ABSTRACT

Title of dissertation: Restoration and Domain Adaptation for Unconstrained Face Recognition

Jie Ni, Ph.D. Oral Examination, 2014

Dissertation directed by: Professor Rama Chellappa
Department of Electrical and Computer Engineering

Face recognition (FR) has received great attention and tremendous progress has been made during the past two decades. While FR at close range under controlled acquisition conditions has achieved a high level of performance, FR at a distance under unconstrained environment remains a largely unsolved problem. This is because images collected from a distance usually suffer from blur, poor illumination, pose variation etc. In this dissertation, we present models and algorithms to compensate for these variations to improve the performance for FR at a distance.

Blur is a common factor contributing to the degradation of images collected from a distance, e.g., defocus blur due to long range acquisition, motion blur due to movement of subjects. For this purpose, we study the image deconvolution problem. This is an ill-posed problem, and solutions are usually obtained by exploiting prior information of desired output image to reduce ambiguity, typically through the Bayesian framework. In this dissertation, we consider the role of an example driven manifold prior to address the deconvolution problem. Specifically, we incorporate unlabeled image data of the object class in the form of a patch manifold to effectively regularize the inverse problem. We propose both parametric and non-parametric approaches to implicitly estimate the manifold prior from the given unlabeled data. Extensive experiments show that our method performs better than many competitive image deconvolution methods.

More often, variations from the collected images at a distance are difficult to address through physical models of individual degradations. For this problem, we utilize domain adaptation methods to adapt recognition systems to the test data. Domain adaptation addresses the problem where data instances of a source domain have different distributions from that of a target domain. We focus on the unsupervised domain adaptation problem where labeled data are not available in the target domain. We propose to interpolate subspaces through dictionary learning to link the source and target domains. These subspaces are able to capture the intrinsic domain shift and form a shared feature representation for cross domain recognition. Experimental results on publicly available datasets demonstrate the effectiveness of our approach for face recognition across pose, blur and illumination variations, and cross dataset object classification.

Most existing domain adaptation methods assume homogeneous source domain which is usually modeled by a single subspace. Yet in practice, oftentimes we are given mixed source data with different inner characteristics. Modeling these source data as a single domain would potentially deteriorate the adaptation performance, as the adaptation procedure needs to account for the large within class variations in the source domain. For this problem, we propose two approaches to mitigate the heterogeneity in source data. We first present an approach for selecting a subset of source samples which is more similar to the target domain to avoid negative knowledge transfer. We then consider the scenario that the heterogenous source data are due to multiple latent domains. For this purpose, we derive a domain clustering framework to recover the latent domains for improved adaptation. Moreover, we formulate submodular objective functions which can be solved by an efficient greedy method. Experimental results show that our approaches compare favorably with the state-of-the-art.

Restoration and Domain Adaptation for Unconstrained Face Recognition

by

Jie Ni

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Professor Rama Chellappa, Chair/Advisor
Professor Min Wu
Professor Piya Pal
Professor Wojciech Czaja
Professor David Jacobs, Dean's Representative

Dedication

To my parents.

Acknowledgments

First and foremost, I owe my gratitude to my advisor, Professor Rama Chellappa for accepting me as his graduate student and supporting me to work on this interesting topic over the past several years. His critical insights and constructive advice have helped me greatly to reach this milestone. Besides, his pursuit of excellence in research and other aspects of life has been a great inspiration to me, which I will continue to benefit from in my future endeavors.

I would like to thank Prof. Min Wu, Prof. David Jocabs, Prof. Wojciech Czaja and Prof. Piya Pal for agreeing to serve on my thesis committee and offering valuable time to review this manuscript.

It has been a pleasure to work with my fellow friends and colleagues in Prof. Chellappa's group, among whom I should particularly mention Prof. Aswin Sankaranarayanan, Prof. Pavan Turaga, Dr. Vishal Patel, Dr. Ruonan Li, Dr. Ming Du and Dr. Qiang Qiu for fruitful collaborations and many other help along the way.

Finally, I am forever indebted to my parents, for their unconditional love and support. I dedicate this dissertation to them.

Table of Contents

List of Tables vii

List of Figures viii

1 Introduction 1
- 1.1 Image Restoration . . . 4
- 1.2 Domain Adaptation . . . 7
- 1.3 Contributions of This Dissertation . . . 10
- 1.4 Dissertation Outline . . . 12

2 Background on Remote Identification of Faces 14
- 2.1 Face Recognition At a Distance . . . 15
- 2.2 Long Range Facial Image Quality . . . 24
- 2.3 Re-identification . . . 27
- 2.4 Remote Face Database . . . 28
- 2.5 Evaluation of Face Recognition Algorithms . . . 29
  - 2.5.1 Baseline Algorithm . . . 29
  - 2.5.2 Sparse Representation-based Algorithm . . . 30
  - 2.5.3 Experimental Results . . . 32
    - 2.5.3.1 Results on Remote Re-identification . . . 36
- 2.6 Discussions . . . 37

3 Example-Driven Patch Manifold for Image Deconvolution 39
- 3.1 Introduction . . . 39
- 3.2 Manifold Learning Techniques . . . 44
- 3.3 Manifold Modeling of Image Classes . . . 47
  - 3.3.1 Regularizing the deblurring problem with the manifold prior . 48
- 3.4 Sampling and learning the patch-manifold . . . 51
  - 3.4.1 Non-parametric manifold learning . . . 51
  - 3.4.2 Parametric manifold learning . . . 53
- 3.5 Generalized Cross Validation (GCV) . . . 55
- 3.6 Experimental Results . . . 58

3.6.1 Blind Deconvolution . . . 68
3.6.2 Computational Complexity . . . 70
3.7 Discussions and Conclusion . . . 70

4 Subspace Interpolation via Dictionary Learning for Unsupervised Domain Adaptation 72
4.1 Introduction . . . 72
4.2 Prior work . . . 77
4.3 Sparse Representation and Dictionary Learning . . . 78
4.3.1 Sparse Coding . . . 79
4.3.2 Design of Dictionaries . . . 81
4.4 Learning Intermediate Domains for Unsupervised Domain Adaptation 84
4.4.1 Learning Intermediate Domain Dictionaries . . . 84
4.4.2 Recognition Under Domain Shift . . . 87
4.4.3 Quantification of Domain Shift . . . 89
4.5 Nonlinear Dictionary Learning for Unsupervised Domain Adaptation 90
4.5.1 Learning Nonlinear Intermediate Domain Dictionaries . . . 90
4.5.2 Nonlinear Recognition Under Domain Shift . . . 92
4.6 Experiments . . . 94
4.6.1 Face Recognition Under Pose Variation . . . 94
4.6.2 Face Recognition Across Blur and Illumination Variations . . . 96
4.6.3 Face Re-identification . . . 97
4.6.4 Cross Dataset Object Recognition . . . 100
4.7 Conclusions . . . 105

5 Submodular Optimization for Robust Domain Adaptation 107
5.1 Introduction . . . 107
5.2 Related Work . . . 110
5.3 Submodular Sample Selection . . . 112
5.3.1 Preliminaries . . . 112
5.3.2 Domain Similarity Function . . . 113
5.3.3 Class Balance Function . . . 115
5.3.4 Objective Function . . . 118
5.4 Submodular Latent Domain Discovery . . . 119
5.4.1 Graph Representation . . . 119
5.4.2 Entropy Rate . . . 119
5.4.3 Domain Balancing Function . . . 121
5.4.4 Objective function . . . 124
5.5 Experiments . . . 125
5.5.1 Pivot Sample selection . . . 126
5.5.2 Latent domain discovery . . . 130
5.6 Conclusion . . . 135

6 Summary and Directions for Future Work 136
6.1 Summary 136
6.2 Future Research Directions 136

Bibliography 139

List of Tables

3.1 Algorithm for patch-manifold based regularization for deblurring. . . 50

3.2 ISNR for different experiments. The highest ISNR for each experiment is shown in bold. 65

4.1 Face recognition under pose variation on CMU-PIE dataset [1] 96

4.2 Face recognition across illumination and blur variations on CMU-PIE dataset [1] 98

4.3 Face re-identification with the Baltimore dataset as the source domain and the Comcast dataset as the target domain 100

4.4 Cross dataset object recognition in unsupervised setting 101

4.5 Cross dataset object recognition in semi-supervised setting 102

4.6 QDS values between Amazon/DSLR/Webcam/Caltech datasets . . . 105

5.1 Cross dataset object recognition in unsupervised setting 128

5.2 Face recognition across pose and illumination variations on CMU-PIE dataset [1] 129

5.3 Recognition performance using the original and recovered latent domains 132

List of Figures

2.1 Results of albedo estimation for remotely acquired images. Left: Original images; Right: Estimated albedo images. 17

2.2 Pose normalization. Left column: Original input images. Middle column: Recovered albedoes corresponding to frontal face images. Right column: Pose normalized relighted images. 19

2.3 Some occluded face images in remote face dataset. 20

2.4 Albedo recovery result in presence of blur. (a) Original image. (b) Noisy blurred image. (c) Recovered albedo. 22

2.5 A typical low-resolution face image in remote face dataset. 23

2.6 Extreme illumination conditions caused by the sun. 25

2.7 Typical original images from our remote face dataset. 26

2.8 Cropped face images with different variations from the remote face database. 29

2.9 Comparison of intensity images and albedo maps using baseline. . . . 32

2.10 Performance of the baseline algorithm as the condition of probe varies. 33

2.11 Comparison between SRC and baseline algorithms. 34

2.12 Cropped face images with different variations from the second remote dataset. 34

2.13 Re-identification performance of the baseline algorithm as the condition of the probe set varies. 35

2.14 Comparison of baseline and sparse representation for re-identification. 36

3.1 Model for the deconvolution problem. 40

3.2 Locally-linear parametrization of a densely sampled manifold. 51

3.3 Some of the natural images used to learn the patch-manifold of natural images. 59

3.4 Images used in this chapter for different experiments. (a) *Barbara* image, (b) *Tiger* image, (c) a *face* image, (d) *Koala* image, (e) *Flowers* image and (f) *Boat* image. 60

3.5 Details of the image deconvolution experiment with a *Barbara* image. (a) Original image. (b) Noisy blurred image. (c) Hyper-Laplacian [2] estimate (ISNR 5.19 dB). (d) Wavelet domain sparsity-based estimate [3] (ISNR 6.24 dB).(e)LPA-ICI [4] estimate (ISNR 7.88 dB) (f) Parametric manifold-based estimate (ISNR 7.98 dB) suggested in this chapter. 61

3.6 Details of the image deconvolution experiment with a *Tiger* image. (a) Original image. (b) Noisy blurred image. (c)Hyper-laplacian [2] estimate (ISNR 8.14 dB). (d) Wavelet domain sparsity-based estimate [3] estimate (ISNR 8.28 dB). (e) LPA-ICI [4] estimate (ISNR 9.14 dB) (f) Parametric manifold-based estimate suggested in this chapter (ISNR 9.02 dB). (g) Blur kernel. 63

3.7 Details of the image deconvolution experiment with a *face* image. (a) Original image. (b) Noisy blurred image. (c) Hyper-Laplacian [2] estimate (ISNR 5.16 dB). (d) Wavelet domain sparsity-based estimate [3] estimate (ISNR 6.1 dB). (e) LPA-ICI estimate [4] (ISNR 7.4 dB) (f) Parametric manifold-based estimate (ISNR 8.49 dB) suggested in this paper. 64

3.8 GCV function for regularization with manifold prior. (a) *Barbara* Experiment. (b) *Koala* Experiment. (c) *flowers* Experiment. 67

3.9 ISNR performance of manifold-based algorithm compared to other methods as a function of BSNR 67

3.10 $\frac{1}{MN}\|\mathbf{y} - \mathbf{H}\mathbf{x}^{(k)}\|_2^2$ vs. number of iterations to determine the stopping criteria. 68

3.11 Details of the blind deconvolution experiment with a *Boat* image. (a) Original image. (b) Blurred noisy image. (c) Result obtained by applying a blind deconvolution method in [5] (ISNR -0.19 dB). (d) Result obtained by applying the parametric manifold deconvolution method using blur kernel estimated from [5] (ISNR 1.59 dB). (e) Estimated kernel. 69

4.1 Examples of dataset shifts. Each column contains two images of the same subject collected under different conditions. 73

4.2 Given labeled data in the source domain and unlabeled data in the target domain, our DA procedure learns a set of intermediate domains (represented by dictionaries $\{\mathbf{D}_k\}_{k=1}^{K-1}$) and the target domain (represented by dictionary $\mathbf{D}_K$) to capture the intrinsic domain shift between two domains. $\{\Delta\mathbf{D}_k\}_{k=0}^{K-1}$ characterize the gradual transition between these subspaces. . 74

4.3 Synthesized intermediate representations between frontal face images and face images at pose $c11$. The first row shows the transformed images from a source image (in red box) to the target domain. The second row shows the transformed images from a target image (in green box) to the source domain. 95

4.4 Synthesized intermediate representations from face recognition across blur and illumination variations (motion blur with length of 9). The first row shows the transformed images from a source image (in red box) to the target domain. The second row shows the transformed images from a target image (in green box) to the source domain. (The left most image in the second row is an approximation to the blur-free image in the source domain.) 99

4.5 Example images of the *bike* category from the (a) Caltech (b) Webcam (c) Amazon (d) DSLR dataset. (Images best viewed in color) 103

4.6 Average reconstruction error of the target domain decomposed with the source and intermediate domains. The combinations of source and target domains are (a) frontal face images v.s. face images at pose $c29$ (b) DSLR v.s. Webcam (c) Caltech v.s. Amazon, respectively. 104

5.1 Example images of pivot source samples with Caltech as the source domain and DSLR as the target domain from: (a) the *bike* category (b) the *laptop* category. 131

5.2 Estimation of the number of latent domains using cross validation. (a) Amazon and Caltech datasets (b) Action videos taken from camera 2,3,4 (c) Caltech, DSLR and Webcam datasets 133

5.3 Example images of latent domains from : (a) Amazon and Caltech datasets (b) Webcam and DSLR datasets. 134

Chapter 1: Introduction

Face recognition (FR) has been one of the most successful applications in image analysis and computer vision. Tremendous progress has been made in the field of FR during the past two decades. Active research activities in FR have led to a wide range of practical applications in biometrics, information security, access control, surveillance systems and social networks. For example, in access control applications, a face recognition system is used to monitor continuously who is in front of a computer terminal. It allows the user to log on or continue a previous session if the user is recognized. Otherwise, the user who tries to log on without authorization is denied. Besides, as security is of primary concern at public places such as airports and train stations, surveillance systems that use face recognition technology will become a reality soon. Such a system can send out alerts whenever someone matching the appearance of a known terrorist suspect enters a security checkpoint.

The general problem of FR can be described as follows: identify or verify one or more persons from a given still or video images of a scene, using a stored database of faces. The design of a generic FR system usually involves three steps: 1) detecting faces from cluttered scenes, 2) extracting features from face regions and

3) recognition.

Detection: Face detection is the first step in automatic face recognition. A successful face detection algorithm is able to correctly identify the presence and the rough location of a face in the image. There are different categories of face detection techniques. Template matching methods compute the correlation between an input image and the stored patterns for detection. Feature invariant methods aim to find structural features which exist under varying lighting, pose or viewpoint conditions. These features are then used for face localization. Appearance-based approaches learn models from a set of training images which capture the representative variations of facial appearances, and then the learned models are used for detection.

Feature Extraction: Extracting reliable facial features are very important for FR. Even holistic approaches need key facial features to normalize the detected faces. Typical features under consideration include eyebrows, eyes, nostrils, mouth, cheeks, chin and geometric constraints on the features.

Recognition: In this dissertation, we focus on recognition from intensity images. Typical methods fall into two categories: holistic and feature-based methods. Holistic approaches try to identify faces using representations based on the entire image rather than local features. One of the most wildly used representations is eigenfaces [6], which assumes that any face can be approximately reconstructed using just a small number of eigenfaces and the corresponding projection coefficients along each eigenface. Later on, Linear Discriminant Analysis [7] was proposed which maximizes the ratio of the between-class scatter and within-class scatter, and is

better suited for classification than eigenfaces. More recently, a compact face representation learned from a deep neural network was proposed in [8], which closely approaches human-level performance on the LFW benchmark dataset. In feature-based methods, local features such as eyes, nose and mouth are extracted and local statistics of these features are fed into a structural classifier. One well-known approach is the Elastic Bunch Graph Matching (EBGM) system [9]. Besides, high level visual features (gender, race, age, hair color, etc.) are also exploited to train an attribute classifier for face verification [10].

The difficulties of the FR problem arise due to variations among the face images of the same individual which can be larger than the variations resulting from changes in identity. The sources of variations can be categorized into two groups: intrinsic and extrinsic factors [11]. Intrinsic factors are purely due to the physical nature of the faces, e.g., facial expression, glasses, cosmetics, ethnicity etc, while extrinsic factors usually include illumination, pose, resolution etc.

Numerous datasets are now available for the development of FR algorithms, e.g., the CMU PIE dataset, the FRGC/FRVT dataset, the FERET dataset, Extended Yale B dataset, AR dataset and the LFW dataset. Some of these datasets are collected at close range (less than a few meters) and under different levels of controlled acquisition conditions. For instance, studio lights are used to control the illumination and pose variations are controlled by cooperative subjects, etc.

While FR systems on these datasets have reached high levels of recognition performance, research in unconstrained FR field is still at a nascent stage. In this dissertation, we are interested in studying and developing more robust algorithms

for FR at a distance in unconstrained environment. Various artifacts can occur in face images as a result of long range acquisition. First, as the subjects may not be cooperative, the pose of the face and body relative to the sensor is likely to vary greatly. Second, the lighting is uncontrolled and could be extreme in its variation. Third, the effects of scattering and high magnification resulting from long distance contribute to the blurriness of face images.

This dissertation focuses on investigating models and algorithms to compensate for FR in unconstrained environments. In particular, this dissertation focuses on addressing the following aspects: image restoration for reliable feature extraction, domain adaptation methods for handling more complicated variations between training and test data, and submodular optimization frameworks for tackling heterogenous source training data.

1.1 Image Restoration

The purpose of image restoration is to compensate for defects which degrade an image. There are several manifestations of the restoration problem. For instance, deblurring tries to estimate clear images from blurred and noisy inputs. Inpainting is the process of reconstructing lost or deteriorated parts of images. Super-resolution aims to enhance the resolution of a down-sampled image. These problems are ill-posed due to inadequate (noisy) observations. Solutions are usually obtained by exploiting the structure of the desired output image to reduce ambiguity. In the first part of this dissertation, we specifically study the deblurring problem, as blurriness

is a common problem in long range acquisition conditions. Blurriness can be due to camera motion, defocusing as well as atmospheric turbulence. Effectively restoring a blurred image is important for subsequent feature extraction.

The blurring process can usually be described using a convolution model: an observed image is produced as the convolution of an unknown desired image with a linear time-invariant point spread function, and then contaminated by additive or multiplicative white or colored noise. The act of restoring the unknown clear image is typically an under-constrained problem. Prior knowledge about natural images is usually employed for achieving improved results. For instance, Tikhonov regularization [12] is one of the most commonly used methods for regularizing the desired smoothness of the recovered image. Yet it often creates Gibbs oscillations in the neighborhood of discontinuities in the image. Alternatively, sparsity-based priors [13, 3] have been successfully designed to improve the visual quality of the recovered image. Another popular prior exploits the heavy-tailed characteristics of an image's gradient distribution, which is often parameterized as a mixture of Gaussian distributions [5]. Yet recent studies suggest that using priors learned from examples usually lead to improved performance compared with pre-specified ones. For example, priors based on image gradients may not work well for face images where the majority region is smooth.

In the first part of this dissertation, we propose a learning-based patch manifold prior to effectively constrain the ill-posed deconvolution problem. We consider the problem of exploiting extra information in the form of prior knowledge of the object class to regularize the inverse problem. This approach can be broadly termed

as example-based image enhancement [14]. An important step in our work involves learning the appropriate image representation. Images are formed through the interaction of light with surfaces. Surface properties such as geometry and reflectance can give rise to varied appearances, which are then imaged by a projective camera. For simple scenes, we can use a clear model of each of these factors to characterize the image space. However, it would be much more difficult to extend the factorization for more general classes of objects. Alternatively, manifold learning tools which are usually used to learn the underlying embedding space from high dimensional data would become less helpful here due to the extreme high dimensionality of image space. While image manifolds are very difficult to model in the general case, we exploit a far weaker requirement instead. We exploit patch manifold for image representation and assume that small patches from a given class lie on a manifold, which is far easier than the image manifold assumption. Since we do not have an analytical characterization of the patch manifold, we learn the manifold through dense sampling of patches using training images from the class of images under consideration. The goodness of fit between a given image and the manifold is then measured by averaging the distance of each patch of the image to the manifold. The proposed regularization term enforces the condition that the restored image traces a curve close to the manifold, so that the restored image has statistics similar to clear natural images. Significant computations may be involved in finding the closest point on the manifold to a given patch; for this purpose, we derive both non-parametric and parametric manifold learning methods to efficiently implement the projection operator on the manifold. Experiments demonstrate that the proposed method is

very competitive with state-of-the-art deconvolution methods.

1.2 Domain Adaptation

Typical pattern recognition systems often face a major challenge when applied "in the wild": conditions under which the system was developed are usually different from the actual conditions in which the system may be employed. For example, face recognition systems trained under constrained laboratory environments may be used to recognize face images acquired in unconstrained environment where the images suffer from a variety of degradations. One way to handle the variations in the test data which are not seen in the training data is to acquire labels in each new environment. Yet in many scenarios it is very expensive and impractical to collect new labels and rebuild the recognition system from scratch. Therefore it is essential to leverage the original "out-of-domain" data to transfer the classification knowledge to the new domain.

Typical methods to address this problem usually fall into two categories: domain adaptation (DA) and transfer learning (TL). DA addresses the problem where the conditional distributions of labels are similar while the marginal distributions of data in the training and test are different. For example, in unconstrained face recognition settings, the marginal distribution shift can be due to pose, illumination, resolution etc. On the other hand, TL handles the scenario where the marginal distributions are similar while the conditional distributions differ in the training and test domains. For instance, in object detection, to learn a detector for a new cate-

gory with insufficient training data, TL is used to leverage the detectors that have previously been learnt for similar categories by regularizing the distance between the new model and the source models.

Although DA and TL are fundamental problems in machine learning, they have not received much attention in the field of computer vision until recently. In this dissertation, we focus on the DA problem. We call the training data with plenty of labels as the source domain while the target domain is defined as data samples collected from a different distribution.

Based on the availability of labels in the target domain, DA methods can be broadly classified into two categories: semi-supervised DA and unsupervised DA. Semi-supervised DA is usually performed by utilizing the correspondence between source and target domains or a few labels in the target domain to learn the similarity among data instances across domains. In unsupervised DA, oftentimes prior assumption is needed to relate the source and target domains. For example, the structural correspondence learning method [15] induces correspondence among features from two domains by modeling their correlations with pivot features which appear frequently in both domains. Manifold alignment-based [16] DA methods attempt to compute the similarity between data points in different domains through the local geometry of data points within each domain, where the local geometry is defined by the distance between a data instance and the samples in its neighborhood.

Inspired by incremental learning, recent works in DA attempt to gradually learn a smooth transition path between the source and target domains in order to model the underlying domain shift. Incremental learning refers to using newly

obtained information to refine the existing knowledge of a certain subject. This self-adaptation process is a pre-requisite for many general learning tasks. One major reason is that information is often received in a sequential manner, and sometimes a learning process is needed long before all the information is available, and then the knowledge structure is constantly revised based on newly acquired information. The methodology of incremental learning has been applied in many computer vision applications. For instance, in object tracking, due to the drastic appearance changes of a target object, it is important to adapt the appearance model incrementally so as to produce a robust tracker. In a recent Grassmannian manifold-based [17] DA method, potential intermediate domains between the source and target are identified by gradually following the geodesics between the two domains, so as to discover the underlying domain shift. In the second part of this dissertation, we focus on learning the intermediate domains using dictionary models. We make use of the good reconstruction property of dictionaries to gradually reduce the reconstruction residue of target data while learning the intermediate dictionaries. The learned transition path is then used to form a new feature space for subsequent classification.

Most existing DA methods assume that the source data contain a single domain with very similar inner characteristics. Yet with the deluge of data from sources such as internet search engines and surveillance videos, this simplified assumption may not be valid in many realistic applications. For example, face images collected from the web usually consist of a mixture of illumination, expression and pose variations. Modeling these source data as a single domain would potentially result in negative knowledge transfer. Therefore, in the third part of this dissertation, we investigate

methods to mitigate the heterogeneity in the source data to facilitate the following adaptation task.

We first propose to select *pivot* samples which are a subset of the source data distributed most similar to the samples in the target domain. Identifying these samples can reduce the divergence between the two domains and boost subsequent adaptation performance. Alternatively, we consider the scenario that the heterogeneity in the source data is attributed to the presence of multiple latent domains without specific domain labels. This is different from previous approaches that deal with multiple source datasets where the partitions among the source domains are known a priori. For this problem, we adopt an entropy rate-based clustering framework which separates the heterogeneous source data into compact and homogeneous latent domains. More importantly, both of our objective functions are submodular which enables us to derive efficient optimization algorithms with guaranteed performance of at least $1 - \frac{1}{e}$ approximation to the optimum.

1.3 Contributions of This Dissertation

We make the following contributions in this dissertation:

• We investigate the problems and challenges that are present for FR in remote and unconstrained environments. We describe a face database collected in remote acquisition conditions, and evaluate a subset of still image-based FR algorithms on this dataset.

• We study the image deconvolution problem as image blurriness is a common

problem due to long range acquisition. Specifically, we consider the role of prior knowledge of the object class in the form of a patch manifold for regularizing the deconvolution problem. We implicitly estimate the manifold prior from the given unlabeled data using both non-parametric and parametric methods. Furthermore, we derive a generalized cross validation technique for automatically determining the regularization parameter at each iteration without explicitly knowing the noise variance.

• We consider the concept of DA to handle the large variations between training and test data in unconstrained FR. We propose to interpolate subspaces through dictionary learning to link the source and target domains. These subspaces are able to capture the intrinsic domain shift and form a shared feature representation for cross domain recognition. We then introduce a quantitative measure to characterize the shift between two domains, which enables us to select the optimal domain to adapt, given multiple source domains. Further, we extend our work to learn the set of intermediate dictionaries in a high dimensional feature space to handle the non-linearities in the data. We present experiments on FR across pose, illumination and blur variations, face re-identification, cross dataset object recognition, and report improved performances over the state-of-the-art.

• We investigate the problem of DA with heterogeneous source data. We first propose to select pivot source samples which are distributed more similar to the samples in the target domain. We derive a domain similarity function which encourages the selected source samples to be most representative of the target data. Further, in order to preserve the discrimination power of the source domain, we

derive a class balance function which ensures the labels of each class in the selected subset to follow the distribution in the original source domain. We then tackle the scenario that the heterogenous source data contain different latent domains and utilize an entropy rate-based domain clustering approach to obtain compact and homogeneous latent domains. Besides, we incorporate a domain balancing function which enforces the constraint that the distribution of class labels within each latent domain follow the prior label distribution in the original source domain. As our objective functions are submodular, we exploit the diminishing return property of submodularity to solve the problems efficiently. Experimental results demonstrate the advantage of our approaches compared to the state-of-the-art.

1.4 Dissertation Outline

The rest of the dissertation is organized as follows.

In chapter 2, we present the prospects and progress of the remote and unconstrained FR problem. We introduce a face dataset collected at a distance in outdoor environment and report the experimental results of two representative FR algorithms on this dataset. Discussions and conclusions from the corresponding recognition results are also provided.

Chapter 3 presents the work on using example-driven patch manifold prior for the deconvolution problem. We review different regularization methods used in prior work and then present our approach in detail. Extensive experiments demonstrate that our method is very competitive with state-of-the-art deconvolution methods.

In chapter 4, we review some representative DA methods, and then propose a novel unsupervised DA method based on dictionary learning models. We use FR across pose variations, blur and illumination variations, face re-identification and 2D object recognition to demonstrate the effectiveness of the proposed method.

Further, in chapter 5, we introduce our submodular optimization approaches for handling heterogeneous source data. We first present a pivot sample selection algorithm, and then describe a latent domain recovery method. We evaluate our methods for cross dataset object recognition, face recognition across pose and illumination variations, cross view activity recognition, and report competitive performance with the state-of-the-art.

Finally, conclusions and directions for future work are discussed in chapter 6.

Chapter 2: Background on Remote Identification of Faces

During the past two decades, FR has received great attention and tremendous progress has been made [18]. FR has a wide range of practical applications in access control, identification systems, surveillance, pervasive computing and social networks etc. Numerous image-based algorithms [6, 19, 7, 9, 20, 21, 22, 18] and video-based algorithms [23, 24] have been developed in the FR community. Currently, most of the existing FR algorithms have been evaluated using databases which are collected at close range (less than a few meters) and under different levels of controlled acquisition conditions. Some of the most extensively used face datasets such as CMU PIE [1], FERET [25] and YaleB [26] were captured in constrained settings. For instance, studio lights are used to control the illumination and pose variations are controlled by cooperative subjects etc.

While FR techniques on these datasets have reached a high level of recognition performance over the years, research in remote unconstrained FR field is still at a nascent stage. Recently a new database called "Labeled Faces in the Wild" (LFW) [27] whose images are collected from the web, has been widely used to address some of the issues in unconstrained FR problem. Yet concerns have been raised that these images are typically posed and framed by photographers and there is no guarantee

that such a set accurately captures the range of variations found in the real world settings [28]. Yao et al. [29] describe a face video database, UTK-LRHM, which is acquired from long distances with high magnifications. The magnification blur is described as a major source of degradation in their data.

In the following, we address some of the issues related to the problem of FR when face images are captured in unconstrained and remote setting. As one has very little control of the acquisition process, the images one gets often suffer from low resolution, poor illumination, blur, pose variation and occlusion etc. These variations present serious challenges to existing FR algorithms. We provide a brief review of developments and progress in the field of remote FR. We then introduce the re-identification problem and address the difficulties of this problem coupled with other inherent variations in remote acquisition conditions. Further, we introduce a new dataset which was collected in a remote maritime environment. We provide some preliminary experimental studies on this dataset and offer insights and suggestions for the remote FR problem.

