Spring 2017

Investigation of new learning methods for visual recognition

Qingfeng Liu
New Jersey Institute of Technology

ABSTRACT

INVESTIGATION OF NEW LEARNING METHODS FOR VISUAL RECOGNITION

by
Qingfeng Liu

Visual recognition is one of the most difficult and prevailing problems in computer vision and pattern recognition due to the challenges in understanding the semantics and contents of digital images. Two major components of a visual recognition system are discriminatory feature representation and efficient and accurate pattern classification. This dissertation therefore focuses on developing new learning methods for visual recognition.

Based on the conventional sparse representation, which shows its robustness for visual recognition problems, a series of new methods is proposed. Specifically, first, a new locally linear K nearest neighbor method, or LLK method, is presented. The LLK method derives a new representation, which is an approximation to the ideal representation, by optimizing an objective function based on a host of criteria for sparsity, locality, and reconstruction. The novel representation is further processed by two new classifiers, namely, an LLK based classifier (LLKc) and a locally linear nearest mean based classifier (LLNc), for visual recognition. The proposed classifiers are shown to connect to the Bayes decision rule for minimum error. Second, a new generative and discriminative sparse representation (GDSR) method is proposed by taking advantage of both a coarse modeling of the generative information and a modeling of the discriminative information. The proposed GDSR method integrates two new criteria, namely, a discriminative criterion and a generative criterion, into the conventional sparse representation criterion. A new generative and discriminative sparse representation based classification (GDSRc) method is then presented based on the derived new representation. Finally, a new Score space based multiple Metric Learning (SML) method is presented for a challenging visual recognition application, namely, recognizing kinship relations or kinship verification. The proposed

SML method, which goes beyond the conventional Mahalanobis distance metric learning, not only learns the distance metric but also models the generative process of features by taking advantage of the score space. The SML method is optimized by solving a constrained, non-negative, and weighted variant of the sparse representation problem.

To assess the feasibility of the proposed new learning methods, several visual recognition tasks, such as face recognition, scene recognition, object recognition, computational fine art analysis, action recognition, fine grained recognition, as well as kinship verification are applied. The experimental results show that the proposed new learning methods achieve better performance than the other popular methods.

INVESTIGATION OF NEW LEARNING METHODS
FOR VISUAL RECOGNITION

by
Qingfeng Liu

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

May 2017

APPROVAL PAGE

INVESTIGATION OF NEW LEARNING METHODS FOR VISUAL RECOGNITION

Qingfeng Liu

Dr. Chengjun Liu, Dissertation Advisor Date
Professor of Computer Science, NJIT

Dr. James Geller, Committee Member Date
Professor of Computer Science, NJIT

Dr. Ali Mili, Committee Member Date
Professor of Computer Science, NJIT

Dr. Taro Narahara, Committee Member Date
Associate Professor of Architecture and Design, NJIT

Dr. Zhi Wei, Committee Member Date
Associate Professor of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author:	Qingfeng Liu
Degree:	Doctor of Philosophy
Date:	May 2017

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science
  New Jersey Institute of Technology, Newark, NJ, 2017
- Master of Software Engineering
  Peking University, Beijing, China, 2012
- Bachelor of Software Engineering
  Wuhan University, Wuhan, China, 2010

Major: Computer Science

Publications:

Q. Liu and C. Liu, "A Novel Locally Linear KNN Method with Application to Visual Recognition", ***IEEE Transactions on Neural Networks and Learning Systems***. Accepted as a regular paper, Early Access in IEEE Xplore, 2017.

A. Puthenputhussery, **Q. Liu**, and C. Liu, "A Sparse Representation Model Using the Complete Marginal Fisher Analysis Framework And Its Applications to Visual Recognition", ***IEEE Transactions on Multimedia***. Accepted as a regular paper, to appear in 2017.

A. Puthenputhussery, **Q. Liu**, and C. Liu, "Sparse Representation Based Complete Kernel Marginal Fisher Analysis Framework for Computational Art Painting Categorization", the European Conference on Computer Vision 2016 (**ECCV** 2016).

Q. Liu, A. Puthenputhussery, and C. Liu, "A Novel Inheritable Color Space with Application to Kinship Verification", the IEEE Winter Conference on Applications of Computer Vision 2016 (**WACV** 2016).

A. Puthenputhussery, **Q. Liu** and C. Liu, "Color Multi-Fusion Fisher Vector Feature for Computational Painting Categorization", the IEEE Winter Conference on Applications of Computer Vision 2016 (**WACV** 2016).

A. Puthenputhussery, **Q. Liu**, and C. Liu, "SIFT Flow Based Genetic Fisher Vector Feature for Kinship Verification", the IEEE International Conference on Image Processing 2016 (**ICIP** 2016).

Q. Liu and C. Liu, "A Novel Locally Linear KNN Model for Visual Recognition", the IEEE Conference on Computer Vision and Pattern Recognition 2015 (**CVPR** 2015).

Q. Liu, A. Puthenputhussery, and C. Liu, "Inheritable Fisher Vector Feature for Kinship Verification", the IEEE International Conference on Biometrics: Theory, Applications and Systems 2015 (**BTAS** 2015).

Q. Liu, A. Puthenputhussery, and C. Liu, "Learning the Discriminative Dictionary for Sparse Representation by a General Fisher Regularized Model", the IEEE International Conference on Image Processing 2015 (**ICIP** 2015).

Q. Liu, A. Puthenputhussery, and C. Liu, "Novel General KNN Classifier and General Nearest Mean Classifier for Visual Classification", the IEEE International Conference on Image Processing 2015 (**ICIP** 2015).

Q. Liu and C. Liu, "A Novel Hierarchical Interaction Model and HITS Map for Action Recognition in Static Images", the IEEE International Conference on Systems, Man, and Cybernetics 2014 (**SMC** 2014), (Best Paper Award Nomination).

Q. Liu and C. Liu, "A New Locally Linear KNN Method with an Improved Marginal Fisher Analysis for Image Classification", the IEEE International Joint Conference on Biometrics 2014 (**IJCB** 2014).

Q. Liu, Y.Lavinia, A.Verma, J.Lee, L.Spasovic and C. Liu, "Feature Representation and Extraction for Image Search and Video Retrieval", in Recent Advances in Intelligent Image Search and Video Retrieval, Springer, 2017.

Q. Liu and C. Liu, "Inheritable Color Space (InCS) and Generalized InCS Framework with Applications to Kinship Verification", in Recent Advances in Intelligent Image Search and Video Retrieval, Springer, 2017.

Q. Liu and C. Liu, "Improved Soft Assignment Coding for Image Classification", in Recent Advances in Intelligent Image Search and Video Retrieval, Springer, 2017.

Papers Under Reviews:

Q. Liu, A. Puthenputhussery, and C. Liu, "A New Generative and Discriminative Sparse Representation Method and its Application to Visual Recognition," ***IEEE Transactions on Neural Networks and Learning Systems***, under review.

Q. Liu, A. Puthenputhussery, and C. Liu, "Exploiting the Generative and Discriminative Information in Sparse Representation Model for Visual Recognition Applications", the IEEE International Conference on Computer Vision 2017, under review.

In memory of my Mother

I do not think of you lying in the wet clay
Of a Monaghan graveyard; I see
You walking down a lane among the poplars
On your way to the station, or happily

Going to second Mass on a summer Sunday –
You meet me and you say:
“Don’t forget to see about the cattle – ”
Among your earthiest words the angels stray.

And I think of you walking along a headland
Of green oats in June,
So full of repose, so rich with life –
And I see us meeting at the end of a town.

On a fair day by accident, after
The bargains are all made and we can walk
Together through the shops and stalls and markets
Free in the oriental streets of thought.

O you are not lying in the wet clay,
For it is a harvest evening now and we
Are piling up the ricks against the moonlight
And you smile up at us – eternally.

ACKNOWLEDGMENT

I would like to express my sincere appreciation to those people, whose insight, advice and support have helped me go through all the challenges in my research and my life during these years in the Department of Computer Science at NJIT. First, I would like to thank my dissertation advisor, Dr. Chengjun Liu. He has always provided invaluable insight, immense encouragement and constant support for my research. Second, I express my gratitude to Dr. James Geller, Dr. Ali Mili, Dr. Taro Narahara and Dr. Zhi Wei for serving on my committee. I want to thank them for the time they have spent to provide me with their valued feedback and suggestions on my research.

I would also like to thank all my lab mates Ajit Puthenputhussery, Atreyee Sinha, Sugata Banerji, Shaobo Liu, and Hao Liu for their support and assistance. I must also thank Ms. Angel Butler and Dr. George Olsen in the Computer Science department who has helped me in academic issues that I have had during these years.

Third, I would like to take this opportunity to thank my girl-friend, Jiaojiao Chen, who has helped me a lot in both life and research. Her kindness, insights as well as her value of life have influenced me much. I must also thank all my friends during these years inside and outside United States, whose names I cannot show since there are too many to list here.

Finally, I am grateful to my family for their support of every decision I have made. Without their faith in me, I could not have accomplished my goal of being a PhD at NJIT.

TABLE OF CONTENTS

Chapter Page

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS (Continued)

Chapter Page

LIST OF TABLES

LIST OF TABLES
(Continued)

Table Page

LIST OF FIGURES

LIST OF FIGURES
(Continued)

CHAPTER 1

INTRODUCTION

Due to the availability of massive databases of digital images and videos, content based image and video analysis has been a challenging and active research topic for many years in computer vision and pattern recognition. One important and challenging sub-area is visual recognition. Visual recognition, which is a representative problem in computer vision, deals with classifying an image or a video into a predefined category. From robotics to information retrieval, many real world applications depend on the capability to recognize object categories, scene places, and faces. As shown in Figure 1.1, there are many visual recognition tasks applied in our daily life, such as face recognition, scene recognition, object recognition, computational fine art analysis, action recognition, fine grained recognition, as well as kinship verification. In particular, computational fine art analysis contains two sub-tasks: (i) artist classification, namely, classifying a given fine art painting to its author; and (ii) style classification, namely, classifying a given fine art painting to a pre-defined style. Besides, fine grained recognition refers to the task of distinguishing more specific categories, for example, distinguishing a Husky dog from an Alaskan dog. In comparison, object recognition is the task of distinguishing a dog from a bird. Kinship verification is to recognize the kinship relations given a pair of parent and child images.

The intrinsic difficulty of visual recognition lies in the understanding of the semantics and contents of images and videos with high variations in noise, scale, pose, view, illumination as well as occlusion. As shown in Figure 1.2, the major challenges of visual recognition contain occlusion, scale, deformation, clutter, varying illumination, changing viewpoint, and object pose variation [29], [121].

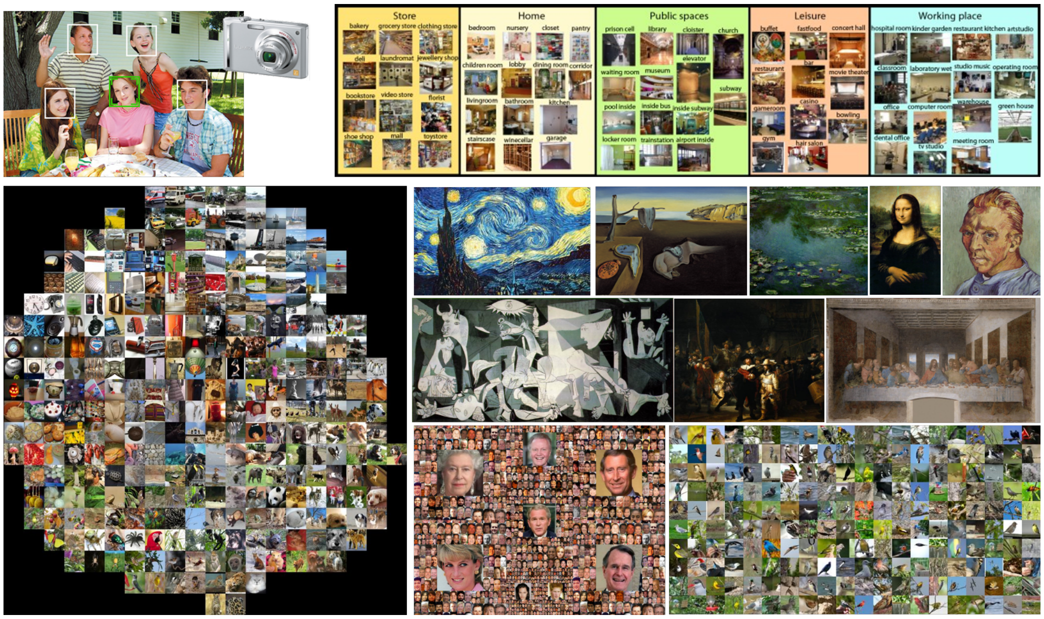


Figure 1.1 Some real world visual recognition applications.

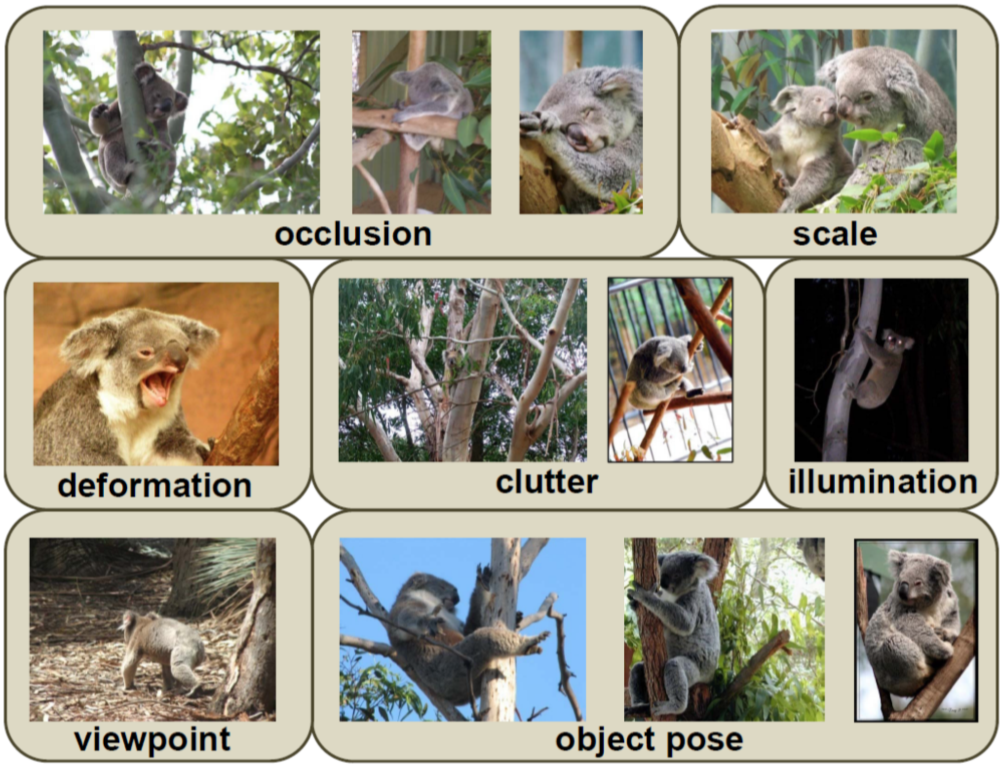


Figure 1.2 Some challenges of visual recognition.

A key step towards building a good visual recognition system includes addressing the key issue of discriminatory feature representation. Besides, an efficient and accurate classification based on the feature representation is indispensable to visual recognition.

Current state-of-the-art methods for visual recognition generally can be categorized into two main frameworks, namely, the bag of features (BoF) framework and the global feature learning framework. Figure 1.3 illustrates the system architecture of both frameworks.

