Large scale online learning of image similarity through ranking. *The Journal of Machine Learning Research*, 11:1109–1135.

Chen, D., Cao, X., Wang, L., Wen, F., and Sun, J. (2012). Bayesian face revisited: A joint formulation. In *Computer Vision–ECCV 2012*, pages 566–579. Springer.

Chen, D.-R., Wu, Q., Ying, Y., and Zhou, D.-X. (2004). Support vector machine soft margin classifiers: error analysis. *The Journal of Machine Learning Research*, 5:1143–1175.

Cinbis, R. G., Verbeek, J., and Schmid, C. (2011). Unsupervised metric learning for face identification in tv video. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1559–1566. IEEE.

Clemençon, S., Lugosi, G., and Vayatis, N. (2008). Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, pages 844–874.

Cox, D. and Pinto, N. (2011). Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 8–15. IEEE.

Cucker, F. and Zhou, D. X. (2007). *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press.

Cui, Z., Li, W., Xu, D., Shan, S., and Chen, X. (2013). Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3554–3561. IEEE.

Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. (2007). Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM.

De la Pena, V. and Giné, E. (1999). *Decoupling: from dependence to independence*. Springer.

Deng, W., Hu, J., and Guo, J. (2005). Gabor-eigen-whiten-cosine: A robust scheme for face recognition. In *Analysis and Modelling of Faces and Gestures*, pages 336–349. Springer.

Dikmen, M., Akbas, E., Huang, T. S., and Ahuja, N. (2011). Pedestrian recognition with a learned metric. In *Computer Vision–ACCV 2010*, pages 501–512. Springer.

Dollár, P., Tu, Z., Tao, H., and Belongie, S. (2007). Feature mining for image classification. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

Everingham, M., Sivic, J., and Zisserman, A. (2006). Hello! my name is... buffy–automatic naming of characters in tv video.

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

Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Academic press.

Globerson, A. and Roweis, S. (2005). Metric learning by collapsing classes. In *NIPS*, volume 18, pages 451–458.

Gray, D., Brennan, S., and Tao, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International workshop on performance evaluation of tracking and surveillance*. Citeseer.

Gray, D. and Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Computer Vision–ECCV 2008*, pages 262–275. Springer.

Guillaumin, M., Verbeek, J., and Schmid, C. (2009). Is that you? metric learning approaches for face identification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 498–505. IEEE.

Hariharan, B., Malik, J., and Ramanan, D. (2012). Discriminative decorrelation for clustering and classification. In *Computer Vision–ECCV 2012*, pages 459–472. Springer.

Hazan, E. (2008). Sparse approximate solutions to semidefinite programs. In *LATIN 2008: Theoretical Informatics*, pages 306–316. Springer.

Heikkilä, M., Pietikäinen, M., and Schmid, C. (2006). Description of interest regions with center-symmetric local binary patterns. In *Computer Vision, Graphics and Image Processing*, pages 58–69. Springer.

Horn, R. and Johnson, C. Topics in matrix analysis, 1991. *Cambridge University Presss, Cambridge*.

Hu, J., Lu, J., and Tan, Y.-P. (2014). Discriminative deep metric learning for face verification in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1875–1882. IEEE.

Hu, W., Hu, M., Zhou, X., Tan, T., Lou, J., and Maybank, S. (2006). Principal axis-based correspondence between multiple cameras for people tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):663–671.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst.

JacobGoldberger, S. and GeoffHinton, R. (2004). Neighbourhood components analysis. *NIPS04*.

Jain, P., Kulis, B., and Dhillon, I. S. (2010). Inductive regularized learning of kernel functions. In *Advances in Neural Information Processing Systems*, pages 946–954.

Jin, R., Wang, S., and Zhou, Y. (2009). Regularized distance metric learning: Theory and algorithm. In *Advances in neural information processing systems*, pages 862–870.

Kan, M., Shan, S., Xu, D., and Chen, X. (2011). Side-information based linear discriminant analysis for face recognition. In *BMVC*, pages 1–12.

Kan, M., Xu, D., Shan, S., Li, W., and Chen, X. (2013). Learning prototype hyperplanes for face recognition in the wild.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole dEté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media.

Koltchinskii, V. and Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, pages 1–50.

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

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. (2009). Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer.

Li, H., Hua, G., Lin, Z., Brandt, J., and Yang, J. (2013). Probabilistic elastic matching for pose variant face verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3499–3506. IEEE.

Li, P., Fu, Y., Mohammed, U., Elder, J. H., and Prince, S. J. (2012). Probabilistic models for inference about identity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1):144–157.