2.1 Face Recognition At a Distance

Reliable extraction and matching of biometric signatures from faces acquired at a distance is a challenging problem [30]. First, as the subjects may not be cooperative, the pose of the face and body relative to the sensor is likely to vary greatly. Second, the lighting is uncontrolled and could be extreme in its variation. Third, when the subjects are at a long distance, the effects of a scattering media

(static: fog and mist, dynamic: rain, sleet, or sea spray) are greatly amplified. Fourth, the relative motion between the subjects and the sensors produce jitter and motion blur in the images. In this section, we investigate various factors that can affect long range FR system performance, which can be summarized into four types [30]: (1) technology (dealing with the quality of face images, heterogeneous face images, etc.), (2) environment (lighting, etc.), (3) user (expression, facial hair, facial ware etc.), and (4) user-system (pose, height, etc.). In what follows, we discuss some of these factors in detail.

Illumination: Variation in illumination conditions is one of the major challenges in remote FR. In particular, when images are captured from long ranges, one does not have control over lighting conditions. As a result, the captured images often suffer from extreme (due to sun) or low light conditions (due to shadow, bad weather, evening, etc.).

The performance of most existing FR algorithms is highly sensitive to illumination variations. Changes induced by illumination can usually render face images of the same subject farther apart than those of different subjects. Various methods have been introduced to deal with this problem in FR. Among them are methods based on the illumination cone [26, 31], spherical harmonics [32, 33, 34], quotient images [35, 36], gradient faces [37], logarithmic total variation [38], albedo estimation [39], photometric stereo [40], dictionaries [41, 42] etc.

Estimates of albedo are often used to mitigate the illumination effect. Albedo is the fraction of light that a surface point reflects when it is illuminated. It is an intrinsic property that depends on the material properties of the surface and

Figure 2.1: Results of albedo estimation for remotely acquired images. Left: Original images; Right: Estimated albedo images.

it is invariant to changes in illumination. Assuming that the facial surface can be described using the Lambertian reflectance model, one can relate the surface normals, albedo and the intensity image by an image formation model. The diffused component of the surface reflection is given by

$$x_{i,j} = \rho_{i,j} \max(\mathbf{n}_{i,j}^T \mathbf{s}, 0) \tag{2.1}$$

where x is the pixel intensity, $\mathbf{s}$ is the light source direction, $\rho_{i,j}$ is the surface albedo at position (i, j), $\mathbf{n}_{i,j}$ is the surface normal of the corresponding surface point and $1 \leq i, j \leq N$. The max function in (2.1) accounts for the formation of attached shadows. Neglecting the attached shadows, (2.1) can be linearized as

$$x_{i,j} = \rho_{i,j} \max(\mathbf{n}_{i,j}^T \mathbf{s}, 0) \approx \rho_{i,j} \mathbf{n}_{i,j}^T \mathbf{s}. \tag{2.2}$$

Let $\mathbf{n}_{i,j}^{(0)}$ and $\mathbf{s}^{(0)}$ be the initial values of the surface normal and illumination direction, which can be domain dependent average values. The Lambertian assumption imposes the following constraints on the initial albedo

$$\rho_{i,j}^{(0)} = \frac{x_{i,j}}{\mathbf{n}_{i,j}^{(0)} . \mathbf{s}^{(0)}} \tag{2.3}$$

where . is the standard dot product operator. Using (2.2), (2.3) can be re-written as

$$\rho_{i,j}^{(0)} = \rho_{i,j}\frac{\mathbf{n}_{i,j}.\mathbf{s}}{\mathbf{n}_{i,j}^{(0)}.\mathbf{s}^{(0)}} = \rho_{i,j} + \frac{\mathbf{n}_{i,j}.\mathbf{s} - \mathbf{n}_{i,j}^{(0)}.\mathbf{s}^{(0)}}{\mathbf{n}_{i,j}^{(0)}.\mathbf{s}^{(0)}}\rho_{i,j} \tag{2.4}$$

$$= \rho_{i,j} + \omega_{i,j}, \tag{2.5}$$

where $\omega_{i,j} = \frac{\mathbf{n}_{i,j}.\mathbf{s}-\mathbf{n}_{i,j}^{(0)}.\mathbf{s}^{(0)}}{\mathbf{n}_{i,j}^{(0)}.\mathbf{s}^{(0)}}\rho_{i,j}$. This can be viewed as a signal estimation problem where $\rho_{i,j}$ is the original signal, $\rho^{(0)}$ is the degraded signal and ω is the signal dependent noise. Based on this model, the albedo map can be estimated as the linear minimum mean square error estimate of the true albedo [39]. The illumination insensitive albedo image can then be used as the input for recognition. Figure 2.1 shows the results of albedo estimation for two face images acquired at a distance using the method presented in [39].

Pose variation: Pose variation can be considered as one of the most important and challenging problems in FR. Magnitudes of variations of innate characteristics, which distinguish one face from another, are often smaller than magnitudes of image variations caused by pose variations [43]. Popular frontal FR algorithms, such as Eigenfaces [6] or Fisherfaces [19, 7], usually have low recognition rates under pose changes as these holistic appearance-based methods are very sensitive to misalignment.

Existing methods for FR across poses can be roughly divided into two categories: techniques that rely on 3D models [44, 45] and 2D techniques which do not require 3D prior information [46, 47, 48]. Image patch-based approaches have also received significant attention in recent years [49, 50, 51, 52, 53], as modeling

Figure 2.2: Pose normalization. Left column: Original input images. Middle column: Recovered albedoes corresponding to frontal face images. Right column: Pose normalized relighted images.

face images as a collection of patches is more robust to pose changes than using the holistic appearance.

As the pose change is often coupled with other parameters such as illumination variation, it is desirable to compensate for joint pose and lighting variations. One solution is to estimate pose-robust albedo maps which can be considered as an extension of the approach in [39]. Let $\bar{\mathbf{n}}_{i,j}$, $\bar{\mathbf{s}}$ and $\bar{\Theta}$ be some initial estimates of the surface normals, illumination direction and initial estimate of surface normals in pose Θ, respectively. Then, the initial albedo at pixel (i, j) can be obtained by

$$\bar{\rho}_{i,j} = \frac{x_{i,j}}{\bar{\mathbf{n}}_{i,j}^{\bar{\Theta}}.\bar{\mathbf{s}}}, \tag{2.6}$$

where $\bar{\mathbf{n}}_{i,j}^{\bar{\Theta}}$ denotes the initial estimate of surface normals in pose $\bar{\Theta}$. Using this model, we can re-formulate the problem of recovering albedo as a signal estimation problem. Using arguments similar to (2.3), we get the following formulation for the

albedo estimation problem in the presence of pose variation:

$$\bar{\rho}_{i,j} = \rho_{i,j} h_{i,j} + \omega_{i,j}, \tag{2.7}$$

where $w_{i,j} = \frac{\bar{\mathbf{n}}^{\Theta}_{i,j}.\mathbf{s} - \bar{\mathbf{n}}^{\Theta}_{i,j}.\bar{\mathbf{s}}}{\bar{\mathbf{n}}^{\Theta}_{i,j}.\bar{\mathbf{s}}} \rho_{i,j}$, $h_{i,j} = \frac{\bar{\mathbf{n}}^{\Theta}_{i,j}.\bar{\mathbf{s}}}{\bar{\mathbf{n}}^{\Theta}_{i,j}.\bar{\mathbf{s}}}$, $\rho_{i,j}$ is the true unknown albedo and $\bar{\rho}_{i,j}$ is the rough estimate of albedo. Then a stochastic filtering framework which iterates between updating the albedo and pose estimates is performed to output a frontal albedo image. Figure 2.2 shows some examples of pose normalized images using this method. These normalized images can then be utilized for illumination and pose robust FR.

Occlusion: Another challenge in remote FR is that since face images are usually captured from non-cooperative subjects, acquired images are often contaminated by occlusion. The occlusion may be the result of subject wearing sunglasses, scarf, hat or a mask. Some representative techniques for recognizing subjects in the presence of occlusion include the principal component pursuit method [54], and the sparse representation-based method [20]. They are based on the fact that errors due to occlusion are often sparse with respect to the given basis. Figure 2.3 shows some images with occlusion from the remote face dataset.

Figure 2.3: Some occluded face images in remote face dataset.

Blur: In remote FR, the distance between the subject and the sensor results in producing degraded face images. Motion blur is another phenomenon that occurs when the subject is moving rapidly or the camera is shaking. [55, 56] are some of the methods that attempt to address this issue in FR. In [56], blurred face images are recognized using local phase quantization, which is based on quantizing the Fourier transform phase in local neighborhoods. It is shown that the quantized phase is blur invariant when certain conditions are met. [55] proposes a method to infer the point spread function (PSF) by using the prior information derived from a training set of blurred faces, such that the ill-posed problem becomes more tractable.

In remote acquisition settings, oftentimes blur is coupled with illumination variations. It might be desirable to develop an algorithm that can restore an image free from blur and illumination variations simultaneously. For this problem, one possible solution is to estimate the intrinsic albedo in the presence of blur. This turns out to be an inverse problem which is bilinear in the unknown albedo and blur. Given the $N \times N$ arrays y and x, representing the observed image and the image to be estimated, respectively, the image deconvolution problem can be described as

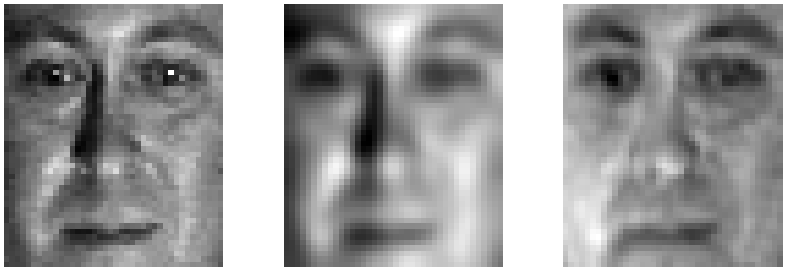
$$\mathbf{y} = \mathbf{H}\mathbf{x} + \gamma, \tag{2.8}$$

where $\mathbf{y}$, $\mathbf{x}$, and γ are $N^2 \times 1$ column vectors representing the arrays y, x, and γ lexicographically ordered, $\mathbf{H}$ is the $N^2 \times N^2$ matrix that models the blur operator and γ denotes an $N \times N$ array of noise signals. Using the Lambertian model in

(2.2), (2.8) can be re-written as

$$\begin{aligned}\mathbf{y} &= \mathbf{H}\mathbf{x} + \gamma = \mathbf{H}\Phi\rho + \gamma \\ &= \mathbf{G}\rho + \gamma, \end{aligned} \tag{2.9}$$

where $\Phi = diag(\mathbf{n}_{i,j}^T \mathbf{s})$ of size $N^2 \times N^2$, ρ is $N^2 \times 1$ vector representing ρ and $\mathbf{G} = \mathbf{H}\Phi$. Having observed $\mathbf{y}$, the general inverse problem is to estimate ρ with incomplete information of $\mathbf{G}$. It is well-known that regularization is often used to find a unique and stable solution to the ill-posed inverse problem. One learning based regularization method using the patch-manifold prior was developed in [57]. Figure 2.4 shows an example of recovered albedo using the method in [57].



(a) (b) (c)

Figure 2.4: Albedo recovery result in presence of blur. (a) Original image. (b) Noisy blurred image. (c) Recovered albedo.

Low resolution: Image resolution is an important parameter in remote face acquisition, where there is no control over the distance of the subject from the camera. Figure 2.5 illustrates a practical scenario where one is faced with a challenging problem of recognizing humans when the captured face images are of very low resolution (LR). Many methods have been proposed in the literature to deal with this

problem for FR. Most of these methods are based on some application of super-resolution (SR) techniques to increase the resolution of images so that the recovered higher-resolution (HR) images can be used for recognition. One of the major drawbacks of applying the SR techniques is that the recovered HR images may contain serious artifacts. This is often the case when the resolution of the image is very low. As a result, these recovered images may not look like the images of the same person and the recognition performance may degrade significantly.

Figure 2.5: A typical low-resolution face image in remote face dataset.

An Eigen-face domain SR method for FR was proposed in [58]. This method proposes to perform FR at LR by applying super-resolution (SR) on multiple LR images using their PCA domain representation. Given a LR face image, [59] proposes to directly compute a maximum likelihood identity parameter vector in the HR tensor space that can be used for SR and recognition. A Tikhonov regularization method that can combine SR and recognition in one step was proposed in [60].

As LR images are not directly suitable for the purpose of FR and the problem of recognition is not the same as SR, therefore, different approaches which do not require SR before recognition have been suggested. Coupled Metric Learning [61] attempts to solve this problem by mapping the LR and HR images to a joint subspace, where the distance measure is more ideal for recognition. A similar approach for improving the matching performance of the LR and HR images using multidimensional scaling was recently proposed in [62]. Additional methods for LR FR include a log-polar domain-based method [63], a correlation filter-based approach [64], a support vector data description-based method [65], a dictionary-based method [66], and 3D face modeling-based techniques [67, 68].

Atmospheric and weather artifacts: Most of the current vision algorithms and applications are applied to the images that are captured under clear and nice weather conditions. However, oftentimes in outdoor applications, one faces adverse weather conditions such as extreme illumination, fog, haze, rain and snow [30, 69, 70]. These extreme conditions can also present additional difficulties in developing robust algorithms for FR. [71] proposes to recover pertinent scene properties, such as the 3-D structure, from images taken under poor weather conditions. Yet the manifestations of weather on face images is still rarely explored in the literature.

2.2 Long Range Facial Image Quality

As discussed in the previous section, various factors could affect the quality of remotely acquired images. It is therefore essential to derive an image quality

Figure 2.6: Extreme illumination conditions caused by the sun.

measurement to study the relation between the image quality and recognition performance. To this end, a blind signal-to-noise ratio estimator has been defined for determining the qualify of facial images [30]. It is based on the concept that the statistics of the edge intensities of an image are correlated with the noise level of the image [72].

Suppose the pdf $f_{\|\nabla I\|}(r)$ of the edge intensity image $\|\nabla \mathbf{I}\|$ can be calculated as a mixture of Rayleigh pdfs, we define the following quantity

$$Q = \int_{2\mu}^{\infty} f\|\nabla \mathbf{I}\|(r)\,\mathrm{d}r,$$

where μ is the mean of $\|\nabla \mathbf{I}\|$. It has been shown that the value of Q for a noisy image is always smaller than that for an image with no noise [72]. Then, the face

Figure 2.7: Typical original images from our remote face dataset.

image quality is defined as

$$Q' = \frac{\sum \textit{edge above 2 } \mu\textit{'s pixels}}{\sum \textit{edge pixels}} \simeq \int_{2\mu}^{\infty} f\|\nabla \mathbf{I}\|(r)\,\mathrm{d}r.$$

It has been experimentally verified that the estimator Q' is well correlated with the recognition performance in FR [30]. Hence, setting up a comprehensive metric to evaluate the quality of face images is essential in remote FR. Also, these measurements can be used to reject images of low quality.

2.3 Re-identification

In re-identification, one has to identify a subject initialized at one location with a feasible set of candidates at other locations and over time. We define the remote face re-identification problem as follows.

Definition 1 *(Remote re-identification)* Given a probe set acquired at location L_p, *remote re-identification* aims to match them with the subjects in a gallery set, which were collected at a different location L_g and at a different time. Both gallery and probe sets are collected in remote and unconstrained setting.

Note that the data capture process of the gallery and probe sets may not be the same. That is, facial hair and ware of the subjects, the weather condition and illumination effect can be quite different, which might cause a large information gap between the face images collected at two different locations. In particular, this information gap is coupled with the variations we discussed before, which makes the remote face re-identification problem intrinsically difficult.

2.4 Remote Face Database

In this section, we introduce a remote face database in which a significant number of images are taken from long distances and under an unconstrained outdoor maritime environment. As discussed previously, the quality of the images differs in following aspects: the illumination is not controlled and is often severe; there are pose variations and occluded faces due to non-cooperative subjects; finally, the effects of scattering and high magnification resulting from long distance contribute to the blurriness of face images.

The distance from which the face images were taken varies from 5m to 250m under different scenarios. Since we could not reliably extract all the faces in the data set using existing state-of-the-art face detection algorithms and the faces only occupied small regions in large background scenes, we manually cropped the faces and rescaled them to a fixed size. The resulting database for still color face images contains 17 different individuals and 2106 face images in total.

We manually labeled the faces according their type (i.e. different illumination conditions, occlusion, blur etc.). In total, the database contains 688 clear images, 85 partially occluded images, 37 severely occluded images, 540 images with medium blur, 245 with sever blur, and 244 in poor illumination conditions. The remaining images have two or more coupled conditions, such as coupled poor lighting and blur, coupled occlusion and blur etc. Figure 2.7 shows two sample images acquired in a remote maritime setting. Some of the extracted images from the database are shown in Figure 2.8.

Figure 2.8: Cropped face images with different variations from the remote face database.

2.5 Evaluation of Face Recognition Algorithms

In this section, we first describe two state-of-the-art FR algorithms and then present and compare the recognition performance of these two algorithms on the remote face database.

2.5.1 Baseline Algorithm

The baseline recognition algorithm used in this chapter performs Principle Component Analysis (PCA) [73] followed by Linear Discriminate Analysis (LDA) [19, 7] for dimension reduction and a Support Vector Machine (SVM) [74] for classification.

LDA is a well-known feature extraction method for pattern recognition and classification tasks. It finds projection matrix $\mathbf{A}$ in such a way that the ratio of the between-class scatter and the within-class scatter is maximized [7]. The objective

function is defined as

$$\mathbf{A}_{opt} = \arg\max_{\mathbf{A}} \frac{|\mathbf{A}^T \Sigma_B \mathbf{A}|}{|\mathbf{A}^T \Sigma_W \mathbf{A}|}$$

where $|.|$ denotes the determinant of a matrix, Σ_B and Σ_W are between-class and within-class scatter matrices, respectively.

The within-class scatter matrix becomes singular when the dimension of the input data is larger than the number of training samples. To deal with this issue, we first use PCA to project the raw data onto an intermediate feature space with much lower dimension. Then, LDA is applied on the features from this intermediate space.

It is well known that LDA is not feasible when there is only one image per subject. To further mitigate this small sample size problem, we impose a regularization term in the objective function

$$\mathbf{a}_{opt} = \arg\max_{\mathbf{a}} = \frac{\mathbf{a}^T \Sigma_B \mathbf{a}}{\mathbf{a}^T \Sigma_W \mathbf{a} + \alpha J(\mathbf{a})}$$

where the resulting solutions form the columns of the optimal projection matrix $\mathbf{A}_{opt}$. We choose the Tikhonov regularizer $J(\mathbf{a}) = \|\mathbf{a}\|_2^2$ in our experiments. The resulting method is often known as Regularized Discriminate Analysis (RDA) [75]. Then, the low-dimensional discriminant features from RDA are fed into a linear SVM for classification.

2.5.2 Sparse Representation-based Algorithm

A state-of-the-art sparse representation-based classification (SRC) algorithm for FR was proposed in [20]. It demonstrates that if sparsity in the recognition

problem is properly harnesses, the choice of feature extraction method is no longer critical. Besides, the proposed framework can handle errors due to occlusion and corruption uniformly by exploiting the fact that these errors are often sparse with respect to the standard (pixel) basis.

Let each image be represented as a vector in $\mathbb{R}^n$, $\mathbf{D}$ be the training dictionary and $\mathbf{y}$ be the test image. The SRC algorithm is as follows:

1. Create a matrix of training samples $\mathbf{D} = [\mathbf{D}_1, ..., \mathbf{D}_k]$ for k classes, where $\{\mathbf{D}_i\}, i = 1, ..., k$ are the set of images of each class.

2. Reduce the dimensionality of the training images and the test image by any feature extraction method. Denote the resulting dictionary and the test vector as $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{y}}$, respectively.

3. Normalize the columns of $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{y}}$.

4. Solve the following ℓ_1 minimization problem

$$\hat{\alpha} = \arg\min_{\alpha'} \| \alpha' \|_1 \quad \text{subject to} \quad \tilde{\mathbf{y}} = \tilde{\mathbf{D}}\alpha', \tag{2.10}$$

5. Calculate the residuals

$$r_i(\tilde{\mathbf{y}}) = \|\tilde{\mathbf{y}} - \tilde{\mathbf{D}}\delta_i(\hat{\alpha})\|_2,$$

   for $i = 1, ..., k$ where δ_i a characteristic function that selects the coefficients associated with the ith class.

6. Identity($\mathbf{y}$)=$\arg\min_i r_i(\tilde{\mathbf{y}})$.

The assumption made in this method is that given sufficient training samples $\mathbf{D}_i$ of the ith class, any new test image $\mathbf{y}$ that belongs to the same class will approximately lie in the linear span of $\mathbf{D}_i$. This implies that most of the coefficients not associated with class i in $\hat{\alpha}$ will be close to zero. Hence, α' is a sparse vector. Further, a method of rejecting invalid test samples can also be incorporated within this framework. In particular, the notion of Sparsity Concentration Index (SCI) [20] has been proposed to decide whether a given test sample is a valid sample or not.

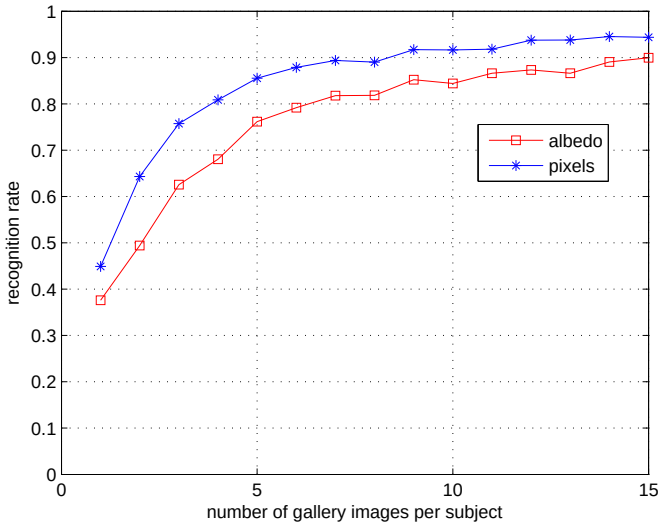


Figure 2.9: Comparison of intensity images and albedo maps using baseline.

2.5.3 Experimental Results

In the following, we report experimental results using the algorithms described earlier in this section.

The first set of experiments was designed to test the effectiveness of albedo maps [39]. We select the gallery set from clear images, and gradually increase the

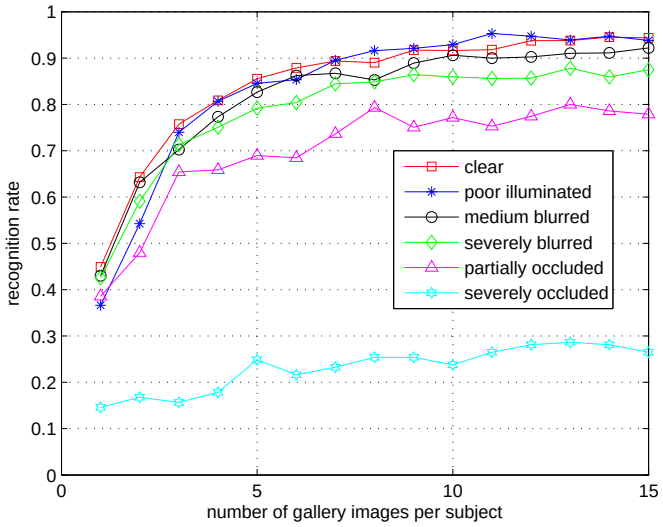


Figure 2.10: Performance of the baseline algorithm as the condition of probe varies.

number of gallery images from one to fifteen images per subject, and all the remaining clear images are selected for testing. We choose the gallery images randomly, and repeat five different trials to obtain the average recognition result. We compare the input of albedo maps with the intensity images using the baseline algorithm. All the parameters of PCA, LDA and SVM are fine tuned. The results are shown in Figure 2.9. We observe that intensity images outperform albedo maps although the albedo images are not sensitive to illumination variations. One possible reason is that, some face images in the database are a bit away from frontal. As albedo estimation needs a good alignment between the observed image and the ensemble mean, the resulting albedo map becomes erroneous. These artifacts are also seen in Figure 2.1.

In the second set of tests, the same gallery is chosen as the first set of experiments, while the test images are chosen to be clear, poorly illuminated, medium

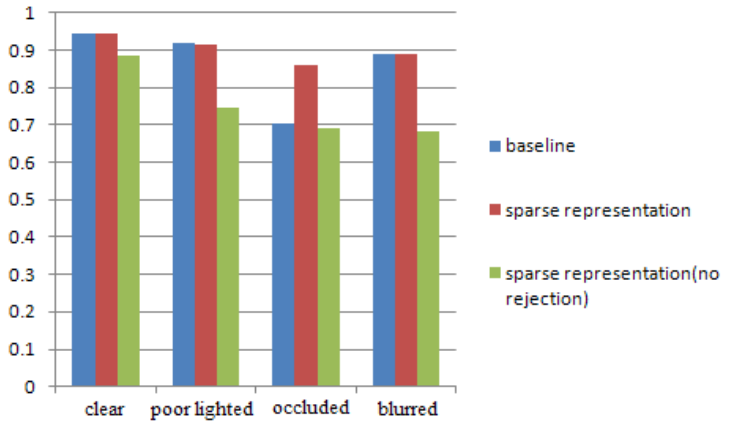


Figure 2.11: Comparison between SRC and baseline algorithms.

Figure 2.12: Cropped face images with different variations from the second remote dataset.

blurred, severely blurred, partially occluded and severely occluded respectively. The intensity images are used as input. The rank-1 recognition results using the baseline algorithm are given in Figure 2.10. We observe that the degradations in the conditions of test images decrease the performance, especially when the faces are occluded and severely blurred.

In the third set of experiments, we compare the performance of the SRC method and the baseline algorithm. We selected 14 subjects with 10 clear images per subject to form the gallery set. The test images are selected to contain clear, blurred, poorly illuminated and occluded images respectively. For the SRC method, we compute the SCI value of each image which can be used as a criteria to reject images of low quality.

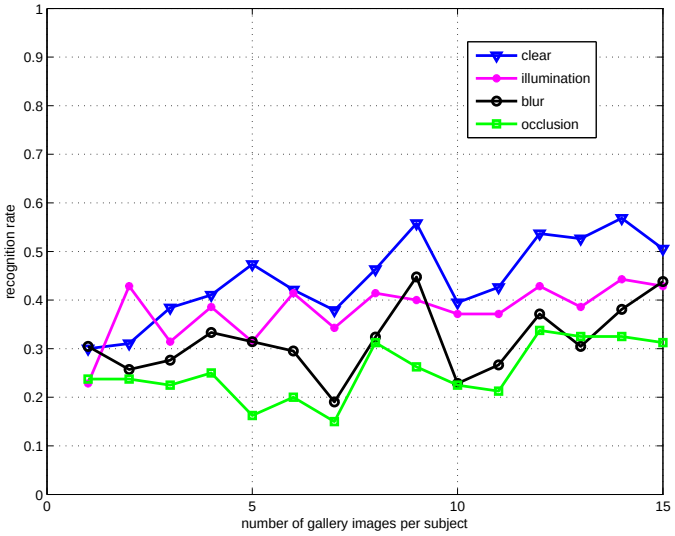


Figure 2.13: Re-identification performance of the baseline algorithm as the condition of the probe set varies.

From the comparison results reported in Figure 2.11, we observe that when no rejection of test images is allowed, the recognition accuracy of the baseline algorithm is superior to the SRC method. One possible reason is that when gallery images do not contain variations that occurred in the test images, the SRC method can not approximate the test images correctly through linear span of the gallery images. However, when rejection of test images is allowed, we remove those images with lower

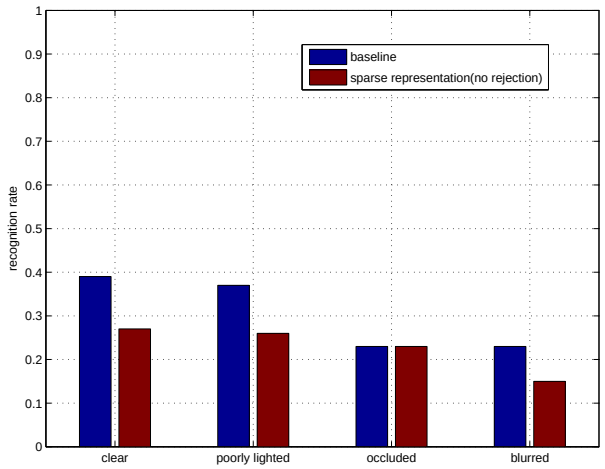


Figure 2.14: Comparison of baseline and sparse representation for re-identification.

SCI values so that test images become closer to the linear span of training images, and the performance of the SRC method improves accordingly. The rejection rates in Figure 2.11 are $6\%, 25.11\%, 38.46\%$ and 17.33% when the test images are clear, poorly lighted, occluded and blurred, respectively. Besides, the advantage of the SRC method for handling occluded images is also observed.

2.5.3.1 Results on Remote Re-identification

To study the difficulty of remote face re-identification, we present some results using the datasets we collected. The above remote dataset is used as the gallery set, and another outdoor remote dataset which was collected at a distance around 200 meters is used as the probe set. The time gap between these two datasets is more than two years. Five subjects which appear in both datasets are selected for the re-identification experiments. Figure 2.12 shows some of the cropped face images with different variations from the second remote dataset.

In the fist set of experiments on re-identification, we gradually increase the number of gallery images in the first remote dataset from one to fifteen per subject. The probe images from the second remote dataset is partitioned into four different subsets: clear, blurred, occluded and with illumination variation. Figure 2.13 shows the rank-1 recognition result using the baseline algorithm.

In the second set of experiments, we select 10 clear images per subject from the first remote dataset as gallery, and the same set of images as in previous experiment from the second dataset are used as probe. The comparison between the baseline algorithm and sparse representation-based method is reported in Figure 2.14.

Comparing Figure 2.10 and Figure 2.13, we see that the performance drops significantly in the remote re-identification case. Note that in both cases, the gallery settings are very similar except the number of subjects. This may be the result of large variations in facial appearances between these two datasets. Similarly, the decrease in the recognition performance can also be found by comparing Figure 2.11 and Figure 2.14.

2.6 Discussions

In this chapter, we briefly discussed some of the key issues in remote FR and introduced the remote re-identification problem. We then described a remote face database collected by UMD researchers and reported the performance of state-of-the-art FR algorithms on it. The results demonstrate that recognition rate decreases as the remotely acquired face images are affected by illumination variation, blur,

occlusion, pose variation etc. The coupling among different variation factors makes the remote FR problem extremely difficult. Therefore, it is essential to develop robust recognition algorithms under these conditions, as well as finding features that are robust to these variations. In the mean time, the re-identification problem raises an interesting new challenge for FR: how to make the FR system self-adaptive at a different location and over time is also worth investigating.