The bag-of-features (BoF) framework generally consists of five major steps for classification. First, feature descriptors, such as the well-known scale invariant feature transform (SIFT) descriptor [87], or histogram of oriented gradients (HOG) [16], are extracted from the image and represented as a vector for further processing. The SIFT descriptor method applies the local accumulation of the magnitude of pixel gradients for each orientation, and finally derives a histogram vector with 128 dimensions. Second, a dictionary is learned from the local feature descriptors by using some learning methods, such as k-means [52], [77] and sparse coding [149], [75]. Third, a feature coding method, such as sparse coding [149], Fisher vector coding [40], and the soft-assignment coding[79] is applied to represent each feature descriptor as a new vector using the learned dictionary. Fourth, the pooling approach is applied by integrating all the feature codings on each dictionary item into one value. Classical pooling methods include the average pooling and the max pooling [51]. Finally, a classification method, such as the sparse representation based classification method [138] or the support vector machine, is applied for recognition.

In comparison, the global feature learning framework, such as the EigenFaces [123], the FisherFaces [4] or the deep convolutional neural networks [50], takes the pattern vector of the data as an input. Then a learning method, such as the principal component analysis, the discriminant analysis or the deep learning [32], [50], [117], [120], [156] is applied to derive the representation. Finally, a classification method is applied for recognition. Please

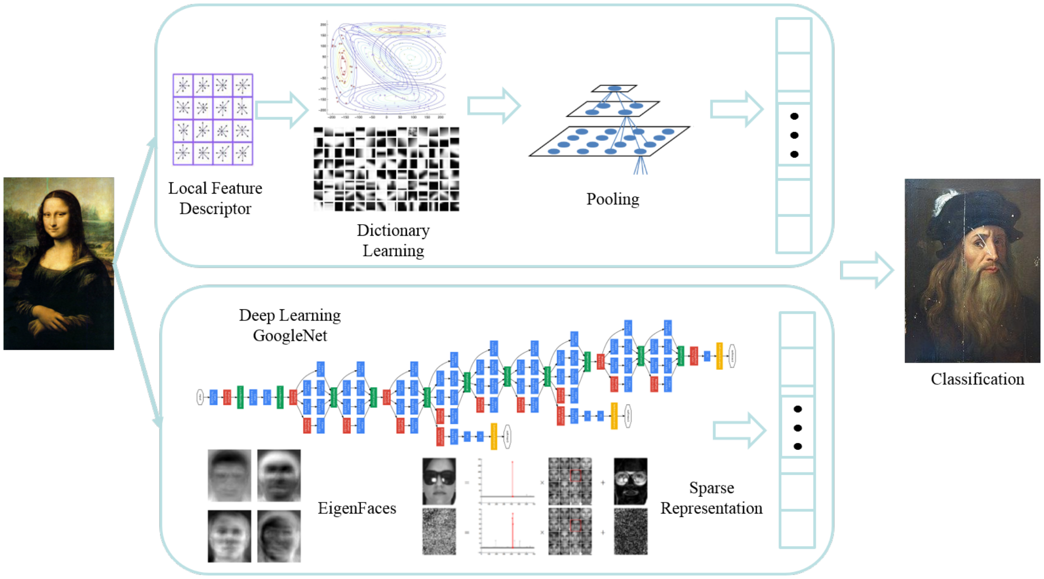


Figure 1.3 Two visual recognition frameworks.

note that the global feature learning framework can take input from the BoF framework as well.

This dissertation presents two new methods for visual recognition by developing a new feature representation method and new classification methods based on the global feature learning framework. Specifically, first, a new locally linear KNN method, or LLK method, is presented. The LLK method derives a new representation, which is an approximation to the ideal representation, by optimizing an objective function based on a host of criteria for sparsity, locality, and reconstruction. The novel representation is further processed by two classifiers, namely, an LLK based classifier (LLKc) and a locally linear nearest mean based classifier (LLNc), for visual recognition. The proposed classifiers are shown to connect to the Bayes decision rule for minimum error. Second, a new generative and discriminative sparse representation (GDSR) method is proposed by taking advantage of both a coarse modeling of the generative information and a modeling of the discriminative information. This hybrid method provides new insights and leads to a new effective representation and classification schema for improving the visual recognition performance. The proposed GDSR method integrates two new criteria, namely, a discriminative criterion and a generative criterion, into the conventional sparse representation criterion. Please note that the term "discriminative" means that the dictionary is learned from the training data in order to improve the performance of classification, while the term "generative" means the process how the dictionary is generated, namely, the class conditional probability density function, e.g., $p(\mathbf{d}_j|c)$. Finally, a new score space-based multiple metric learning (SML) method is presented for a challenging visual verification application, namely, the kinship verification. The proposed SML, which goes beyond the conventional Mahalanobis distance metric learning [144], not only learns the distance metric but also models the generative process of features by taking advantage of the score space. The SML is optimized by solving a constrained, non-negative, and weighted variant of the sparse representation problem.

Then the proposed methods are evaluated on several visual recognition tasks, namely, face recognition, scene recognition, object recognition, computational fine art analysis, action recognition, fine grained recognition and kinship verification. The experimental results show the effectiveness of the proposed methods.

This dissertation is organized in the following manner. First, Chapter 2 presents the background and the related work of several areas related to this dissertation. Second, Chapter 3 presents a new Locally Linear KNN (LLK) method for visual recognition and and conducts extensive experiments on various popular visual recognition data sets. Third, Chapter 4 proposes a new Generative and Discriminative Sparse Representation (GDSR) method for visual recognition and applies it to several popular visual recognition tasks. Then, Chapter 5 demonstrates a novel Score space based multiple Metric Learning (SML) method for kinship verification and shows its feasibility on challenging kinship verification data sets. Finally, Chapter 6 discusses the future work for research.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Sparse Representation

Sparse representation or sparse coding has received much attention in recent years. Both the local feature based sparse coding and the global feature based sparse coding methods are broadly applied for object recognition, scene recognition, and action recognition [149], [133], [5], [38], [138], [19], [20], [151], [152]. Yang et al. [149] used the sparse coding to learn a dictionary and a vector of coefficients to represent the local features. Wang et al. [133] proposed the locality-constrained linear coding method that considers the local information in the feature coding process. Gao et al. [26] further proposed the Laplacian sparse coding that preserves both the similarity and the locality information among the local features.

One global feature based sparse coding method [138] was proposed as well for robust face recognition and further applied to other tasks such as object recognition, scene recognition, and action recognition. Some authors [10], [160], [15] proposed the collaborative representation method, which interprets the sparse representation from other viewpoints. To further incorporate discriminative information into the sparse representation methods, some authors seek to model the intra-class variations within a dictionary to improve the performance for face recognition [19], [20]. An excellent idea is proposed to combine the sparse representation method with the linear discriminant analysis [25], [4], [147], [68] to obtain the discriminative ability for achieving a good sparse representation for signal classification [36]. This method, however, assumes a fixed dictionary, as the dictionary is not discriminatively learned. Furthermore, this method uses Fisher's discrimination factor based on a Frobenius norm a.k.a. the matrix norm or the

HilbertSchmidt norm instead of the popular discriminant analysis and criteria based on the scatter matrices.

Recently, three categories of dictionary learning methods have been proposed for sparse representations. The first category co-trains the discriminative dictionary, sparse representation and the linear classifier together. Mairal et al. [93] proposed to co-train the discriminative dictionary, sparse representation as well as the linear classifier using a combined objective function. Zhang et al. [161] proposed a similar objective function and applied a discriminative singular value decomposition (D-KSVD) method to learn the discriminative dictionary and the classifier simultaneously. Jiang et al. [42] improved upon the method introduced in [161] by introducing a label consistent regularization term.

The second category combines the sub-dictionaries to utilize their discriminative power. Zhou et al. [164] presented a Joint Dictionary Learning (JDL) method that jointly learns both a commonly shared dictionary and class-specific sub-dictionaries to enhance the discrimination of the dictionaries. Yang et al. [151], [152] proposed a Fisher Discrimination Dictionary Learning (FDDL) method, which learns a structured dictionary that consists of a set of class-specific sub-dictionaries.

The third category learns the dictionary by modeling the relation between the dictionary and each class. Yang et al. [150] proposed a latent dictionary learning (LDL) method by jointly learning a latent vector which indicates the relation between the dictionary and labels. Naveed et al. [1] proposed the discriminative Bayesian dictionary learning (DBDL) method by inferring the distribution of the dictionary using an approximation of the Beta process [1].

Besides, many approaches to sparse representation focus on developing efficient learning algorithms to derive the sparse representation and the dictionary [3], [53], [92], [23], [143], [142], [132], [131], [58] or focus on exploring the data manifold structures for representation [26], [133], [162].

In comparison, the proposed LLK method differs from the discriminative dictionary learning methods in the following aspects. First, the proposed LLK method, which does not incorporate any discriminative term, focuses on establishing the relation between a representation method and its classifiers based on the Bayes decision rule for minimum error. The discriminative dictionary methods cannot achieve such a relation. Second, the proposed LLK method, which does not learn a dictionary, avoids the time consuming iterative process of updating the dictionary and the sparse coefficients. Third, some discriminative dictionary methods are based on sub-dictionary learning for each class, which may result in deteriorated performance when the number of the training samples for each class is small. The proposed LLK method does not learn the sub-dictionary. Fourth, some discriminative dictionary methods contain linear classifiers in their objective functions, which excludes other nonlinear classifiers that may achieve better performance. The proposed LLK method develops new classifiers for improving the performance.

It is also worth noting that although the proposed LLK method and LLC [133] both intend to derive the best coding scheme for image classification, the differences between them are two-fold. First, the proposed LLK method focuses on the global representation of an image by means of establishing the relation between image representation and classification. The LLC method in contrast derives a coding scheme for local features, such as the SIFT features. These local features may be used to further derive a global image representation, but additional methods are required. Second, we are able to show that the LLK method approximates the Bayes decision rule for minimal error. The LLC method, in contrast, mainly focuses on the feature codings.

The proposed GDSR method differs from the discriminative dictionary learning methods in the following aspects. First, the proposed GDSR method not only considers the discriminative information by utilizing the underlying topology of the sparse representations but also models the generative information of the dictionary in comparison with other methods. Second, the proposed GDSR method does not depends on any assumption about the probability distribution, such as Bernoulli distributions in DBDL [1]. Third, the proposed GDSR method leads to a new classification method GDSRc, which shows its connection to the Bayes ideal feature, which has the minimum Bayes error for classification. Finally, the proposed GDSR method does not depend on the sub-dictionary, which might lead to over-fitting and deteriorate the performance when the training data of each class is not sufficient.

2.2 Kinship Verification

The pioneer work of kinship analysis originated from the anthropology and psychology communities. Bressnan et al. [9] evaluated the phenotype matching on facial features and claimed that parents have correlated visual resemblance with their offspring. Studies [2] in anthropology have confirmed that children resemble their parent more than other people and they may resemble a particular parent more at different ages. Later work [24] by Fang et al. shows the feasibility of applying computer vision techniques for kinship verification. Xia et al. [141] proposed a transfer subspace learning-based algorithm by using the young parents' set as an intermediate set to reduce the significant divergence in the appearance distributions between children and old parents' facial images. Lu et al. [89] proposed neighborhood repulsed metric learning (NRML) in which the intraclass samples within a kinship relation are pulled as close as possible and interclass samples are pushed as far as possible for kinship verification. The term "neighborhood repulsed" means that the neighborhood kinship relations are pulled as close as possible while the neighborhood non-kinship relatioins are pused as far as possible. Dehghan et al. [18] proposed to apply the generative and the discriminative gated autoencoders to learn the genetic features and metrics together for kinship verification. Yan et al. [145] proposed a multimetric learning method to combine different complementary feature descriptors for kinship verification and later [146] proposed to learn the discriminative mid-level features by constructing a reference data set instead of using hand-crafted descriptors. Lu et al.

[88] presented the results of various teams on the FG 2015 Kinship Verification in the Wild challenge. Zhang et al. [159] presented a deep convolutional neural network based method for kinship verification and achieved good performance.

2.3 Metric Learning

Metric learning methods have gained a lot of attention for computer vision and machine learning applications. Earlier work by Xing et al. [144] applied the semi-definite programming to learn a Mahalanobis metric [144]. Goldberger et al. [28] proposed the neighborhood component analysis (NCA) by minimizing the cross validation error of the kNN classifier. Weinberger et al. proposed the large margin nearest neighbor (LMNN) method [136], which uses the hinge loss to encourage the related neighbors to be at least one distance unit closer than points from other classes. Davis et al. proposed the information-theoretic metric learning (ITML) method [17] by formulating the problem as minimizing the differential relative entropy between two multivariate Gaussian distributions parameterized by the learned metric space and a prior known metric space. Hieu and Li [97] proposed the cosine similarity metric learning (CSML) which utilizes the favorable properties of cosine similarity. Wang et al. [134] proposed a metric learning method using multiple kernels and learned the Mahalanobis distance metric on the kernel feature space. Lu et al. [89] proposed the neighborhood repulsed metric learning (NRML) for kinship verification which pays more attention to neighborhood samples.

2.4 Features

Liu et al. [62] show the effectiveness of Gabor features for face recognition. Wolf et al. [137] proposed the three-patch Local Binary Pattern (LBP) which is computed by comparing the values of three patches. Cao et al. [11] proposed the learning-based (LE) descriptor, which is learned by unsupervised learning techniques and achieves a good trade-off between discriminative power and invariance. Chen et al. [13] proposed the high-dimensional LBP feature by extracting multi-scale patches centered at dense facial landmarks. Together with a rotated sparse regression based compression, the high-dimensional LBP is able to achieve superior performance. Simonyan et al. [116] proposed to apply the Fisher vectors [40] for face verification which achieves very good performance. Color information contributes significantly to the discriminative power of image representation. Conventional color spaces such as RGB, YUV, YIQ, YCbCr have shown their ability for improving the performance of face recognition [115, 86, 66, 67]. For a detailed comparison among different color spaces, refer to [115]. Liu [65] proposed the uncorrelated, independent, and discriminating color spaces for the face recognition problem. Van de Sande et al. [124] show that color information along with shape features yield excellent results on image classification system. Khan et al. [46] proposed the use of color attributes as an explicit color representation for object detection. Zhang et al. [158] proposed a new biologically inspired color image descriptor that uses a hierarchical non-linear spatio-chromatic operator yielding spatial and chromatic opponent channels. Khan et al. [47] show that better results can be obtained for object recognition by explicitly separating the color cue to guide attention by means of a top-down category-specific attention map. Yang et al. [153] proposed a new color model - the $g_1g_2g_3$ model based on the log chromacity color space, which preserves the relationship between R, G and B in the model. Khan et al. [48] cluster color values together based on their discriminative power such that the drop of mutual information of the final representation is minimized.

CHAPTER 3

THE LOCALLY LINEAR KNN METHOD FOR VISUAL RECOGNITION

3.1 Introduction

Feature extraction and classification are two fundamental issues that have received a lot of attention over the past years in computer vision and pattern recognition. To address these issues, many representative methods have been proposed. For example, for feature extraction, some broadly used feature extraction methods are the linear methods [123], [4], [69], [33], [72], [148], [71] and the non-linear methods [63], [154]. Besides, the sparse representation based method [138] has been successfully applied to face recognition by addressing the problem of robust feature representation and classification. The method derives the sparse representation and applies the minimal residual classifier for classification. The representation and the classification are separately developed and no discriminative information is utilized. However, for some specific tasks, e.g., image classification, the representation often needs to serve and facilitate the subsequent classification method. Therefore, many other methods [161], [42], [152] are proposed to learn a discriminative dictionary and the sparse representation by utilizing the label information. Moreover, these methods lead to other issues such as classifier restriction, and higher computational complexity.