Liu, C. and Wechsler, H. (1998). Enhanced fisher linear discriminant models for face recognition. In *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, volume 2, pages 1368–1372. IEEE.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Maurer, A. (2008). Learning similarity with operator-valued large-margin classifiers. *The Journal of Machine Learning Research*, 9:1049–1082.

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188.

McFee, B. and Lanckriet, G. (2011). Learning multi-modal similarity. *The Journal of Machine Learning Research*, 12:491–523.

Mendez-Vazquez, H., Martinez-Diaz, Y., and Chai, Z. (2013). Volume structured ordinal features with background similarity measure for video face recognition. In *Biometrics (ICB), 2013 International Conference on*, pages 1–6. IEEE.

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

Moghaddam, B., Jebara, T., and Pentland, A. (2000). Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782.

Nemirovski, A. (1994). *Efficient methods in convex programming*. Lecture Notes, FACULTY OF INDUSTRIAL ENGINEERING & MANAGEMENT.

Nesterov, Y. (2004). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer.

Nguyen, H. V. and Bai, L. (2011). Cosine similarity metric learning for face verification. In *Computer Vision–ACCV 2010*, pages 709–720. Springer.

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

Rosales, R. and Fung, G. (2006). Learning sparse metrics via linear programming. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 367–373. ACM.

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

Shalit, U., Weinshall, D., and Chechik, G. (2010). Online learning in the manifold of low-rank matrices. In *NIPS*, pages 2128–2136.

Shechtman, E. and Irani, M. (2007). Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

Shen, C., Kim, J., Wang, L., and Van Den Hengel, A. (2009). Positive semidefinite metric learning with boosting. In *NIPS*, volume 22, pages 629–633.

Swets, D. L. and Weng, J. (1996). Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on pattern analysis and machine intelligence*, 18(8).

Taigman, Y., Wolf, L., and Hassner, T. (2009). Multiple one-shots for utilizing class label information. In Cavallaro, A., Prince, S., and Alexander, D. C., editors, *BMVC*, pages 1–12. British Machine Vision Association.

Torresani, L. and Lee, K.-c. (2007). Large margin component analysis. *Advances in neural information processing systems*, 19:1385.

Tsang, I. W., Kwok, J. T., Bay, C., and Kong, H. (2003). Distance metric learning with kernels. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 126–129. Citeseer.

Turk, M. and Pentland, A. (1991a). Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86.

Turk, M. A. and Pentland, A. P. (1991b). Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE.

Vapnik, V. (2000). *The nature of statistical learning theory*. Springer Science & Business Media.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.

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

Wang, X. and Tang, X. (2004). A unified framework for subspace face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(9):1222–1228.

Weinberger, K., Blitzer, J., and Saul, L. (2006). Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems*, 18:1473.

Weinberger, K. Q. and Saul, L. K. (2008). Fast solvers and efficient implementations for distance metric learning. In *Proceedings of the 25th international conference on Machine learning*, pages 1160–1167. ACM.

Wolf, L., Hassner, T., and Maoz, I. (2011a). Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534. IEEE.

Wolf, L., Hassner, T., and Taigman, Y. (2009a). The one-shot similarity kernel. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 897–902. IEEE.

Wolf, L., Hassner, T., and Taigman, Y. (2009b). Similarity scores based on background samples. In Zha, H., ichiro Taniguchi, R., and Maybank, S. J., editors, *ACCV (2)*, volume 5995 of *Lecture Notes in Computer Science*, pages 88–97. Springer.

Wolf, L., Hassner, T., and Taigman, Y. (2011b). Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(10):1978–1990.

Wolf, L., Hassner, T., Taigman, Y., et al. (2008). Descriptor based methods in the wild. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*.

Wolf, L. and Levy, N. (2013). The svm-minus similarity score for video face recognition. In *CVPR*, pages 3523–3530. IEEE.

Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. (2003). Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, pages 521–528.

Xu, W. (2011). Towards optimal one pass large scale learning with averaged stochastic gradient descent. *arXiv preprint arXiv:1107.2490*.

Ying, Y. and Campbell, C. (2009). Generalization bounds for learning the kernel.

Ying, Y. and Campbell, C. (2010). Rademacher chaos complexities for learning the kernel problem. *Neural computation*, 22(11):2858–2886.

Ying, Y., Huang, K., and Campbell, C. (2009). Sparse metric learning via smooth optimization. In *Advances in neural information processing systems*, pages 2214–2222.

Ying, Y. and Li, P. (2012). Distance metric learning with eigenvalue optimization. *The Journal of Machine Learning Research*, 13:1–26.

Zheng, W.-S., Gong, S., and Xiang, T. (2011). Person re-identification by probabilistic relative

distance comparison. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 649–656. IEEE.