Chapter 3: Example-Driven Patch Manifold for Image Deconvolution

3.1 Introduction

Image deconvolution is a classical inverse problem where we observe a two-dimensional image y that consists of an unknown desired image x degraded by a point spread function (PSF) h (often assumed to be known) and then corrupted by zero-mean additive white Gaussian noise (AWGN) γ with variance σ^2 (see Fig. 3.1). Assuming that the images are of size $M \times M$, this model can be expressed as

$$y(n_1, n_2) = (x \circledast h)(n_1, n_2) + \gamma(n_1, n_2), \tag{3.1}$$

where $0 \leq n_1, n_2 \leq M - 1$. Using matrix notation, this model can be written as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \gamma, \tag{3.2}$$

where $\mathbf{y}, \mathbf{x}$, and γ are $M^2 \times 1$ lexicographically ordered column vectors representing the arrays y, x and γ, respectively and $\mathbf{H}$ is the $M^2 \times M^2$ matrix that models the point spread function. In the discrete Fourier transform (DFT) domain, we have for (3.1)

$$Y(k_1, k_2) = H(k_1, k_2)X(k_1, k_2) + \Gamma(k_1, k_2), \tag{3.3}$$

where $Y(k_1, k_2), H(k_1, k_2), X(k_1, k_2)$ and $\Gamma(k_1, k_2)$ are the 2D DFTs of y, h, x, and γ, respectively, for $-M/2 \leq k_1, k_2 \leq M/2 - 1$. Given y and h, we seek to estimate x. Such linear inverse problems often arise in many image processing applications such as radiometry, satellite imaging, optical systems, magnetic resonance imaging and seismic processing.

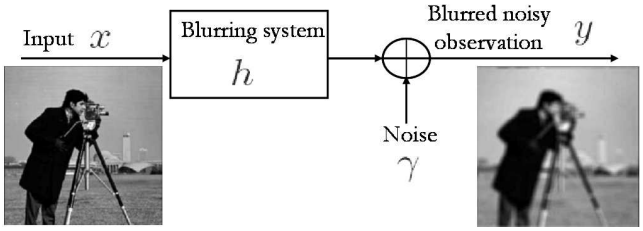


Figure 3.1: Model for the deconvolution problem.

It is well known that the deconvolution problem is ill-posed. To find a unique and stable solution, regularization is often used. A popular way to estimate the unknown image x is to use Tikhonov regularization [12] which consists of minimizing the following term

$$J_T(x) = \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 + \lambda E(\mathbf{x}), \tag{3.4}$$

where $E(\mathbf{x}) = \|\mathbf{C}\mathbf{x}\|_2^2$ and $\mathbf{C}$ is an $M^2 \times M^2$ matrix operator, known as the regularizing operator (e.g. Laplacian). The first term in (3.4) expresses the fidelity to x, and the second term expresses the desired smoothness of the restored image. Here, λ is the regularization parameter that represents the trade-off between fidelity to the data and the smoothness of the recovered image. The solution to the Tikhonov regularization problem can be obtained directly in the Fourier space

$$\tilde{X}(k_1, k_2) = \frac{H^*(k_1, k_2)Y(k_1, k_2)}{|H(k_1, k_2)|^2 + \lambda |C(k_1, k_2)|^2}. \tag{3.5}$$

The Tikhonov method offers computational advantages. However, it often creates Gibbs oscillations in the neighborhood of discontinuities in the image [76]. As a result, the visual quality of the recovered image often degrades.

Recently, considerable efforts have been spent on designing alternative sparsity constraints which preserve such features. Methods based on these sparsity constraints have been successfully used for image deconvolution (c.f. [13, 77, 78, 79, 80, 3, 81, 82]). Among various signal transformations, transformations based on wavelets, curvelets [83], contourlets [84, 85] and shearlets [86] are popular for image representation and are often used for image restoration. This is because wavelet transformations provide economical representations for a diverse class of signals, including signals with singularities. In fact, among all orthogonal transformations, the wavelet transformation can capture the maximum signal energy using any fixed number of coefficients for the worst-case Besov space signal [78].

Another popular deconvolution method is based on total variation [87], where $E(\mathbf{x})$ in (3.4) is set equal to $\|\mathbf{Cx}\|_1$, where $\|\mathbf{Cx}\|_1$ is the ℓ_1-norm of gradients of $\mathbf{x}$. Variations of this method have also been proposed [88, 89]. A local polynomial approximation method that uses intersecting confidence intervals was proposed in [4]. In [90], a locally adaptive kernel regression method was proposed to solve (3.4).

However, it has been shown recently that for image restoration, learning a representation from examples instead of using pre-specified ones, usually leads to improved results. For instance, [14] proposes an example-based image super-resolution method. As the richness of real world images is difficult to be captured analytically, a training set is used to learn the fine details that correspond to different image

regions observed at a low resolution. Then a Markov network is used to model the probabilistic relationships between high and low resolution patches, and between neighboring high resolution patches. Finally, fine details in high resolution images are predicted by exploiting the learned relationships. The reason this category of generic learning algorithm works is that the collection of image pixels are special signals that have much less variability than the corresponding set of completely random variables. These regularities can be utilized to create plausible image information.

Following this line of pursuit, in this chapter, we take a learning-based approach to the problem of image deconvolution by exploiting extra information in the form of prior knowledge of the object class to regularize the inverse problem [91]. Specifically, we use image data of the object class, as the available extra information. The proposed method assumes that the set of all patches (e.g. 3×3) from a given class of images - say faces, or natural images - live on a *manifold*. We shall define this in more precise terms as we progress. First, let us motivate the role of patch-manifolds in representing images. Images are formed by the interaction of light with surfaces. Surface properties such as geometry and reflectance give rise to varied appearances, which are then imaged by a projective camera. To characterize the space of images thus formed, one needs to have a clear model for each of these factors. For example, under variations in lighting conditions, with fixed viewing angle and pose, the set of face images obtained live on a 'cone' [31]. However, it is difficult to extend these results to more general classes of objects and scenes. Alternately, vision researchers have explored the tools of 'manifold learning' in such cases when one may have access to a large set of examples from each class. Manifold

learning algorithms such as Isomaps [92], LLE [93] etc, have proven useful in many cases and have been used to estimate the manifold of faces under pose variations. However, image manifolds are extremely high dimensional in the general case, since real images result from all of the above factors playing out simultaneously instead of in isolation. The situation gets much more complicated when several objects are present in the scene, each with its own surface properties. Since the number of samples needed to estimate even relatively low-dimensional manifolds is quite high (c.f. [94]), this makes the estimation of image-manifold in a general unconstrained setting, a difficult proposition.

On the other hand, assuming that small patches from a given class lie on a manifold is a far weaker requirement. It can be shown that even simple patch-manifold models give rise to complex imagery. For example, by assuming that each patch consists of small binary line segments, one can span the set of all 'cartoon' images. Similarly, the patch-manifold of locally parallel textures gives rise to complex finger-print type images [91]. Locally parallel textures can be analytically described by 2D sinusoidal functions, whereas the global manifold of images thus obtained are hard to describe in closed-form. When one does not have an analytical form for the patch-manifold, patch-manifold learning is still far easier than image-manifold learning. Since even a single image gives rise to an abundance of patches, and this affords a large set of samples on the patch manifold from unlabeled data. Coupled with the fact that the space of patches is far smaller than the space of images, this makes estimating the patch-manifold far easier.

Learning and using the patch-manifold often requires expensive computations

as the patch-manifold consists of a large number of samples. Hence we propose efficient parameterizations for computing the parameters of the manifold. Further, as the debluring performance is dependent on the regularization parameter, we drive a closed-form generalized cross validation function to automatically find a value of the regularizer λ without explicitly calculating the noise variance. We present experimental results on a wide variety of images and also discuss the computational expenses.

3.2 Manifold Learning Techniques

In this section, we introduce the fundamentals of manifold and several common manifold learning techniques.

A *manifold* $\mathcal{M}$ is a topological space that is locally Euclidean, i.e, around every point of $\mathcal{M}$ is a neighborhood which is topologically the same as the open unit ball in $\mathbb{R}^D$ [95]. A manifold is usually represented by an embedding in a certain space, e.g., $\mathbb{R}^D$ so that its topological properties are preserved in the embedded space [95]. Over the years, several manifold learning techniques have been raised to learn the underlying low dimensional manifold. We list several popular techniques in the following.

Principal Component Analysis (PCA): PCA is probably the most known and widely used method for analyzing high-dimensional data. It transforms a number of possibly correlated data into a smaller number of uncorrelated data called principal components.

Let the input data be $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_M$, and $\mathbf{\Phi}_i = \mathbf{Y}_i - 1/M \sum_{i=1}^{M} \mathbf{Y}_i$. Then by performing an eigen-decomposition of the covariance matrix $\mathbf{C} = 1/M \sum_{i=1}^{M} \mathbf{\Phi}_i \mathbf{\Phi}_i^T$, one can get the principal components $\mathbf{x}_k$ and the associated eigenvalues λ_k.

In many practical applications, one cannot overlook the intrinsic nonlinearity of the data. From the historic perspective, preserving distance is the first criteria proposed for manifold learning in a nonlinear way. Intuitively, as any manifold can be described by pairwise distances, the low-dimensional representation can be learned so that the initial distances are preserved [95].

Metric Multidimensional scaling (MDS): Metric MDS preserves the pairwise distance

$$E_{mMDS} = \sum_{i,j=1}^{N} w_{i,j}(d_y(i,j) - d_x(i,j))^2$$

where $d_y(i,j), d_x(i,j)$ are the Euclidean distances in the high and low dimensional spaces, respectively. Non-degenerative weights $w_{i,j}$ are often equal to one.

Isomap: Isomap [92] is a simple nonlinear dimension reduction method which shares similarity with metric MDS. The difference is that it uses the graph distance to approximate the geodesic distance.

Another category of manifold learning methods uses the topology of the data instead of pairwise distances. Topology, i.e., the neighborhood relationship is an important characteristic of a manifold. To some extent, distances give too much information, while comparative information between distances, like inequalities or ranks, suffice to characterize a manifold for any embedding [95].

Local Linear Embedding (LLE): LLE [93] proposes to preserve topology

based on a conformal mapping which is a transformation that preserves local angles. It first represents each data point $y(i)$ as a linear combination of its neighbor points

$$\varepsilon(\mathbf{W}) = \sum_{i=1}^{N} \|\mathbf{y}(i) - \sum_{j \in \mathcal{N}(i)} w_{i,j}\mathbf{y}(j)\|^2$$

where $\mathcal{N}(i)$ is the neighbor set of the ith data instance, and $w_{i,j}$ represents the weights of the neighbor data points. LLE assumes that such geometry also stands for the underlying low-dimensional manifold and optimizes the following objective function

$$\Phi(\mathbf{X}) = \sum_{i=1}^{N} \|\mathbf{x}(i) - \sum_{j \in \mathcal{N}(i)} w_{i,j}\mathbf{x}(j)\|^2$$

Alternatively, manifold learning has been cast as an inference problem on two special manifolds: the Grassmannian and Stefiel manifold. The Grassmannian manifold is the space of $d-$dimensional subspaces in $\mathbf{R}^n$ and the Stiefel manifold is the space of d orthonormal vectors in $\mathbf{R}^n$. Statistical modeling of these two special manifolds have been derived by the Riemannian geometric properties [96]. Studies of these manifolds have been used for face recognition [97], shape analysis [98], human activity recognition [99] and dynamic textures [100].

In the next section, we describe the details of our approach for manifold modeling of a given image class for deblurring. We show the limitations of traditional manifold learning methods in our patch-manifold setting, and propose efficient parameterizations suitable for learning the patch-manifold.

3.3 Manifold Modeling of Image Classes

In the following, we use x to denote the unknown image to be solved for, and $\mathbf{x}$ as the vector representation of the image x. We follow the theoretical foundations set forth in [91] for modeling images using a patch-manifold. We briefly review the required preliminaries before describing how we employ it for the deblurring problem. Let us denote a patch extracted from the image x, at location $q \in [0,1]^2$ of width $\tau > 0$ by $p_q(x)(t) = x(q+t), \forall t \in [-\tau/2, \tau/2]^2$. Further, we have $x \in L^2[0,1]^2$, which denotes the set of 2-dimensional finite energy signals. The class dependent image-ensemble is then denoted as $\Theta \subset L^2[0,1]^2$. The patch-manifold associated with this ensemble is denoted as $\mathcal{M} = \{p_q(x) | q \in [0,1]^2, x \in \Theta\} \subset L^2[-\tau/2, \tau/2]^2$. An image x is now represented as a surface traced on the manifold $\mathcal{M}$ given as

$$c_x : q \mapsto p_q(x) \in \mathcal{M}. \tag{3.6}$$

Given an image and the manifold representation, one can now measure the goodness of fit between them. To do this, first one needs a way to compute the closest point on the manifold. This is done in two stages. First, patches from an image are projected onto the patch-manifold. This step is denoted by $c(q) = Proj_{\mathcal{M}}(p_q(x))$, which assigns closest patches from the manifold to the given image patches. Thus, $Proj_{\mathcal{M}}(p) = \underset{t \in \mathcal{M}}{\arg\min} \|p - t\|$. The distance of a patch from the manifold is then given by $d(p, \mathcal{M}) = \|p - Proj_{\mathcal{M}}(p)\|$. Then, the goodness of fit of a given image is measured by averaging the distance of each patch from the patch-manifold

$$E_{\mathcal{M}}(x) = \int_{[0,1]^2} d(p_q(x), \mathcal{M})^2 dq \tag{3.7}$$

$$= \int_{[0,1]^2} \|p_q(x) - Proj_{\mathcal{M}}(p_q(x))\|^2 \, dq. \tag{3.8}$$

An image x has low-energy $E_{\mathcal{M}}(x)$ if it traces a curve $c_x = \{p_q(x)\}$ close to the manifold. This curve can be projected onto the manifold by means of the $Proj$ operator. The projected curve is thus represented as

$$\widetilde{c}_x(q) = Proj_{\mathcal{M}}(p_q(x)) \in \mathcal{M}. \tag{3.9}$$

Now, from this projected curve one can compute the projection of the image x onto the set of images generated by the patch manifold. Reconstruction is achieved by means of averaging overlapping patches. Specifically, the projection of the image x is represented by $Proj_{\mathcal{M}}(x) = Aver(\widetilde{c}_x)$, where

$$Aver(c_x) = \frac{1}{\tau^2} \int_{\|q-z\| \leq \tau/2} p_z(x - z) dz, \text{with } p_z(c) = c(z). \tag{3.10}$$

3.3.1 Regularizing the deblurring problem with the manifold prior

The optimization problem for deblurring is now recast by introducing a new variable c^* which is a manifold-valued function. The optimization is rewritten as finding an optimal $\mathbf{x}^*$, given an observation y and the manifold prior as

$$(x^*, c^*) = E(x, c) \tag{3.11}$$

$$= \arg\min_{x,c} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 + \lambda \int_{[0,1]^2} \|p_q(x) - c(q)\|^2 dq \tag{3.12}$$

where λ controls the relative weightage between the data and prior terms. A stationary point is obtained by means of an iterative procedure that alternates between solving for $\mathbf{x}^*$ and c^*. Given a current estimate of the image $x^{(k)}$, $c^{(k)}$ is obtained as

$$c^{(k+1)} = Proj_{\mathcal{M}}(x^{(k)}). \tag{3.13}$$

Next, given $c^{(k+1)}$, we solve for $\mathbf{x}$ as

$$\mathbf{x}^{(k+1)} = (\mathbf{H^T H} + \lambda \mathbf{I})^{-1}(\mathbf{H^T y} + \lambda \mathbf{vec}(Aver(c^{(k+1)})), \tag{3.14}$$

where $Aver(c)$ is as defined in (3.10), and **vec**() returns the vectorized version of its argument. This procedure is repeated till convergence and it is summarized in Table 3.1.

As (3.12) is non-convex, the algorithm in Table 3.1 may not converge to the global optimum. However, for a smooth manifold $\mathcal{M}$, the iterates $(x(k), c(k))$ of our algorithm will converge to a stationary point (x^*,c^*) [91]. Note that the matrix inversions involved in the optimization steps in Table 3.1 are all implemented implicitly using the properties of the PSF matrix $\mathbf{H}$ [101].

Table 3.1: Algorithm for patch-manifold based regularization for deblurring.

1. Set $\mathbf{x}^{(0)} = \mathbf{H}^{\mathbf{T}}\mathbf{y}$ and $k \leftarrow 0$.

2. Update the manifold-valued function as

$$\forall q \in [0,1]^2, c^{(k+1)}(q) = Proj_{\mathcal{M}}(p_q(\mathbf{x}^{(k)})).$$

3.Update the current estimate of $\mathbf{x}$ as

$$\mathbf{x}^{(k+1)} = (\mathbf{H}^{\mathbf{T}}\mathbf{H} + \lambda\mathbf{I})^{-1}(\mathbf{H}^{\mathbf{T}}\mathbf{y} + \lambda\mathbf{vec}(Aver(c^{(k+1)}))).$$

4. Repeat till convergence or till maximum iterations are reached.

3.4 Sampling and learning the patch-manifold

In actual implementation, we do not have an analytical characterization of the patch manifold. An analytical characterization would lead to a closed-form version of the *Proj* operator. We instead learn the manifold using training examples of images from the class of images under consideration, e.g. faces or natural images. The *Proj* operation then amounts to searching for the closest point to a given patch in the learnt manifold. We explore two ways to solve this problem - non-parametric and parametric. We describe these two approaches in the following.

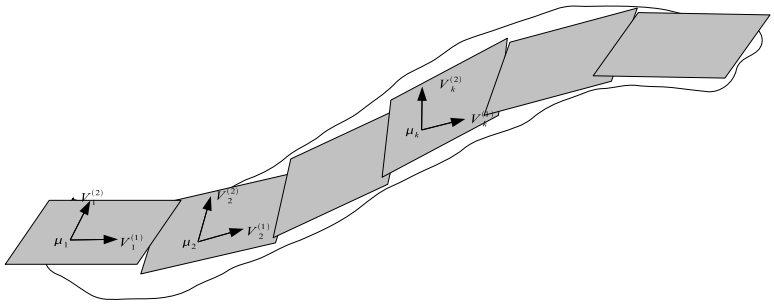


Figure 3.2: Locally-linear parametrization of a densely sampled manifold.

3.4.1 Non-parametric manifold learning

In the non-parametric case, we assume that we have a large number of samples from the underlying patch-manifold. In experiments we find that the assumption of a dense sampling is in fact very well justified given the easy availability of a large number of patches. With this, the *Proj* operation is efficiently implemented using approximate nearest neighbor search strategies. We choose locality sensitive hashing (LSH) [102] for this task due to its sub-linear search efficiency. Given a training set

of images, patches centered at all pixel locations are extracted from every image. The set of patches thus obtained constitutes the sampling of the manifold. This set is then indexed using LSH.

Here, we briefly review the basic concepts of LSH. LSH attempts to solve a problem called the (r, ϵ)-NN problem. The problem is described as follows: given a database of points $D = \{x_i\}$ in $\mathbb{R}^n$ and a query x_q, if there exists a point $x \in D$ such that $d(x, x_q) \leq r$, then with high probability, a point $x' \in D$ is retrieved such that $d(x', x_q) \leq (1 + \epsilon)r$. Now, LSH solves this problem by constructing a family of hash functions $\mathcal{F}$ over $\mathbb{R}^n$. These functions are called locality sensitive, if for any $u, v \in D$

$$d(u, v) \leq r \Rightarrow Pr(f(u) = f(v)) \geq p_1 \tag{3.15}$$

$$d(u, v) \geq (1 + \epsilon)r \Rightarrow Pr(f(u) = f(v)) \leq p_2 \tag{3.16}$$

Popular choices of f include random projections, i.e. $f(v) = sgn(v.r)$ where r is a randomly chosen unit vector, and sgn is the signum function. In this case, f is a binary variable taking values in $\{+1, -1\}$. A generalization of this is termed random projections using 'p-stable' distributions [103], with $f(v) = \lfloor \frac{v.r+b}{w} \rfloor$ where r is a randomly chosen direction whose entries are chosen independently from a stable distribution, and b is a random number chosen between $[0, w]$. In this case, the hash function takes on integer values. A $k-$bit hash is constructed by appending k randomly chosen hash functions as follows

$$F(x) = [f_1(x), f_2(x), \ldots f_k(x)] \tag{3.17}$$

where $F \in \mathcal{F}^k$. Then, L hash tables are constructed by randomly choosing $F_1, F_2 \ldots F_L \in \mathcal{F}^k$. All the training examples (patches) are hashed into the L hash tables. For a query point x_q, an exhaustive search is carried out among the examples in the union of the L hash buckets indexed by q. Appropriate choices of k and L ensure that the algorithm succeeds in finding a (r, ϵ)-NN of the query x_q with a high probability. In our work, we used random projections based hashing, i.e. the hash function is $f(v) = sgn(v.r)$.

3.4.2 Parametric manifold learning

Even though a dense sampling of the patch-manifold appears to be a reasonable assumption, implementing the $Proj$ operation involves significant computation for the entire image, as we need to hash and search for every patch in the given image. Also, the $Proj$ operator implemented in this manner is susceptible to noise in the dataset. Further, if the sampling density is reduced, the quality of reconstructions can be significantly affected. To deal with these situations, we explore a parametric way for modeling the patch manifold. While several parameterizations of the patch-manifold are possible, we choose the one that leads to computationally efficient algorithms for implementing the $Proj$ operation. Note that one could potentially use algorithms such as LLE [93] and Isomaps [92] to estimate the manifold, but there are a few considerations which make their use prohibitive in the current setting. To

begin with, these algorithms have a high computational complexity for estimating the manifold when the number of samples is high. Further, out-of-sample extension, i.e. finding the parameters of a new patch which is not in the training database, is a non-trivial task [104]. Here, we propose a much simpler parametrization of the patch-manifold which is computationally efficient to learn when a dense sampling of the patch-manifold is available, and has a graceful out-of-sample extension when the sampling density reduces.

We assume that the patch manifold can be decomposed into a union of subspaces, i.e. $\mathcal{M} = \bigcup_{i=1}^{K} S_i$, where each S_i is a d-dimensional affine subspace in $\mathbb{R}^n$, represented by its offset μ_i and orthonormal basis vectors $\mathbf{V}_i$ (written in matrix form). Each patch on the manifold is then parameterized by the index of the subspace on which it lies and the coefficients of its projection on the appropriate subspace as follows

$$\psi(p) = (\hat{i}, \hat{\alpha}) = \arg\min_{i,\alpha} \|p - \mu_i - \mathbf{V}_i \alpha\| . \tag{3.18}$$

Figure 3.2 presents a graphical illustration of the locally linear parametrization of the manifold. To learn this manifold from the training data, we adopt a two stage approach. In the first stage, given the training set of patches $D = \{x_i\}$, we cluster all the patches into K distinct clusters. Each cluster center is associated with the offset of the subspaces μ_i. Within each cluster, we then estimate the optimal basis vectors using principal component analysis (PCA). Given a new patch, the closest patch on the manifold is estimated in two stages. In the first stage, the closest cluster center is computed by comparing it with all the cluster centers. Once the closest cluster

center is found, the patch is projected onto the subspace of that cluster. Therefore, given a new patch p, we obtain the parameterizations as follows

$$\hat{i} = \min_i \|p - \mu_i\|, \quad \hat{\alpha} = \mathbf{V}_{\hat{i}}^T (p - \mu_{\hat{i}}) \tag{3.19}$$

Then, the $Proj$ operation is easily implemented as

$$Proj_{\mathcal{M}}(p) = \mu_{\hat{i}} + \mathbf{V}_{\hat{i}} \hat{\alpha}, \tag{3.20}$$

where $(\hat{i}, \hat{\alpha})$ are as defined in (3.19).

3.5 Generalized Cross Validation (GCV)

Note that the deblurring cost function in (3.12) and thereby the solution in (3.14) depends on the value of λ. The deblurred image depends greatly on the degree of regularization which is determined by the regularization parameter [101]. In this section, we describe a generalized cross validation (GCV) function [105, 106] to compute the regularization parameter automatically. The GCV method is based on statistical considerations, namely, a good value of the regularization parameter should predict missing data values [107]. One of the main advantages of this GCV method is that it can obtain the regularization parameter without knowing the noise variance.

First, we define the singular value decomposition (SVD) of the blur matrix $\mathbf{H}$ as $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}^T$ are orthogonal matrices, satisfying $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{M^2}$

and $\mathbf{V}^T\mathbf{V} = \mathbf{I}_{M^2}$, and $\mathbf{\Sigma} = diag(\sigma_i)$ is a diagonal matrix. Let $\mathbf{u}_i$ and $\mathbf{v}_i$ be the columns of $\mathbf{U}$ and $\mathbf{V}$, respectively.

In the principle of minimizing the predictive mean-square error, [107] defines the GCV function as

$$G(\lambda) = \frac{\|\mathbf{Hx} - \mathbf{y}\|_2^2}{(trace(\mathbf{I} - \mathbf{HH}^\sharp))^2} \tag{3.21}$$

where $\mathbf{x}$ is the restored image and $\mathbf{H}^\sharp$ is the regularized inverse given by

$$\begin{aligned} \mathbf{H}^\sharp &= (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1}\mathbf{H}^T \\ &= (\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T + \lambda\mathbf{I})^{-1}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T \\ &= \mathbf{V}(\mathbf{\Sigma}^2 + \lambda\mathbf{I})^{-1}\mathbf{\Sigma}\mathbf{U}^T. \end{aligned} \tag{3.22}$$

Let $\phi_i = \frac{\sigma_i^2}{\sigma_i^2+\lambda}$, $\mathbf{\Phi} = diag(\phi_i)$, then (3.22) can be written as

$$\mathbf{H}^\sharp = \mathbf{V}\mathbf{\Phi}\mathbf{\Sigma}^{-1}\mathbf{U}^T. \tag{3.23}$$

Substituting (3.23) into (3.21), the GCV function becomes

$$G(\lambda) = \frac{\|\mathbf{y} - \mathbf{Hx}\|_2^2}{(trace(\mathbf{I} - \mathbf{HV}\mathbf{\Phi}\mathbf{\Sigma}^{-1}\mathbf{U}^T))^2}. \tag{3.24}$$

By replacing $\mathbf{x}$ in (3.24) with the manifold-based solution, we obtain the GCV function of our proposed algorithm.

We split the manifold solution into two parts: $\mathbf{x} = \mathbf{x}_\lambda + \tilde{\mathbf{x}}$, where $\mathbf{x}_\lambda = (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1}\mathbf{H}^T\mathbf{y} = \mathbf{H}^\sharp\mathbf{y} = \mathbf{V}\mathbf{\Phi}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{y}$, $\tilde{\mathbf{x}} = (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1}\lambda\mathbf{vec}(Aver(c^{(k)}))$. Hence, $\mathbf{y} - \mathbf{Hx} = (\mathbf{y} - \mathbf{Hx}_\lambda) - \mathbf{H}\tilde{\mathbf{x}}$, where

$$\begin{aligned} \mathbf{y} - \mathbf{Hx}_\lambda &= \mathbf{y} - \mathbf{HV}\mathbf{\Phi}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{y} \\ &= \mathbf{y} - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Phi}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{y} \\ &= \mathbf{y} - \mathbf{U}\mathbf{\Phi}\mathbf{U}^T\mathbf{y}. \end{aligned} \tag{3.25}$$

Also,

$$\begin{aligned}\mathbf{H}\tilde{\mathbf{x}} &= \mathbf{H}(\mathbf{H}^T\mathbf{H}+\lambda\mathbf{I})^{-1}\lambda\mathbf{vec}(Aver(c^{(k)}))\\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}(\mathbf{\Sigma}^2+\lambda\mathbf{I})^{-1}\mathbf{V}^T\lambda\mathbf{vec}(Aver(c^{(k)}))\\ &= \mathbf{U}\mathbf{\Sigma}(\mathbf{\Sigma}^2+\lambda\mathbf{I})^{-1}\mathbf{V}^T\lambda\mathbf{vec}(Aver(c^{(k)}))\end{aligned} \tag{3.26}$$

Since the 2-norm is invariant under orthogonal transformation, $\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2 = \|\mathbf{U}^T(\mathbf{y}-\mathbf{H}\mathbf{x})\|_2^2$, so we can work in the coordinates of the SVD. From (3.25) and (3.26), we have

$$\begin{aligned}&\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2\\ &= \|\mathbf{U}^T(\mathbf{y}-\mathbf{H}\mathbf{x}_\lambda-\mathbf{H}\tilde{\mathbf{x}})\|_2^2\\ &= \|\mathbf{U}^T(\mathbf{y}-\mathbf{U}\mathbf{\Phi}\mathbf{U}^T\mathbf{y}-\\ &\mathbf{U}\mathbf{\Sigma}(\mathbf{\Sigma}^2+\lambda\mathbf{I})^{-1}\mathbf{V}^T\lambda\mathbf{vec}(Aver(\mathbf{c}^{(k)})))\|_2^2\\ &= \|(\mathbf{I}-\mathbf{\Phi})\mathbf{U}^T\mathbf{y}-\mathbf{\Sigma}(\mathbf{\Sigma}^2+\lambda\mathbf{I})^{-1}\mathbf{V}^T\lambda\mathbf{vec}(Aver(\mathbf{c}^{(k)}))\|_2^2\\ &= \sum_{i=1}^{M^2}\left(\frac{\lambda\mathbf{u}_i^T\mathbf{y}-\lambda\sigma_i\mathbf{v}_i^T\mathbf{vec}(Aver(c^{(k)}))}{\sigma_i^2+\lambda}\right)^2.\end{aligned} \tag{3.27}$$

Further,

$$\begin{aligned}&(trace(\mathbf{I}-\mathbf{H}\mathbf{V}\mathbf{\Phi}\mathbf{\Sigma}^{-1}\mathbf{U}^T))^2\\ &\qquad= (trace(\mathbf{I}-\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Phi}\mathbf{\Sigma}^{-1}\mathbf{U}^T))^2\\ &\qquad= (trace(\mathbf{U}(\mathbf{I}-\mathbf{\Phi})\mathbf{U}^T))^2\\ &\qquad= (trace(\mathbf{I}-\mathbf{\Phi}))^2\\ &\qquad= \left(\sum_{i=1}^{M^2}\frac{\lambda}{\lambda+\sigma_i^2}\right)^2.\end{aligned} \tag{3.28}$$

Hence, substituting the expressions from (3.27) and (3.28) into (3.24), we obtain

the GCV function for our manifold-based algorithm

$$G^{(k)}(\lambda) = \frac{\sum_{i=1}^{M^2} \left(\frac{\mathbf{u}_i^T \mathbf{y} - \sigma_i \mathbf{v}_i^T \mathbf{vec}(Aver(c^{(k)}))}{\sigma_i^2 + \lambda} \right)^2}{\left(\sum_{i=1}^{M^2} \frac{1}{\sigma_i^2 + \lambda} \right)^2}. \tag{3.29}$$

Note that the GCV function changes with every iteration and is thus indexed with k. This means that the optimal value of λ changes with every iteration. Hence, at each iteration we need to compute the best λ by evaluating the GCV function for various values of λ and choosing one that minimizes the GCV function. Thus,

$$\lambda_{optimal}^{(k)} = \arg\min_{\lambda} G^{(k)}(\lambda), \tag{3.30}$$

where $G^{(k)}(\lambda)$ is as given in (3.29).