We, therefore, present a novel Locally Linear KNN (LLK) method in this chapter. Specifically, the LLK method first applies the criteria of reconstruction, locality, and sparsity to represent each test data. As a result, the new representation shows the grouping effect of the nearest neighbors, which strengthens the coefficients of training samples in the same class as the test sample. Then the new representation is processed by an LLK based classifier (LLKc) and a locally linear nearest mean based classifier (LLNc), respectively. The effectiveness of the proposed LLKc is revealed by its connection to

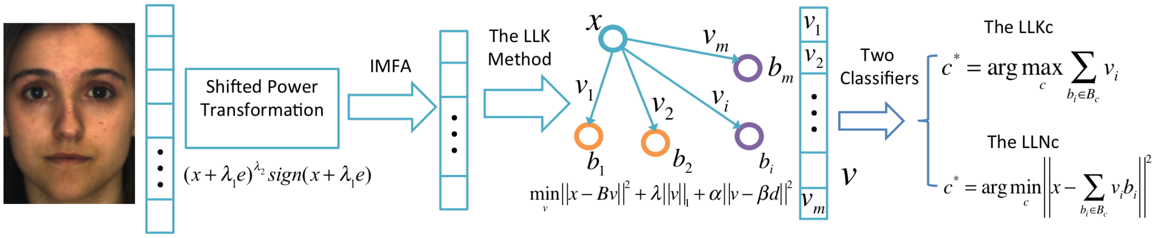

Shifted Power Transformation
$(x+\lambda_1 e)^{\lambda_2} sign(x+\lambda_1 e)$
IMFA
The LLK Method
x
v_m
b_m
v_1
v_2
v_i
b_1
b_2
b_i
$\min_v \|x-Bv\|^2+\lambda\|v\|_1+\alpha\|v-\beta d\|^2$
v_1
v_2
v_m
v
Two Classifiers
The LLKc
$c^* = \arg\max_c \sum_{b_i \in B_c} v_i$
The LLNc
$c^* = \arg\min_c \left\| x - \sum_{b_i \in B_c} v_i b_i \right\|^2$

Figure 3.1 The pipeline of the proposed LLK method. The input pattern vector is first processed by the shifted power transformation. The IMFA method then extracts discriminative features. The proposed LLK method further derives a new representation **v**. The LLKc and the LLNc are finally applied for robust visual recognition.

the Bayes decision rule for minimum error from the kernel density estimation point of view. The proposed LLNc is also shown to be a special case of the Bayes decision rule for minimum error when the conditional density is a Gaussian distribution with weighted mean and diagonal covariance matrix. Besides, other theoretical issues of the proposed LLK method are also discussed such as the performance when non-negative constraint is applied, the performance when group regularization is applied and the computational efficiency of the LLK method using a screening rule method. An improved marginal Fisher analysis (IMFA), which integrates an eigenvalue spectrum analysis for determining the appropriate dimensionality, is further proposed for feature extraction. The shifted power transformation and a coefficient truncating method are also applied for improving the performance. The proposed LLK method is then evaluated for four visual recognition tasks on eight representative data sets. Extensive experimental results show that the proposed LLK method outperforms other popular methods. The pipeline of the proposed LLK method is shown in Figure 3.1.

3.2 Locally Linear KNN Method

The motivation of our LLK method comes from the "ideal representation" (defined in Definition 3.2.1), which represents a test sample as a linear combination of all the training samples while it constrains the corresponding coefficients to be non-zero if the training samples are in the same class as the test sample, and otherwise zero.

Definition 3.2.1 ***The ideal representation***. *The ideal representation of a test sample* $\boldsymbol{x} \in \mathbb{R}^n$ *from the* c*-th class is the coefficient vector* $\boldsymbol{v} = [v_1, v_2, ..., v_m]^t \in \mathbb{R}^m$ *defined as*

$$\boldsymbol{x} = \sum_{i=1}^{m} v_i \boldsymbol{b}_i \tag{3.1}$$

where $\boldsymbol{b}_i \in \mathbb{R}^n (i = 1, 2, ..., m)$ *is the* i*-th training sample and*

$$v_i = \begin{cases} \textit{non-zero}, & \textit{if } \boldsymbol{b}_i \textit{ belongs to the } c\textit{-th class} \\ 0, & \textit{otherwise} \end{cases} \tag{3.2}$$

As a result, the ideal representation is highly sparse, which leads to the development of the conventional sparse representation based methods [138]. However, the representation and the classifier derived from the conventional sparse representation based methods have the following inherent issues that are still ignored.

- First, the training samples that are in the same class as the test sample are often highly correlated and the conventional sparse representation method often tends to derive a representation that only activates one of them with non-zero coefficient [165]. Such a representation violates the condition of the ideal representation, which activates a group of training samples with non-zero coefficients.

- Second, the minimal residual classifier [138] is not directly related to the Bayes decision rule for minimum error [22].

In order to address these two issues, we propose a new Locally Linear KNN (LLK) method defined as follows:

$$\min_{\mathbf{v}} ||\mathbf{x} - \mathbf{B}\mathbf{v}||^2 + \lambda||\mathbf{v}||_1 + \alpha||\mathbf{v} - \beta\mathbf{d}||^2 \tag{3.3}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the test sample, $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_m] \in \mathbb{R}^{n \times m}$ represents all the training samples and $\mathbf{v} \in \mathbb{R}^m$ is the representation that is derived by the proposed LLK method.

The vector $\mathbf{d} = [d_1, d_2, ..., d_m]^t \in \mathbb{R}^m$ represents the distance measure between each training sample and the test sample as follows.

$$d_i = exp\{-\frac{1}{2\sigma^2}||\mathbf{x} - \mathbf{b}_i||^2\} \tag{3.4}$$

where the parameter σ controls the decay speed of the distance measure.

The first term in Equation 3.3 represents the reconstruction criterion, the second term represents the sparsity criterion using the L_1 norm for robustness, and the third term maintains the relation between the coefficients and the distances for the locality property. The three positive parameters λ, α and β weigh the contributions of these three terms.

The proposed LLK method has the following two properties. First, our LLK method, as shown in Section 3.2.1, exhibits the grouping effect of the nearest neighbors (GENN), where training samples are inclined to obtain similar and large coefficients if they are highly correlated and close to the test sample. Furthermore, it is highly probable that these training samples share the same class label with the test sample. As a result, our LLK method is inclined to derive a representation that has large and similar coefficients for the training samples in the same class as the test sample, which is consistent with the case of ideal representation. Second, as shown in Section 3.2.2, the LLKc approximates the Bayes decision rule for minimum error from the kernel density estimation point of view. The LLNc is a special case of the Bayes decision rule for minimum error when the conditional density is a Gaussian distribution with weighted mean and diagonal covariance matrix.

3.2.1 A New Representation that Approximates the Ideal Representation

The fast iterative shrinkage thresholding algorithm or the FISTA algorithm [3] is applied to derive the new representation **v** of our LLK method by optimizing the objective function defined in Equation 3.3. For convergence, the selected step size should be bounded by the maximal step size $\frac{1}{L}$ for the FISTA algorithm, where $L = 2\lambda_{max}(\mathbf{B}^t\mathbf{B} + \alpha\mathbf{I})$, and $\lambda_{max}(\cdot)$ means the largest eigenvalue.

The new representation exhibits the grouping effect of the nearest neighbors (GENN) as defined in Proposition 3.2.1.

Proposition 3.2.1 ***Grouping Effect of the Nearest Neighbors***. *Let* $\boldsymbol{v}^* = [v_1^*, v_2^*, ..., v_m^*]^t$ *be the solution of optimizing the objective function defined in Equation 3.3,* $\rho = \boldsymbol{b}_i^t\boldsymbol{b}_j$ *be the sample correlation of two training samples* $\boldsymbol{b}_i$ *and* $\boldsymbol{b}_j$ $(i, j = 1, 2, ..., m)$*, and* $M(i,j) = |v_i^* - v_j^*|$ *be the difference between two coefficients, if the signs of* v_i^* *and* v_j^* *are the same, then the grouping effect of the nearest neighbors (GENN) is proposed as follows:*

$$M(i,j) \leq \frac{C}{\alpha}\sqrt{2(1-\rho)} + \beta|d_i - d_j| \tag{3.5}$$

where all the samples, namely $\boldsymbol{x}$ *and* $\boldsymbol{b}_i (i = 1, 2, ..., m)$*, are normalized using the* L_2 *normalization and* $C = \sqrt{(1 + \alpha\beta^2||\boldsymbol{d}||^2)}$ *is a constant.*

The GENN property demonstrates an intuition that if the training samples are highly correlated ($\rho \approx 1$) and close to the test sample ($d_i \approx d_j$ and d_i, d_j are large), then the coefficients of the training samples are similar ($v_i^* \approx v_j^*$). This is consistent with the ideal representation. The experiments conducted in Section 3.3.7 further show the effectiveness of the GENN.

The proof of Proposition 3.2.1 is shown as follows.

First, the Equation 3.3 is equal to

$$\min_{v_i} L(\mathbf{v}) = ||\mathbf{x} - \sum_{i=1}^{m} v_i\mathbf{b}_i||^2 + \lambda\sum_{i=1}^{m}|v_i| + \alpha\sum_{i=1}^{m}(v_i - \beta d_i)^2 \tag{3.6}$$

If $\mathbf{v}^* = [v_1^*, v_2^*, ..., v_m^*]^t$ is the derived representation by the LLK method, then take the first derivative of v_i^* and v_j^* as

$$\begin{aligned} \frac{\partial L}{\partial v_i^*} &= -2\mathbf{b}_i^t(\mathbf{x} - \mathbf{B}\mathbf{v}^*) + \lambda sign(v_i^*) + 2\alpha(v_i^* - \beta d_i) = 0 \\ \frac{\partial L}{\partial v_j^*} &= -2\mathbf{b}_j^t(\mathbf{x} - \mathbf{B}\mathbf{v}^*) + \lambda sign(v_j^*) + 2\alpha(v_j^* - \beta d_j) = 0 \end{aligned} \tag{3.7}$$

Please note that if we continue taking the second derivative of the objective function $L(\mathbf{v})$, the second derivative of v_i^* and v_j^* will be $2 + 2\alpha$, which is larger than 0. Then we know that the objective function has a global minimum value.

Since $sign(v_i^*) = sign(v_j^*)$, then $\frac{\partial L}{\partial v_i^*} - \frac{\partial L}{\partial v_j^*}$ is

$$\alpha(v_i^* - v_j^*) = (\mathbf{b}_i - \mathbf{b}_j)^t(\mathbf{x} - \mathbf{B}\mathbf{v}^*) + \alpha\beta(d_i - d_j) \tag{3.8}$$

take the absolute value on both sides and make use of $L(\mathbf{v}^*) \leq L(\mathbf{0})$ and $||\mathbf{x}||^2 = 1$.

$$\begin{aligned}
|v_i^* - v_j^*| &= |\frac{1}{\alpha}(\mathbf{b}_i - \mathbf{b}_j)^t(\mathbf{x} - \mathbf{B}\mathbf{v}^*) + \beta(d_i - d_j)| \\
&\leq \frac{1}{\alpha}|(\mathbf{b}_i - \mathbf{b}_j)^t(\mathbf{x} - \mathbf{B}\mathbf{v}^*)| + \beta|(d_i - d_j)| \\
&\leq \frac{1}{\alpha}||\mathbf{b}_i - \mathbf{b}_j|| * ||\mathbf{x} - \mathbf{B}\mathbf{v}^*|| + \beta|d_i - d_j| \\
&= \frac{1}{\alpha}\sqrt{(||\mathbf{b}_i||_2^2 + ||\mathbf{b}_j||_2^2 - 2 * \mathbf{b}_i^t\mathbf{b}_j)}||\mathbf{x} - \mathbf{B}\mathbf{v}^*|| + \beta|d_i - d_j| \\
&= \frac{1}{\alpha}\sqrt{2(1-\rho)}||\mathbf{x} - \mathbf{B}\mathbf{v}^*|| + \beta|d_i - d_j| \\
&\leq \frac{1}{\alpha}\sqrt{2(1-\rho)}\sqrt{(||\mathbf{x}||^2 + \alpha||\beta\mathbf{d}||^2)} + \beta|(d_i - d_j)| \\
&= \frac{C}{\alpha}\sqrt{2(1-\rho)} + \beta|d_i - d_j|
\end{aligned} \tag{3.9}$$

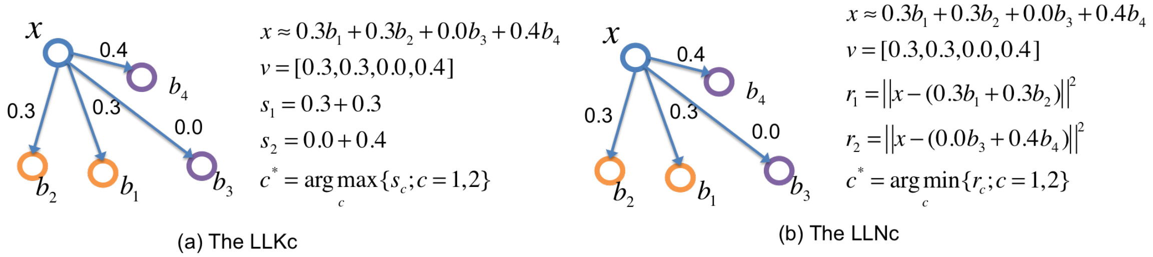
where $C = \sqrt{(1 + \alpha\beta^2||\mathbf{d}||^2)}$.

3.2.2 LLKc and LLNc

The following LLK method based classifier (LLKc) is then proposed to classify the new representation $\mathbf{v}$ derived by our LLK method for the test sample $\mathbf{x}$.

$$\begin{aligned}
c^* &= \arg\max_c s_c \\
&= \arg\max_c \sum_{\mathbf{b}_i \in \mathbf{B}_c} v_i
\end{aligned} \tag{3.10}$$

where $c = 1, 2, ..., w$ denotes the class label, $\mathbf{B}_c$ represents the set of training samples in the c-th class and s_c is the sum of the coefficients of the training samples in the c-th class.


x
0.4
0.3
0.3
0.0
b_4
b_2
b_1
b_3
$x \approx 0.3b_1 + 0.3b_2 + 0.0b_3 + 0.4b_4$
$v = [0.3, 0.3, 0.0, 0.4]$
$s_1 = 0.3 + 0.3$
$s_2 = 0.0 + 0.4$
$c^* = \arg\max_c \{s_c; c = 1,2\}$
x
0.4
0.3
0.3
0.0
b_4
b_2
b_1
b_3
$x \approx 0.3b_1 + 0.3b_2 + 0.0b_3 + 0.4b_4$
$v = [0.3, 0.3, 0.0, 0.4]$
$r_1 = \|x - (0.3b_1 + 0.3b_2)\|^2$
$r_2 = \|x - (0.0b_3 + 0.4b_4)\|^2$
$c^* = \arg\min_c \{r_c; c = 1,2\}$

(a) The LLKc

(b) The LLNc

Figure 3.2 Examples of the LLKc and the LLNc. In the Figure (a), the test sample **x** is assigned to class 1 because s_1 is larger than s_2. In the Figure (b), the test sample **x** is assigned to class that has the minimum value of $r_i (i = 1, 2)$.

The LLK method based classifier (LLKc) thus classifies the test sample by counting the sum of coefficients of all the training samples in each class. For example, as shown in Figure 3.2(a), there are four training samples $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ and $\mathbf{b}_4$ with class labels 1, 1, 2, 2. The representation $\mathbf{v} = [0.3, 0.3, 0.0, 0.4]^t$ is derived by means of the proposed method for the test sample **x**. Then the LLK method based classifier will classify the test sample to class 1 since $0.3 + 0.3 > 0.0 + 0.4$.

The effectiveness of the LLKc is revealed in Proposition 3.2.2, where a theoretical connection is expressed between the proposed LLKc and the Bayes decision rule for minimum error with limited assumptions. Please note that the reason why we does not apply Bayes decision rule directly is that the Bayes decision rule is based on class conditional probability, which is very difficult to estimate in practice. As an alternative way, we apply a kernel density estimation using the sparse representation for a better estimation.

Proposition 3.2.2 *Given the test sample **x** and its new representation **v** derived by the proposed LLK method, if each class has the same prior distribution $p(c)$, then the connection between the proposed LLKc and the Bayes decision rule is as follows:*

$$\begin{aligned} c^* &= \arg\max_c \sum_{\boldsymbol{b}_i \in \boldsymbol{B}_c} v_i \\ &\propto \arg\max_c p(c|\boldsymbol{x}) \end{aligned} \tag{3.11}$$

In order to relate $\sum_{\mathbf{b}_i \in \mathbf{B}_c} v_i$ to the posterior probability, the following transformations can be applied to normalize $\sum_{\mathbf{b}_i \in \mathbf{B}_c} v_i$ to the range of $[0, 1]$, without any influence on the final classification results of the LLKc.

$$\begin{aligned} v_{ni} &= \frac{v_i - v_{min}}{v_{max} - v_{min}} \\ v_{ni2} &= \frac{v_{ni}}{\sum_{i=1}^{m} v_{ni}} \end{aligned} \tag{3.12}$$

where v_{min} and v_{max} are the minimal and maximal value among all the elements of the vector **v**, respectively.

The proof of Proposition 3.2.2 is shown as follows.

If **v** is the derived representation by the proposed LLK method, then $L(\mathbf{v}) \leq L(\mathbf{0})$, which means

$$||\mathbf{x} - \mathbf{B}\mathbf{v}||^2 + \lambda ||\mathbf{v}||_1 + \alpha ||\mathbf{v} - \beta \mathbf{d}||^2 \leq ||\mathbf{x}||^2 + \alpha ||\beta \mathbf{d}||^2 \tag{3.13}$$

then $||\mathbf{v} - \beta\mathbf{d}||_2 \leq \sqrt{\frac{1}{\alpha} + \beta^2 ||\mathbf{d}||^2}$, which means $||\mathbf{v} - \beta\mathbf{d}||_2$ is bounded by a small positive constant so that $\mathbf{v} \approx \beta\mathbf{d} + \mathbf{const}$. the transformations in Equation 3.12 guarantees that each $0 \leq v_i \leq 1$ and $\sum_{i=1}^{m} v_i = 1$. It is worth noting that the transformations do not affect the classification results. Thus, the LLKc is approximated as follows

$$\begin{aligned} c^* &= \arg\max_c \sum_{\mathbf{b}_i \in \mathbf{B}_c} v_i \\ &\approx \arg\max_c \sum_{\mathbf{b}_i \in \mathbf{B}_c} \beta d_i + const \\ &\propto \arg\max_c \sum_{\mathbf{b}_i \in \mathbf{B}_c} exp\{-\frac{1}{2\sigma^2} ||\mathbf{x} - \mathbf{b}_i||^2\} \\ &\propto \arg\max_c \frac{1}{hm_c} \sum_{\mathbf{b}_i \in \mathbf{B}_c} k(\frac{\mathbf{x} - \mathbf{b}_i}{h}) \end{aligned} \tag{3.14}$$

where m_c is the number of training samples in the c-th class and h is the bandwidth that controls the degree of smoothing and it is fixed for all the classes. We use the zero mean and unit variance Gaussian kernel $k(u) = \frac{1}{\sqrt{2\pi}} exp(-\frac{u^2}{2})$ here. Then the Equation 3.14 can be used for the kernel density estimation [25] of the conditional probability $p(\mathbf{x}|c)$.

Therefore, if the prior probability $p(c)$ is equal for all the classes, then

$$\begin{aligned} c^* &= \arg\max_c \sum_{\mathbf{b}_i \in \mathbf{B}_c} v_i \\ &\approx \arg\max_c p(\mathbf{x}|c) \\ &\propto \arg\max_c p(c|\mathbf{x}) \quad (Bayes\ classifier) \end{aligned} \tag{3.15}$$

In summary, the LLKc approximates the Bayes decision rule for minimum error in the view of density estimation, and the approximation error mainly comes from the $\mathbf{v} \approx \beta\mathbf{d} + \mathbf{const}$ (relatively large α and small β are preferred) and the kernel density estimation error.

We also present another classifier, the Locally Linear Nearest mean classifier (LLNc), which is defined as follows:

$$\begin{aligned} c^* &= \arg\min_c ||\mathbf{x} - \mathbf{m}_c||_2^2 \\ &= \arg\min_c ||\mathbf{x} - \sum_{\mathbf{b}_i \in \mathbf{B}_c} v_i \mathbf{b}_i||_2^2 \end{aligned} \tag{3.16}$$

where $\mathbf{m}_c = \sum_{\mathbf{b}_i \in \mathbf{B}_c} v_i \mathbf{b}_i$ is the “mean” of the c-th class.

For example, as shown in Figure 3.2(b) there are four training samples $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ and $\mathbf{b}_4$ with class labels 1, 1, 2, 2. The representation $\mathbf{v} = [0.3, 0.3, 0.0, 0.4]$ is derived by means of the proposed method for the test sample $\mathbf{x}$. Then the LLNc will compute $r_1 = ||\mathbf{x} - (0.3\mathbf{b}_1 + 0.3\mathbf{b}_2)||^2$ and $r_2 = ||\mathbf{x} - (0.0\mathbf{b}_3 + 0.4\mathbf{b}_4)||^2$ and choose the smaller one.

An interesting property of the proposed LLNc is that it can be regarded as a special case of Bayes decision rule when the conditional density function of the c-th class $p(\mathbf{x}|c)$ is a Gaussian distribution with the “mean” of the c-th class $\mathbf{m}_c$ and a diagonal covariance matrix. As a result, the proposed LLNc is formulated as $c^* = \arg\min_c - \log p(\mathbf{x}|c)$.

3.2.3 Shifted Power Transformation for Improving Reliability

Note that a global window width may deteriorate the robustness of the kernel density estimation of the real world data, whose underlying density distribution often has different degrees of smoothing at different locations. In the proposed LLK method, the parameter σ in Equation 3.4 plays the role of the global window width, and thus may deteriorate the performance if not carefully selected. In order to alleviate the sensitivity of our LLK method to the parameter σ, the following shifted power transformation [129] is applied.

$$T(\mathbf{x}) = |\mathbf{x} + \lambda_1 \mathbf{e}|^{\lambda_2} sign(\mathbf{x} + \lambda_1 \mathbf{e}) \tag{3.17}$$

where $sign(\mathbf{x})$ represents the vector of the sign of each element in the vector **x** with the value 0, 1 and -1, $0 < \lambda_1, \lambda_2 \leq 1$ and $e = [1, 1, ..., 1]^t$.

The shifted power transformation is capable of transforming the data into a near Gaussian shape, for which the kernel density estimation may work well [129]. Different from [129] where the estimate of the density of the new data can be “back-transformed” to an estimate of the density of the original data, our goal is not to provide the density estimation of the original data, but to provide the density estimation in the transformed space for better approximation to the Bayes decision rule for minimum error. Thus, we do not operate “back-transformation”, which has a negative impact in practice. Moreover, we also normalize **d** using the L_2 normalization to further reduce the sensitivity to the value of σ in the new transformed space. As a result, robustness is achieved for various values of σ in practice (see Section 3.3.6).

3.2.4 Coefficient Truncating for Enhancing Generalization Performance

Note that the noise introduced by some distant data samples with trailing coefficients in the new transformed space may deteriorate the generalization performance of the LLK method. In practice, not all the training samples in the c-th class ($\mathbf{B}_c$) are required and some farther away noisy samples that have trailing coefficients may adversely influence the performance of kernel density estimation. Our solution is a coefficient truncating method, where only the k largest coefficients v_i of each class are required for classification.

As a result, both the LLKc and the LLNc are redefined as follows, respectively.

$$c^* = \arg\max_c \sum_{\substack{(\mathbf{b}_i \in \mathbf{B}_c) \wedge \\ (v_i \in T(k))}} v_i \tag{3.18}$$

$$c^* = \arg\min_c ||\mathbf{x} - \sum_{\substack{(\mathbf{b}_i \in \mathbf{B}_c) \wedge \\ (v_i \in T(k))}} v_i \mathbf{b}_i||_2^2 \tag{3.19}$$

where T(k) represents the set of k largest v_i for each class. In practice, the parameter k plays an important role in the performance (see Section 3.3.6).

3.2.5 Improved Marginal Fisher Analysis for Feature Extraction

Before we apply the proposed LLK method for deriving new representation and classification, we propose to apply the principal component analysis (PCA) and the improved marginal Fisher analysis [76] (IMFA) to reduce the dimension of the data and extract features.

The proposed IMFA improves upon the marginal Fisher analysis [147] by integrating an eigenvalue spectrum analysis to determine the proper dimensionality of PCA. Specifically, the IMFA keeps a proper balance between the energy criterion and the magnitude criterion [69] to alleviate the over-fitting to an improper dimensionality reduced by PCA. The energy criterion prefers high-dimensional spaces to preserve the spectral energy of the original data by including more eigenvalues, while the magnitude criterion favors low-dimensional spaces and discards small trailing eigenvalues that often encode noise. In order to balance these two criteria, an eigenvalue spectrum analysis is applied to the covariance matrix Σ of the original space that deals with the energy criterion and the intraclass compactness matrix $\mathbf{A}_c$ constructed from the intrinsic graph as defined in [147] for the magnitude criterion.

Particularly, two eigenvalue spectrums are first computed in terms of the relative magnitude $g_i / \sum_{k=1}^{m} g_k$ as y-axis and the number of principal components as x-axis (see Figure 3.3 for an example), where g_i is the i-th eigenvalue of the corresponding

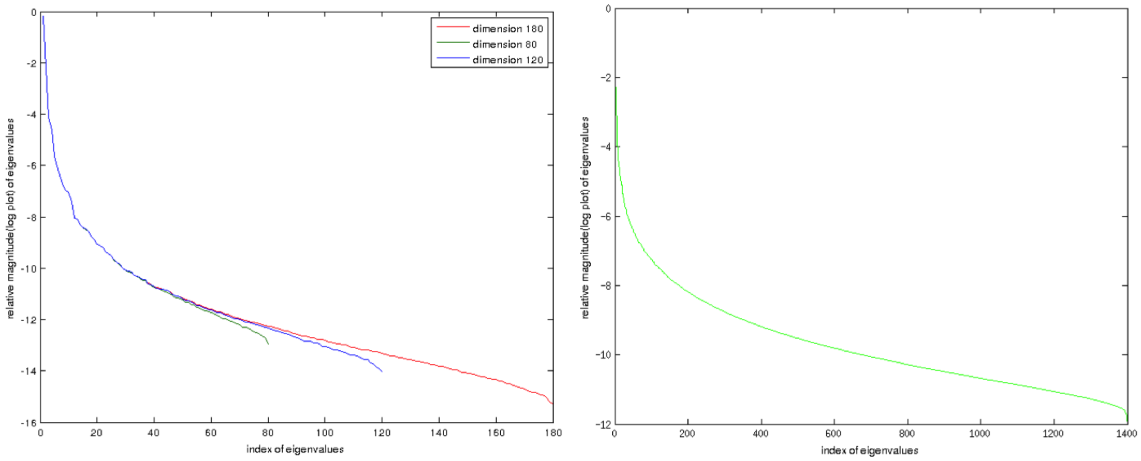


Figure 3.3 The spectrum of the covariance matrix Σ (left side) for three dimensions: 80, 120, 180 and the spectrum of the matrix $\mathbf{A}_c$ (right side) when the dimensionality is determined as 180 for the AR face data set.

matrix. Then the eigenvalue spectrum is truncated at an appropriate point and the number of principal components is determined by considering the above two criteria. After the dimensionality is determined, we can reduce the original dimension to the selected dimension using PCA with good generalization performance and derive the discriminatory features with the same dimension using the proposed IMFA.

3.3 Experiments

The performance of our proposed LLK method is assessed on four visual recognition tasks on eight representative data sets. Particularly, the AR Face Database [94] and the Extended Yale Face Database B [54] are evaluated for face recognition; the 15 Scenes data set [52] and the MIT-67 Indoor Scenes data set [109] are applied for scene recognition; as for object recognition, the Caltech 101 data set [55] and the Caltech 256 [30] are utilized; and the UIUC Sports Event data set [41] and the UCF50 data set [110] are assessed for action recognition. Please see sample images of different datasets in Figure 3.4. The size of each data set is also presented in Table 3.1. To further investigate the properties of the proposed

Table 3.1 The Data Sets used and their Sizes

data set	Size	data set	Size
Caltech 256 Object Category [30]	30,607	MIT-67 Indoor Scenes [109]	15,620
Caltech 101 Object Category [55]	9,144	UCF50 data set [110]	6,676
Scenes data set [52]	4,485	AR Face Database [94]	4,000
Extended Yale face B [54]	2,414	UIUC Sports Event [41]	1,574

method, we also conducted a comprehensive analysis of some critical issues concerning the performance.

The implementation details are as follows. The data is first represented as a pattern vector. For fair comparison or comparable results to the state-of-the-art methods, different pattern vectors are derived for different data sets. As for the two face databases, the pattern vector is formed as the concatenation of the column pixels. Also, to prove the robustness of the proposed method, the random faces [138], which is the row vectors of a randomly generated transformation matrix from a zero-mean normal distribution, is applied to project the face pattern vector into a dimension of 540 representation vector. Each row of the transformation matrix is normalized to unit length. For the 15 Scenes data set, the spatial pyramid feature provided by [42] is applied for fair comparison. For the MIT-67 Indoor Scenes data set, the Fisher vector feature [112] is used to represent the image. For the Caltech 101 and Caltech 256 data set, the proposed method is built upon the 4096 dimension image features that are extracted by using a pre-trained convolutional neural network CNN-M [12]. As for the UIUC Sports Event data set, the locally constrained linear coding (LLC) [133] method is used to represent the image. The SIFT feature used to construct the LLC representation follows the common settings [52] with a fixed step size as 8 pixels and patch size as 16 pixels for a single scale. As for the UCF50 data set [110],

Table 3.2 Comparison with other Popular Methods on AR Face Database under Experimental Setting 1

Experimental setting 1	Accuracy $\%$
D-KSVD [161]	85.40
LC-KSVD [42]	89.7
JDL [164]	91.7
FDDL [152]	92.00
SRC [138]	94.99
The proposed LLNc	96.14
The proposed LLKc	**97.00**

the action bank feature [111] is extracted to represent the video for fair comparison. All the model parameter are selected using 5-fold cross validation.

3.3.1 Face Recognition

AR Face Database As for the AR Face Database, a subset of the data [94] is often used for evaluation, where the images are cropped into the size of 165*120. The widely used experimental setting [138] is applied first. The original color space is replaced with the DCS color space [65], and the dimension of the face vector is reduced to 180. The model parameters are as follows: $\sigma = 1$, $\lambda = 0.02$, $\alpha = 0.1$, and $\beta = 1.5$ for the proposed LLK method, $\lambda_1 = 0.0$ and $\lambda_2 = 0.9$ for the shifted power transformation, $k = 5$ for LLKc and $k = 7$ for the LLNc. The results presented in Table 3.2 show significant improvements over other popular algorithms.