3.6 Experimental Results

In this section, we present the results of our algorithm and compare them with various state-of-the-art methods: deconvolution based on sparsity prior in wavelet domain [3], hyper-Laplacian prior-based deconvolution [2], Fourier-Wavelet Regularized deconvolution (ForWaRD) [78], Anisotropic nonparametric image resotoration (LPA-ICI) [4] and Tikhonov deconvolution [101]. The regularization parameters for these methods are either chosen from a set of values within a wide range or set to be the optimal value reported in the corresponding papers. In the following experiments, we use the improvement in signal-to-noise-ratio (ISNR) as an criteria to compare the different methods. The ISNR is defined as

$$ISNR = 10 \log_{10} \left(\frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2} \right).$$

For an image of size $M \times N$, the BSNR is defined in decibels as

$$BSNR = 10 \log_{10} \left(\frac{\|\mathbf{Hx} - \mu(\mathbf{Hx})\|_2^2}{MN\sigma^2} \right),$$

where $\mu(\mathbf{Hx})$ represents the mean of $\mathbf{Hx}$.

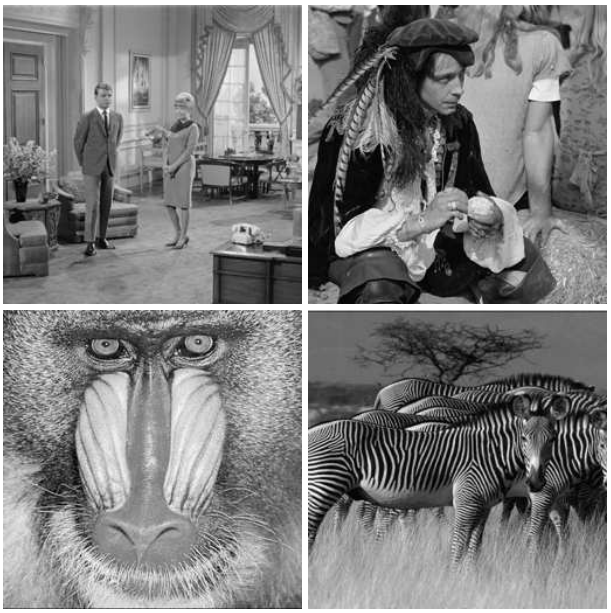
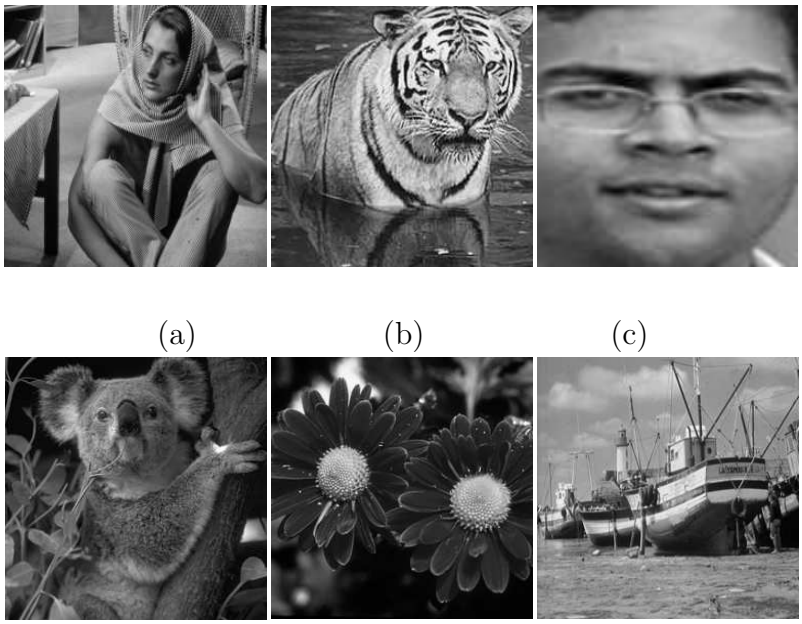


Figure 3.3: Some of the natural images used to learn the patch-manifold of natural images.

Fig. 3.3 shows some of the images used to learn the patch manifold for our algorithm. We randomly sample $22,500$ patches of size 4×4 from each image. So we have $112,500$ patches in total to learn the patch manifold. In Fig. 3.4, we display the test images used for different experiments in this paper.

In the first set of experiments, a *Barbara* image, shown in Fig. 3.4(a), is blurred by the following point spread function: $h(n_1, n_2) = (1 + n_1^2 + n_2^2)^{-1}$, for $n_1, n_2 = -7, ..., 7$. The AWGN variance σ^2 is chosen with a BSNR of 40 dB. The ISNR values obtained by different methods are compared in Table 3.2 under the Experiment 1



(a) (b) (c)

(d) (e) (f)

Figure 3.4: Images used in this chapter for different experiments. (a) *Barbara* image, (b) *Tiger* image, (c) a *face* image, (d) *Koala* image, (e) *Flowers* image and (f) *Boat* image.

column. The parametric and non-parametric manifold-based methods yield ISNR values of 7.98 dB and 7.95 dB respectively, which are better than the values obtained by any of the other methods. A portion of the image is zoomed in to reveal the visual detail of the results obtained by the different methods, and are shown in Fig. 3.5(a)-(f). As can be seen from the figure, our manifold-based method recovers details better than the other methods.

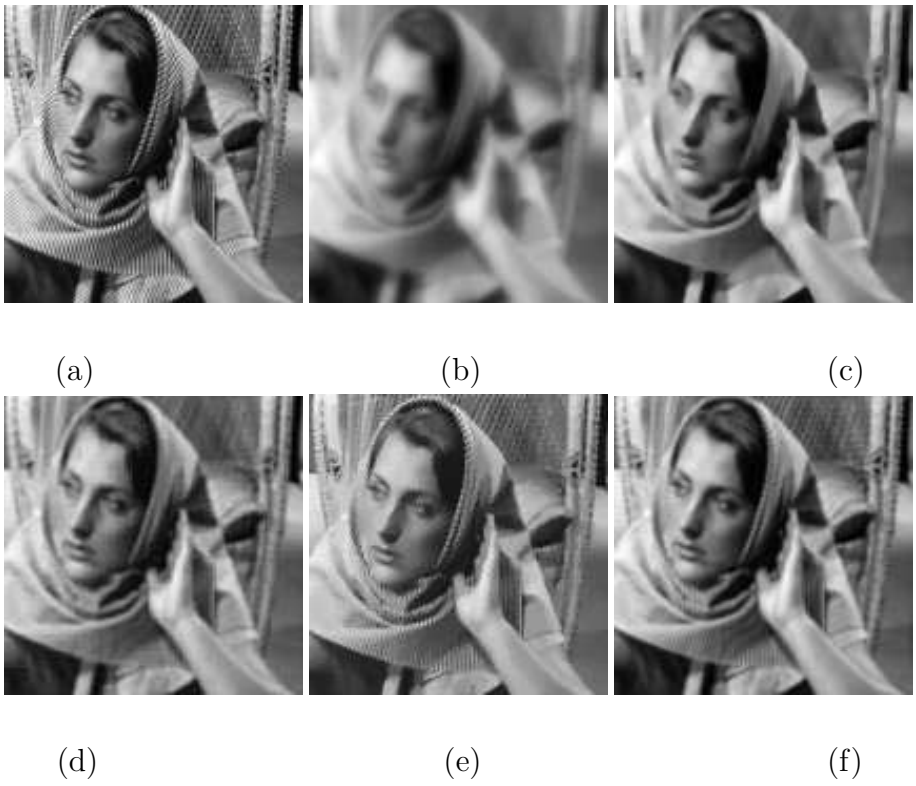


Figure 3.5: Details of the image deconvolution experiment with a *Barbara* image. (a) Original image. (b) Noisy blurred image. (c) Hyper-Laplacian [2] estimate (ISNR 5.19 dB). (d) Wavelet domain sparsity-based estimate [3] (ISNR 6.24 dB).(e)LPA-ICI [4] estimate (ISNR 7.88 dB) (f) Parametric manifold-based estimate (ISNR 7.98 dB) suggested in this chapter.

In the second set of experiments, the *Tiger* image, shown in Fig. 3.4(b), is blurred by a real-world camera shake kernel [108]. In this experiment, we choose the noise variance σ^2 with a BSNR of 30 dB. The simulation results are reported under the Experiment 2 column of Table 3.2. The deblurred image details obtained by the different methods are shown in Fig. 3.6(a)-(f). The blur PSF used in this experiment is shown in Fig. 3.6(g). The LPA-ICI method gives an ISNR value of 9.14 dB which is slightly better than our method. Note that the LPA-ICI method obtains the initial estimate using a local polynomial approximation method. To further enhance their performance, a regularized Wiener filtering (RW) is applied to the initial estimate. Similarly, we can enhance the performance of our algorithm by adapting RW filtering as a postprocessing step as was done in [78] and [80].

In the third set of tests, a *face* image is blurred by a Gaussian PSF defined as

$$h(i,j) = De^{-\frac{i^2+j^2}{2\eta^2}}$$

for $i, j = -5, ..., 5$, where D is a normalizing constant ensuring that the blur is of unit mass, and η^2 is the variance that determines the severity of the blur. Noise is added with a BSNR of 40 dB. The results are summarized under the Experiment 3 column of Table 3.2. Again, our manifold-based algorithm performs the best in terms of ISNR. A portion of the deblurred images from different methods are shown in Fig. 3.7(a)-(f).

In the fourth set of tests, the image of *Koala* is blurred by a separable filter [4] with weights $[1, 4, 6, 4, 1]/16$ in both the horizontal and vertical directions and the AWGN is added such that the BSNR value equals to 30 dB. The simulation

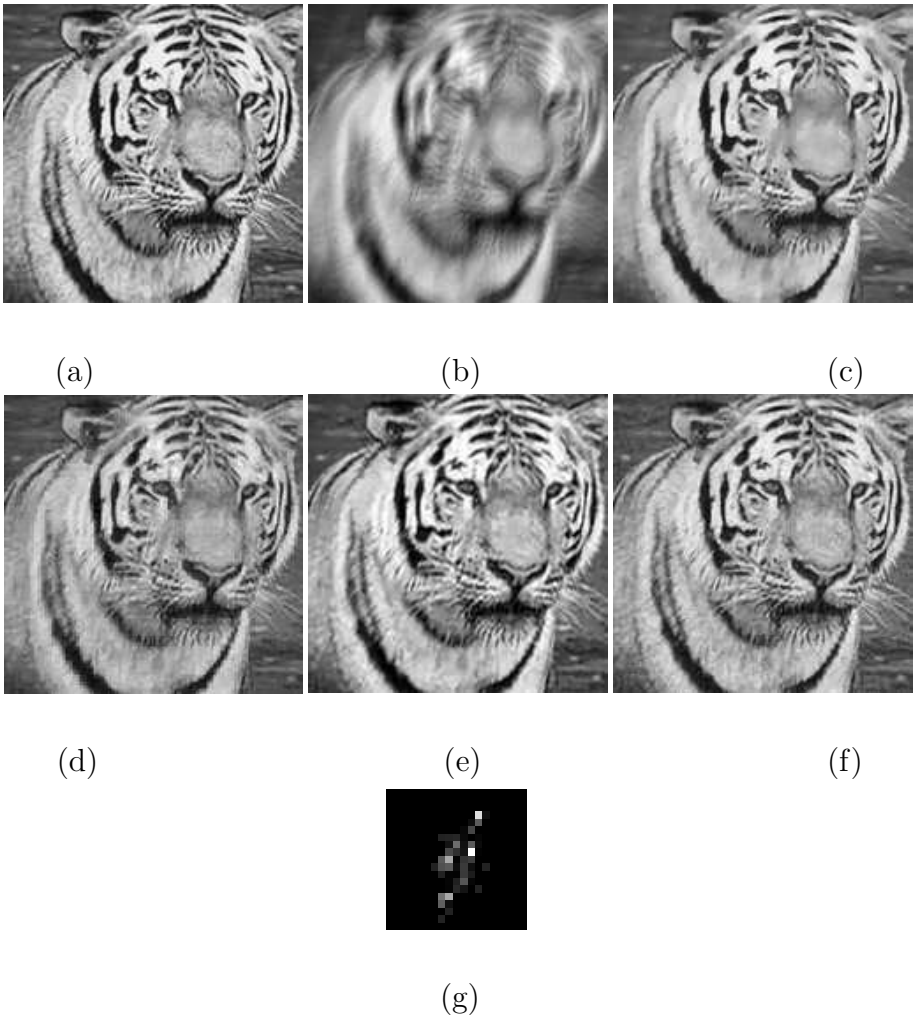
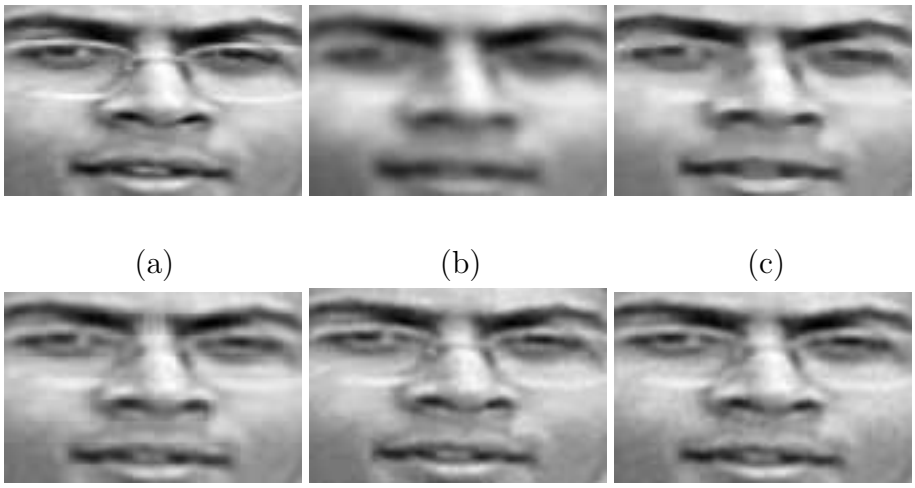


Figure 3.6: Details of the image deconvolution experiment with a *Tiger* image. (a) Original image. (b) Noisy blurred image. (c)Hyper-laplacian [2] estimate (ISNR 8.14 dB). (d) Wavelet domain sparsity-based estimate [3] estimate (ISNR 8.28 dB). (e) LPA-ICI [4] estimate (ISNR 9.14 dB) (f) Parametric manifold-based estimate suggested in this chapter (ISNR 9.02 dB). (g) Blur kernel.



(a) (b) (c)

(d) (e) (f)

Figure 3.7: Details of the image deconvolution experiment with a *face* image. (a) Original image. (b) Noisy blurred image. (c) Hyper-Laplacian [2] estimate (ISNR 5.16 dB). (d) Wavelet domain sparsity-based estimate [3] estimate (ISNR 6.1 dB). (e) LPA-ICI estimate [4] (ISNR 7.4 dB) (f) Parametric manifold-based estimate (ISNR 8.49 dB) suggested in this paper.

Table 3.2: ISNR for different experiments. The highest ISNR for each experiment is shown in bold.

	Experiments				
Method	Barbara	Tiger	Face	Koala	Flowers
Non-parametric Manifold-based deconvolution	7.95	8.96	8.27	**3.48**	**7.65**
Parametric Manifold-based deconvolution	**7.98**	9.02	**8.49**	3.21	**7.65**
Anisotropic Nonparametric Image Restoration	7.88	**9.14**	7.40	3.38	5.13
Fourier-Wavelet Regularized Deconvolution	7.6	9.02	7.74	3.04	7.4
Wavelet domain sparsity-based deconvolution	6.24	8.28	6.1	3.25	6.03
Hyper-laplacian prior-based deconvolution	5.19	8.14	5.16	2.74	5.39
Tikhonov deconvolution	3.04	4.26	4.39	1.02	4.64

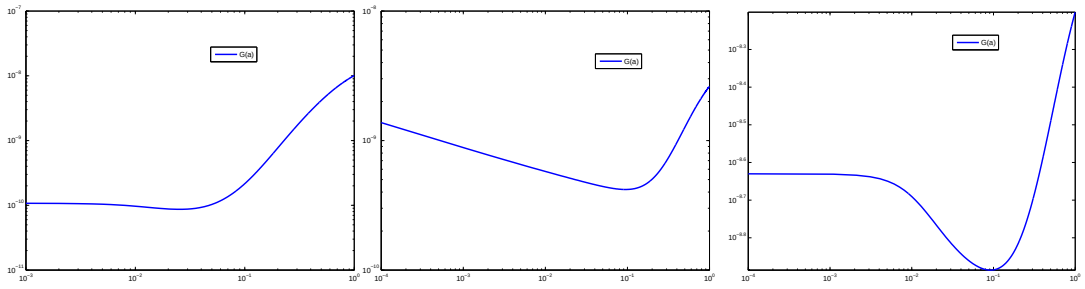
results are reported under the Experiment 4 column of Table 3.2. Both wavelet domain sparsity-based method and parametric manifold-based method perform approximately the same with ISNR values of 3.25 dB and 3.21 dB respectively. In this experiment, the non-parametric manifold-based algorithm performs the best with an ISNR value of 3.48 dB.

In the fifth experiment, we apply a horizontal motion blur kernel with length 7 on a *Flowers* image. For this experiment, the BNSR value is set to be 25 dB. The deconvolution results obtained by different methods are reported under the Experiment 5 column of Table 3.2. Both parametric and non-parametric manifold-based methods perform the same yielding an ISNR value of 7.65 dB. This experiment shows that, even in the case of low BSNR, our manifold-based method can provide better reconstruction than some of the competitive deconvolution methods.

In Fig. 3.8(a)-(c), we display a few of the GCV curves obtained from Experiment 1, 4 and 5, respectively. The minimizers of these GCV curves are chosen to be the regularization parameters in each experiment. Hence, unlike some of the other deconvolution algorithms such as [78], our method does not require the explicit knowledge of noise variance and it automatically determines the regularization parameter at each iteration.

In Fig. 3.9, we compare the values of ISNR of different methods as a function of the value of BSNR. For this experiment, we used a Gaussian blur on the *Barbara* image. As it is seen from the figure, the performance of the manifold-based method decreases slower than other methods when the noise level increases.

The stopping criterion for our method is usually that the norm of the difference



(a) (b) (c)

Figure 3.8: GCV function for regularization with manifold prior. (a) *Barbara* Experiment. (b) *Koala* Experiment. (c) *flowers* Experiment.

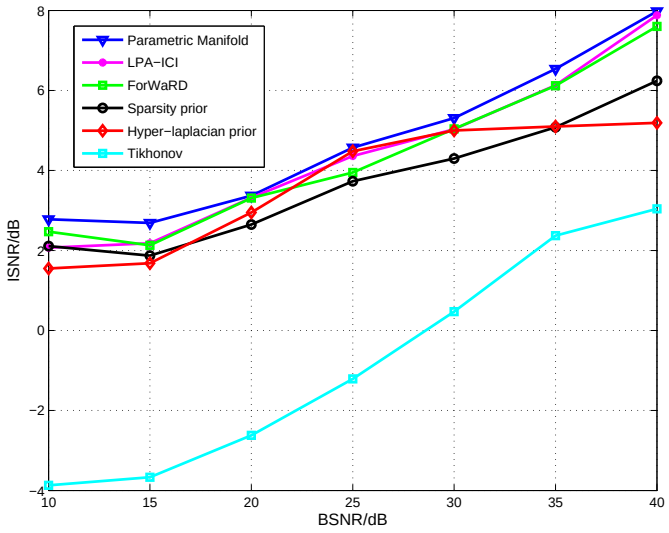


Figure 3.9: ISNR performance of manifold-based algorithm compared to other methods as a function of BSNR

between two successive estimates falls below a pre-specified threshold. That is, we stop when $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}\|_2^2 < 10^{-3}$. Empirical results show that our manifold-based methods typically converge in about 3 to 5 iterations. In Fig. 3.10, we plot the value of the data fidelity term as the number of iterations increases, for the case when a Gaussian blur is applied on the image shown in Fig. 3.4(c) with a BSNR value of 35 dB. As it is seen from the figure, our method converges in about 3 iterations and the difference between two successive estimates after 3 iterations is very small.

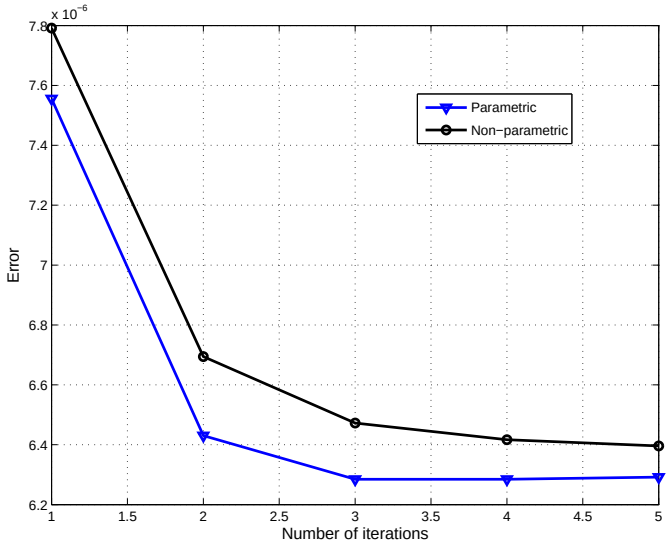


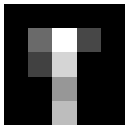
Figure 3.10: $\frac{1}{MN}\|\mathbf{y} - \mathbf{H}\mathbf{x}^{(k)}\|_2^2$ vs. number of iterations to determine the stopping criteria.

3.6.1 Blind Deconvolution

In many realistic applications, we don't have the form of the blur kernel. Hence, this requires blind deconvolution methods. It is stated in [108] that a robust blind deconvolution strategy is to first use the maximum a-posterior (MAP) estimate

to recover the blur kernel, and then use the recovered kernel to solve for the sharp image using a non-blind deconvolution algorithm. In this experiment, we employ this strategy and test the robustness of our method to small errors in blur-kernel estimation. We apply a 5×5 box-car blur on an image, as shown in Fig. 3.4(f), with BSNR of 35 dB. Fig. 3.11 shows the details of blind deconvolution result using the method proposed in [5] and the deconvolution result using parametric manifold method based on the blur kernel estimated by [5]. The ISNR values are -0.19 dB and 1.59 dB, respectively. We observe that our method can suppress the ringing artifacts and is more robust when the estimated kernel is not accurate enough.

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



(e)

Figure 3.11: Details of the blind deconvolution experiment with a *Boat* image. (a) Original image. (b) Blurred noisy image. (c) Result obtained by applying a blind deconvolution method in [5] (ISNR -0.19 dB). (d) Result obtained by applying the parametric manifold deconvolution method using blur kernel estimated from [5] (ISNR 1.59 dB). (e) Estimated kernel.

3.6.2 Computational Complexity

In our deconvolution method, the most computationally intensive part is to find the projection on the manifold. Using Matlab on a linux system with Intel Core 2.00 GHz/2.00 GB processor, projecting one patch onto a manifold formed by 112,500 patches using non-parametric manifold learning takes around 2.5e-2 seconds, while parametric manifold learning reduces the computation time to 5e-3 seconds. On average our algorithm takes about 3 minutes to process an image of size 256×256.

Based on the experimental results, we observe that using the parametric manifold gives similar performance as the non-parametric case, while the former is much more computationally efficient. Further, the computation can be made more efficient by making the sampling of the patch manifold more compact.

3.7 Discussions and Conclusion

In this chapter, we have presented a way of utilizing unlabeled image data to regularize the deconvolution problem. We formalized this via a patch-manifold prior for image classes which was shown to work very well for a wide variety of image content. This paves the way for interesting new directions of work. For example, using image formation models for specific cases, one could ask if there exist closed form expressions for the patch manifold. Further, several other inverse problems such as super-resolution, recovery of compressed signals, etc can be explored using example-driven priors. Finally, it would be interesting to fuse the example data with multi-view geometric constraints to better estimate the patch manifold with fewer

examples.

Chapter 4: Subspace Interpolation via Dictionary Learning for Unsupervised Domain Adaptation

4.1 Introduction

Traditional classification problems often assume that training and testing data are captured from the same underlying distribution. Yet this assumption is often violated in many real life applications. For instance, images collected from an internet search engine are compared with those captured from real life [109, 110]. Face recognition systems trained on frontal and high resolution images, are applied to probe images with non-frontal poses and low resolution [111]. Human actions are recognized from an unseen target view using training data taken from source views [112, 113]. We show some examples of dataset shifts in Figure 4.1.

In these scenarios, magnitudes of variations of innate characteristics, which distinguish one class from another, are oftentimes smaller than the variations caused by distribution shift between training and testing datasets. Directly applying the classifier from the training set to testing set will result in degraded performance. Therefore, it is essential to adapt classification systems to new environments. This is often known as the *domain adaptation* problem which has recently drawn much

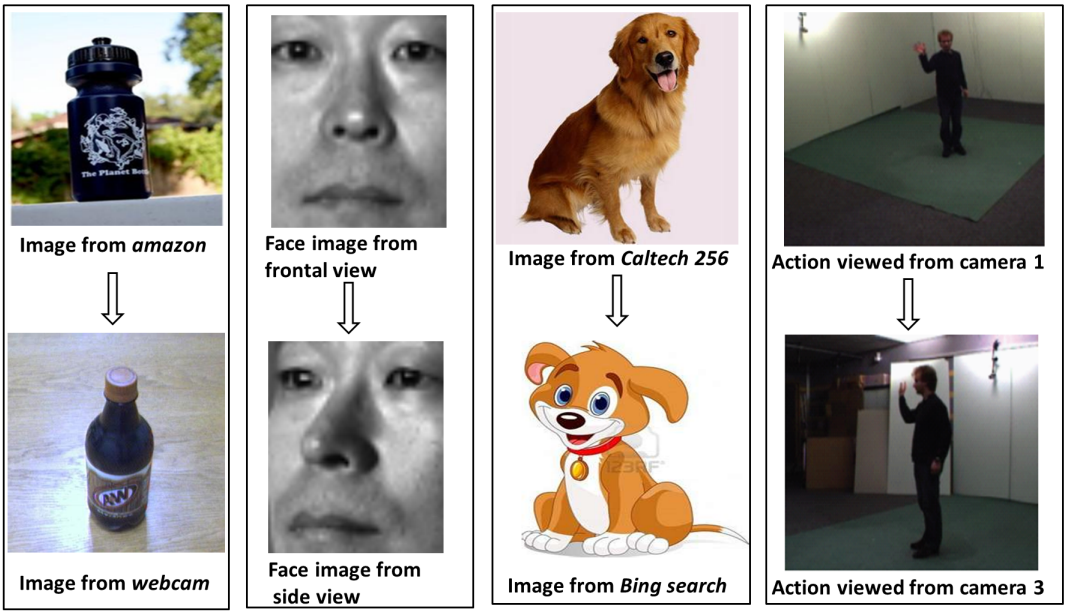


Figure 4.1: Examples of dataset shifts. Each column contains two images of the same subject collected under different conditions.

attention in the computer vision community [109, 17, 114, 115].

Domain Adaptation (DA) aims to utilize a *source domain* with plenty of labeled data to learn a classifier for the *target domain* which is collected from a different distribution. Based on the availability of labeled data in the target domain, DA methods can be classified into two categories: *semi-supervised* and *unsupervised* DA. Semi-supervised DA leverages few labels in the target data or correspondence between the source and target data to reduce the difference between two domains. Unsupervised DA is inherently a more challenging problem without any labeled target data to build association between the two domains. On the other hand, unsupervised DA is more representative of real world scenarios. For instance, face recognition systems trained under constrained laboratory environments will en-

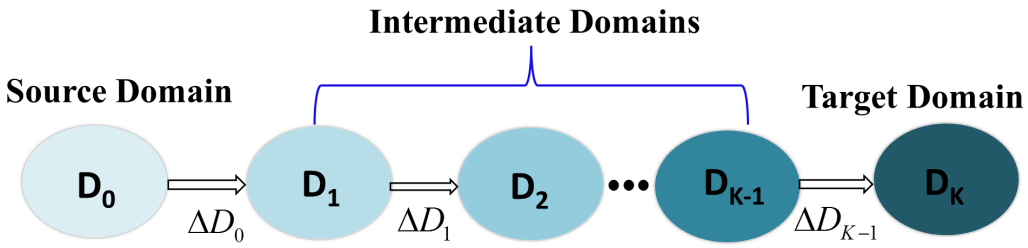


Figure 4.2: Given labeled data in the source domain and unlabeled data in the target domain, our DA procedure learns a set of intermediate domains (represented by dictionaries $\{\mathbf{D}_k\}_{k=1}^{K-1}$) and the target domain (represented by dictionary $\mathbf{D}_K$) to capture the intrinsic domain shift between two domains. $\{\Delta\mathbf{D}_k\}_{k=0}^{K-1}$ characterize the gradual transition between these subspaces.

counter great challenges when applied to faces 'in the wild', where the acquired face images suffer from a variety of degradations such as low resolution, poor illumination, blur, pose variation, occlusion etc [116]. Sometimes the coupling effects among these different factors give rise to more variations. As it is very costly to collect labels for target data under various acquisition conditions 'in the wild', it is more desirable that the recognition system be able to adapt in an unsupervised fashion.

An important class of unsupervised DA methods attempts to find suitable representations whose characteristics are shared between the two domains. In this chapter, we use subspace representations to model the source and target domains. Subspace modeling has been ubiquitous in the field of computer vision. This is due to the fact that data of high dimensionality usually lie on an intrinsically low-dimensional subspace. In this work, we use a dictionary to represent one domain, as dictionary learning based methods [117, 118] have recently become very popular for

subspace modeling. It is based on the fact that data signals in the same subspace can be linearly decomposed with a small number of atoms from an over-complete dictionary. Unlike traditional subspace modeling using PCA, these atoms are not constrained to be orthogonal, which allows more flexibility to better adapt to the given data signals [119]. The resulting sparse codes are usually leveraged as a feature representation for classification. Effectively learned dictionaries have seen state-of-the-art performance in reconstruction and recognition tasks [120, 20, 121].

Yet the issue of dictionary learning under distribution shifts has received less attention. Specifically, the presence of domain shifts violates the assumption that test data lie in the linear span of training data. As the dictionary atoms learned for one domain are not optimal for a different domain, and only a small subset of the atoms are allowed for representation, it will incur large reconstruction errors for the target data. Further, signals of the same class in the target domain will not have similar sparse codes as those from the source domain. These factors will cause inferior performance for both reconstruction and recognition tasks. Therefore, effectively leverage unlabeled target data to adapt the dictionary from one domain to another while maintaining certain invariant representation becomes crucial for successful DA.

In this chapter, we propose a novel unsupervised DA framework by interpolating subspaces through dictionary learning. We hypothesize the existence of a virtual path which smoothly connects the source and target domains. Imagine the source domain consists of face images in the frontal view while the target domain contains those in the profile view. Intuitively, face images which gradually transform from

the frontal to profile view will form a smooth transition path. Recovering intermediate representations along the transition path allows us to more likely capture the underlying domain shift, as well as to build meaningful feature representations which are preserved across different domains. We encapsulate this intuition into our approach. Specifically, we sample several intermediate domains along a virtual path between the source and target domains, and represent each intermediate domain using a dictionary. We then utilize the good reconstruction property of dictionaries, and learn the set of intermediate domain dictionaries which incrementally reduce the reconstruction residue of the target data. In the mean time, we constrain the magnitude of changes between dictionaries for adjacent intermediate domains to ensure the smoothness of the transition path (refer to Figure 4.2 for an illustration). We then apply invariant sparse codes across the source, intermediate and target domains to render intermediate representations, which convey a smooth transition in the data signal space. It also provides a shared feature representation where the sample differences caused by distribution shifts are reduced, and we utilize this new feature representation for cross domain recognition. Sometimes, we may be faced with multiple source domains. In order to select the optimal source domain to perform adaptation, we provide a quantitative measure of domain shift by measuring the similarity between the source and target domain dictionaries which are learned using our proposed DA approach. Further, we extend our framework to nonlinear cases by learning the set of intermediate domain dictionaries in the high dimensional feature space. We demonstrate the wide applicability of our approach for face recognition across pose, illumination and blur variations, cross dataset object

recognition, and report improved performance over existing DA methods.

4.2 Prior work

Several DA methods have been discussed in the literature. We briefly review the relevant work below.

Semi-supervised DA methods rely on labeled target data to perform cross domain classification. Daume [122] proposed a feature augmentation technique such that data points from the same domain are more similar than those from different domains. The Adaptive-SVM method introduced in [123] selects the most effective auxiliary classifiers to adapt to the target dataset. The method in [124] designed a cross-domain classifier based on multiple base kernels. Metric learning approaches [109, 125] were also proposed to learn a cross domain transformation to link two domains. Recently, Jhuo et al. [115] utilized low-rank reconstructions to learn a transformation so that the transformed source samples can be linearly reconstructed by the target samples.

Given no labels in the target domain to learn the similarity measure between data instances across domains, unsupervised DA is more difficult to tackle. Therefore it usually enforces certain prior assumptions to relate source and target data. Structural correspondence learning [15] induces correspondence among features from two domains by modeling their relations with *pivot* features, which appear frequently in both domains. Manifold alignment based DA [16] computes similarity between data points in different domains through the local geometry of data points

within each domain. The techniques in [126, 127] reduce the distance across the two domains by learning a latent feature space where domain similarity is measured through maximum mean discrepancy. Shi and Sha [128] define an information-theoretic measure which balances between maximizing domain similarity and minimizing expected classification error on the target domain. Two recent approaches [17], [114] are more relevant to our methodology, where the source and target domains are linked by sampling finite or infinite number of intermediate subspaces on the Grassmannian manifold. These intermediate subspaces appear to be able to capture the intrinsic domain shift. Compared to their abstract manifold walking strategies, our approach emphasizes on synthesizing intermediate subspaces in a manner which gradually reduces the reconstruction residue of the target data.

Also related is the recent work presented in [129], which jointly learns aligned dictionaries from multiple domains with correspondence available in those domains. Domain invariant sparse codes are designed for cross domain recognition, alignment and synthesis. Our DA approach differs in that we can operate in the unsupervised mode where no correspondence is available.

4.3 Sparse Representation and Dictionary Learning

As discussed in previous sections, we use dictionaries and sparse representation for signal representation in this work. A simple but important property of sparse representation is that: in many applications, data of high dimensionality exhibits *degenerate structure*, i.e., they lie on or near low-dimensional subspaces,

sub-manifolds, or stratifications [20]. Therefore, given a collection of representative data, a typical data is expected to have a sparse representation with respect to the given basis. In this section, we present some preliminaries and introduce some common techniques in sparse representation.

Sparse and redundant representation aims to represent a data signal y as linear combinations of a few atoms from an over-complete dictionary $\mathbf{D} \in \mathcal{R}^{n \times m}$,$m > n$. The representation can be either exact:

$$y = \mathbf{D}x$$

or be an approximation:

$$\|y - \mathbf{D}x\|_p \leq \epsilon$$

4.3.1 Sparse Coding

A fundamental step in sparse representation is sparse coding, which finds the representation coefficients x given the data signal y and the dictionary $\mathbf{D}$:

$$\min_x \|x\|_0, s.t. \|y - \mathbf{D}x\|_2 \leq \epsilon$$

This is a NP-hard problem. Classical methods tackle this problem by greedily selecting columns of $\mathbf{D}$ and forming successively better approximations to y. Among them two commonly used methods are Matching Pursuit [130] and the Orthogonal Matching Pursuit [131].

Matching Pursuit (MP): Matching pursuit is an iterative greedy algorithm that decomposes a signal into a linear combination of elements from a dictionary. A

key element in MP is the residue t, which is the as-yet "unexplained" portion of the measurements. In each iteration, a vector from the dictionary which is maximally correlated with the residue is selected. The algorithm stops when the residue is below some quantity. The pseudo-code of MP is provided in Algorithm 1.

Algorithm 1 Matching Pursuit

Require: Dictionary $\mathbf{D}$, data signal y

return Sparse coefficient x

Initialize $r^0 = y, x^0 = 0, n = 0$

while stopping criterion is not met **do**

$n \leftarrow n + 1$

$g^n = \mathbf{D}^T r^{n-1}$

$i^n = \underset{i}{argmax} |g_i^n|$

$x_{i^n}^n = x_{i^n}^{n-1} + g_{i^n}^n$

$r^n = r^{n-1} - \mathbf{D}_{i^n} g_{i^n}^n$

end while

Orthogonal Matching Pursuit (OMP): As the complexity of MP increases linearly with the number of iterations, it can be computationally infeasible for many problems. Orthogonal Matching Pursuit is a simple modification of MP such that the maximum number of iterations can be upper bounded. In each iteration, it computes the projection of residue r onto the orthogonal subspace to the linear span of the currently selected dictionary elements. This quantity better represents the unexplained portion of the residue. Its pseudo-code is summarized in Algorithm 2, where $\mathbf{D}_{\Gamma^n}$ represents a sub-matrix of $\mathbf{D}$ containing only those columns of $\mathbf{D}$ with

Algorithm 2 Orthogonal Matching Pursuit

Require: Dictionary $\mathbf{D}$, data signal y

return Sparse coefficient x

Initialize $r^0 = y, x^0 = 0, \Gamma^0 = \emptyset, n = 0$

while stopping criterion is not met **do**

$n \leftarrow n + 1$

$g^n = \mathbf{D}^T r^{n-1}$

$i^n = \underset{i}{argmax} |g_i^n|$

$\Gamma^n = \Gamma^{n-1} \cup i^n$

$x^n = \mathbf{D}_{\Gamma^n}^+ y$

$r^n = y - \mathbf{D}x^n$

end while

indices in Γ^n, and $\mathbf{D}_{\Gamma^n}^+$ is the pseudo-inverse of $\mathbf{D}_{\Gamma^n}$.

4.3.2 Design of Dictionaries

The choice of a dictionary is crucial for a successful vision application. While off-the-shell/pre-specified dictionaries such as DCT, Gabor and wavelet are simple and efficient, a trained dictionary is more appealing as it can adapt the dictionary to specific applications. Dictionary learning can be formalized as:

$$\min_{\mathbf{D}, \{x_i\}_{i=1}^M} \sum_{i=1}^{M} \|y_i - \mathbf{D}x_i\|_2^2, s.t. \|x_i\|_0 \leq K, 1 \leq i \leq M$$

We list a few representative dictionary learning techniques in the following.

Method of Optimal Directions: The Method of Optimal Directions (MOD)

Algorithm 3 Method of Optimal Directions

Require: Data signals $\mathbf{Y}$

return Optimal Dictionary $\mathbf{D}_{(k)}$, Sparse coefficient $\mathbf{X}$

Initialize $\mathbf{D}_{(0)} \in \mathbf{R}^{n \times m}$ using random entries , $k = 0$

while $\|\mathbf{Y} - \mathbf{D}_{(k)}\mathbf{X}_{(k)}\|_F^2 > \epsilon$ **do**

$k \leftarrow k + 1$

Sparse coding: $\hat{x}_i = \arg\min_x \|y_i - \mathbf{D}_{(k-1)}x\|_2^2, s.t. \|\hat{x}_i\|_0 \leq K, 1 \leq i \leq M$. Form the matrix $\mathbf{X}_{(k)} = [\hat{x}_1, \hat{x}_2, ..., \hat{x}_M]$

Dictionary updating: $\mathbf{D}_{(k)} = \arg\min_{\mathbf{D}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}_{(k)}\|_F^2 = \mathbf{Y}\mathbf{X}_{(k)}^T(\mathbf{X}_{(k)}\mathbf{X}_{(k)}^T)^{-1}$

end while

uses a Block-Coordinate-Relaxation algorithm which was proposed by Engan et.al [132]. It alternates between sparse coding and dictionary updating steps. In each iteration, the dictionary is updated by solving a least squares minimization problem where the error is evaluated using Frobenius norm. The iterations are continued until a convergence criteria is reached. Let $\mathbf{Y} = [y_1, y_2, ..., y_M]$, $\mathbf{X} = [x_1, x_2, ..., x_M]$, Algorithm 3 gives the summarization of MOD [133].

The K-SVD Algorithm: The K-SVD is similar to MOD, except in the dictionary updating stage. Instead of using matrix inversion, K-SVD performs an atom-by-atom updating in a simple and efficient fashion. We present the dictionary updating part [133] of the K-SVD algorithm in Algorithm 4, where $x_T^{j_0}$ is defined as the j_0th row in the sparse coefficient matrix $\mathbf{X}$.

Algorithm 4 K-SVD Algorithm

Require: Data signals $\mathbf{Y}$

return Optimal Dictionary $\mathbf{D}_{(k)}$, Sparse coefficient $\mathbf{X}$

Initialize $\mathbf{D}_{(0)} \in \mathbf{R}^{n \times m}$ using random entries , $J = 1$

while convergence is not reached **do**

Sparse coding: $\hat{x}_i = \arg\min_x \|y_i - \mathbf{D}_{(J-1)}x\|_2^2, s.t. \|\hat{x}_i\|_0 \leq K, 1 \leq i \leq M.$

Column-wise Dictionary updating:

for $j_0 = 1, 2, ..., m$ in $\mathbf{D}_{J-1}$ **do**

Define the group of signals which use the atom d_{j_0}: $\Omega_{j_0} = \{i | 1 \leq i \leq M, x_T^{j_0}(i) \neq 0\}$.

Compute the residual matrix $\mathbf{E}_{j_0} = \mathbf{Y} - \sum_{j \neq j_0} d_j x_j^T$.

Restrict $\mathbf{E}_{j_0}$ by choosing only the columns corresponding to Ω_{j_0} to obtain $\mathbf{E}_{j_0}^R$.

Apply SVD decomposition $\mathbf{E}_{j_0}^R = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathbf{T}}$. Choose the updated dictionary column d_{j_0} to be the first column of $\mathbf{U}$. Update the coefficient vector $x_R^{j_0}$ to be the first column of $\mathbf{V}$ multiplied by $[\Delta](1,1)$.

end for

Set $J = J + 1$

end while

4.4 Learning Intermediate Domains for Unsupervised Domain Adaptation

In this section, we introduce our general framework for unsupervised DA in details. We first describe some notations to facilitate subsequent discussions.

Let $\mathbf{Y}_s \in \mathbb{R}^{n*N_s}$, $\mathbf{Y}_t \in \mathbb{R}^{n*N_t}$ be the data instances from the source and target domain respectively, where n is the dimension of the data instance, N_s and N_t denote the number of samples in the source and target domains. Let $\mathbf{D}_0 \in \mathbb{R}^{n*m}$ be the dictionary learned from $\mathbf{Y}_s$ using standard dictionary learning methods, e.g, K-SVD [117], where m denotes the number of atoms in the dictionary. As introduced in Section 4.1, our approach samples several intermediate domains from a smooth transition path between the source and target domains. We associate each intermediate domain with a dictionary $\mathbf{D}_k, k \in [1, K]$, where K is the number of intermediate domains which will be determined in our DA approach.

4.4.1 Learning Intermediate Domain Dictionaries

Starting from the source domain dictionary $\mathbf{D}_0$, we sequentially learn the intermediate domain dictionaries $\{\mathbf{D}_k\}_{k=1}^{K}$ to gradually adapt to the target data. This is also conceptually similar to incremental learning. The final dictionary $\mathbf{D}_K$ which best represents the target data in terms of reconstruction error is taken as the target domain dictionary. Given the k-th domain dictionary $\mathbf{D}_k, k \in [0, K-1]$, we learn the next domain dictionary $\mathbf{D}_{k+1}$ based on its coherence with $\mathbf{D}_k$ and the remaining

residue of the target data. Specifically, we decompose the target data $\mathbf{Y}_t$ with $\mathbf{D}_k$ and get the reconstruction residue $\mathbf{J}_k$:

$$\mathbf{\Gamma}_k = \arg\min_{\mathbf{\Gamma}} \|\mathbf{Y}_t - \mathbf{D}_k\mathbf{\Gamma}\|_F^2, s.t. \forall i, \|\alpha_i\|_0 \leq T \tag{4.1}$$

$$\mathbf{J}_k = \mathbf{Y}_t - \mathbf{D}_k\mathbf{\Gamma}_k \tag{4.2}$$

where $\mathbf{\Gamma}_k = [\alpha_1, ..., \alpha_{N_t}] \in \mathbb{R}^{m*N_t}$ denote the sparse coefficients of $\mathbf{Y}_t$ decomposed with $\mathbf{D}_k$, and T is the sparsity level. We then obtain $\mathbf{D}_{k+1}$ by estimating $\Delta\mathbf{D}_k$, which is the adjustment in the dictionary atoms between $\mathbf{D}_{k+1}$ and $\mathbf{D}_k$:

$$\min_{\Delta\mathbf{D}_k} \|\mathbf{J}_k - \Delta\mathbf{D}_k\mathbf{\Gamma}_k\|_F^2 + \lambda\|\Delta\mathbf{D}_k\|_F^2 \tag{4.3}$$

Equation (4.3) consists of two terms. The first term ensures that the adjustments in the atoms of $\mathbf{D}_k$ will further decrease the current reconstruction residue $\mathbf{J}_k$. The second term penalizes abrupt changes between adjacent intermediate domains, so as to obtain a smooth path. The parameter λ controls the balance between these two terms. This is a ridge regression problem. By setting the first order derivatives to be zeros, we obtain the following closed form solution:

$$\Delta\mathbf{D}_k = \mathbf{J}_k\mathbf{\Gamma}_k^T(\lambda\mathbf{I} + \mathbf{\Gamma}_k\mathbf{\Gamma}_k^T)^{-1} \tag{4.4}$$

where $\mathbf{I}$ is the identity matrix. The next intermediate domain dictionary $\mathbf{D}_{k+1}$ is then obtained as:

$$\mathbf{D}_{k+1} = \mathbf{D}_k + \Delta\mathbf{D}_k \tag{4.5}$$

Note that when $\lambda = 0$, the Method of Optimal Direction (MOD) [132] becomes a special case of (4.3), where no regularization is enforced.

Starting from the source domain dictionary $\mathbf{D}_0$, we apply the above adaptation framework iteratively, and stop the procedure when the magnitude of $\|\Delta\mathbf{D}_k\|_F$ is below certain threshold, so that the gap between the two domains is absorbed into the learned intermediate domain dictionaries. This stopping criteria also automatically gives the number of intermediate domains to sample from the transition path. We summarize our approach in Algorithm 5. We also show in Proposition 1 that, in each step, the residue $\mathbf{J}_k$ is non-increasing w.r.t the current intermediate domain dictionary and the encoding coefficients. We demonstrate the empirical convergence of our algorithm in Section 4.6.

Proposition 1 *Given the estimate of* $\Delta\mathbf{D}_k$ *using equation* (4.4)*, the residue* $\mathbf{J}_k$ *is non-increasing w.r.t* $\mathbf{D}_k$ *and the corresponding sparse coefficients* $\mathbf{\Gamma}_k$

$$\|\mathbf{J}_k - \Delta\mathbf{D}_k\mathbf{\Gamma}_k\|_F^2 \leq \|\mathbf{J}_k\|_F^2 \tag{4.6}$$

Proof: Substitute (4.4) *into* (4.6)*, we have*

$$\begin{aligned}
&\|\mathbf{J}_k\|_F^2 - \|\mathbf{J}_k - \Delta\mathbf{D}_k\mathbf{\Gamma}_k\|_F^2 \\
=&\|\mathbf{J}_k\|_F^2 - \|\mathbf{J}_k - \mathbf{J}_k\mathbf{\Gamma}_k^T(\lambda\mathbf{I} + \mathbf{\Gamma}_k\mathbf{\Gamma}_k^T)^{-1}\mathbf{\Gamma}_k\|_F^2 \\
=&tr(2\mathbf{\Gamma}_k^T(\lambda\mathbf{I} + \mathbf{\Gamma}_k\mathbf{\Gamma}_k^T)^{-1}\mathbf{\Gamma}_k\mathbf{J}_k^T\mathbf{J}_k) - \\
&tr(\mathbf{\Gamma}_k^T(\lambda\mathbf{I} + \mathbf{\Gamma}_k\mathbf{\Gamma}_k^T)^{-1}\mathbf{\Gamma}_k\mathbf{J}_k^T\mathbf{J}_k\mathbf{\Gamma}_k^T(\lambda\mathbf{I} + \mathbf{\Gamma}_k^T\mathbf{\Gamma}_k)^{-1}\mathbf{\Gamma}_k)
\end{aligned} \tag{4.7}$$

Let us define the Singular Value Decomposition (SVD) of $\mathbf{\Gamma}_k$ *as* $\mathbf{\Gamma}_k = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$*, where* $\mathbf{U}$ *and* $\mathbf{V}$ *are orthogonal matrices, and* $\mathbf{\Sigma} = [\tilde{\mathbf{\Sigma}}, \mathbf{0}]$ *is a rectangular diagonal matrix,*

with $\tilde{\mathbf{\Sigma}} = diag(\sigma_i)$ *being a diagonal matrix. Then*

$$\begin{aligned}
&\mathbf{\Gamma}_k^T(\lambda\mathbf{I} + \mathbf{\Gamma}_k\mathbf{\Gamma}_k^T)^{-1}\mathbf{\Gamma}_k \\
=&\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T(\lambda\mathbf{I} + \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{U}^T)^{-1}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \\
=&[\mathbf{V}_1, \mathbf{V}_2]\mathbf{\Sigma}^T\mathbf{U}^T(\lambda\mathbf{I} + \mathbf{U}\tilde{\mathbf{\Sigma}}^2\mathbf{U}^T)^{-1}\mathbf{U}\mathbf{\Sigma}[\mathbf{V}_1, \mathbf{V}_2]^T \\
=&\mathbf{V}_1\tilde{\mathbf{\Sigma}}(\lambda\mathbf{I} + \tilde{\mathbf{\Sigma}}^2)^{-1}\tilde{\mathbf{\Sigma}}\mathbf{V}_1^T \\
=&\mathbf{V}_1\mathbf{\Phi}\mathbf{V}_1^T
\end{aligned} \tag{4.8}$$

where $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2]$*, with* $\mathbf{V}_1$ *being a square matrix, and* $\mathbf{\Phi} = diag(\frac{\sigma_i^2}{\sigma_i^2+\lambda})$*. Substitute* (4.8) *into* (4.7)*, we have*

$$\begin{aligned}
&\|\mathbf{J}_k\|_F^2 - \|\mathbf{J}_k - \Delta\mathbf{D}_k\mathbf{\Gamma}_k\|_F^2 \\
=&tr(2\mathbf{V}_1\mathbf{\Phi}\mathbf{V}_1^T\mathbf{J}_k^T\mathbf{J}_k) - tr(\mathbf{V}_1\mathbf{\Phi}\mathbf{V}_1^T\mathbf{J}_k^T\mathbf{J}_k\mathbf{V}_1\mathbf{\Phi}\mathbf{V}_1^T) \\
=&tr((2\mathbf{\Phi} - \mathbf{\Phi}^2)\mathbf{V}_1^T\mathbf{J}_k^T\mathbf{J}_k\mathbf{V}_1) \\
=&tr(\mathbf{H}\mathbf{V}_1^T\mathbf{J}_k^T\mathbf{J}_k\mathbf{V}_1\mathbf{H}) \\
=&\|\mathbf{J}_k\mathbf{V}_1\mathbf{H}\|_F^2 \geq 0
\end{aligned} \tag{4.9}$$

where $\mathbf{H} = diag(\frac{\sqrt{\sigma_i^4+2\lambda\sigma_i^2}}{\sigma_i^2+\lambda})$

4.4.2 Recognition Under Domain Shift

Up to now, we have learned a transition path which is encoded with the underlying domain shift. This provides us with rich information to obtain new representations to associate source and target data. Here, we simply apply invariant sparse codes across the source, intermediate, target domain dictionaries $\{\mathbf{D}_k\}_{k=0}^K$.

Algorithm 5 Algorithm for subspace interpolation between source and target domains through dictionary learning (SIDL).

1: Input: Dictionary $\mathbf{D}_0$ trained from the source data, target data $\mathbf{Y}_t$, sparsity level T, stopping threshold δ, parameter λ, $k = 0$.

2: Output: Dictionaries $\{\mathbf{D}_k\}_{k=1}^{K-1}$ for the intermediate domains, dictionary $\mathbf{D}_K$ for the target domain.

3: **while** stopping criteria is not reached **do**

4: Decompose the target data with the current intermediate domain dictionary $\mathbf{D}_k$, get the reconstruction residue $\mathbf{J}_k$ using (4.1) and (4.2)

5: Get an estimate of the adjustment in dictionary atoms $\Delta\mathbf{D}_k$ and the next intermediate domain dictionary $\mathbf{D}_{k+1}$ using (4.4) and (4.5). Normalize the atoms in $\mathbf{D}_{k+1}$ to have unit norm.

6: $k \leftarrow k + 1$

7: check the stopping criteria $\|\Delta\mathbf{D}_k\|_F \leq \delta$

8: **end while**

The new augmented feature representation is obtained as follows:

$$[(\mathbf{D}_0\alpha)^T, (\mathbf{D}_1\alpha)^T, ..., (\mathbf{D}_K\alpha)^T]^T$$

where $\alpha \in \mathbb{R}^m$ is the sparse code of a source data signal decomposed with $\mathbf{D}_0$, or a target data signal decomposed with $\mathbf{D}_K$. This new representation incorporates the smooth domain transition recovered in the intermediate dictionaries into the signal space. It brings the source and target data into a shared feature space where the data distribution shift is mitigated. Therefore, it can serve as a more robust

characteristic across different domains. Given the new feature vectors, we apply PCA for dimension reduction[1], and then employ a SVM classifier for cross domain recognition.

4.4.3 Quantification of Domain Shift

We now introduce a metric, Quantification of Domain Shift (QDS) to compare the similarity of two domains, which has much practical utility. For instance, we may be faced with more than one source domains in some scenarios. QDS will allow us to select the optimal source domain which has the least domain shift w.r.t the target domain to perform adaptation. We propose to obtain QDS by measuring the similarity between the source domain dictionary $\mathbf{D}_0$ and the target domain dictionary $\mathbf{D}_K$ which is learned using Algorithm 5. This similarity measure characterizes the amount of domain shift encoded along the transition path. Specifically, it is defined as $Q_{s,t} = \|\mathbf{D}_K^T \mathbf{D}_0\|_F$, where a higher value indicates higher coherence between $\mathbf{D}_0$ and $\mathbf{D}_K$, and less domain shift along the learned transition path. Similarly, by reversing the role of source and target domain to learn the transition path, we can obtain $Q_{t,s}$ which is the amount of shift from target to source domain. Then the symmetric QDS between two domains is defined as $(1/2)(Q_{s,t} + Q_{t,s})$.

[1]The number of principal components is chosen to preserve 98% of the input data's energy. Alternatively, one can choose any other dimension reduction method for this step.

4.5 Nonlinear Dictionary Learning for Unsupervised Domain Adaptation

The unsupervised DA framework introduced in Section 4.4 uses a linear dictionary to represent a domain, which may not be sufficient to capture the non-linearity presented in the input data. In this section, we extend our DA framework by learning the set of intermediate dictionaries in a high dimensional RKHS induced by a Mercer kernel mapping.

4.5.1 Learning Nonlinear Intermediate Domain Dictionaries

Let $\Phi : \mathbb{R}^n \rightarrow \mathcal{F}$ be a mapping from $\mathbb{R}^n$ into a dot product space $\mathcal{F}$. We adopt the model in [134] to represent the kth intermediate domain dictionary as follows:

$$\Phi(\mathbf{D}_k) = \Phi(\mathbf{Y})\mathbf{A}_k = \Phi(\mathbf{Y}_s)\mathbf{A}_{sk} + \Phi(\mathbf{Y}_t)\mathbf{A}_{tk} \tag{4.10}$$

where $\Phi(\mathbf{Y}) = [\Phi(\mathbf{Y}_s)\ \Phi(\mathbf{Y}_t)]$ serves as a base dictionary, and $\mathbf{A}_k = \left[\begin{smallmatrix}\mathbf{A}_{sk}\\ \mathbf{A}_{tk}\end{smallmatrix}\right]$ is the atom representation matrix for the kth intermediate domain. This model allows the intermediate dictionary lie in the linear span of the samples $\Phi(\mathbf{Y})$ and provides adaptive representation via modification of the matrix $\mathbf{A}_k$. The base dictionary $\Phi(\mathbf{Y})$ implicitly incorporates the prior knowledge that the intermediate domain consists of a mixture representation of the source and target. Further, we penalizes the magnitude of differences between two adjacent atom representation matrixes so that the resulting intermediate domains represent the incremental domain shift by gradually varying the proportions of source and target combinations.

We first initialize source dictionary $\Phi(\mathbf{D}_0)$ by solving the following problem using the kernel KSVD algorithm [134]:

$$(\mathbf{A}_0, \mathbf{\Gamma}_0) = \arg\min_{\mathbf{A},\mathbf{\Gamma}} \|\Phi(\mathbf{Y}_s) - \Phi(\mathbf{Y})\mathbf{A}_0\mathbf{\Gamma}\|_F^2,\ s.t., \forall i, \|\alpha_i\|_0 \leq T \tag{4.11}$$

where $\mathbf{A}_0$ denotes the representation matrix for the source dictionary. Details of optimization can be found in [134]. Using the same notations as in (4.1) and (4.2), we obtain the reconstruction residue of the target data $\Phi(\mathbf{J}_k)$ given the kth intermediate dictionary in the feature space $\mathcal{F}$ as follows:

$$\mathbf{\Gamma}_k = \arg\min_{\mathbf{\Gamma}} \|\Phi(\mathbf{Y}_t) - \Phi(\mathbf{Y})\mathbf{A}_k\mathbf{\Gamma}\|_F^2,\ s.t. \forall i, \|\alpha_i\|_0 \leq T \tag{4.12}$$

$$\Phi(\mathbf{J}_k) = \Phi(\mathbf{Y}_t) - \Phi(\mathbf{Y})\mathbf{A}_k\mathbf{\Gamma}_k \tag{4.13}$$

where (4.12) can be solved using the kernel orthogonal matching pursuit algorithm [134]. Similar to (4.4), we then solve for the next intermediate domain dictionary in the feature space $\mathcal{F}$ by estimating $\Delta\mathbf{A}_k$, the difference between two adjacent atom representation matrixes:

$$\min_{\Delta\mathbf{A}_k} \|\Phi(\mathbf{J}_k) - \Phi(\mathbf{Y})\Delta\mathbf{A}_k\mathbf{\Gamma}_k|_F^2 + \lambda\|\Phi(\mathbf{Y})\Delta\mathbf{A}_k\|_F^2 \tag{4.14}$$

(4.14) has a closed form solution by setting the first order derivatives of $\Delta\mathbf{A}_k$ to be zeroes:

$$\Delta\mathbf{A}_k = (\mathcal{K}(\mathbf{Y},\mathbf{Y})^{-1}\mathcal{K}(\mathbf{Y},\mathbf{Y}_t) - \mathbf{A}_k\mathbf{\Gamma}_k)\mathbf{\Gamma}_k^T(\mathbf{\Gamma}_k\mathbf{\Gamma}_k^T + \lambda\mathbf{I})^{-1} \tag{4.15}$$

where $\mathcal{K}(\mathbf{Y},\mathbf{Y})$ is a kernel matrix whose entries are computed as:

$$k(i,j) = \Phi(\mathbf{y}_i)^T \Phi(\mathbf{y}_j)$$

The kernel matrix only requires dot products, which can be computed using the Mercer kernel function, instead of explicitly carrying out the mapping Φ. We then obtain the next intermediate dictionary in the kernel space as follows:

$$\Phi(\mathbf{D}_{k+1}) = \Phi(\mathbf{Y})(\mathbf{A}_k + \Delta\mathbf{A}_k) \tag{4.16}$$

We summarize our nonlinear intermediate dictionary learning framework in Algorithm 6.