Moreover, we also follow another experimental setting [42] and [161] for assessing the robustness of the proposed method, where 20 images are randomly selected for training and 6 for testing for each person for 10 iterations. In this experimental setting, the random face [138] is applied to prove the robustness of the proposed method, and its dimension

Table 3.3 Comparison with other Popular Methods on the AR Face Database under Experimental Setting 2

Experimental setting 2	Accuracy $\%$
D-KSVD [161]	95.00
SRC [138]	97.50
LC-KSVD [42]	97.80
The proposed LLNc	**98.32**
The proposed LLKc	98.28

is reduced to 200. The model parameters are as follows: $\sigma = 1$, $\lambda = 0.02$, $\alpha = 0.1$, and $\beta = 1.5$ for the proposed method, $\lambda_1 = 0.0$ and $\lambda_2 = 0.9$ for the shifted power transformation, and $k = 5$ for both the LLKc and the LLNc. The experimental results shown in Table 3.3 demonstrate the robustness of the proposed method.

Extended Yale Face Database B As for the Extended Yale Face Database B, a cropped version [54] is often used that all the images are manually aligned, cropped, and then resized to 168 × 192. We follow the experimental setting [151] that 20 images are randomly selected for training for each subject, and the remaining images for testing for 10 iterations. Note that this experimental setting is more difficult than that in [161]. The image is first scaled to 42 × 48 and then the random faces [138] is applied to obtain the pattern vector to prove the robustness of the proposed method. The dimension of the pattern vector is further reduced to 350. The model parameters are as follows: $\sigma = 1$, $\lambda = 0.02$, $\alpha = 0.1$, and $\beta = 0.5$ for the LLK method, $\lambda_1 = 0.0$ and $\lambda_2 = 0.8$ for the shifted power transformation, and $k = 5$ for both the LLKc and the LLNc. The final results shown in Table 3.4 demonstrate that our proposed method significantly improves upon the other popular methods by more than 4 percentage points.

Table 3.4 Comparison with other Popular Methods on the Extended Yale Face Database B

Methods	Accuracy %
D-KSVD [151]	75.30
SRC [151]	90.00
FDDL [151]	91.90
The proposed LLNc	95.35
The proposed LLKc	**95.39**

3.3.2 Scene Recognition

The 15 Scenes data set As for the 15 Scenes data set [52], the experimental protocol in [52] and [149] is followed, where 100 images from each class are randomly selected for training and the remaining for testing for 10 iterations. The dimensionality of the feature vector is reduced to 500. The model parameters are as follows: $\lambda = 0.05$, $\alpha = 0.1$, and $\beta = 1.0$ for the LLK method, $\lambda_1 = 0.0$ and $\lambda_2 = 0.5$ for the shifted power transformation, $k = 1$ for the LLKc and $k = 2$ for the LLNc. The results in Table 3.5 show that our proposed method significantly outperform both the compared methods.

The MIT-67 Indoor Scenes data set As for the MIT-67 Indoor Scenes data set [109], the commonly used experimental setting [109] is followed. The Fisher vector feature [112] with dimensionality of 2*256*80 = 40960 is extracted from the SIFT descriptors of 80 dimension and a codebook with 256 visual words, and its dimensionality is further reduced to 2000. The model parameters are chosen as follows: $\lambda = 0.01$, $\alpha = 0.1$, and $\beta = 1.5$ for the LLNc while $\beta = 0.5$ for the LLKc, $\lambda_1 = 0.0$ and $\lambda_2 = 0.5$ for the shifted power transformation, and $k = 20$ for both the LLKc and the LLNc. Note that we derive the Fisher vector without learning the part detectors [43] and applying data augmentation [113].

Table 3.5 Comparison with other Popular Methods on the 15 Scenes Data Set

Methods	Accuracy %
KSPM [52]	81.40 ±0.50
ScSPM [149]	80.28 ±0.93
LLC [133]	80.57 ±−
KC [126]	76.67 ±0.93
D-KSVD [161]	89.10
LC-KSVD [42]	90.40
LaplacianSC [26]	89.7
DHVFC [27]	86.4
The proposed LLNc	**97.45** ±0.27
The proposed LLKc	93.54 ±0.45

However, as shown in Table 3.6, our proposed method still achieves the near state-of-the-art results when compared with the recent representative methods [43].

We further designed new experiments to comparatively evaluate our proposed LLKc and LLNc classifiers, as well as the support vector machine (SVM) for the new LLK representation derived from our LLK method. The LLK representations for both the training images and the test images are derived directly by optimizing Equation 3.3. For the MIT-67 Indoor Scenes data set, the SVM classifier achieves 57.31% classification accuracy, while our LLKc and LLNc methods achieve 58.18% and 59.12% classification accuracy, respectively. These results thus validate that our proposed LLKc and LLNc classifiers perform better than the SVM classifier for the learned LLK representations.

Table 3.6 Comparison with other Popular Methods on the MIT-67 Indoor Scenes Data Set

Methods	Mean Accuracy $\%$
ROI + Gist [109]	26.1
DPM [101]	30.4
Object Bank [56]	37.6
miSVM [57]	46.4
D-Parts [119]	51.4
DP + IFV [43]	**60.8**
CNN-SVM no Aug [113]	58.4
The proposed LLNc with FV	59.12
The proposed LLKc with FV	58.18

3.3.3 Object Recognition

The Caltech 101 data set As for the Caltech 101 data set [55], [133], we partition the whole data set randomly into 5, 10, 15, 20, 25, 30 training images per class and no more than 50 test images per class, and measure the performance using the average accuracy over 102 classes. For achieving comparable results to the state-of-the-art methods, the proposed method is built upon the 4096 dimension image representation features that are extracted by using a pre-trained convolutional neural network CNN-M [12]. Then we reduce the dimension to 1000 except the case with 5 training images, where the dimension is reduced to 500. The model parameters are as follows: $\sigma = 1.5$, $\lambda = 0.05$, $\alpha = 0.1$, and $\beta = 1.5$ for the LLK method, $\lambda_1 = 0.0$ and $\lambda_2 = 0.5$ for the shifted power transformation, and $k = 20$ if the training image size is larger than 20, otherwise k is the value of the training image size for both the LLKc and the LLNc.

Table 3.7 Comparison with other Popular Methods on the Caltech 101 Data Set

training images	5	10	15	20	25	30
SVM-KNN [157]	46.60	55.8	59.10	62.00	-	66.20
SPM [52]	-	-	56.40	-	-	64.60
ScSPM [149]	-	-	67.00	-	-	73.20
LLC [133]	51.15	59.77	65.43	67.74	70.16	73.44
SRC [138]	48.80	60.10	64.90	67.70	69.20	70.70
D-KSVD [161]	49.60	59.50	65.10	68.60	71.10	73.00
LC-KSVD [42]	54.00	63.10	67.70	70.50	72.30	73.60
Zeiler [156]	-	-	83.80	-	-	86.5
CNN-M + Aug [12]	-	-	-	-	-	87.15
The proposed LLNc	**76.96**	**82.71**	**84.79**	85.96	86.62	87.68
The proposed LLKc	76.49	82.34	84.76	**86.11**	**86.77**	**87.74**

Although our proposed method does not rely on the data augmentation and fine-tuning techniques [12], the results in Table 3.7 demonstrate that the proposed method still achieves at least comparable results to the state-of-the-art methods [156], [12].

Caltech 256 data set As for the Caltech 256 data set [30], the widely applied experimental setting [133] is followed, where 15, 30, 45, 60 images per category are randomly selected for training and no more than 25 images for testing for 3 iterations. The 4096 dimension CNN features are extracted by using a pre-trained convolutional neural network CNN-M [12] and their dimension are further reduced to 1000. The model parameters are as follows: $\sigma = 1.5$, $\lambda = 0.01$, $\alpha = 0.1$, and $\beta = 1.5$ for the LLK method, $\lambda_1 = 0.0$ and $\lambda_2 = 0.5$ for the shifted power transformation, and $k = 15$ for both the LLKc and the LLNc.

Table 3.8 Comparison with other Popular Methods on the Caltech 256 Data Set

training images	15	30	45	60
SAC [126]	-	27.17	-	-
ScSPM [149]	27.73	34.02	37.46	40.14
LLC [133]	34.36	41.19	45.31	47.68
IFV [104]	34.70	40.80	45.00	47.90
Bo et al. [6]	40.50	48.00	51.90	55.20
Zeiler [156]	65.70	70.60	72.70	74.20
The proposed LLNc	68.32	71.89	**74.13**	**75.47**
The proposed LLKc	**68.55**	**72.09**	74.07	75.36

Although our proposed method does not depend on the data augmentation and fine-tuning techniques [12], the results in Table 3.8 demonstrate that the proposed method still achieves at least comparable results to the state-of-the-art methods [156], [12].

3.3.4 Action Recognition

The UIUC Sports Event data set As for the UIUC Sports Event data set [41], we randomly select 70 images for training and 60 images for testing in each event class and report the average accuracy on 10 random training/testing splits in Table 3.9. We apply the LLC method [133] to derive the image vector with the dimension of 21504 and then reduce the dimension to 500. The model parameters are as follows: $\sigma = 1.5$, $\lambda = 0.01$, $\alpha = 0.1$, and $\beta = 0.5$ for the LLK method, $\lambda_1 = 0.0$ and $\lambda_2 = 0.5$ for the shifted power transformation, and $k = 30$ for both the LLKc and the LLNc. As shown in Table 3.9, the proposed method is able to outperform these popular methods that use support vector machine with linear or non-linear kernel.

Next we present new experimental results on evaluating the time used by our LLK method. In particular, we first report the time used by the LLC method [133] on processing the images in the UIUC Sports Event data set. We then report the time used by our LLK method for deriving the LLK representation. Specifically, Table X shows that the LLC method uses 730.18 seconds to process the images in the UIUC Sports Event data set, and our LLK method uses 6.28 seconds to derive the LLK representation. These experimental results reveal that the time used by our LLK method accounts for less than 1% of the time used by the LLC method. Therefore, the time incurred by our LLK method is negligible when compared with the time consumed by the LLC method. The reason is that the time used by LLC method depends on the number of local features sampled from all the images, such as the SIFT features. The time used by our LLK method in comparison depends on the number of images. Apparently, the number of local features is much larger than the number of images.

The UCF50 data set The UCF50 data set [110] is a challenge dataset that contains large variations in camera motion, object appearance, view point, etc. The experimental setting [111] is followed, where the data set is divided into 5 groups with similar size and the 5-fold group-wise cross-validation is used to measure the performance. This setting is more challenge than the leave-one-out-cross-validation with 25 folds in [110]. The action bank feature [111] is obtained and reduced from 14965 dimension to 500 dimension. The model parameters are as follows: $\sigma = 1.5$, $\lambda = 0.05$, $\alpha = 0.1$, and $\beta = 1.5$ for the LLK method, $\lambda_1 = 0.01$ and $\lambda_2 = 0.8$ for the shifted power transformation, and $k = 10$ for both the LLKc and the LLNc. We compare the proposed method with other popular methods, which is shown in Table 3.11. The results demonstrate the effectiveness of the proposed method.

Table 3.9 Comparison with other Popular Methods on the UIUC Sports Event Data Set

Methods	Accuracy %
SIFT + GGM [41]	73.4
OB [56]	76.3
CA-TM [98]	78.0
SAC [127]	82.04 ±2.37
ScSPM [149]	82.74 ±1.46
LLC [133]	81.41 ±1.84
LSAC [74]	82.29 ±1.84
LSC [26]	85.18 ±0.46
HMP [5]	85.7 ±1.3
The proposed LLNc	**86.79** ±1.33
The proposed LLKc	86.44 ±1.44

Table 3.10 The Time Used by the LLC Method and our Proposed LLK Method on the UIUC Sports Event Data Set

Method	Time (in seconds)
LLC [133]	730.18
LLK	6.28

Table 3.11 Comparison with other Popular Methods on the UCF 50 Action Recognition Data Set

Methods	Accuracy %
GIST [100]	38.8
Wang et al. [130]	47.9
Action bank [111]	57.9
SRC [138]	59.6
D-KSVD [161]	38.6
LC-KSVD [42]	53.6
JDL [164]	53.5
FDDL [152]	61.1
The proposed LLNc	62.42
The proposed LLKc	**62.66**

3.3.5 Effectiveness of the Shifted Power Transformation

The effectiveness of the shifted power transformation (SPT) is assessed by comparing the results of LLKc (already applying SPT), LLKc without SPT, LLNc (already applying SPT) and LLNc without SPT with different values of the parameter k. The results illustrated in Figure 3.5 show the improvement of performance using the shifted power transformation for all the values of k.

To further investigate the property of the shifted power transformation, we also compare the results of LLKc and LLNc for different values of the parameter λ_2, which is the power exponent in the shifted power transformation. All the other parameters are fixed and the value of $\lambda_1 = 0.0$. As shown in Figure 3.6, the value of λ_2 has large impact on the performance of the proposed method.

3.3.6 Sensitivity Analysis

Theoretical analysis in Section 3.2.3 shows that the shifted power transformation and the L_2 normalization are capable of reducing the sensitivity of our LLK method to the parameter σ for robust kernel density estimation. This section presents empirical evidence as well by comparing the results of LLKc and LLNc for different values of the parameter σ.

As shown in Figure 3.7, the performance is indeed not sensitive to the parameter σ after the application of both the shifted power transformation and the L_2 normalization. In comparison, the performance of our LLK method depends heavily on the value of σ without the shifted power transformation. The performance will drop below 10% without the necessary L_2 normalization.

Theoretical analysis in Section 3.2.4 demonstrates that the coefficient truncating method is able to discard the distant samples to avoid their adverse impact on the performance. To provide empirical evidence, the performance of our LLK method for different values of the parameter k is presented. The results in Figure 3.5 prove that neither a very small value nor a very large value of k are necessary for better performance. On the contrary, a careful selection of k is critical to the performance.

3.3.7 Grouping Effect of the Nearest Neighbors

The grouping effect of the nearest neighbors are shown to be an important property for deriving an approximation to the ideal representation. To evaluate the tightness of the bound and the degree of approximation to the ideal representation, experimental results are presented in this section for different values of the parameter λ, α and β in terms of three variables: the classification accuracy, the true activation ratio (TAR) and the false activation ratio (FAR).

The true activation ratio (TAR) is defined as follows:

$$TAR = \frac{\sum_{i=1}^{N_{test}} t_i}{N_{test}} \tag{3.20}$$

where N_{test} is the number of test samples and t_i represents the number of the correctly activated (non-zero) coefficients for the i-th test sample.

The false activation ratio (FAR) is similarly defined as follows:

$$FAR = \frac{\sum_{i=1}^{N_{test}} f_i}{(c-1)N_{test}} \tag{3.21}$$

where c is the number of classes and f_i represents the number of the mistakenly activated (non-zero) coefficients for the i-th test sample.

Taking the MIT-67 Indoor Scenes data set for example, if there are 80 training samples for a class and the test sample comes from the same class, then the TAR of the ideal representation is 80, which is the size of the training samples in this class, and the FAR of the ideal representation is 0.

The results in Table 3.12 not only show the effectiveness of the grouping effect of the nearest neighbors but also demonstrate the necessity of the trade-off between the value of FAR and the value of TAR to achieve the best performance due to the observation that the increasing (decreasing) of TAR sometimes may result in the increasing (decreasing) of FAR, which may deteriorate the performance.

3.3.8 Stability Discussion

This section provides the stability discussion for different parameters and explains the reasons why our method is stable when some parameters change. Specifically, we consider the following parameters: λ, α, β, σ, and k.

First, as shown in Table 3.12, the performance drops significantly when the value of λ increases. The reason is mainly due to the importance of the parameter λ in the sparse representation related problems. As a comparison, the performance is very stable when the value of α and value of β change within a reasonable range.

Second, as shown in Figure 3.7, the performance is stable under different values of σ. To further show the stability of our method to the parameter σ, some extreme cases are also

Table 3.12 The Classification Accuracy, FAR and TAR for Different Values of the Parameter λ, α and β on the MIT-67 Indoor Scenes Data Set

λ	α	β	LLKc (%)	LLNc (%)	TAR	FAR
0.01	0.1	1.5	57.65	59.12	30.60	15.35
0.05	0.1	1.5	57.13	56.06	11.05	1.44
0.10	0.1	1.5	49.15	48.60	4.35	0.16
0.01	0.1	1.5	57.65	59.12	30.60	15.35
0.01	0.3	1.5	56.91	58.52	35.92	18.33
0.01	0.5	1.5	55.95	57.88	39.91	20.63
0.01	0.1	0.5	58.18	58.60	30.45	15.34
0.01	0.1	1.0	57.80	58.82	30.52	15.34
0.01	0.1	1.5	57.65	59.12	30.60	15.35
0.01	0.1	2.0	57.95	59.12	30.66	15.36

considered, where the values of σ such as 20, 30 are also evaluated. The performance of LLKc is 58.40 (even better than the reported results above) and the performance of LLNc is 58.49 for both values. The main reason why our method is stable under different values of σ is that both the shifted power transformation and the L_2 normalization are applied.

Third, as shown in Figure 3.5, the performance becomes stable when k exceeds a certain point. The reason might be that if k is too small, the kernel density estimation is not reliable, which may harm the performance.

3.3.9 Computational Efficiency

This section assesses the computational efficiency of our proposed method. The major computational cost of our method is due to the size of the dictionary, as the larger the size of the dictionary is, the more time the computation of the step size and the optimization require. Thus we propose two methods to improve the computational efficiency of our proposed method.

First, we extend the safe screening rule [132] for the conventional sparse representation problem to our proposed method. Please note that the safe screening rule is a process before optimizing the objective function of the sparse representation so that we can discard the training samples with zero coefficients for reducing the computational time.

The extended safe screening rule is defined as follows:

Proposition 3.3.1 ***A New Safe Screening Rule.***
if $|\boldsymbol{x}^t\boldsymbol{b}_i + \alpha\beta\boldsymbol{d}^t\boldsymbol{e}_i| < \lambda_m - \sqrt{(1+\alpha)(1+\alpha\beta^2||\boldsymbol{d}||^2)}(\frac{\lambda_m}{\lambda} - 1)$*, then* $v_i = 0$*, where* $\boldsymbol{e}_i$ *is a vector whose elements are all zeros except for the* i*-th element, whose value is 1, and* $\lambda_m = \max_i |\boldsymbol{x}^t\boldsymbol{b}_i + \alpha\beta\boldsymbol{d}^t\boldsymbol{e}_i|$.

Second, we apply the FISTA algorithm with backtracking [3] instead of the basic FISTA algorithm to avoid the large scale eigenvalue decomposition problem. A good initialization is required for fast convergence.

Table 3.13 Comparison among Different Techniques for Computational Efficiency

Methods	Time (s)
Basic implementation	6.55
Extended screening rule	3.12
FISTA with backtracking [3]	3.13

Table 3.13 shows the experimental results measured in terms of the average optimization time, which is defined as the total time for optimization divided by the number of the test images. Specifically, the experimental results in Table 3.13 reveal that both the extended safe screening rule and the FISTA algorithm with backtracking can improve the computational efficiency for our method.

3.3.10 Non-negative Constraint

In this section, we present the evaluation of the non-negative constraint ($v_i \geq 0$) for the objective function defined in Equation 3.3. Given such a new constrained optimization problem, the structure of the FISTA algorithm remains the same but with the only difference of the proximal operator as our method becomes an extension to the non-negative lasso problem [118]. We replace the original soft thresholding operator with an efficient projection operator [21] considering the non-negative constraint. The experimental result on the MIT-67 Indoor Scenes data set achieves 56.60 for LLKc and 56.49 for LLNc, which shows that the non-negative constraint does not necessarily contribute to better performance.

3.3.11 Group Regularization

In this section, we demonstrate the evaluation when group regularization (more than one regularization term) is incorporated into our method. Specifically, a simple form of group regularization is evaluated, namely the $L_1 + L_2$ grouping regularization by adding a quadratic term $\lambda_3||\mathbf{v}||^2$ ($\lambda_3 = 0.05$ in our experiment for best performance) to the objective function defined in Equation 3.3. The results obtained on the MIT scene 67 data set are 58.49 and 58.41 percent for LLKc and LLNc, respectively when using this grouping regularization term. It can be seen that the results of group regularization are similar to our method.

3.3.12 Discussion of the Improved Marginal Fisher Analysis

The proposed LLK method relies on the assumption that the nearest training samples are highly probable to share the same class label to the test sample. That is the reason why the L_2 term $||\mathbf{v} - \beta\mathbf{d}||^2$ is applied in Equation 3.3. However, this assumption not always holds in the real world application. Therefore, we can improve the proposed LLK method by introducing a metric learning method.

We find out that the proposed improved marginal fisher analysis method not only plays the role of feature extraction, but also plays the role of metric learning. Because it is derived by pulling close training samples in the same class, while push away training samples from different classes. Consequently, **d** can be computed by using the features extracted by the improved marginal Fisher analysis and the assumption can be realized for the LLK method.

3.3.13 Double L_1 Norm

The proposed LLK method is based on a combination of both L_1 norm and L_2 norm. In the final classification stage, the coefficient truncating method is applied for the LLKc classification method. It is well-known that the L_1 norm can also have the effect of the coefficient truncating since it can make the coefficients sparse.

Therefore we can extend the proposed LLK method by using double L_1 norms as follows:

$$\min_{\mathbf{v}} ||\mathbf{x} - \mathbf{B}\mathbf{v}||^2 + \lambda||\mathbf{v}||_1 + \alpha||\mathbf{v} - \beta\mathbf{d}||_1 \tag{3.22}$$

The new introduced L_1 norm $||\mathbf{v} - \beta\mathbf{d}||_1$ is a relative sparsity criterion, which enforces a specific structure of the sparse representation **v**. The conventional sparse representation **v** is sparse with respect to the origin, namely a zero vector that resides in the m dimensional space, which leads to many zero elements in **v**. In contrast, such an extension constrains the sparsity with respect to a more informative vector in addition to the origin vector to incorporate more structural and discriminative information.

The experiment on the MIT-67 Indoor Scenes dataset using both the LLKc and LLNc can achieve an accuracy of 59.02 and 59.31, respectively. It shows that the double L_1 norms method can achieve slightly better result and meantime, the coefficient truncating method is not required any more.

3.4 Conclusions

This chapter presents a novel LLK method for robust visual recognition. The proposed method derives a new representation that has the grouping effect of the nearest neighbors using the criteria of reconstruction, locality, and sparsity. Then two classifiers, namely the Locally Linear KNN based classifier (LLKc) and the Locally Linear Nearest mean based classifier (LLNc), are proposed. Both the LLKc and LLNc are related to the Bayes decision rule for minimum error. Besides, some other theoretical issues of the proposed LLK method are also addressed such as the non-negative constraint, the group regularization and the computational efficiency. Furthermore, the improved marginal Fisher analysis (IMFA) is proposed for feature extraction, the shifted power transformation and a coefficient truncating method are applied as well. The feasibility of the proposed LLK method is successfully evaluated for several visual recognition tasks.

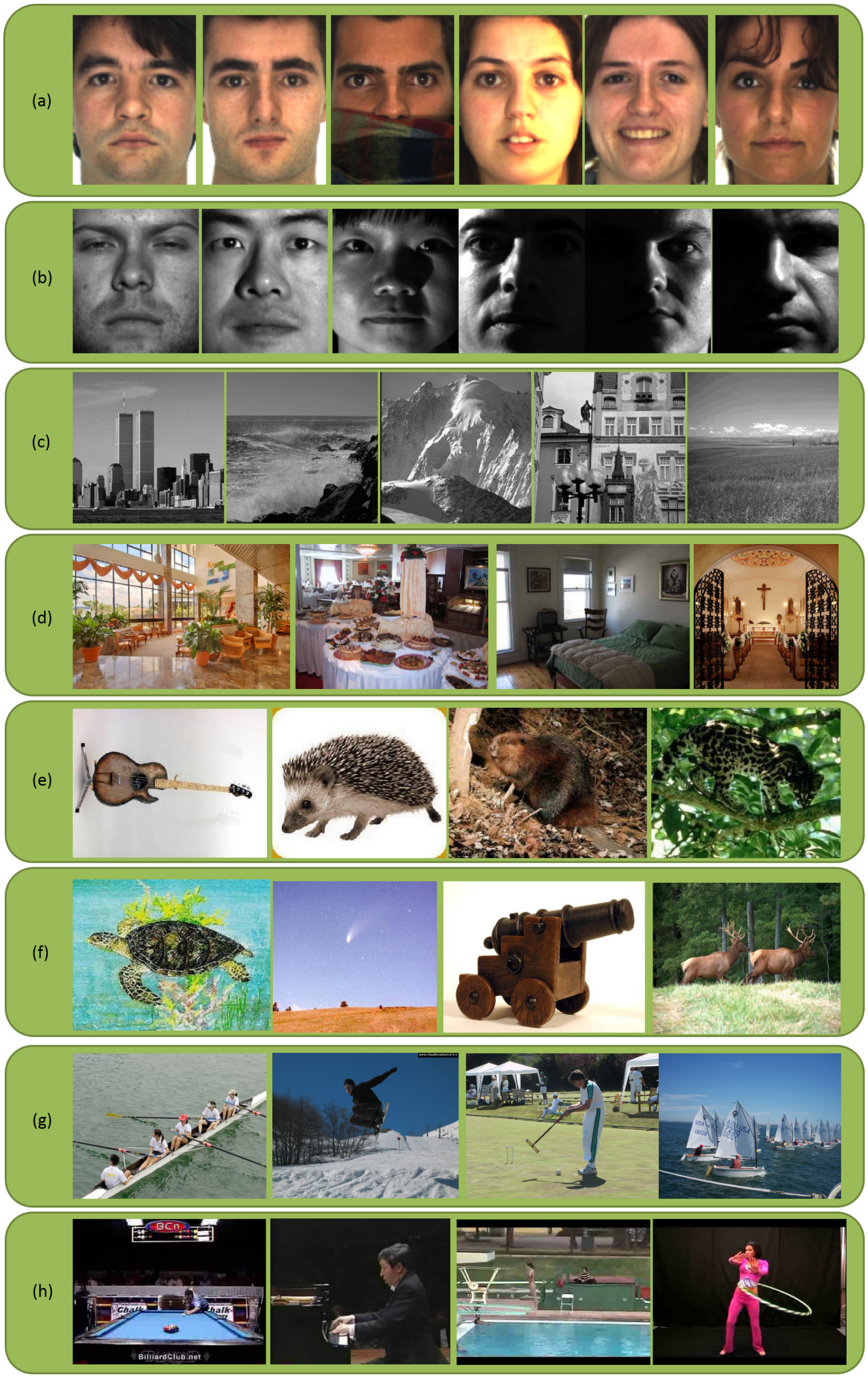


Figure 3.4 Sample images of the data sets: (a) the AR Face Database, (b) the Extended Yale Face Database B, (c) the 15 Scenes data set, (d) the MIT-67 Indoor Scenes data set, (e) the Caltech 101 data set, (f) the Caltech 256 data set (g) the UIUC Sports Event data set and (h) the UCF50 data set.

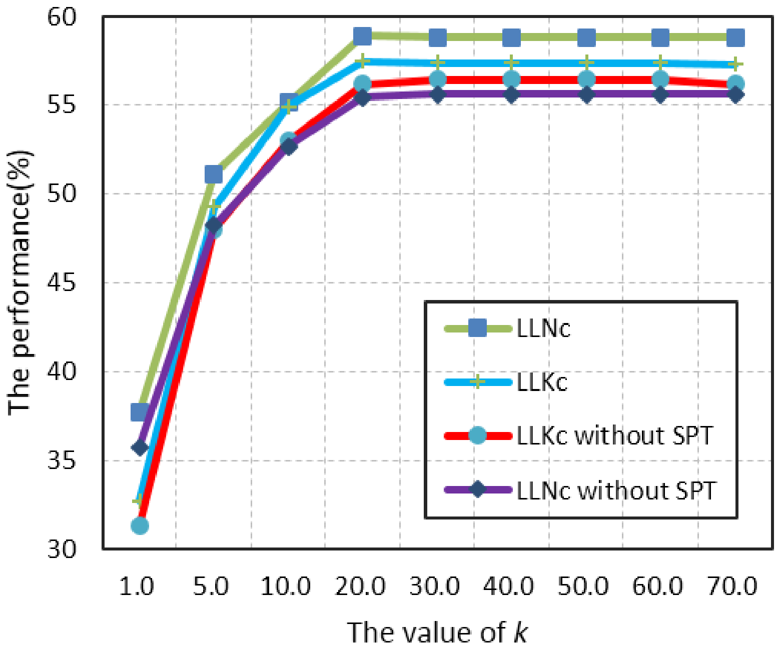


Figure 3.5 The evaluation of the shifted power transformation for different values of the parameter k on the MIT-67 Indoor Scenes data set.

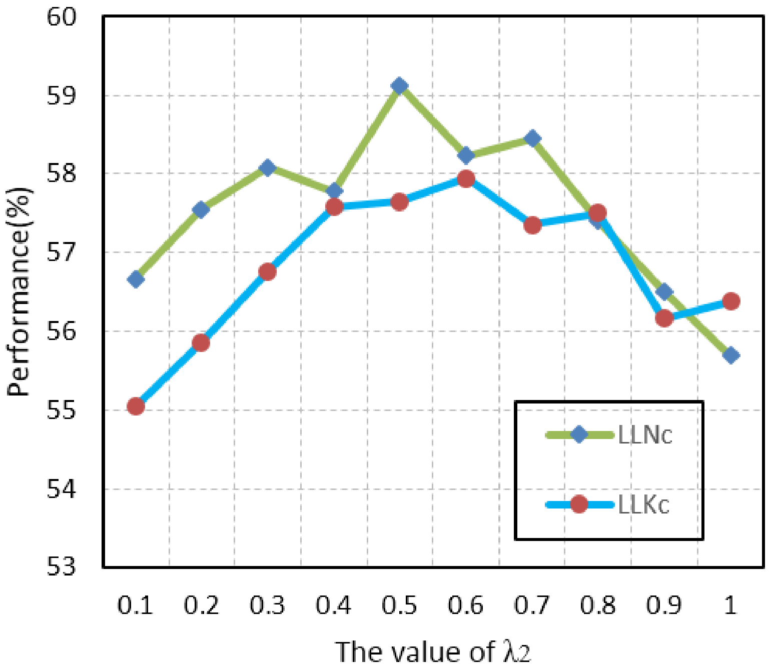


Figure 3.6 The evaluation of the shifted power transformation for different values of the parameter λ_2 on the MIT-67 Indoor Scenes data set

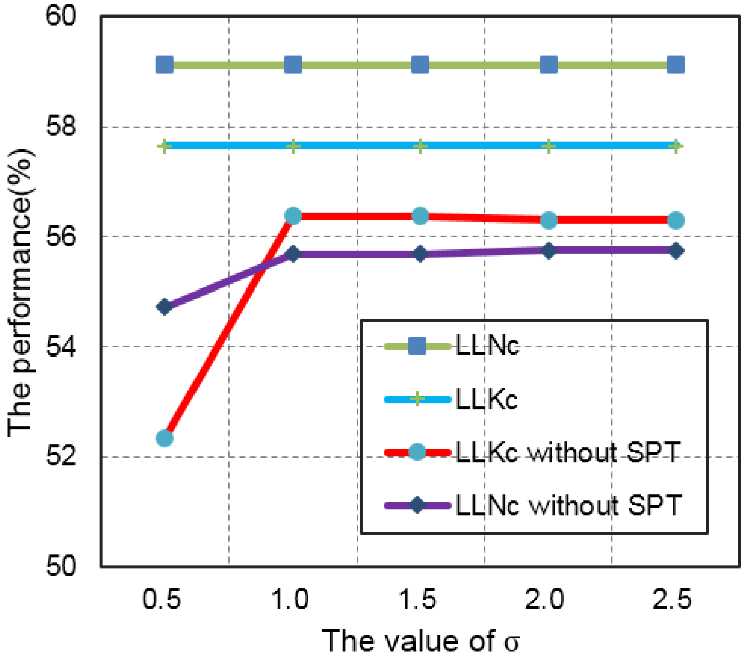


Figure 3.7 The evaluation of our LLK method for different values of the parameter σ on the MIT-67 Indoor Scenes data set

CHAPTER 4

GENERATIVE AND DISCRIMINATIVE SPARSE REPRESENTATION

4.1 Introduction

Although the sparse representation method achieves impressive results in various challenging tasks, such as face recognition [138], and scene recognition [78], [81], [84], [76], it lacks the support of discriminative information as it is only derived from the representation criterion. Efforts have been made recently for incorporating the discriminative criterion into the sparse representation criterion. These methods can be classified into two major categories. One category is to devise a discriminative dictionary by using a set of sub-dictionaries for each class [91], [135], [164], [151], [152], [83]. The other way is to add constraints to the sparse representation for learning the dictionary [161], [42]. Despite their success, the generative information of the dictionary, namely the class conditional probability of each dictionary item, is still disregarded. The complementary nature of the discriminative and the generative approaches, which has been well studied in the past few years [96], [39], has demonstrated the effectiveness of a combination of both the discriminative information and the generative information.