4.5.2 Nonlinear Recognition Under Domain Shift

After we obtain the set of nonlinear intermediate dictionaries, we are now able to form the augmented feature vectors in the kernel space $\mathcal{F}$ as follows:

$$\mathbf{f}(\alpha) = [((\Phi(\mathbf{Y})\mathbf{A}_0\alpha)^T, (\Phi(\mathbf{Y})\mathbf{A}_1\alpha)^T, ..., (\Phi(\mathbf{Y})\mathbf{A}_K\alpha))^T]^T$$

where α is the sparse code of a source (target) data instance decomposed with the source (target) dictionary using KOMP. Then for two original data signals with corresponding sparse codes α_i and α_j, we compute the inner product between their augmented feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ as follows:

$$\mathbf{f}^T(\alpha_i)\mathbf{f}(\alpha_j) = \alpha_i^T \sum_{k=0}^{K} \mathbf{A}_k^T \mathcal{K}(\mathbf{Y},\mathbf{Y})\mathbf{A}_k \alpha_j = \alpha_i^T \mathbf{H} \alpha_j \tag{4.17}$$

with $\mathbf{H} = \sum_{k=0}^{K} \mathbf{A}_k^T \mathcal{K}(\mathbf{Y},\mathbf{Y})\mathbf{A}_k \in \mathbb{R}^{m\times m}$ denoting a positive semi-definite matrix. The kernel matrix $\mathbf{H}$ is then used to train a SVM classifier for cross domain

Algorithm 6 Algorithm for subspace interpolation between source and target domains through nonlinear dictionary learning (KerSIDL).

Input: Source data $\mathbf{Y}_s$, target data $\mathbf{Y}_t$, sparsity level T, stopping threshold ϵ, parameter λ, $k = 0$.

Output: Atom representation matrixes $\{\mathbf{A}_k\}_{k=0}^{K}$ for the source, intermediate and target domains.

Obtain the atom representation matrix for the source domain $\mathbf{A}_0$ using (4.11).

while stopping criteria is not reached **do**

    Decompose the target data with the current intermediate domain dictionary $\Phi(\mathbf{D}_k)$, and get the reconstruction residue $\Phi(\mathbf{J}_k)$ using (4.12) and (4.13).

    Estimate the difference of the two adjacent atom representation matrixes $\Delta\mathbf{A}_k$ and the next intermediate domain dictionary $\Phi(\mathbf{D}_{k+1})$ using (4.15) and (4.16).

    $k \leftarrow k + 1$

    check the stopping criteria $\|\Delta\mathbf{A}_k\|_F \leq \epsilon$

end while

classification. Our nonlinear domain adaptation framework has the advantage that inner products between cross domain feature vectors can be computed efficiently in closed-form, instead of explicitly carrying out the high dimensional augmented feature representations as in Section 4.4.

4.6 Experiments

In this section, we evaluate our DA approach on face recognition across pose, lighting and blur variations, face re-identification and 2D cross dataset object recognition.

4.6.1 Face Recognition Under Pose Variation

We carried out the first experiment on face recognition across pose variation on the CMU-PIE dataset [1]. We included 68 subjects under 5 different poses in this experiment. Each subject has 21 images at each pose, with variations in lightings. We selected the frontal face images as the source domain, with a total of 1428 images. The target domain contains images at different poses, which are denoted as $c05$ and $c29$ (yawn angle about $\pm 22.5^o$), $c37$ and $c11$ (yawn angle bout $\pm 45^o$) respectively. We chose the front-illuminated source images to be the labeled data in the source domain. The task is to determine the identity of the images in the target domain with the same illumination condition. The classification results are in Table 4.1. We compare our SIDL and KerSIDL frameworks with the following methods. 1) Baseline K-SVD [117], where target data is directly decomposed with the dictionary learned from the source domain, and the resulting sparse codes are compared using a nearest neighbor classifier. 2) GFK [114] and SGF [17], which perform subspace interpolation via infinite or finite sampling on the Grassmann manifold. 3) Eigen light-field [135], where eigen light-field is used as the set of features for pose invariant recognition. 4) SMD [136], which uses stereo matching to

compare the similarity of two faces seen from different poses. We observe that the baseline is heavily biased under domain shift, and all DA methods improve upon it. Both SIDL and KerSIDL demonstrate their advantages over SGF and GFK, the Grassmannian manifold based DA methods. Overall, SMD has the highest average recognition rate, while our KerSIDL method ranks the second. Besides, our approaches do not rely on a generic training set to build pose specific models as the Eigen light-field method, or use feature points to exploit the epipolar geometry of face images as the SMD method. We believe that the incorporation of pose specific knowledge into our framework can further improve the performance. We also show some of the synthesized intermediate images using the SIDL method in Figure 4.3 for illustration. As our DA approach gradually updates the dictionary learned from frontal face images using non-frontal images, these transformed representations thus convey the transition process in this scenario. These transformations could also provide additional information for certain applications, e.g. face reconstruction across different poses.

4.6.2 Face Recognition Across Blur and Illumination Variations

Next, we present the results of a face recognition experiment for dealing with blur and illumination variations. We chose the frontal images of 34 subjects under 21 lighting conditions from the CMU-PIE dataset [1] in this experiment. We selected images of each subject under 11 different illumination conditions to form the source domain. The remaining images with the other 10 illumination conditions

Table 4.1: Face recognition under pose variation on CMU-PIE dataset [1]

	c11	c29	c05	c37	average
KerSIDL	79.4	**100.0**	**98.5**	89.7	91.9
SIDL	76.5	98.5	**98.5**	88.2	90.4
GFK [114]	63.2	92.7	92.7	76.5	81.3
SGF [17]	58.8	89.7	89.7	72.1	77.6
SMD [136]	**97.0**	99.0	97.0	**99.0**	**98.0**
Eigen light-field [135]	78.0	91.0	93.0	89.0	87.8
K-SVD [117]	48.5	76.5	80.9	57.4	65.8

were convolved with a blur kernel to form the target domain. Experiments were performed with the Gaussian kernels with standard deviations of 3 and 4, and motion blurs with lengths of 9 (angel $\theta = 135^o$) and 11 (angel $\theta = 45^o$), respectively. We compare the performance of SIDL and KerSIDL with those of K-SVD [117], GFK [114] and SGF [17]. Besides, we also compare with the Local Phase Quantization (LPQ) [56] method, which is a blur insensitive descriptor, and the method in [39], which estimates an albedo map (Albedo) as an illumination robust signature for matching. We report the results in Table 4.2.

It is observed that KerSIDL achieves the highest recognition rate, while SIDL gives the second best performance and slightly improves upon GFK [114]. Since the domain shift in this experiment consists of both illumination and blur variations,

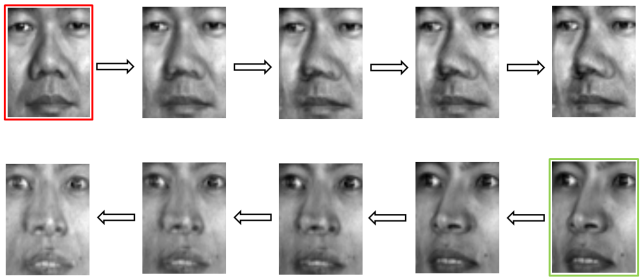


Figure 4.3: Synthesized intermediate representations between frontal face images and face images at pose $c11$. The first row shows the transformed images from a source image (in red box) to the target domain. The second row shows the transformed images from a target image (in green box) to the source domain.

traditional methods which are only illumination insensitive or robust to blur are not able to fully handle both variations. DA methods are useful in this scenario as they do not rely on the knowledge of physical domain shift. We also show transformed intermediate representations along the transition path using the SIDL approach in Figure 4.4, which clearly captures the transition from clear to blur images and vice versa. Particularly, we believe that the transformation from blur to clear conditions is useful for blind deconvolution, which is a highly under-constrained problem [108].

4.6.3 Face Re-identification

Next, we perform experiments on face re-identification using the dataset described in Section 2. Face re-identification aims to match one subject's face image collected at one location with candidate sets acquired at a different location and over time. Re-identification is a fundamentally challenging problem due to the large visual appearance changes caused by variations in view angles, lighting and

Table 4.2: Face recognition across illumination and blur variations on CMU-PIE dataset [1]

	$\sigma = 3$	$\sigma = 4$	$L = 9$	$L = 11$
KerSIDL	**86.47**	**82.65**	**89.71**	**83.24**
SIDL	80.29	77.94	85.88	81.18
GFK [114]	78.53	77.65	82.35	77.65
SGF [17]	70.88	60.29	72.35	67.94
LPQ [56]	66.47	32.94	73.82	62.06
Albedo [39]	50.88	36.76	60.88	45.88
K-SVD [117]	40.29	25.59	42.35	30.59

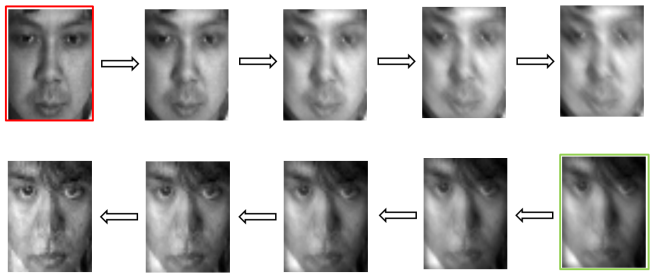


Figure 4.4: Synthesized intermediate representations from face recognition across blur and illumination variations (motion blur with length of 9). The first row shows the transformed images from a source image (in red box) to the target domain. The second row shows the transformed images from a target image (in green box) to the source domain. (The left most image in the second row is an approximation to the blur-free image in the source domain.)

background clutter etc during the data acquisition process. We formulate the face re-identification problem as a domain adaptation problem.

We use the face dataset collected at Baltimore Inner Harbor as the source domain, and another face dataset collected at Comcast Center as the target domain, where the time gap between two datasets is more than two years. Five subjects that appear in both datasets are used in the experiments, with 75 images in the source domain, and 150 images in the target domain. We compare SIDL and KerSIDL with 1) a baseline which performs PCA followed by a SVM classifier 2) a sparse representation based method [20] 3) SGF [17] 4) GFK [114], and report the results in Table 4.3. It is observed that the sparse representation-based method performs ineffectively in this experiment setting, as the complicated variations presented in the target domain severely violates the assumption that the test data lie in the linear

Table 4.3: Face re-identification with the Baltimore dataset as the source domain and the Comcast dataset as the target domain

Method	SVM	Sparse representation [20]	SGF [17]	GFK [114]	SIDL	KerSIDL
Accuracy	32.00	20.0	27.33	29.33	46.0	**59.33**

span of the training data. Both SIDL and KerSIDL outperform other approaches by a large margin, which demonstrate that our intermediate domain dictionaries can better capture the underlying domain shift in the re-identification setting.

4.6.4 Cross Dataset Object Recognition

Following the experiment setting in [114], we evaluated our DA approach for 2D object recognition on four datasets, with a total of 2533 images from 10 categories. The first three datasets were collected by [109], which include images from *amazon.com* (Amazon), collected with a *digital SLR* (DSLR) and a *webcam* (Webcam). The fourth dataset is Caltech-256 (Caltech) [137]. Each dataset constitutes one domain. We used a SURF detector [138] to extract interest points. Then a randomly chosen subset of the interest point descriptors from the Amazon dataset were quantized to visual words by k-means clustering. Each image was represented as a histogram over the quantized visual words of dimension 800. Based on this data representation, we applied our DA approach.

We report performance on eight different pairs of source and target combinations. In the source domain, we randomly selected 8 labeled images per category for

Table 4.4: Cross dataset object recognition in unsupervised setting

Domain		Unsupervised				
source	target	K-SVD [117]	SGF [17]	GFK [114]	SIDL	KerSIDL
Caltech	Amazon	20.5±0.8	36.8±0.5	40.4±0.7	45.4±0.3	**48.2**±1.0
Caltech	DSLR	19.8±1.0	32.6±0.7	41.1±1.3	42.3±0.4	**44.7**±1.5
Amazon	Caltech	20.2±0.9	35.3±0.5	37.9±0.4	40.4±0.5	**41.3**±0.5
Amazon	webcam	16.9±1.0	31.0±0.7	35.7±0.9	37.9±0.9	**38.6**±1.0
webcam	Caltech	13.2±0.6	21.7±0.4	29.3±0.4	36.3±0.3	**36.6**±1.1
webcam	Amazon	14.2±0.7	27.5±0.5	35.5±0.7	38.3±0.3	**38.4**±0.8
DSLR	Amazon	14.3±0.3	32.0±0.4	36.1±0.4	39.1±0.5	**40.6**±1.3
DSLR	webcam	46.8±0.8	66.0±0.5	79.1±0.7	86.2±1.0	**86.7**±1.4

Table 4.5: Cross dataset object recognition in semi-supervised setting

Domain		Semi-supervised				
source	target	K-SVD [117]	SGF [17]	GFK [114]	SIDL	KerSIDL
Caltech	Amazon	31.2±1.0	40.2±0.7	46.1±0.6	50.0±0.5	**53.4**±0.8
Caltech	DSLR	34.6±1.0	36.6±0.8	55.0±0.9	57.1±0.4	**58.0**±1.2
Amazon	Caltech	25.2±0.7	37.7±0.5	39.6±0.4	41.5±0.8	**44.3**±0.8
Amazon	webcam	42.7±0.6	37.9±0.7	56.9 ±1.0	57.8±0.5	**60.9**±0.9
webcam	Caltech	23.4±0.4	29.2±0.7	32.8±0.7	40.6±0.4	**41.1**±1.3
webcam	Amazon	32.9±0.7	38.2±0.6	46.2±0.7	**51.5**±0.6	51.0±0.7
DSLR	Amazon	31.2±1.2	39.2±0.7	46.2±0.6	50.3±0.2	**52.9**±1.9
DSLR	webcam	49.9±1.4	69.5±0.9	80.2±0.4	87.8±1.0	**88.5**±1.6

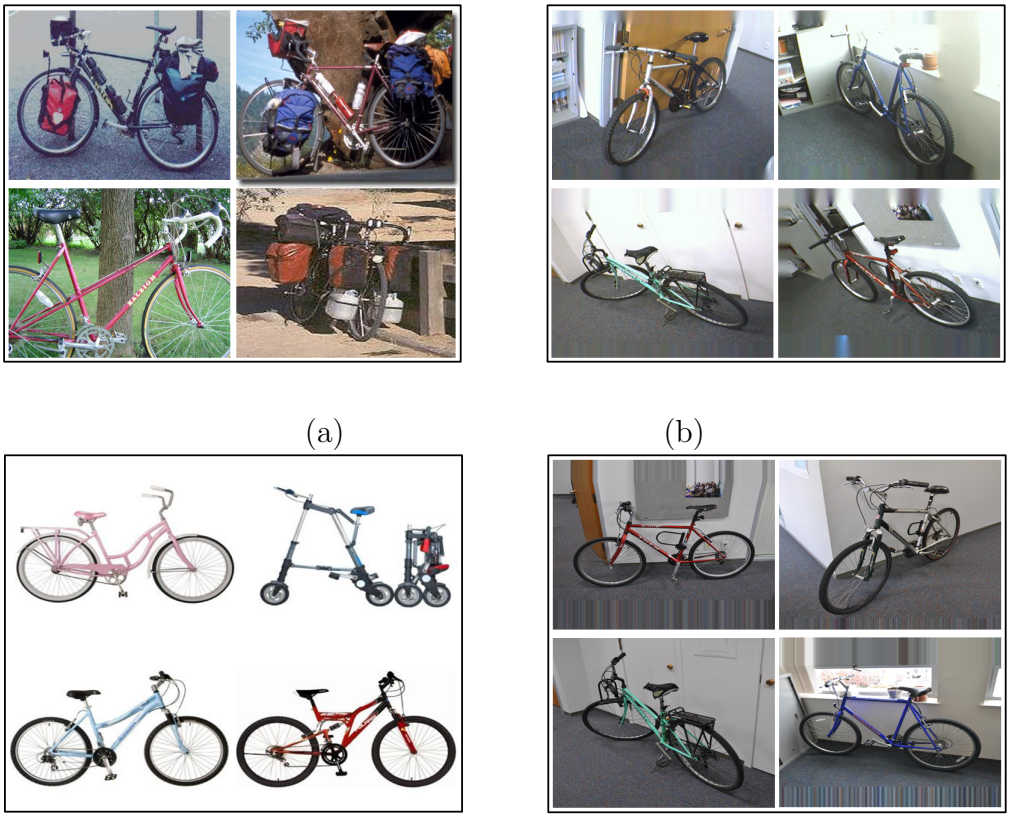
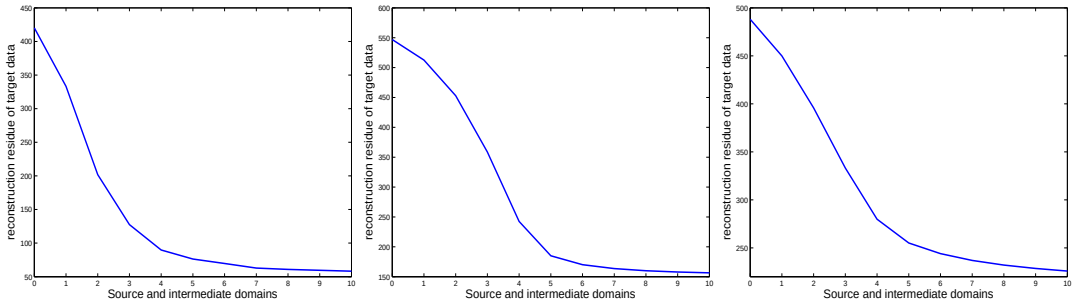


Figure 4.5: Example images of the *bike* category from the (a) Caltech (b) Webcam (c) Amazon (d) DSLR dataset. (Images best viewed in color)

Webcam/DSLR/Caltech and 20 for Amazon. Our SIDL and KerSIDL approaches are compared with K-SVD [117], GFK [114] and SGF [17]. To draw complete comparison with existing DA methods, we also carried out experiments in the semi-supervised setting where we additionally sampled 3 labeled images per category from the target domain. We ran 20 different trials corresponding to different selections of labeled data from the source and target domains. The average recognition rate and standard deviation was reported in Table 4.4 and Table 4.5 for both unsupervised

and supervised settings. It is seen that baseline K-SVD has the lowest recognition rate except for one pair of source and target combination in the semi-supervised setting. Overall, our methods, both SIDL and KerSIDL, consistently demonstrate better performance over state-of-the-art methods.



(a) (b) (c)

Figure 4.6: Average reconstruction error of the target domain decomposed with the source and intermediate domains. The combinations of source and target domains are (a) frontal face images v.s. face images at pose $c29$ (b) DSLR v.s. Webcam (c) Caltech v.s. Amazon, respectively.

Choice of parameters: In our experiments, the regularization parameter λ varies from 1000 to 2000, and the stopping threshold δ is chosen to be between 0.2 to 0.8.

Decrease of reconstruction residue along the transition path: Figure 4.6 shows the average reconstruction residue of target data decomposed with the source, and intermediate domain dictionaries $\{\mathbf{D}_k\}_{k=0}^{K}$ along the transition path which were learned using Algorithm 5. We provide results on three pairs of source and target combinations: frontal face images v.s. face images at pose c29, DSLR v.s.

Table 4.6: QDS values between Amazon/DSLR/Webcam/Caltech datasets

	Amazon	DSLR	Webcam	Caltech
Amazon	NA	8.13	9.03	**9.78**
DSLR	8.13	NA	**9.60**	8.25
Webcam	9.03	**9.60**	NA	8.96
Caltech	**9.78**	8.25	8.96	NA

Webcam dataset, Caltech v.s. Amazon, respectively. We observe that the residue is gradually reduced along the transition path, and Algorithm 5 generally stops within five to ten iterations in our experiments, which demonstrates that our framework is able to bridge the gap between two domains.

QDS values: In Table 4.6, we provide QDS values discussed in Section 4.4.3 between the Amazon/DSLR/Webcam/Caltech datasets. These quantitative values of domain shift are in line with our experimental performance, i.e., higher QDS values indicate less domain shift, and a higher recognition rate between the corresponding two domains.

4.7 Conclusions

We presented a fully unsupervised DA method by incrementally learning intermediate domain dictionaries to capture the underlying domain shift. This allows us to transform original data instances from different modalities into a shared feature

representation, which serves as a robust signature for cross domain classification. We further extended our framework to handle the non-linearities in the data by learning the intermediate dictionaries in a high dimensional RKHS. We evaluated our method on public available datasets and obtain improved performance upon the state of the art. We believe our synthesized intermediate representations are also beneficial for certain applications, e.g, face reconstruction across different poses, blur removal etc.

Chapter 5: Submodular Optimization for Robust Domain Adaptation

5.1 Introduction

Supervised learning usually needs rich labeled data to learn an accurate classification model. Yet it may be very expensive and impractical to obtain sufficient labels for new visual domains, e.g., object recognition from fast-growing online images, person re-identification across camera views from surveillance videos etc. A feasible solution in these scenarios is to leverage related *out-of-domain* labeled data so as to transfer the classification knowledge to the new domain. This is known as the *domain adaptation* problem which has received increasing attention in computer vision. Applications of domain adaptation have been seen in image categorization [109, 17], object detection [139] and activity recognition [113] etc.

Domain adaptation methods utilize a source domain with plenty of labels to learn a classifier for a target domain which is collected from a different distribution. In this work, we are interested in unsupervised domain adaptation where no labels are available in the target domain. A key step in domain adaptation is to find suitable representations such that the distribution difference between two domains is

minimized. One category of popular approaches aim to learn a transformation such that source and target data are projected to a shared latent feature space, where Maximum Mean Discrepancy (MMD) is commonly used to compare the distance between two domains [127, 140, 141]. Another line of research is based on learning intermediate representations [17, 114, 142] to smoothly connect the source and target data.

One limitation of these existing approaches is the assumption that the source data have the same (similar) inner characteristics, usually modeled by a single subspace. Yet with the deluge of data from sources such as internet search engines and surveillance videos, this simplified assumption may not be valid in many realistic applications. For example, face images collected from the web consist of variations in lighting, pose, expression, and usually a coupling among these different variation factors. Such variations in the source data will have the following effect on domain adaptation: 1) The large variations in visual properties in the source domain would likely increase the divergence between the source and target, which could result in negative knowledge transfer. 2) The adaptation algorithm may be less effective to explore the important portion of the source domain for adaptation, as it needs to account for the large within-class variations of the source data. Therefore, it is essential to mitigate the heterogeneity in the source domain to facilitate subsequent adaptation.

We make the following contributions in this chapter.

1) To reduce the divergence between source and target which may be caused by the large variations in properties in the source domain, we aim to identify *pivot*

samples which are a subset of the source domain that are most similar to the target domain. For example, in object recognition with large appearance changes, those source images with similar background and lighting conditions as the target images are more amenable for knowledge transfer. Identifying those pivot samples helps to form a more homogenous domain closer to the target domain, and boost subsequent adaptation performance. For this purpose, we propose a domain similarity function which encourages the selected source samples to be most representative of the target data. Further, in order to preserve the discrimination power of the source domain, we derive a class balance function which ensures that the labels of each class in the selected subset follow the distribution in the original source domain. To this end, we formulate a submodular objective function for our source sample selection algorithm which combines the domain similarity term and the class balance term. By exploiting the diminishing return property of the submodular function, we obtain an efficient greedy algorithm with guaranteed performance of at least $1 - \frac{1}{e}$ approximation to the optimum.

2) We consider the scenario that the heterogeneity in the source data are due to multiple latent domains. For example, images downloaded from the web can contain images of low noise captured using a digital SLR camera as well as those of high noise recorded using a simple webcam. More often, we are not able to define clear visual characteristics to separate those source data. Our goal is to cluster these source data into homogeneous latent domains, where the within-class variations are reduced in each latent domain. This is different from previous approaches dealing with multiple source datasets where the partitions among the source data are known

a prior. The problem is challenging, as a standard clustering algorithm such as K-means would separate data based on their visual similarities only and are prone to forming clusters pertaining to category labels. To this end, we formulate this problem as a constrained clustering problem. We utilize an entropy rate clustering framework [143] which maps the source data to a graph, with vertices denoting the samples and edges representing pairwise similarities among data samples. We use the entropy rate of the random walk over the graph to obtain compact and homogeneous latent domains. Further, we incorporate a domain balancing function which ensures that the distribution of class labels within each latent domain follow the prior label distributions in the original source domain, so that consistent discriminative ability is preserved within each latent domain. By combining the entropy rate function and the domain balancing function, we obtain an objective function which is submodular and enables efficient greedy optimization algorithm.

3) We demonstrate the wide applicability of our source sample selection and latent domain recovery framework on cross dataset object recognition, face recognition across pose and illumination variations, cross view activity recognition, and report improved performance over the state-of-the-art.

5.2 Related Work

Domain Adaptation: Existing domain adaptation algorithms can be roughly classified into three categories: *feature transformation*, *sample re-weighting* and *parameter adaptation*. Feature transformation-based methods focus on discovering

a shared feature space to reduce the distribution difference. A popular distance metric is the Maximum Mean Discrepancy (MMD) which is used to compare the distribution difference between two domains in the Reproducing Kernel Hilbert Space (RKHS) [127]. Different methods to learn domain invariant features based on the MMD criteria have been proposed [140, 141, 144]. Another line of research is based on learning the intermediate representations to form a common feature space. [17, 114] propose to represent subspaces as points on the Grassmannian manifold and identify intermediate domains by sampling along the geodesics path. More recently, A dictionary-based subspace interpolation approach is proposed in [142] to bridge the gap between the source and target domains. Sample re-weighting based methods account for the domain shift by assigning weights to the source data such that the distance between re-weighted source and target distribution is close [145]. Parameter adaptation-based methods use pre-learned models from the source domain as a prior to constrain the classifier learned on the target domain [123].

One relevant work is the constrained assignment algorithm proposed in [146] which separates heterogeneous training data into latent clusters. Yet the assumption on the mixture of Gaussian distributions of data instances may not be satisfied. More closely related is the landmark selection method presented in [147] which discovers source samples close to the target domain, and the framework presented in [148] which reshapes datasets into latent domains based on the *maximum distinctiveness* and *maximum learnability* properties. While [147] uses the MMD criterion to select landmark samples, our proposed domain similarity function for sample selection encourages the selected source data to be most representative of the target domain.

Besides, instead of solving for the binary weights of source samples as in [147, 148], both of our objective functions are submodular which enables more efficient greedy optimization.

Submodularity: Submodularity is the discrete analogue of convexity in continuous domains [149]. Maximizing a submodular function is in general a hard combinatorial problem. Nevertheless, a desirable property of submodularity is that we can obtain $1 - \frac{1}{e}$ approximation through efficient greedy methods. Optimization of submodular functions has been explored in a large spectrum of computer vision applications, such as image segmentation [143], sparse representation [150], anisotropic diffusion [151], attribute selection [152] etc. We differ from previous approaches in that we exploit submodularity in the context of domain adaptation.

5.3 Submodular Sample Selection

In this section, we describe our pivot sample selection algorithm from the source domain with large inner characteristic variations in order to reduce the divergence between source and target distributions.

5.3.1 Preliminaries

Submodularity: Let $\mathcal{V}$ be a finite set. A set function $F : 2^{\mathcal{V}} \rightarrow \mathcal{R}$ is submodular if and only if

$$F(\mathcal{A} \cup k) - F(\mathcal{A}) \geq F(\mathcal{B} \cup k) - F(\mathcal{B}) \tag{5.1}$$

for all subsets $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ and $k \in \mathcal{V} \backslash \mathcal{B}$. This property is refered as the diminishing return property, which states that the marginal gain of adding element k is higher than adding it to any larger set [153].

Let $\mathcal{S} = \{\mathbf{x}_i^s\}_{i=1}^{N_s} \in \mathbb{R}^d$ denote samples from the source domain, with $y_i \in \{1, 2, ..., M\}$ denoting the label of $\mathbf{x}_i^s$. Let $\mathcal{T} = \{\mathbf{x}_i^t\}_{i=1}^{N_t} \in \mathbb{R}^d$ represent unlabeled samples in the target domain. We propose the following domain similarity function to measure the distance between selected source subset $\mathcal{A} \subseteq \mathcal{S}$ and the target domain $\mathcal{T}$.

5.3.2 Domain Similarity Function

We define $s_{j,k}$ as the similarity between source sample $\mathbf{x}_j$ and target sample $\mathbf{x}_k$. We aim to select at most K source samples, such that the sum of maximum similarity between each target sample and the selected source samples in set $\mathcal{A}$ is maximized. Specifically, we define our domain similarity function $\mathbf{V}(\mathcal{A})$ as follows:

$$\begin{aligned} \mathbf{V}(\mathcal{A}) &= \sum_{k \in \mathcal{T}} \max_{j \in \mathcal{A}} s_{j,k} \\ s.t. \mathcal{A} &\subseteq \mathcal{S}, N_{\mathcal{A}} \leq K \end{aligned} \tag{5.2}$$

where $N_{\mathcal{A}}$ is the number of samples in set $\mathcal{A}$. (5.2) favors the selected sample $\mathbf{x}_j$ to be similar to the elements in the target domain, such that the final selected set $\mathcal{A}$ is representative of the target domain. Note that when $\{\mathbf{x}_j\}_{j=1}^{N_{\mathcal{A}}}$, $\{\mathbf{x}_k\}_{k=1}^{N_t}$ are considered within the same domain, (5.2) becomes the well studied facility location problem [154].

The domain similarity function is a submodular function as shown in Proposi-

tion 2. Monotonicity is easily observed because the addition of any source sample to $\mathcal{A}$ does not decrease the value of $\max_{j\in\mathcal{A}} s_{j,k}$ for each target sample $\mathbf{x}_k$. The diminishing return property comes from the fact that the increase in the value of $\max_{j\in\mathcal{A}} s_{j,k}$ from adding a source sample is less in a later stage because the value of $\max_{j\in\mathcal{A}} s_{j,k}$ may have become larger from previously added source samples.