This chapter therefore presents a new generative and discriminative sparse representation (GDSR) method by integrating the conventional sparse representation, a new generative criterion and a new discriminative criterion. The proposed GDSR method intrinsically models a hybrid paradigm of both the generative information and the discriminative information. It applies the generative model as a regularization for the discriminative model to avoid over-fitting from the regularization point of view.

Figure 4.1 illustrates the system architecture of the proposed GDSR method. Specifically, the generative criterion plays the role of generative modeling by representing each dictionary item as a linear combination of the training samples and emphasizing the coefficients of the nearest training samples. Theoretical analysis shows that these coefficients, termed as dictionary distribution coefficients, are capable of approximately modeling the class conditional probability of each dictionary item. The discriminative criterion utilizes the underlying topology of the sparse representations by considering only the k nearest neighbors for defining a new discriminant analysis criterion with newly defined within-class and between-class scatter matrices. In addition, a new classification method, namely the generative and discriminative sparse representation based classification (GDSRc) method, is proposed by utilizing both the new sparse representation and the dictionary distribution coefficients. Theoretical analysis shows that the GDSRc method is related to the Bayes ideal feature [25] which has the minimum error for classification.

The optimization issue of the proposed GDSR method is discussed as well by using a coordinate descent method, which iteratively updates the sparse representation, the dictionary and the dictionary distribution coefficients. In particular, the sparse representation is derived by using the fast iterative shrinkage thresholding (FISTA) algorithm [3], where the issues of initialization and step size are discussed. The dictionary is obtained by using a fast approximation and the optimization of the Lagrange dual problem. The dictionary distribution coefficients are derived by solving a variant of the ridge regression problem.

The effectiveness of the proposed GDSR method is evaluated on various visual recognition tasks, such as face recognition on the AR Face Database [94] and the Extended Yale Face Database B [54], computational fine art analysis on the Painting-91 dataset [45], scene recognition on the 15 Scenes dataset [52] and the MIT-67 Indoor Scenes dataset [109], as well as fine grained recognition on the CUB-200-2011 dataset [128]. The experimental results show the feasibility of the proposed method.

To summarize, the proposed GDSR method makes the following major contributions.

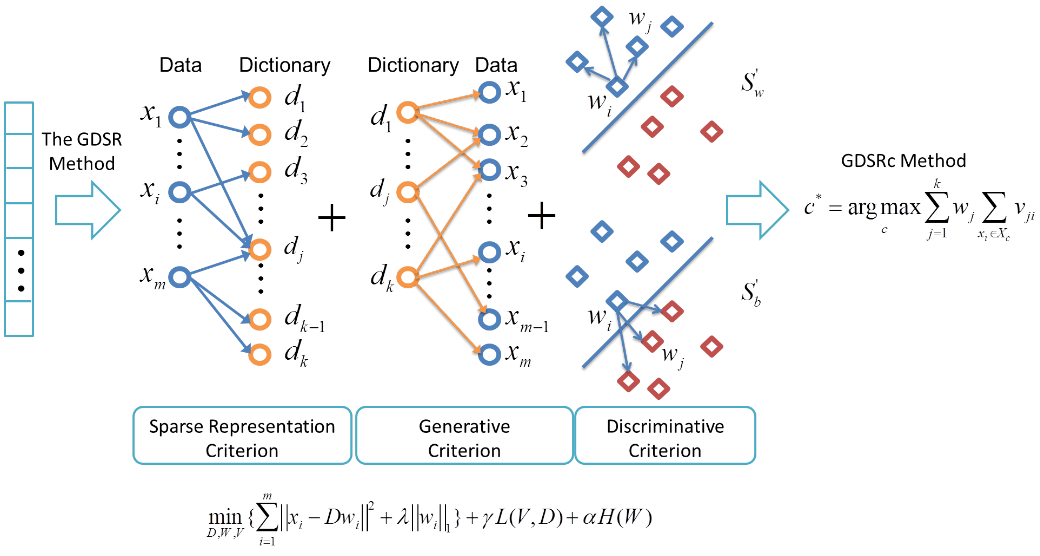


Figure 4.1 The system architecture of the proposed GDSR method.

- First, the GDSR method explicitly models the class conditional probability of each dictionary item by using a new generative criterion and reveals the generative structure of the dictionary.

- Second, the GDSR method achieves its discriminatory power by introducing a new option for defining the scatter matrices for discriminant analysis.

- Third, a new GDSRc method is proposed for classification and the power of the GDSRc method is revealed by its connection to the Bayes ideal feature.

4.2 Generative and Discriminative Sparse Representation

Dictionary learning plays an important role in the conventional sparse representation method. An important question arising from the issue of dictionary learning is how the dictionary can help classify testing data, which has been well studied in many papers. Another equally important question, which has received much less attention, is how a dictionary item is generated given a specific category, namely the generative information of the dictionary. One naive answer is to construct a dictionary that consists of carefully selected training samples [138]. In this scenario, each dictionary item corresponds to a training sample from a specific category. Such a dictionary might achieve good results, however, the performance of this method relies heavily on the selection of the training samples and the number of the selected training samples.

Our solution to both questions is the proposed GDSR method, which explicitly models the class conditional probability of each dictionary item $p(\mathbf{d}_j|c)$, where $\mathbf{d}_j$ is the j-th dictionary item and c is the class label, and introduces a new discriminative criterion for enhancing the discriminative power of the dictionary. Suppose the training sample data matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ is composed of m samples $[\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m]$ and each sample resides in an n dimensional space. The dictionary $\mathbf{D} \in \mathbb{R}^{n\times k}$ can be represented as $[\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_k]$, where each dictionary item $\mathbf{d}_j (j = 1, 2, ..., k)$ also resides in the n dimensional space. Then our GDSR method derives the sparse representations $\mathbf{w}_i \in \mathbb{R}^{k\times 1} (i = 1, 2, ..., m)$ for each training sample $\mathbf{x}_i$, and the dictionary distribution coefficients $\mathbf{v}_j \in \mathbb{R}^{m\times 1} (j = 1, 2, ..., k)$ for each dictionary item $\mathbf{d}_j$.

Specifically, the GDSR method is defined as follows:

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{W},\mathbf{V}} \sum_{i=1}^{m} ||\mathbf{x}_i - \mathbf{D}\mathbf{w}_i||^2 + \lambda||\mathbf{w}_i||_1 + \gamma L(\mathbf{V}, \mathbf{D}) + \alpha H(\mathbf{W}) \\ &s.t. \quad ||\mathbf{d}_j|| \leq 1, (j = 1, 2, ..., k) \end{aligned} \tag{4.1}$$

The first term in Equation 4.1 is the conventional sparse representation criterion, where the parameter λ controls the L_1 normalization. The second term $L(\mathbf{V}, \mathbf{D})$, which represents the generative criterion, is defined as follows:

$$L(\mathbf{V}, \mathbf{D}) = \sum_{j=1}^{k} ||\mathbf{d}_j - \mathbf{X}\mathbf{v}_j||^2 + \sigma||\mathbf{v}_j - \eta\mathbf{p}_j||^2 \tag{4.2}$$

where $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k]$ is the matrix that consists of the dictionary distribution coefficients vector $\mathbf{v}_j = [v_{j1}, v_{j2}, ..., v_{jm}]^t$. The vector $\mathbf{p}_j = [p_{j1}, p_{j2}, ..., p_{jm}]^t \in \mathbb{R}^m$ represents the distance measure between the dictionary item $\mathbf{d}_j$ and the training sample

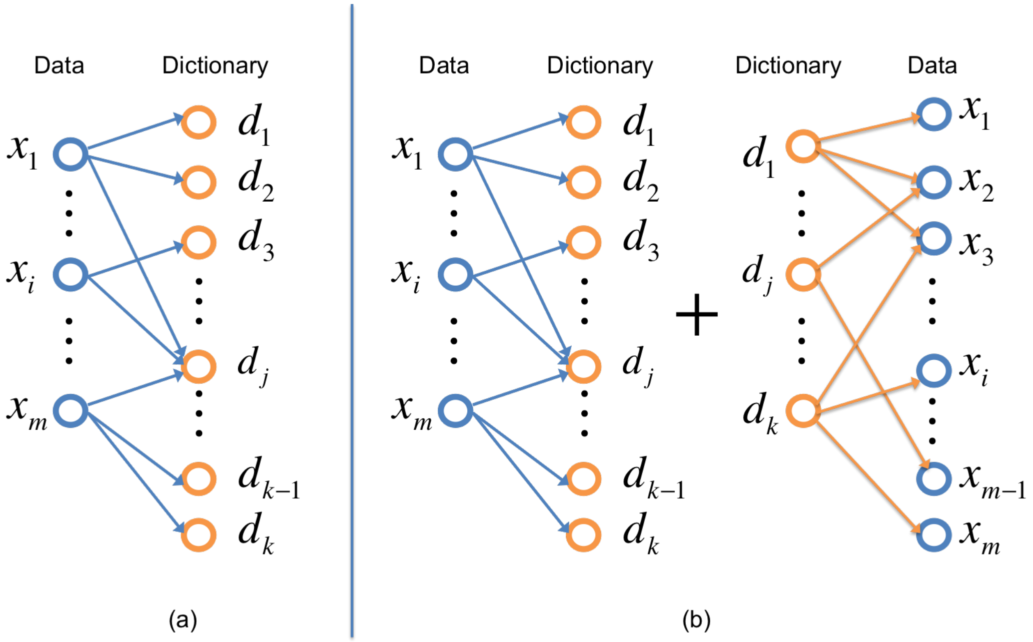


Figure 4.2 (a) Conventional view represents the data as a linear combination of the dictionary items. (b) The proposed GDSR represents each dictionary item as a linear combination of the data as well.

$\mathbf{x}_i$ as follows:

$$p_{ji} = exp\{-\frac{1}{2h^2}||\mathbf{d}_j - \mathbf{x}_i||^2\} \tag{4.3}$$

where the parameter h controls the decay speed. Please note that $p_{ji} \leq 1$ and $||\mathbf{p}_j||^2$ can be normalized.

As illustrated in Figure 4.2, traditional view of the dictionary learning is to represent the training sample as a linear combination of the dictionary items. In comparison, our generative criterion demonstrates a reciprocal viewpoint as well, which represents each dictionary item as a linear combination of the training samples. As a matter of fact, the dictionary items and the training samples consist of a bipartite graph and they influence each other mutually. In addition, the generative criterion also makes a constraint on the dictionary distribution coefficients vector $\mathbf{v}_j$, where the coefficients are proportional to the distance between the dictionary item and the training sample, in order to estimate the class conditional probability of each dictionary item $p(\mathbf{d}_j|c)$ by using $\mathbf{v}_j$ (Proposition 4.2.1).

The third term is the discriminative criterion, which is defined as follows:

$$H(\mathbf{W}) = \mathbf{tr}(\beta \mathbf{S}_w^{'} - (1-\beta)\mathbf{S}_b^{'}) \tag{4.4}$$

The new within-class scatter matrix is defined as $\mathbf{S}_w^{'} = \sum_{i=1}^{m} \sum_{(\mathbf{w}_i,\mathbf{w}_j)\in T_k^w} (\mathbf{w}_i - \mathbf{w}_j)(\mathbf{w}_i - \mathbf{w}_j)^t$, where T_k^w represents the set of $(\mathbf{w}_i, \mathbf{w}_j)$ pairs where the sample $\mathbf{x}_i$ and sample $\mathbf{x}_j$ are among their k nearest neighbors respectively in the same class. The new between-class scatter matrix is then defined as $\mathbf{S}_b^{'} = \sum_{i=1}^{m} \sum_{(\mathbf{w}_i,\mathbf{w}_j)\in T_k^b} (\mathbf{w}_i - \mathbf{w}_j)(\mathbf{w}_i - \mathbf{w}_j)^t$, where T_k^b represents the set of the k nearest $(\mathbf{w}_i, \mathbf{w}_j)$ pairs among all the $(\mathbf{w}_i, \mathbf{w}_j)$ pairs between sample $\mathbf{x}_i$ and sample $\mathbf{x}_j$ from the different classes.

This discriminative criterion utilizes the underlying topology of the sparse representation of the training samples for defining new within-class and between-class scatter matrices by considering only the k nearest neighbors. The new discriminative criterion can be further transformed into $H(\mathbf{W}) = \mathbf{tr}(\mathbf{W}\mathbf{L}\mathbf{W}^t)$, where $\mathbf{L} = 2\beta(\mathbf{D}_w - \mathbf{W}_w) - 2(1-\beta)(\mathbf{D}_b - \mathbf{W}_b)$. In particular, let $\mathbf{W}_w$ be a matrix, whose elements $W_w(i,j) = 1$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ are among the k nearest neighbors of each other in the same class, and $W_w(i,j) = 0$, otherwise. Let $\mathbf{W}_b$ be a matrix, whose elements $W_b(i,j) = 1$ if the pair $(\mathbf{w}_i, \mathbf{w}_j)$ is among the k nearest pairs from all the pairs among the samples of the different classes, and $W_b(i,j) = 0$, otherwise. Let $\mathbf{D}_w$ and $\mathbf{D}_b$ be diagonal matrices, whose main diagonal elements are $D_w(i,i) = \sum_{j\neq i} W_w(i,j)$ and $D_b(i,i) = \sum_{j\neq i} W_b(i,j)$, respectively.

One important property of the proposed GDSR method is its modeling of class conditional probability of each dictionary item $p(\mathbf{d}_j|c)$ stated as the following generative property 4.2.1.

Proposition 4.2.1 ***Generative Property*** *Given that* **V** *is the derived dictionary distribution coefficients by the proposed GDSR method, the class conditional probability of each* *dictionary item $p(\boldsymbol{d}_j|c)$ is modeled as follows.*

$$p(\boldsymbol{d}_j|c) \propto \sum_{\boldsymbol{x}_i \in \boldsymbol{X}_c} v_{ji} \tag{4.5}$$

where $\mathbf{X}_c$ is the set of training samples in the c-th class.

The proof of Proposition 4.2.1 is shown as follows.