Proposition 2 *The domain similarity function* $\mathbf{V} : 2^{n_s} \rightarrow R$ *is a monotonically increasing submodular function.*

Proof. ***Monotonicity:*** *We prove that* $\mathbf{V}(\mathcal{A})$ *is monotonically increasing by showing that for all* $a \in \mathcal{S}, a \notin \mathcal{A}$

$$\mathbf{V}(\mathcal{A} \cup a) \geq \mathbf{V}(\mathcal{A}) \tag{5.3}$$

$$\mathbf{V}(\mathcal{A} \cup a) - \mathbf{V}(\mathcal{A}) = \sum_{k\in\mathcal{T}} (\max_{j\in\mathcal{A}\cup a} s_{j,k} - \max_{j\in\mathcal{A}} s_{j,k}) \geq 0. \tag{5.4}$$

Submodularity: *We prove* $\mathbf{V}(A)$ *is submodular by showing that*

$$\mathbf{V}(A \cup \{a_1\}) - \mathbf{V}(A) \geq \mathbf{V}(A \cup \{a_1, a_2\}) - \mathbf{V}(A \cup \{a_2\}) \tag{5.5}$$

$$\begin{aligned}
&\mathbf{V}(A \cup \{a_1\}) - \mathbf{V}(A) - \mathbf{V}(A \cup \{a_1, a_2\}) + \mathbf{V}(A \cup \{a_2\}) \\
=&\sum_{k\in\mathcal{T}} (\max_{j\in\mathcal{A}\cup\{a_1\}} s_{j,k} - \max_{j\in\mathcal{A}} s_{j,k} - \max_{j\in\mathcal{A}\cup\{a_1,a_2\}} s_{j,k} + \max_{j\in\mathcal{A}\cup\{a_2\}} s_{j,k}) \\
=&\sum_{k\in\mathcal{T}} \{\max(\max_{j\in\mathcal{A}} s_{j,k}, s_{a_1,k}) - \max_{j\in\mathcal{A}} s_{j,k} - \max(\max_{j\in\mathcal{A}} s_{j,k}, s_{a_1,k}, s_{a_2,k}) \\
&+ \max(\max_{j\in\mathcal{A}} s_{j,k}, s_{a_2,k})\}
\end{aligned} \tag{5.6}$$

We show that

$$V_k = \max(\max_{j\in\mathcal{A}} s_{j,k}, s_{a_1,k}) - \max_{j\in\mathcal{A}} s_{j,k} - \max(\max_{j\in\mathcal{A}} s_{j,k}, s_{a_1,k}, s_{a_2,k}) + \max(\max_{j\in\mathcal{A}} s_{j,k}, s_{a_2,k}) \geq 0, \forall k \in \mathcal{T} \tag{5.7}$$

Case 1: *Assume* $\max_{j\in\mathcal{A}} s_{j,k} \geq s_{a_1,k}$,$\max_{j\in\mathcal{A}} s_{j,k} \geq s_{a_2,k}$,

$$V_k = \max_{j\in\mathcal{A}} s_{j,k} - \max_{j\in\mathcal{A}} s_{j,k} - \max_{j\in\mathcal{A}} s_{j,k} + \max_{j\in\mathcal{A}} s_{j,k} = 0. \tag{5.8}$$

Case 2: *Assume* $s_{a_1,k} \geq \max_{j\in\mathcal{A}} s_{j,k}$,$s_{a_1,k} \geq s_{a_2,k}$,

$$V_k = s_{a_1,k} - \max_{j\in\mathcal{A}} s_{j,k} - s_{a_1,k} + \max(\max_{j\in\mathcal{A}} s_{j,k}, s_{a_2,k}) \geq 0. \tag{5.9}$$

Case 3: *Assume* $s_{a_2,k} \geq \max_{j\in\mathcal{A}} s_{j,k}$,$s_{a_2,k} \geq s_{a_1,k}$,

$$V_k = \max(\max_{j\in\mathcal{A}} s_{j,k}, s_{a_1,k}) - \max_{j\in\mathcal{A}} s_{j,k} - s_{a_2,k} + s_{a_2,k} \geq 0. \tag{5.10}$$

From above three cases, we conclude that $\mathbf{V}(\mathcal{A})$ *is a submodular function.*

5.3.3 Class Balance Function

Further, to preserve the discrimination power in the selected pivot samples, we add the constraint that the proportions of samples per class in the set $\mathcal{A}$ follow the distribution in the original source domain. Let $N(c)$, $N_{\mathcal{A}}(c)$ denotes the number of samples of class $c \in \{1, 2, ..., M\}$ in the source domain and in the subset $\mathcal{A}$ respectively. Let M denotes the number of classes. We define the class balance function $\mathbf{B}(\mathcal{A})$ as follows:

$$\mathbf{B}(\mathcal{A}) = -\sum_{c=1}^{M} \frac{N_{\mathcal{A}}(c)}{\mu N(c)} \log \frac{N_{\mathcal{A}}(c)}{\mu N(c)} \tag{5.11}$$

where μ is a constant. From the log sum inequality, the maximum of $\mathbf{B}(\mathcal{A})$ is achieved when $\frac{N_{\mathcal{A}}(c)}{N(c)}$ are equal, $\forall c \in \{1, 2, ..., M\}$, i.e., when the percentage of samples per class is preserved in the subset $\mathcal{A}$. Therefore, the class balancing function encourages that each class is well represented in the subset $\mathcal{A}$ for the following classification task. $\mathbf{B}(\mathcal{A})$ is submodular as shown in the Proposition 3. The diminishing return property comes from the observation that adding a labeled sample of one class helps more if we have observed less labels of that class so far.

Proposition 3 *The class balancing function* $\mathbf{B} : 2^{n_s} \rightarrow R$ *is a monotonically increasing submodular function.*

Proof. ***Monotonicity:*** *We prove that* $\mathcal{B}(\mathcal{A})$ *is monotonically increasing by showing that for all* $a \in \mathcal{S}, a \notin \mathcal{A}$

$$\mathcal{B}(\mathcal{A} \cup a) \geq \mathcal{B}(\mathcal{A}) \tag{5.12}$$

Assume that a belongs to the k*th class.*

$$\mathcal{B}(\mathcal{A} \cup a) - \mathcal{B}(\mathcal{A}) \tag{5.13}$$

$$= -\sum_{c=1}^{nc} \frac{N_{\mathcal{A}\cup\{a\}}(c)}{\mu N(c)} \log \frac{N_{\mathcal{A}\cup\{a\}}(c)}{\mu N(c)} + \sum_{c=1}^{nc} \frac{N_{\mathcal{A}}(c)}{\mu N(c)} \log \frac{N_{\mathcal{A}}(c)}{\mu N(c)} \tag{5.14}$$

$$= -\frac{n_k + 1}{\mu N(k)} \log \frac{n_k + 1}{\mu N(k)} - (-\frac{n_k}{\mu N(k)} \log \frac{n_k}{\mu N(k)}) \tag{5.15}$$

$$= f(\frac{n_k + 1}{\mu N(k)}) - f(\frac{n_k}{\mu N(k)}) \tag{5.16}$$

$$\geq 0. \tag{5.17}$$

where n_k *denotes the number of selected samples of the* k*th class, and* $f(x)$ *is*

defined as

$$f(x) = -x \log x$$

As $f(x)$ is a strictly increasing function when $x \in (0, 0.36)$, hence the inequality in (5.17) holds when $\mu \in [\frac{1}{0.36}, \infty)$.

Submodularity: *We prove $\mathbf{B}(A)$ is submodular by showing that*

$$\mathbf{B}(A \cup \{a_1\}) - \mathbf{B}(A) \geq \mathbf{B}(A \cup \{a_1, a_2\}) - \mathbf{B}(A \cup \{a_2\}) \tag{5.18}$$

Without loss of generality, we assume that a_1 belongs to class k and a_2 belongs to class m.

Case 1: $k \neq m$.

$$\begin{aligned} &\mathbf{B}(A \cup \{a_1\}) - \mathbf{B}(A) = \mathbf{B}(A \cup \{a_1, a_2\}) - \mathbf{B}(A \cup \{a_2\}) \\ =& -\frac{n_k + 1}{\mu N(k)} \log \frac{n_k + 1}{\mu N(k)} + \frac{n_k}{\mu N(k)} \log \frac{n_k}{\mu N(k)} \end{aligned} \tag{5.19}$$

Case 2: $k = m$.

$$\begin{aligned} &\mathbf{B}(A \cup \{a_1\}) - \mathbf{B}(A) - \mathbf{B}(A \cup \{a_1, a_2\}) + \mathbf{B}(A \cup \{a_2\}) \\ =& -\frac{n_k + 1}{\mu N(k)} \log \frac{n_k + 1}{\mu N(k)} + \frac{n_k}{\mu N(k)} \log \frac{n_k}{\mu N(k)} \\ &+ \frac{n_k + 2}{\mu N(k)} \log \frac{n_k + 2}{\mu N(k)} - \frac{n_k + 1}{\mu N(k)} \log \frac{n_k + 1}{\mu N(k)} \\ =& h(\frac{n_k + 1}{\mu N(k)}) - h(\frac{n_k}{\mu N(k)}) \\ \geq& 0. \end{aligned} \tag{5.20}$$

Where $h(x) = (x + \delta) \log(x + \delta) - x \log x$. The last inequality is obtained by utilizing the strictly increasing property of the function $h(x)$.

From above two cases, we conclude that $\mathbf{B}(\mathcal{A})$ is a submodular function.

5.3.4 Objective Function

We combine the above two criteria and obtain our final objective function $\mathbf{J}(\mathcal{A})$:

$$\mathbf{J}(\mathcal{A}) = \mathbf{V}(\mathcal{A}) + \lambda \mathbf{B}(\mathcal{A}) \tag{5.21}$$

where λ controls the weight of the class balancing term. As a linear combination of submodular functions with nonnegative weights preserve submodularity [153], $\mathbf{J}(\mathcal{A})$ is still a submodular function. Directly maximizing (5.21) is an NP hard problem, therefore, we exploit the submodularity property and adopt a greedy algorithm to obtain a $(1 - \frac{1}{e})$ approximation bound on the optimality of the solution [153]. We start from an empty set $\mathcal{A} = \emptyset$, and adds the sample to the set which has the maximum marginal gain of $\mathbf{J}(\mathcal{A})$ at each iteration. The algorithm terminates when the number of selected samples reaches a pre-specified number or the value of $\mathbf{J}(\mathcal{A})$ decreases. We present our pivot sample selection procedure in Algorithm 7.

Algorithm 7 Algorithm for pivot sample selection.

1: Input: labeled source domain data $\mathcal{S}$,target domain data $\mathcal{T}$,regularization parameter λ, constraint on the number of selected samples K.

2: Output: $\mathcal{A}$.

3: **while** $N_A \leq K$ and $\mathbf{J}(\mathcal{A} \cup a) - \mathbf{J}(\mathcal{A}) \geq 0$ **do**

4: $\quad a^* = \max_a \mathbf{J}(\mathcal{A} \cup a) - \mathbf{J}(\mathcal{A})$

5: $\quad \mathcal{A} \leftarrow a^*$.

6: **end while**

5.4 Submodular Latent Domain Discovery

In this section, we present our algorithm to separate source data into latent domains, so that the within-class variations are reduced in each latent domain. We adopt an entropy-rate based graph partition framework to perform domain clustering.

5.4.1 Graph Representation

We map the source data to an undirected k-nearest neighbor graph $G(V, E)$, where $V = \{v_i\}$ is the vertice set denoting the data points and $E = \{e_{i,j}\}$ is the edge set. The edge weights $\{w_{i,j}\}$ denoting the pairwise similarities between data points are defined as:

$$w_{i,j} = \begin{cases} e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}, & \text{if } \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases} \tag{5.22}$$

where $\mathcal{N}_k(\mathbf{x})$ represents the set of k-nearest neighbors of $\mathbf{x}$, and σ is a normalization constant. Our goal is to select a subset A of the edge set E ($A \subseteq E$) which results in K connected subgraphs, each corresponding to one latent domain.

5.4.2 Entropy Rate

We use the entropy rate of the random walk over the graph G to obtain homogeneous and compact latent domains. The entropy rate is used to measure the uncertainty of a stochastic process Z. For a stationary 1st-order Markov chain, it

is defined as $\mathcal{H}(Z) = \lim_{t\to\infty} H(Z_t|Z_{t-1}) = \lim_{t\to\infty} H(Z_2|Z_1) = H(Z_2|Z_1)$. We then define the transition probability $p_{i,j}$ from vertex v_i to vertex v_j as follows:

$$p_{i,j}(A) = \begin{cases} \frac{w_{i,j}}{w_i}, & \text{if } i \neq j, e_{i,j} \in A \\ 0 & \text{if } i \neq j, e_{i,j} \notin A \\ 1 - \frac{\sum_{j:e_{i,j}\in A} w_{i,j}}{w_i} & \text{if } i = j. \end{cases}$$

where $w_i = \sum_{j:e_{i,j}\in E} w_{i,j}$ is the sum of incident weights of the vertex v_i, and the stationary distribution is obtained as follows:

$$\mu = (\mu_1, \mu_2, ...\mu_{|V|})^T = (\frac{w_1}{w_a}, \frac{w_2}{w_a}, ..., \frac{w_{|V|}}{w_a})^T$$

where $w_a = \sum_{i=1}^{|V|} w_i$ is the sum of incident weights of all vertices. The entropy rate of the random walk can then be computed in the following:

$$\mathcal{H}(A) = -\sum_i \mu_i \sum_j p_{i,j}(A) \log(p_{i,j}(A)) \tag{5.23}$$

The entropy rate of the random walk has been proved to be a monotonically increasing submodular function under the proposed graph construction. Monotonicity is due to that the inclusion of any edge to $\mathcal{A}$ increases the uncertainty of a jump in a random walk. Submodularity is based on the observation that the increase in uncertainty by selecting edge $e_{i,j}$ is less in a later stage as it has been shared with more edges connected to v_i or v_j. For more details, please refer to [143].

5.4.3 Domain Balancing Function

To encourage consistent discrimination in the latent domains, we then propose a domain balancing function to constrain the distribution of class labels within each latent domain. Let M be the number of classes, C_A be the number of connected components in the graph. Our domain balancing function $\mathcal{D}(A)$ is defined as:

$$\mathcal{D}(A) = -\frac{1}{M}\sum_{l=1}^{C_A}\sum_{c=1}^{M}\frac{N_{l,c}}{N_c}\log(\frac{N_{l,c}}{N_c}) - C_A \tag{5.24}$$

where N_c denotes the number of samples from class c in the source domain, and $N_{l,c}$ specifies the number of samples from class c within the lth connected component.

From the log sum inequity, $-\sum_{c=1}^{M}\frac{N_{l,c}}{N_c}\log\frac{N_{l,c}}{N_c}$ achieves maximum when $\frac{N_{l,c}}{N_c}$ is equal, $\forall c \in \{1, 2, ..., M\}$. Hence, the domain balancing function favors that the number of samples per class within each latent cluster follow the prior distribution from the original source domain, such that consistent discriminative ability is preserved. In the mean time, the term $-C_A$ favors fewer number of clusters. Similarly, the domain balancing function is a submodular function as shown in the following proposition.

Proposition 4 *The domain balancing function $\mathcal{D} : 2^E \rightarrow \mathcal{R}$ is a monotonically increasing submodular function under the proposed graph construction.*

Proof. ***Monotonicity:*** *We show that for all $a \in E, a \notin A$*

$$\mathcal{D}(A \cup a) \geq \mathcal{D}(A) \tag{5.25}$$

We are interested in nontrivial cases where the vertices of a belong to different

clusters. Without loss of generality we assume that $a = e_{1,2}$, v_1 *and* v_2 *be in the clusters* S_i *and* S_j *respectively. Clusters* S_i *and* S_j *are merged into cluster* S_k.

$$\mathcal{D}(A \cup \{a = e_{1,2}\}) - \mathcal{D}(A) \tag{5.26}$$

$$= -\frac{1}{C}\sum_{l=1}^{nc-1}\sum_{y=1}^{C}\frac{N_{l,y}}{N_y}\log\frac{N_{l,y}}{N_y} - (nc-1) - (-\frac{1}{C}\sum_{i=1}^{nc}\sum_{y=1}^{C}\frac{N_{l,y}}{N_y}\log\frac{N_{l,y}}{N_y} - nc) \tag{5.27}$$

$$=\frac{1}{C}(\sum_{y=1}^{C}\frac{N_{i,y}}{N_y}\log\frac{N_{i,y}}{N_y} + \sum_{y=1}^{C}\frac{N_{j,y}}{N_y}\log\frac{N_{j,y}}{N_y} - \sum_{y=1}^{C}\frac{N_{i,y}+N_{j,y}}{N_y}\log\frac{N_{i,y}+N_{j,y}}{N_y}) + 1 \tag{5.28}$$

$$=\frac{1}{C}\sum_{y=1}^{C}(\frac{N_{i,y}}{N_y}\log\frac{N_{i,y}}{N_{i,y}+N_{j,y}} + \frac{N_{j,y}}{N_y}\log\frac{N_{j,y}}{N_{i,y}+N_{j,y}}) + 1 \tag{5.29}$$

$$\geq\sum_{y=1}^{C}\frac{N_{i,y}+N_{j,y}}{N_y}\log\frac{N_{i,y}+N_{j,y}}{2(N_{i,y}+N_{j,y})} + 1 \tag{5.30}$$

$$= -\frac{1}{C}\sum_{y=1}^{C}\frac{N_{i,y}+N_{j,y}}{N_y} + 1 \tag{5.31}$$

$$\geq 0. \tag{5.32}$$

Note that the inequality in (5.30) *is obtained by using the log-sum inequity. So far we have completed the proof of the monotonically increasing property of* $\mathcal{D}(A)$.

Submodularity: *We prove* $\mathcal{D}(A)$ *is submodular by showing that*

$$\mathcal{D}(A \cup \{a_1\}) - \mathcal{D}(A) \geq \mathcal{D}(A \cup \{a_1, a_2\}) - \mathcal{D}(A \cup \{a_2\}) \tag{5.33}$$

Without loss of generality, we assume that $a_1 = e_{1,2}$,$a_2 = e_{3,4}$,a_1 *combines clusters* S_i *and* S_j, a_2 *combines clusters* S_m *and* S_n. *We are only interested in nontrivial cases that* a_1 *combines two different clusters. For the case that* $i = j$, $\mathcal{D}(A \cup \{a_1\}) - \mathcal{D}(A) = \mathcal{D}(A \cup \{a_1, a_2\}) - \mathcal{D}(A \cup \{a_2\} = 0$, *hence the submodular property trivially holds.*

Depends on the relationship among i,j,m and n, we discuss the following four cases.

***Case 1**: $\{m,n\} = \{i,j\}$, therefore the addition of a_1 has no effect on the graph partition, hence $\mathcal{D}(A \cup \{a_1\}) - \mathcal{D}(A) \geq \mathcal{D}(A \cup \{a_1, a_2\}) - \mathcal{D}(A \cup \{a_2\}) = 0$.*

***Case 2**: $\{m,n\} \cap \{i,j\} = \emptyset$. Assume that clusters S_i, S_j and S_m,S_n are merged into clusters S_{k_1}, S_{k_2}, respectively.*

$$\begin{aligned}
&\mathcal{D}(A \cup \{a_1, a_2\}) - \mathcal{D}(A \cup \{a_2\}) \\
=&\frac{1}{C}(\sum_{y=1}^{C} \frac{N_{i,y}}{N_y} \log \frac{N_{i,y}}{N_y} + \sum_{y=1}^{C} \frac{N_{j,y}}{N_y} \log \frac{N_{j,y}}{N_y} - \sum_{y=1}^{C} \frac{N_{i,y}+N_{j,y}}{N_y} \log \frac{N_{i,y}+N_{j,y}}{N_y}) + 1 \\
=&\mathcal{D}(A \cup \{a_1\}) - \mathcal{D}(A)
\end{aligned} \tag{5.34}$$

***Case 3**:$m \notin \{i,j\}$,$n \in \{i,j\}$. Assume that the addition of a_2 combines the clusters S_i and S_m.*

$$\mathcal{D}(A \cup \{a_1\}) - \mathcal{D}(A) - (\mathcal{D}(A \cup \{a_1, a_2\}) - \mathcal{D}(A \cup \{a_2\})) \tag{5.35}$$

$$=\frac{1}{C}\sum_{y=1}^{C}\{\frac{N_{i,y}+N_{j,y}+N_{m,y}}{N_y} \log \frac{N_{i,y}+N_{j,y}+N_{m,y}}{N_y} - \frac{N_{i,y}+N_{j,y}}{N_y} \log \frac{N_{i,y}+N_{j,y}}{N_y} \tag{5.36}$$

$$- (\frac{N_{i,y}+N_{m,y}}{N_y} \log \frac{N_{i,y}+N_{m,y}}{N_y} - \frac{N_{i,y}}{N_y} \log \frac{N_{i,y}}{N_y})\} \tag{5.37}$$

$$=\frac{1}{C}\sum_{y=1}^{C}(f(\frac{N_{i,y}+N_{j,y}}{N_y}) - f(\frac{N_{i,y}}{N_y})) \tag{5.38}$$

$$\geq 0. \tag{5.39}$$

Function $f(x)$ *in* (5.38) *is defined as*

$$f(x) = (x + \delta)\log(x + \delta) - x \log x \tag{5.40}$$

We utilize the strictly increasing property of (5.40) *to arrive at the last inequality.*

Case 4*: m=n, i.e., the addition of* a_2 *does not combine any clusters. Therefore* $\mathcal{D}(A \cup \{a_1\}) - \mathcal{D}(A) = \mathcal{D}(A \cup \{a_1, a_2\}) - \mathcal{D}(A \cup \{a_2\})$.

From above four cases, we arrive at the conclusion that $\mathcal{D}(A)$ *is a submodular function.*

5.4.4 Objective function

We combine the entropy rate term and the domain balancing term, and obtain our final objective function $\mathcal{F}(A)$:

$$\begin{aligned} &\max_A \mathcal{F}(A) = \max_A \mathcal{H}(A) + \alpha \mathcal{D}(A) \\ &s.t. A \subseteq E, N_A \geq K \end{aligned} \tag{5.41}$$

where α controls the contribution of the domain balancing term. Similarly, we adopt a greedy algorithm to select the edge which gives the largest gain of $\mathcal{F}$ at each iteration, and stops the algorithm when the number of connected component reaches a pre-specified number. We summarize our latent domain discovery framework in Algorithm 8. While a naive implementation of the algorithm has complexity $O(|E|^2)$, instead, we adopt a lazy greedy approach [155] to speed up the optimization.

Algorithm 8 Algorithm for discovering latent domains

1: Input: $G = (V, E)$,regularization parameter α.

2: Output:A

3: **while** $N_A \leq K$ **do**

4: $a^* = \max_a \mathcal{F}(A \cup a) - \mathcal{F}(A)$

5: $A \leftarrow a^*$.

6: **end while**

5.5 Experiments

In this section, we evaluate our pivot sample selection algorithm and latent domain discovery approach respectively.

Datasets: For 2D object recognition, we use the Office-Caltech datasets [114], which consists of a total of 2533 images from 10 categories. These datasets include images from *amazon.com* (Amazon), images collected with a digital SLR (DSLR), a webcam camera (Webcam), and the Caltech-256 (Caltech) dataset. We use a SURF detector [138] to extract interest points. Then a randomly chosen subset of the interest point descriptors from the Amazon dataset were quantized to generate a code book of size 800. Each image was then represented as a 800 bin histogram.

For face recognition, we evaluate on the CMU-PIE dataset [1] which includes 41,368 images of 68 subjects. We choose the first 34 subjects under 9 pose variations (Pose ID $1 \sim 9$) and 21 lighting conditions. The nine poses range from approximately a full left profile to a full right profile, with neighboring pair of poses about 22.5° apart. All images are 64 by 48 pixels and pixel intensities are used for feature

representation.

For action recognition from videos, we use the IXMAS multi-view action dataset [156] which contains eleven actions categories. Each action is performed three times by twelve actors captured from five different views (Camera 0,1,2,3,4). As suggested in [148], we keep the first five actions (check watch, cross arms, scratch head, sit down, get up) performed by *alba*, *andreas*, *daniel*, *hedlena*, *julien* and *nicolas* to remove the irregularly performed actions. We use the shape-flow descriptor [157] and the spatio-temporal interest point descriptor [158] to characterize the global and local motions of each action respectively. We then generate a codebook with 500 clusters for the shape-flow features and a codebook with 1000 clusters for the spatial-temporal interest point features. Finally, each action sequence is represented as a 1500 dimensional histogram by concatenating the global and local features.

5.5.1 Pivot Sample selection

We first evaluate our source sample approach for object and face recognition. For object recognition, we use the Office-Caltech datasets and evaluate on 9 pairs of cross dataset combinations following the protocol introduced in [147]. The DSLR dataset is never used as the source domain as it contains fewer images. For face recognition, we evaluate on the CMU-PIE dataset. The source and target domain are formed with images associated with different sets of Pose IDs. Further, the images in the source domain consist of 11 different illumination conditions at each

pose, while those in the target domain contain the remaining 10 lighting conditions at each pose. The domain shifts are caused by both lighting and pose variations.

To utilize the pivot source samples, we first train a SVM classifier using the selected samples to predict the category labels of target data. Then we initiate a self-paced adaptation process. Namely, we want to identify a few easier target samples whose predicted labels we are more confident of. The confidence of predication of data sample $\mathbf{x}$ is defined as the probability difference between its two most likely associated classes:

$$f(\mathbf{x}) = \max_{c \in \Omega} p(y = c|\mathbf{x}) - \max_{c \in \Omega \setminus c^*} p(y = c|\mathbf{x})$$

where $c^* = \max_{c \in \Omega} p(y = c|\mathbf{x})$ is the class with the highest probability for $\mathbf{x}$, and Ω is the set of c classes. In each iteration, we move the identified target samples to the source domain and retrain the SVM classifier with the augmented training set. This procedure is stopped until the performance gain between two successive iterations falls below certain threshold or till maximum iterations are reached.

We compare our joint pivot sample selection and self-paced adaptation procedure (PSS-SP) with the following methods. 1) GFK [114], a Grassmannian manifold based domain adaptation method. 2) Kernel mean matching (KMM) [145], a sample re-weighting method to match the source and target distributions. 3) the landmark method [147], which selects source landmark samples to bridge the gap between two domains. Then domain invariant features are learned by minimizing the classification error on the landmark samples, which serve as a proxy to the discriminative loss on the target domain. 4) The statistically invariant sample selection (SISS)

Table 5.1: Cross dataset object recognition in unsupervised setting

Methods	A→C	A→D	A→W	C→A	C→D
Noadapt-SVM	41.7	41.4	34.2	51.8	40.8
GFK [114]	42.2	42.7	40.7	44.5	43.3
KMM [145]	42.2	42.7	42.4	48.3	53.5
Landmark [147]	**45.5**	47.1	46.1	56.7	**57.3**
SISS [141]	44.4	49.0	46.8	55.1	54.8
SP	42.4	42.7	44.1	51.4	43.3
PSS-SP	44.5	**50.3**	**48.8**	**57.4**	52.9
Methods	C→W	W→A	W→C	W→D	average
Noadapt-SVM	42.0	31.1	31.5	70.7	42.8
GFK [114]	44.7	31.8	30.8	75.6	44.0
KMM [145]	45.8	31.9	29.0	72	45.3
Landmark [147]	49.5	40.2	35.4	75.2	50.3
SISS [141]	54.9	39.9	33.7	87.3	51.8
SP	49.8	37.8	34.8	84.7	47.9
PSS-SP	**57.3**	**41.1**	**38.0**	**87.9**	**53.1**

Table 5.2: Face recognition across pose and illumination variations on CMU-PIE dataset [1]

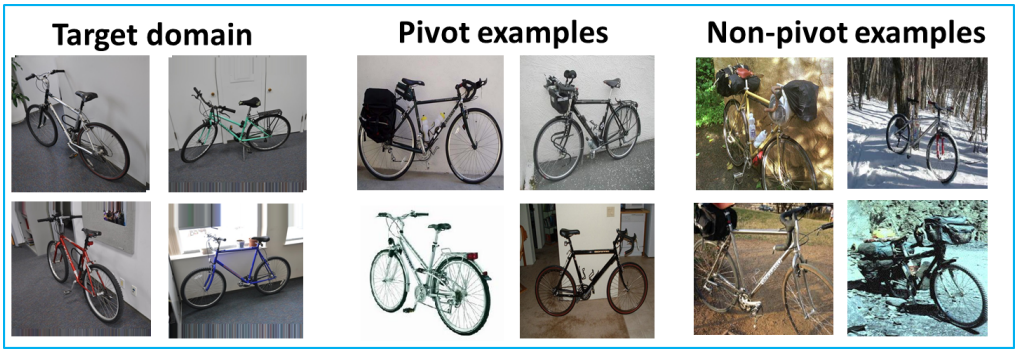
Source	Pose 3,4	Pose 7,8,9	Pose 6,7
Target	Pose 8	Pose 4	Pose 2,3
NoAdapt-SVM	23.5	41.2	31.3
GFK [114]	20.3	40.3	35.9
KMM [145]	22.5	44.7	35.6
Landmark [147]	26.7	35.0	34.1
SP	27.5	48.8	39.6
PSS-SP	**38.2**	**61.8**	**54.3**

method [141], which exploits the Hellinger distance for sample selection. Besides, we also report results using self-paced (SP) adaptation on the whole source data. We show the classification rates for object and face recognition in Table 5.1 and 5.2 respectively. It is seen that for object recognition, our PSS-SP framework performs better than other competing methods on most pairs of source and target. For face recognition, PSS-SP outperforms all other methods significantly in all three cases, which demonstrates that PSS-SP is superior in handling pose and lighting variations. Further, we note that our PSS-SP framework improves upon SP by a large margin in both experiments, which validates that the selected pivot source samples are beneficial to boost the adaptation performance.

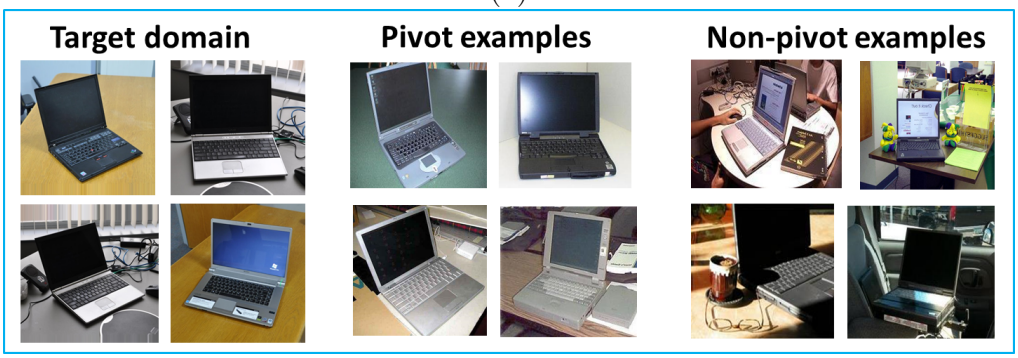
Examples of pivot source samples: In Figure 5.1, we show some exemplar images identified using Algorithm 7 with Caltech as the source domain and DSLR as the target domain. We observe that the selected pivot samples are more similar to the target domain than those non-pivot samples, which validates our assumption.

5.5.2 Latent domain discovery

In this section, we evaluate our latent domain clustering approach for object recognition on the Office-Caltech datasets and for activity recognition on the IXMAS action dataset. Each dataset of the Office-Caltech datasets constitutes one domain, while action videos in the IXMAS dataset from different viewpoints form different domains. For each experiment, we follow the setting in [148], and choose a subset of the domains as the source data while the remaining domains are taken as the target



(a)



(b)

Figure 5.1: Example images of pivot source samples with Caltech as the source domain and DSLR as the target domain from: (a) the *bike* category (b) the *laptop* category.

data. We compare our submodular domain clustering (SDC) algorithm with the following methods: 1) Baseline Union, which merges all source datasets into a single domain to adapt to the target. 2) The domain clustering method in [146] which uses a mixture of Gaussians to model the source data distribution. 3) The dataset reshaping method proposed in [148], where domain assignment is represented using binary weights which are then relaxed into box constraints for optimization.

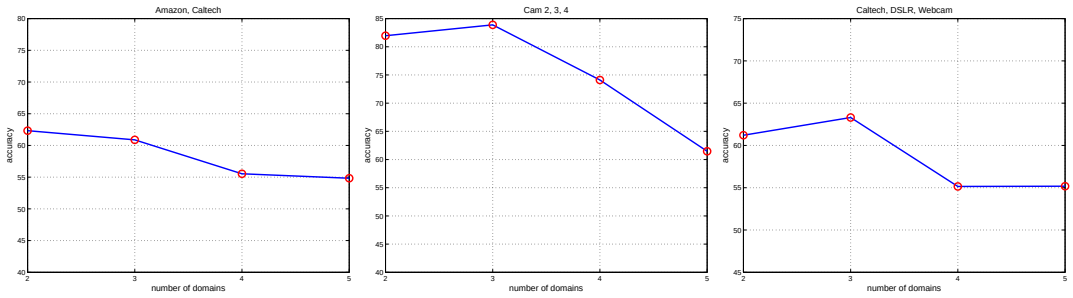
After identifying the latent domains, we use GFK [114] to perform adaptation between each latent domain and the target domain. Then we adopt the *ensemble*

Table 5.3: Recognition performance using the original and recovered latent domains

Source	A,C	D,W	C,D,W	Cam 0,1	Cam 2, 3, 4
Target	D,W	A,C	A	Cam 2, 3, 4	Cam 0, 1
Union	41.7	35.8	41.0	60.7	66.7
Ensemble[146]	31.7	34.4	38.9	60.4	62.2
Matching [146]	39.6	34.0	34.6	56.7	68.3
Ensemble [148]	38.7	35.8	42.8	59.6	71.1
Matching [148]	42.6	35.5	44.6	63.7	71.7
Ensemble-SDC	**46.5**	37.1	**50.6**	62.2	70.6
Matching-SDC	42.7	**38.0**	48.2	**65.9**	**75.0**

and *mathching* strategies to fuse the adaptation results [148]. The ensemble strategy first trains a SVM classifier to predict the domain probabilities of each target sample [146]. Then prediction values from different latent domains are reweighted based on the probabilities that a given test data belongs to each latent domain. For matching strategy, we use the MMD criteria to select the most similar source latent domain to adapt to the target.

We report the comparison results on five different combinations of source and target in Table 5.3. Both SDL and the method in [148] improves the adaptation results over the baseline, which validates the necessity of domain partition for het-



(a) (b) (c)

Figure 5.2: Estimation of the number of latent domains using cross validation. (a) Amazon and Caltech datasets (b) Action videos taken from camera 2,3,4 (c) Caltech, DSLR and Webcam datasets

erogenous source data. Further, our method consistently gives better performance over other methods in all five cases using either the ensemble or matching strategy, which demonstrates the effectiveness of our method in recovering more compact and homogeneous latent domains.

Determine the number of domains: To estimate the optimal number of latent domains, we follow a similar cross-validation procedure as in [148]. Staring from $L = 2$, we use our domain clustering method described in Section 5.4 to separate source data into L domains. We then train SVM classifiers and obtain the five fold cross-validation accuracy z_l for the l-th identified domain. Then accuracy on the whole source data is taken as the weighted average of the accuracies from latent domains: $Z(L) = \sum_{l=1}^{L} \frac{N_l}{N} z_l$, where N_l is the number of samples in the l-th latent domain, and N denotes the total number of samples in the source domain. The optimal number of domains L^* is assigned as the value which gives the highest

(a) (b)

Figure 5.3: Example images of latent domains from : (a) Amazon and Caltech datasets (b) Webcam and DSLR datasets.

cross-validation accuracy: $L^* = \max_L Z(L)$. We plot the cross-validation accuracy using different source training data in Figure 5.2. We observe that the estimated optimal number of domains are in line with the actual number of datasets which the source data contains.

Example images from latent domains: In Figure 5.3, we show some example images from the recovered latent domains. In the left part of Figure 5.3, we provide the domain clustering results from the *backpack* category using the Amazon and Caltech datasets. It is observed that the first latent domain contains more colorful images, while the backpacks in the second domain are mostly dark or gray. The right part of Figure 5.3 demonstrates the results from the *mouse* category using the Webcam and DSLR datasets. It is seen that the majority of the images in the first domain have white background and contain mouses of white color, while the other domain consists of black mouses with wooden background. The different

characteristics of identified domains confirm that our algorithm generates compact and homogeneous latent domains.

5.6 Conclusion

In this chapter, we investigate the problem of domain adaptation with heterogeneous source data. We tackle the problem from two perspectives. We first propose to select pivot source samples that are most similar to the target domain samples to facilitate subsequent adaptation. Alternatively, we derive an entropy rate-based domain clustering framework to separate the source data into homogenous latent domains for improved adaptation. We exploit the submodular property of our objective functions to efficiently solve the NP hard problems. Experimental results on publicly available datasets demonstrate the advantage of our approaches compared to the state-of-the-art. For future research, we plan to investigate selecting informative features for domain adaptation.

Chapter 6: Summary and Directions for Future Work

6.1 Summary

In this dissertation, we investigated the problems and prospects for FR in remote and unconstrained environments. We developed an example-driven manifold prior for regularizing the inverse problem to compensate for the blur variation. In addition, we proposed novel domain adaptation methods for handling more complicated variations between the training and test data. Further, we introduced submodular optimization frameworks to deal with heterogenous source data in domain adaptation. The problems addressed in this dissertation and the methods proposed to solve them lead us to several interesting future research directions.

6.2 Future Research Directions

Detector Adaptation: Following unconstrained FR, the problem of person/face detection in surveillance videos acquired at a distance is also worth investigating, as reliable detection and extraction of robust features are important first steps toward subsequent recognition tasks. Typical person/face detector trained on still images would perform poorly on videos, as videos collected from surveillance

cameras usually suffer from compression artifacts, low resolution, motion blur and low color contrast.

This motivates us to adapt a detector from the image domain to the video domain. We are working toward addressing this problem by building upon boosting-based approaches. We aim to simultaneously minimize the classification error on the labeled source data and the margin violation error of the unlabeled target data. Specifically, during each iteration of the learning procedure, we adjust accordingly the weights of data instances which are wrongly classified in the source domain or lie inside the margin band of the classifier in the target domain. The final classifier learned is expected to have a small generalization error on the target data.

Reference Coding for Person Re-identification: Person re-identification refers to identifying a subject marked at one location with a feasible set of candidates at other locations and over time. It has important applications for recognition tasks in remote and unconstrained scenarios. Yet it is fundamentally challenging due to the large visual appearance changes caused by variations in view angle, lighting, background etc.

As it is difficult to model the variations through parametric formulations, we propose a reference-based method by leveraging a reference set which contains images with different kinds of variations. New feature descriptors of the gallery and probe data are constructed by measuring the similarity between each data instance and the reference images. Re-identification is then performed by comparing the feature descriptors in the reference space. As two images of the same person would look similar to the same set of reference people, therefore they would have

similar reference descriptors. The advantage of using the reference set for feature representation is that it is more robust and consistent than direct comparison in the original feature space under large appearance variations.

Bibliography

[1] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1615–1618, Dec. 2003.

[2] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-Laplacian priors," in *Neural Information Processing Systems*, Vancouver, Canada, Dec. 2009.

[3] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, pp. 586–598, Dec. 2007.

[4] V. Katkovnik, K. Egiazarian, and J. Astola, "A spatially adaptive nonparametric regression image deblurring," *IEEE Transactions on Image Processing*, vol. 14, pp. 1469 –1478, Oct. 2005.

[5] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. Freeman, "Removing camera shake from a single photograph," *ACM Transactions on Graphics*, vol. 25, pp. 787–794, July 2006.

[6] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 71–86, Jan. 1991.

[7] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *Journal of the Optical Society of America*, vol. 14, pp. 1724–1733, Aug. 1997.

[8] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, June 2014.

[9] L. Wiskott, J.-M. Fellous, N. Krger, and C. V. D. Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 775–779, July 1997.

[10] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *International Conference on Computer Vision*, Kyoto, Japan, Oct. 2009, pp. 365–372.

[11] S. Gong, S. J. McKenna, and A. Psarrou, *Dynamic Vision: From Images to Face Recognition*. Imperial College Press, 2000.

[12] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*. Winston, 1977.

[13] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, pp. 1413–1457, Aug. 2004.

[14] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer Graphics and Applications*, vol. 22, pp. 56–65, Mar. 2002.

[15] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Conference on Empirical Methods in Natural Language Processing*, Sydney, Austrialia, July 2006, pp. 120–128.

[16] C. Wang and S. Mahadevan, "Manifold alignment without correspondence," in *International Joint Conference on Artificial Intelligence*, Pasadena, CA, July 2009, pp. 1273–1278.

[17] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *International Conference on Computer Vision*, Barcelona, Spain, Nov. 2011, pp. 999–1006.

[18] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, pp. 399–458, Dec. 2003.

[19] P. Belhumeur, J. Hespanda, and D. Kriegman, "Eigenfaces versus Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711–720, July 1997.

[20] J. Wright, A. Ganesh, A. Yang, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 210–227, Feb. 2009.

[21] B. Moghaddam, "Bayesian face recognition," *Pattern Recognition*, vol. 33, pp. 1771–1782, Nov. 2000.

[22] M. Bartlett, J. Movellan, and T. Sejnowski, "Face recognition by independent component analysis," *IEEE Transactions on Neural Networks*, vol. 13, pp. 1450–1464, Nov. 2002.

[23] S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic recognition of human faces from video," *Computer Vision and Image Understanding*, vol. 91, pp. 214–245, July 2003.

[24] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds," *Computer Vision and Image Understanding*, vol. 99, pp. 303–331, April 2005.

[25] P. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, pp. 295–306, April 1998.

[26] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 643–660, June 2001.

[27] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," *University of Massachusetts, Amherst, Technical Report*, 2007.

[28] N. Pinto, J. DiCarlo, and D. Cox, "How far can you get with a modern face recognition test set using only simple features?" in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Miami, FL, June 2009, pp. 2591–2568.

[29] Y. Yao, B. Abidi, N. Kalka, N. Schmid, and M. Abidi, "Improving long range and high magnification face recognition: database acquisition, evaluation, and enhancement," *Computer Vision and Image Understanding*, vol. 111, pp. 111–125, Aug. 2008.

[30] M. Tistarelli, S. Z. Li, and R. Chellappa, *Handbook of Remote Biometrics: for Surveillance and Security*. Springer, 2009.

[31] P. N. Belhumeur and D. J. Kriegman, "What is the set of images of an object under all possible lighting conditions?" in *IEEE International Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, June 1996, pp. 270–277.

[32] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 218–233, Feb. 2003.

[33] R. Ramamoorthi and P. Hanrahan, "On the relationship between radiance and irradiance: determining the illumination from images of a convex lambertian object," *Journal of the Optical Society of America*, vol. 18, pp. 2448–2459, Oct. 2001.

[34] L. Zhang and D. Samaras, "Face recognition under variable lighting using harmonic image exemplars," *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 19–25, June 2003.

[35] A. Shashua and T. Riklin-Raviv, "The quotient image: Class-based rerendering and recognition with varying iluminations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 129–139, Aug. 2002.

[36] H. Wang, S. Z. Li, and Y. Wang, "Generalized quotient image," *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. Washington, DC, June 2004.

[37] T. Zhang, Y. Y. Tang, B. Fang, Z. Shang, and X. Liu, "Face recognition under varying illumination using gradientfaces," *IEEE Transactions on Imaging Processing*, vol. 18, pp. 2599–2606, Nov. 2009.

[38] T. Chen, W. Yin, X. S. Zhou, D. Comaniciu, and T. S. Huang, "Total variation models for variable lighting face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1519–1524, Sep. 2006.

[39] S. Biswas, G. Aggarwal, and R. Chellappa, "Robust estimation of albedo for illumination-invariant matching and shape recovery," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 884–899, May 2009.

[40] S. K. Zhou, G. Aggarwal, R. Chellappa, and D. W. Jacobs, "Appearance characterization of linear lambertian objects, generalized photometric stereo, and illumination-invariant face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 230–245, Jan. 2007.

[41] V. M. Patel, T. Wu, S. Biswas, P. Phillips, and R. Chellapa, "Dictionary-based face recognition under variable lighting and pose," *IEEE Transactions on Information Forensics and Security*, vol. 7, pp. 954–965, Feb. 2012.

[42] K. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 684–698, May 2005.

[43] X. Zhang and Y. Gao, "Face recognition across pose: A review," *Pattern Recognition*, vol. 42, pp. 2876–2896, Nov. 2009.

[44] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1063–1074, Sep. 2003.

[45] S. Biswas and R. Chellappa, "Pose-robust albedo estimation from a single image," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, June 2010, pp. 2683–2690.

[46] R. Gross, I. Matthews, and S. Baker, "Appearance-based face recognition and light-fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 449 –465, April 2004.

[47] S. Prince, J. Warrell, J. Elder, and F. Felisberti, "Tied factor analysis for face recognition across large pose differences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 970 –984, June 2008.

[48] C. Castillo and D. Jacobs, "Using stereo matching with general epipolar geometry for 2d face recognition across pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 2298 –2304, Dec. 2009.

[49] S. Arashloo and J. Kittler, "Energy normalization for pose-invariant face recognition based on MRF model image matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 1274 –1280, June 2011.

[50] A. Li, S. Shan, X. Chen, and W. Gao, "Maximizing intra-individual correlations for face recognition across pose differences," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Miami, FL, June 2009, pp. 605 –611.

[51] X. Chai, S. Shan, X. Chen, and W. Gao, "Locally linear regression for pose-invariant face recognition," *IEEE Transactions on Image Processing*, vol. 16, pp. 1716 –1725, July 2007.

[52] A. Ashraf, S. Lucey, and T. Chen, "Learning patch correspondences for improved viewpoint invariant face recognition," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Anchorage, AL, June 2008, pp. 1 –8.

[53] T. Kanade and A. Yamada, "Multi-subregion based probabilistic approach toward pose-invariant face recognition," in *International Symposium on Computational Intelligence in Robotics and Automation*, vol. 2, July 2003, pp. 954–959.

[54] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, pp. 1–37, May 2011.

[55] M. Nishiyama, H. Takeshima, J. Shotton, T. Kozakaya, and O. Yamaguchi, "Facial deblur inference to improve recognition of blurred faces," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Miami, FL, June 2009, pp. 1115–1122.

[56] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkilä, "Recognition of blurred faces using local phase quantization," in *International Conference on Pattern Recognition*, Tampa, FL, Dec. 2008, pp. 1–4.

[57] J. Ni, P. Turaga, V. Patel, and R. Chellappa, "Example-driven manifold priors for image deconvolution," *IEEE Transactions on Image Processing*, vol. 20, pp. 3086 –3096, Nov. 2011.

[58] B. Gunturk, A. Batur, Y. Altunbasak, I. Hayes, M.H., and R. Mersereau, "Eigenface-super-resolution for face recognition," *IEEE Transactions on Image Processing*, vol. 12, pp. 597 – 606, May 2003.

[59] K. Jia and S. Gong, "Multi-modal tensor face for simultaneous super-resolution and recognition," in *International Conference Computer Vision*, Beijing, China, Oct. 2005, pp. 1683 –1690.

[60] P. Hennings-Yeomans, S. Baker, and B. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Anchorage, AL, June 2008, pp. 1–8.

[61] B. Li, H. Chang, S. Shan, and X. Chen, "Coupled metric learning for face recognition with degraded images," in *Asian Conference on Machine Learning*, Nanjing, China, Nov. 2009, pp. 220–233.

[62] S. Biswas, K. W. Bowyer, and P. J. Flynn, "Multidimensional scaling for matching low-resolution face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 2019–2030, Oct. 2012.

[63] K. Hotta, T. Kurita, and T. Mishima, "Scale invariant face detection method using higher-order local autocorrelation features extracted from log-polar image," in *International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, April 1998, pp. 70 –75.

[64] R. Abiantun, M. Savvides, and B. Vijaya Kumar, "How low can you go? low resolution face recognition study using kernel correlation feature analysis on the FRGCv2 dataset," in *The Biometric Consortium Conference*, Baltimore, MD, Aug. 2006, pp. 1–6.

[65] S.-W. Lee, J. Park, and S.-W. Lee, "Low resolution face recognition based on support vector data description," *Pattern Recognition*, vol. 39, pp. 1809 – 1812, Sep. 2006.

[66] S. Shekhar, V. Patel, and R. Chellappa, "Synthesis-based recognition of low resolution faces," in *International Joint Conference on Biometrics*, Washington, DC, Oct. 2011, pp. 1–6.

[67] G. Medioni, J. Choi, C.-H. Kuo, A. Choudhury, L. Zhang, and D. Fidaleo, "Non-cooperative persons identification at a distance with 3d face modeling," in *Biometrics: Theory, Applications, and Systems*, Washington, DC, Sep. 2007, pp. 1 –6.

[68] H. Rara, S. Elhabian, A. Ali, M. Miller, T. Starr, and A. Farag, "Distant face recognition based on sparse-stereo reconstruction," in *International Conference on Image Processing*, Cairo, Egypt, Nov. 2009, pp. 4141–4144.

[69] S. Narasimhan and S. Nayar, "Shedding light on the weather," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, WI, June 2003, pp. 665–672.

[70] S. Nayar and S. Narasimhan, "Vision in bad weather," in *International Conference on Computer Vision*, Kerkyra, Greece, Sep. 1999, pp. 820–827.

[71] ——, "Vision in bad weather," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Fort Collins, CO, June 1999, pp. 820–827.

[72] Z. Zhang and R. Blum, "On estimating the quality of noisy images," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, WA, May 1998, pp. 2897–2900.

[73] M.-H. Yang, "Kernel eigenfaces vs. kernel fisherfaces: face recognition using kernel methods," in *IEEE International Conference on Automatic Face and Gesture Recognition*, Washington, DC, Oct. 2002, pp. 215–220.

[74] G. Guo, S. Li, and K. Chan, "Face recognition by support vector machines," in *IEEE International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, Oct. 2000, pp. 196–201.

[75] J. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, pp. 165–175, 1989.

[76] J. Biemond, R. Lagendijk, and R. Mersereau, "Iterative methods for image deblurring," *Proceedings of the IEEE*, vol. 78, pp. 856 –883, May 1990.

[77] J. Kalifa, S. Mallat, and B. Roug, "Deconvolution by thresholding in mirror wavelet bases," *IEEE Transactions on on Image Processing*, vol. 12, pp. 446–457, April 2003.

[78] R. Neelamani, H. Choi, and R. G. Baraniuk, "Forward: Fourier-wavelet regularized deconvolution for ill-conditioned systems," *IEEE Transactions on on Image Processing*, vol. 52, pp. 418–433, Feb. 2004.

[79] D. L. Donoho, "Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition," *Applied and Computational Harmonic Analysis*, vol. 2, pp. 101–126, 1995.

[80] V. M. Patel, G. R. Easley, and D. M. Healy, "Shearlet-based deconvolution," *IEEE Transactions on Image Processing*, vol. 18, pp. 2673–2685, Dec. 2009.

[81] M. Figueiredo and R. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Transactions on Image Processing*, vol. 12, pp. 906–916, Aug. 2003.

[82] J.-L. Starck, M. K. Nguyen, and F. Murtagh, "Wavelets and curvelets for image deconvolution: a combined approach," *Signal Processing*, vol. 83, pp. 2279–2283, Oct. 2003.

[83] E. J. Candès and D. L. Donoho, "Curvelets - a surprisingly effective nonadaptive representation for objects with edges," Vanderbilt University, Tech. Rep., 1999.

[84] M. N. Do and M. Vetterli, "The contourlet transform: an efficient directional multiresolution image representation," *IEEE Transactions on Image Processing*, vol. 14, pp. 2091–2106, Dec. 2005.

[85] A. L. Cunha, J. Zhou, and M. N. Do, "The nonsubsampled contourlet transform: Theory, design, and applications," *IEEE Transactions on Image Processing*, vol. 15, pp. 3089–3101, Oct. 2006.

[86] G. R. Easley, D. Labate, and W. Q. Lim, "Sparse directional image representations using the discrete shearlet transform," *Applied Computational Harmonic Analysis*, vol. 25, pp. 25–46, July 2008.

[87] "Nonlinear total variation based noise removal algorithms," in *International Conference of the Center for Nonlinear Studies on Experimental Mathematics*, Los Alamos, NM, May 1992, pp. 259–268.

[88] Y. Wang, J. Yang, W. Yin, and Y. Zhang, "A new alternating minimization algorithm for total variation image reconstruction," *SIAM Journal on Imaging Sciences*, vol. 1, pp. 248–272, July 2008.

[89] L. He, A. Marquina, and S. Osher, "Blind deconvolution using TV regularization and Bregman iteration," *International Journal of Imaging Systems and Technology*, vol. 15, pp. 74–83, July 2005.

[90] H. Takeda, S. Farsiu, and P. Milanfar, "Deblurring using regularized locally adaptive kernel regression," *IEEE Transactions on Image Processing*, vol. 17, pp. 550 –563, April 2008.

[91] G. Peyre, "Manifold models for signals and images," *Computer Vision and Image Understanding*, vol. 13, pp. 249–260, Feb. 2009.

[92] d.-S. V. Tenenbaum, J. B. and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, Dec. 2000.

[93] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, Dec. 2000.

[94] A. Talwalkar, S. Kumar, and H. A. Rowley, "Large-scale manifold learning," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Anchorage, AL, June 2008, pp. 1–8.

[95] J. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction.* Springer, 2010.

[96] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa, "Statistical computations on grassmann and stiefel manifolds for image and video-based recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 2273–2286, Nov. 2011.

[97] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Madison, WI, June 2003, pp. 313–320.

[98] A. Srivastava, S. H. Joshi, W. Mio, and X. Liu, "Statistical shape analysis: Clustering, learning, and testing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 590–602, April 2005.

[99] A. Veeraraghavan, A. K. Roy-Chowdhury, and R. Chellappa, "Matching shape sequences in video with applications in human movement analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1896–1909, Dec. 2005.

[100] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *International Journal of Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003.

[101] P. C. Hansen, J. G. Nagy, and D. P. O'Leary, *Deblurring Images: Matrices, Spectra, and Filtering.* Society for Industrial and Applied Mathematics, 2006.

[102] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proceedings of the International Conference on Very Large Database Conference*, Edinburgh, Scotland, Sep. 1999, pp. 518–529.

[103] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proceedings of the Symposium on Computational Geometry*, Brooklyn, NY, June 2004, pp. 253–262.

[104] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. L. Roux, and M. Ouimet, "Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering," in *Neural Information Processing Systems*, Quebec, Canada, Dec. 2003, pp. 177–184.

[105] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, pp. 215–223, May 1979.

[106] N. Galatsanos and A. Katsaggelos, "Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation," *IEEE Transactions on Image Processing*, vol. 1, pp. 322 –336, July 1992.

[107] P. C. Hansen, *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion.* Society for Industrial and Applied Mathematics, 1998.

[108] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Understanding and evaluation blind deconvolution algorithms," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Miami, FL, June 2009, pp. 1964–1971.

[109] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European Conference on Computer Vision*, Crete, Greece, Sep. 2010, pp. 213–226.

[110] A. Bergamo and L. Torresani, "Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach." in *Neural Information Processing Systems*, Vancouver, Canada, Dec. 2010, pp. 181–189.

[111] S. Biswas, G. Aggarwal, P. J. Flynn, and K. W. Bowyer, "Pose-robust recognition of low-resolution face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 3037–3049, Dec. 2013.

[112] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011, pp. 3209–3216.

[113] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Providence, RI, June 2012, pp. 2855–2862.

[114] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Providence, RI, June 2012, pp. 2066–2073.

[115] I.-H. Jhuo, D. Liu, D. T. Lee, and S.-F. Chang, "Robust visual domain adaptation with low-rank reconstruction," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Providence, RI, June 2012, pp. 2168–2175.

[116] R. Chellapa, J. Ni, and V. M. Patel, "Remote identification of faces: Problems, prospects, and progress," *Pattern Recognition Letters*, vol. 33, pp. 1849–1859, Oct. 2011.

[117] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD : An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, pp. 4311–4322, Nov. 2006.

[118] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on Image Processing*, vol. 17, pp. 53–69, Jan. 2008.

[119] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *International Conference on Machine Learning*, Quebec, Canada, June 2009, pp. 689–696.

[120] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, pp. 3736–3745, Dec. 2006.

[121] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 791–804, 2012.

[122] H. D. III, "Frustratingly easy domain adaptation," in *Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, June 2007, pp. 256–263.

[123] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *ACM International Conference on Multimedia*, Augsburg, Germany, Sep. 2007, pp. 188–197.

[124] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1667–1680, Sep. 2012.

[125] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms." in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011, pp. 1785–1792.

[126] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," in *AAAI Conference on Artificial intelligence*, Chicago, IL, July 2008, pp. 677–682.

[127] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," in *International Joint Conference on Artifical Intelligence*, Pasadena, CA, July 2009, pp. 1187–1192.

[128] Y. Shi and F. Sha, "Information-theoretical learning of discriminative clusters for unsupervised domain adaptation," in *International Conference on Machine Learning*, Edinburgh, Scotland, June 2012, pp. 1079–1086.

[129] Q. Qiu, V. M. Patel, P. Turaga, and R. Chellappa, "Domain adaptive dictionary learning," in *European Conference on Computer Vision*, Florence, Italy, July 2012, pp. 631–645.

[130] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.

[131] Y. C. Pati, R. Rezaiifar, Y. C. P. R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1993, pp. 40–44.

[132] K. Engan, S. O. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, March 1999, pp. 2443–2446.

[133] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

[134] H. Van Nguyen, V. Patel, N. Nasrabadi, and R. Chellappa, "Design of non-linear kernel dictionaries for object recognition," *Image Processing, IEEE Transactions on*, vol. 22, pp. 5123–5135, Dec 2013.

[135] R. Gross, I. Matthews, and S. Baker, "Appearance-based face recognition and light-fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 449–465, April 2004.

[136] C. Castillo and D. Jacobs, "Using stereo matching with general epipolar geometry for 2d face recognition across pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 2298–2304, Dec. 2009.

[137] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Caltech, Tech. Rep., 2007.

[138] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features," *Computer Vision and Image Understanding*, vol. 110, pp. 346–359, June 2008.

[139] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller, "Shifting weights: Adapting object detectors from image to video," in *Neural Information Processing Systems*, Lake Tahoe, NV, Dec. 2012, pp. 638–646.

[140] M. Long, J. Wang, G. Ding, J. Sun, and P. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, June 2014, pp. 1410–1417.

[141] M. Baktashmotlagh, M. Harandi, B. Lovell, and M. Salzmann, "Domain adaptation on the statistical manifold," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, June 2014, pp. 2481–2488.

[142] J. Ni, Q. Qiu, and R. Chellappa, "Subspace interpolation via dictionary learning for unsupervised domain adaptation," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Portland, OR, June 2013, pp. 692–699.

[143] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy rate superpixel segmentation," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011, pp. 2097–2104.

[144] M. Long, G. Ding, J. W. 0001, J. Sun, Y. Guo, and P. S. Yu, "Transfer sparse coding for robust image representation," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Portland, OR, June 2013, pp. 407–414.

[145] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," in *Neural Information Processing Systems*, British Columbia, Canada, Dec. 2007, pp. 601–608.

[146] J. Hoffman, B. Kulis, T. Darrell, and K. Saenko, "Discovering latent domains for multisource domain adaptation," in *European Conference on Computer Vision*, Florence, Italy, July 2012, pp. 702–715.

[147] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *International Conference on Machine Learning*, Atlanta, GA, June 2013, pp. 222–230.

[148] ——, "Reshaping visual datasets for domain adaptation," in *Neural Information Processing Systems*, Lake Tahoe, NV, 2013, pp. 1286–1294.

[149] L. Lovasz, "Submodular functions and convexity," *Math. Programming-State of the Art*, pp. 235–257, 1983.

[150] A. Das and D. Kempe, "Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection," in *International Conference on Machine Learning*, Bellevue, WA, June 2011, pp. 1057–1064.

[151] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *International Conference on Computer Vision*, Barcelona, Spain, Nov. 2011, pp. 169–176.

[152] J. Zheng, Z. Jiang, R. Chellappa, and J. Phillips, "Submodular attribute selection for action recognition in video," in *Neural Information Processing Systems*, Quebec, Canada, Dec. 2014, pp. 1341–1349.

[153] G. Nemhauser, L. Wolsey, and M. Fisher, "An analysis of approximations for maximizing submodular set functions," *Mathematical Programming*, vol. 14, pp. 265–294, 1978.

[154] Mirchandani and R. Francis, "The uncapacitated facility location problem," *Discrete Location Theory*, 1990.

[155] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, CA, Aug. 2007, pp. 420–429.

[156] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3d exemplars," in *International Conference on Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–7.

[157] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *European Conference on Computer Vision*, Marseille, France, Oct. 2008, pp. 548–561.

[158] J. W. Davis, "Hierarchical motion history images for recognizing human motion," in *IEEE Workshop on Detection and Recognition of Events in Video*, Vancouver, Canada, July 2001, pp. 39–46.