If **V** is the derived dictionary distribution coefficients by the proposed GDSR method, then the objective function in Equation 4.1 satisfies $O(\mathbf{D}, \mathbf{W}, \mathbf{v}_j) \leq O(\mathbf{D}, \mathbf{W}, \mathbf{0})$, which means

$$L(\mathbf{v}_j, \mathbf{D}) \leq L(\mathbf{0}, \mathbf{D}) \tag{4.6}$$

then we have

$$\begin{aligned} ||\mathbf{d}_j - \mathbf{X}\mathbf{v}_j||^2 + \sigma||\mathbf{v}_j - \eta\mathbf{p}_j||^2 &\leq ||\mathbf{d}_j||^2 + \sigma||\eta\mathbf{p}_j||^2 \\ &\leq 1 + \sigma\eta^2 \end{aligned} \tag{4.7}$$

Therefore $||\mathbf{v}_j - \eta\mathbf{p}_j||^2$ is bounded by a small positive constant so that $\mathbf{v}_j \approx \eta\mathbf{p}_j + \mathbf{const}$.

As a result, the GDSRc is approximated as follows

$$\begin{aligned} \sum_{\mathbf{x}_i \in \mathbf{X}_c} v_{ji} &\approx \sum_{\mathbf{x}_i \in \mathbf{X}_c} \eta \times p_{ji} + const \\ &\propto \sum_{\mathbf{x}_i \in \mathbf{X}_c} exp\{-\frac{1}{2h^2}||\mathbf{d}_j - \mathbf{x}_i||^2\} \\ &\propto \frac{1}{s \times m_c} \sum_{\mathbf{x}_i \in \mathbf{X}_c} k(\frac{\mathbf{d}_j - \mathbf{x}_i}{s}) \\ &\propto p(\mathbf{d}_j|c) \end{aligned} \tag{4.8}$$

where m_c is the number of training samples in the c-th class and s is the bandwidth that controls the degree of smoothing and it is fixed for all the classes. We use the zero mean and unit variance Gaussian kernel $k(u) = \frac{1}{\sqrt{2\pi}}exp(-\frac{u^2}{2})$ here for kernel density estimation

[25] of the conditional probability $p(\mathbf{d}_j|c)$. Please note that the prior probability $p(c)$ is assumed to be equal for all the classes.

In summary, the GP is a coarse estimation of the class conditional probability $p(\mathbf{d}_j|c)$ in the view of density estimation, and the approximation error mainly comes from the $\mathbf{v}_j \approx \eta\mathbf{p}_j + \mathbf{const}$ (relatively large σ and small η are preferred) and the kernel density estimation error.

Please note that conventional way to estimate $p(\mathbf{d}_j|c)$ assumes some parametric distribution first, such as Bernoulli distribution. In comparison, the generative property of the proposed method shows that $p(\mathbf{d}_j|c)$ is estimated from the kernel density estimation point of view. The generative property (GP) presents a coarse estimation of the class conditional probability of each dictionary item instead of an accurate estimation since our goal is to accurately classify data instead of accurately estimating the probability, and such a coarse modeling carries sufficient information for improving the classification performance as shown in the experimental section.

Another interesting property of the proposed GDSR method is the grouping property of the dictionary distribution coefficients (GPDDC) for $\mathbf{v}_j$ as shown in Property 4.2.2.

Proposition 4.2.2 ***Grouping Property of the Dictionary Distribution Coefficients*** *Let $\mathbf{v}_j = [v_{j1}, v_{j2}, ..., v_{jm}]^t$ be the solution of optimizing the objective function defined in Equation 4.1, $\rho = \boldsymbol{x}_s^t\boldsymbol{x}_t$ be the sample correlation of two training samples $\boldsymbol{x}_s$ and $\boldsymbol{x}_t$ $(s, t = 1, 2, ..., m)$, and $M(s,t) = |v_{js} - v_{jt}|$ be the difference between two coefficients of $\mathbf{v}_j$, then the grouping property of the dictionary distribution coefficients (GPDDC) is defined as follows.*

$$M(s,t) \leq C\sqrt{1-\rho} + \eta|p_{js} - p_{jt}| \tag{4.9}$$

where all the samples are normalized using the L_2 normalization and $C = \sqrt{\frac{2(1+\sigma\eta^2)}{\sigma^2}}$ is a constant.

The proof of Proposition 4.2.2 is shown as follows.

Let the objective function defined in Equation 4.1 denoted as $O(\mathbf{D}, \mathbf{W}, \mathbf{V})$, then take the derivative of v_{js} and v_{jt} as

$$\begin{aligned} \frac{\partial O}{\partial v_{js}} &= -2\mathbf{x}_s^t(\mathbf{d}_j - \mathbf{X}\mathbf{v}_j) + 2\sigma(v_{js} - \eta p_{js}) = 0 \\ \frac{\partial O}{\partial v_{jt}} &= -2\mathbf{x}_t^t(\mathbf{d}_j - \mathbf{X}\mathbf{v}_j) + 2\sigma(v_{jt} - \eta p_{jt}) = 0 \end{aligned} \tag{4.10}$$

Then $\frac{\partial O}{\partial v_{js}} - \frac{\partial O}{\partial v_{jt}}$ is

$$\sigma(v_{js} - v_{jt}) = (\mathbf{x}_s - \mathbf{x}_t)^t(\mathbf{d}_j - \mathbf{X}\mathbf{v}_j) + \sigma\eta(p_{js} - p_{jt}) \tag{4.11}$$

Take the absolute value on both sides and make use of $O(\mathbf{D}, \mathbf{W}, \mathbf{v}_j) \leq O(\mathbf{D}, \mathbf{W}, \mathbf{0})$ and all the training samples are L_2 normalized.

$$\begin{aligned} |v_{js} - v_{jt}| &= |\frac{1}{\sigma}(\mathbf{x}_s - \mathbf{x}_t)^t(\mathbf{d}_j - \mathbf{X}\mathbf{v}_j) + \eta(p_{js} - p_{jt})| \\ &\leq |\frac{1}{\sigma}(\mathbf{x}_s - \mathbf{x}_t)^t(\mathbf{d}_j - \mathbf{X}\mathbf{v}_j)| + \eta|p_{js} - p_{jt}| \\ &\leq \frac{1}{\sigma}||\mathbf{x}_s - \mathbf{x}_t|| * ||\mathbf{d}_j - \mathbf{X}\mathbf{v}_j|| + \eta|p_{js} - p_{jt}| \\ &= \frac{1}{\sigma}\sqrt{2(1-\rho)} * ||\mathbf{d}_j - \mathbf{X}\mathbf{v}_j|| + \eta|p_{js} - p_{jt}| \\ &\leq \frac{1}{\sigma}\sqrt{2(1-\rho)}\sqrt{1+\sigma\eta^2} + \eta|p_{js} - p_{jt}| \\ &= C\sqrt{1-\rho} + \eta|p_{js} - p_{jt}| \end{aligned} \tag{4.12}$$

where $C = \sqrt{\frac{2(1+\sigma\eta^2)}{\sigma^2}}$.

Intuitively, if two training samples $\mathbf{x}_s$ and $\mathbf{x}_t$ come from the same class (they have high correlation $\rho \approx 1$), and they are close to the dictionary item $\mathbf{d}_j$ ($p_{js} \approx p_{jt}$ and they are large), then their corresponding coefficients are similar and large ($v_{js} \approx v_{jt}$), otherwise the coefficients will be different. In other words, the GPDDC reveals the fact that if a dictionary item has more training sample neighbors from one class, then it will carry more information about this class.

4.3 Optimization Procedure

The objective function in Equation 4.1 now can be optimized by a coordinate descent method, which alternatively updates the sparse representation, the dictionary distribution coefficients, as well as the discriminative dictionary. In order to obtain a better convergence rate, the sparse representation and the dictionary are initialized using the conventional sparse representation method [53], while the dictionary distribution coefficients $\mathbf{v}_j$ are initialized using the value of $\eta\mathbf{p}_j$.

4.3.1 Update the Sparse Representation

First, to obtain the sparse representation $\mathbf{W}$ given the dictionary $\mathbf{D}$ and the dictionary distribution coefficients $\mathbf{V}$, the objective function defined in Equation 4.1 can be optimized by decomposing it into separate objective functions for each training sample $\mathbf{x}_i$ as follows.

$$\min_{\mathbf{w}_i} ||\mathbf{x}_i - \mathbf{D}\mathbf{w}_i||^2 + \alpha L_{ii}\mathbf{w}_i^t\mathbf{w}_i + \alpha\mathbf{w}_i^t\mathbf{h}_i + \lambda||\mathbf{w}_i||_1; \tag{4.13}$$

where $\mathbf{h}_i = \sum_{j \neq i} L_{ij}\mathbf{w}_j = [h_{i1}, h_{i2}, ..., h_{ik}]^t$ and $L_{ij}(i, j = 1, 2, ..., m)$ is the value in the i-th row, j-th column of the matrix $\mathbf{L}$ defined above. For this separate objective function, the FISTA algorithm [3] is applied to learn the sparse representation $\mathbf{w}_i$ for each training sample $\mathbf{x}_i$.

An important quantity to be determined before applying the FISTA algorithm is the step size that guarantees convergence. The largest step size may be theoretically derived for each training sample from Equation 4.13. Specifically, the largest step size that guarantees convergence of the FISTA algorithm for optimizing Equation 4.13 is $\frac{1}{Lip(f)}$, where $Lip(f)$ is the smallest Lipschitz constant of the gradient ∇f for $f(\mathbf{w}_i) = ||\mathbf{x}_i - \mathbf{D}\mathbf{w}_i||^2 + \alpha L_{ii}\mathbf{w}_i^t\mathbf{w}_i + \alpha\mathbf{w}_i^t\mathbf{h}_i$. It can be derived that $Lip(f) = 2E_{max}(\mathbf{D}^t\mathbf{D} + \alpha L_{ii}\mathbf{I})$, namely twice of the maximum eigenvalue of matrix $\mathbf{D}^t\mathbf{D} + \alpha L_{ii}\mathbf{I}$. Note that in practice one has the liberty of choosing a step size that is smaller than the largest step size.

4.3.2 Update the Dictionary Distribution Coefficients

Second, when the dictionary $\mathbf{D}$ and the sparse representation $\mathbf{W}$ are given, the dictionary distribution Coefficients $\mathbf{V}$ can be derived using the following analytical solution.

$$\mathbf{v}_j = (\mathbf{X}^t\mathbf{X} + \sigma\mathbf{I})^{-1}(\mathbf{X}^t\mathbf{d}_j + \sigma\eta\mathbf{p}_j) \tag{4.14}$$

As a matter of fact, $\mathbf{X}^t\mathbf{d}_j$ is the sample correlation between the dictionary item $\mathbf{d}_j$ and all the training samples, and $\mathbf{p}_j$ is the reciprocal of the exponential form of Euclidean distance between $\mathbf{d}_j$ and all the training samples. Therefore, the dictionary distribution coefficient $\mathbf{v}_j$ represents a measurement between the dictionary item and the training samples using a combination of both the correlation information and the distance information. From another perspective, $\mathbf{v}_j$ is a similarity measure using both the angular distance (correlation information) and the Euclidean distance (reciprocal of the exponential form of Euclidean distance). This important property of $\mathbf{v}_j$ will significantly help the derivation of the dictionary as shown in the following sub-section.

4.3.3 Update the Dictionary

Third, given the sparse representation $\mathbf{W}$ and the dictionary distribution coefficients $\mathbf{V}$, the dictionary $\mathbf{D}$ can be derived by optimizing the following objective function.

$$\begin{aligned} &\min_{\mathbf{D}} ||\mathbf{X} - \mathbf{D}\mathbf{W}||^2 + \gamma(||\mathbf{D} - \mathbf{X}\mathbf{V}||^2 + \sigma||\mathbf{V} - \eta\mathbf{P}||^2) \\ &s.t. \quad ||\mathbf{d}_j|| \leq 1, (j = 1, 2, ..., k) \end{aligned} \tag{4.15}$$

where $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_k]$.

The optimization of Equation 4.15 is not a trivial problem due to the exponential form of the vector $\mathbf{p}_j$ with respect to $\mathbf{d}_j$. Instead of using some generic solvers, we seek to a more efficient approximation for deriving the dictionary based on the observation from Equation 4.14 that the coefficients of the nearest neighbors of dictionary items are suffice since the dictionary distribution coefficients vector $\mathbf{v}_j$ represents a similarity measure between training samples and dictionary items. Specifically, the approximation method consists of the following steps.

- First, the influence of distant training samples are diminished by setting the elements in $\mathbf{v}_j$, whose absolute value is less than a threshold, to be zero. The new vector is then denoted as $\bar{\mathbf{v}}_j$.

- Then, the dictionary is derived by solving the following new optimization problem, given that these $\bar{\mathbf{v}}_j$ consist of a new matrix $\bar{\mathbf{V}}$.

$$\begin{aligned} &\min_{\mathbf{D}} ||\mathbf{X} - \mathbf{D}\mathbf{W}||^2 + \gamma||\mathbf{D} - \mathbf{X}\bar{\mathbf{V}}||^2 \\ &s.t. \quad ||\mathbf{d}_j|| \leq 1, (j = 1, 2, ..., k) \end{aligned} \tag{4.16}$$

This problem is a constrained optimization problem with inequality constraints, which may be solved using the Lagrange optimization and the Karush-Kuhn-Tucker condition [53]. Particularly, the primal optimization is solved by taking the first derivative with respect to $\mathbf{D}$ and set it to zero, and then the dual problem may be formulated as follows:

$$\min_{\Lambda} \mathbf{tr}(\mathbf{X}(\mathbf{W}^t + \gamma\bar{\mathbf{V}})(\mathbf{W}\mathbf{W}^t + \gamma\mathbf{I} + \Lambda)^{-1}(\mathbf{W} + \gamma\bar{\mathbf{V}}^t)\mathbf{X}^t + \Lambda - \mathbf{X}^t\mathbf{X}) \tag{4.17}$$

where Λ is a diagonal matrix whose diagonal values are the dual parameters of the primal optimization problem. The dual problem defined by Equation 4.17 can be solved by using the gradient descent method. Finally, assuming Λ^* be the solution of the dual problem, then the dictionary D is updated using the following equation:

$$\mathbf{D} = \mathbf{X}(\mathbf{W}^t + \gamma\bar{\mathbf{V}})(\mathbf{W}\mathbf{W}^t + \gamma\mathbf{I} + \Lambda^*)^{-1} \tag{4.18}$$

4.4 Generative and Discriminative Sparse Representation based Classification

After the dictionary $\mathbf{D}$ and the dictionary distribution coefficients $\mathbf{V}$ are derived, a new generative and discriminative sparse representation based classification (GDSRc) method is presented. Specifically, for a test data $\mathbf{y}$, we derive the generative and discriminative sparse representation by optimizing the following criterion: $\min_{\mathbf{w}} \{||\mathbf{y} - \mathbf{D}\mathbf{w}||^2 + \lambda||\mathbf{w}||_1\}$. Note that as the dictionary **D**, which is learned in the training optimization process, possesses the power of both generative and discriminative information, the representation, $\mathbf{w} = [w_1, w_2, ..., w_k]^t$ is thus called the generative and discriminative sparse representation.

The novel GDSRc method is then applied based on the derived generative and discriminative sparse representation **w** and the dictionary distribution coefficients **V**. In particular, the GDSRc method is defined as follows.

$$c^* = \arg\max_c \sum_{j=1}^{k} w_j \sum_{\mathbf{x}_i \in \mathbf{X}_c} v_{ji} \tag{4.19}$$

Note that we only select the top T largest values of v_{ji} for the GDSRc method. In practice, the value of T plays an important role in the performance.

We now show the connection between the proposed GDSRc method and the Bayes ideal feature [25]. The Bayes ideal feature is known as the optimal feature for classification based on the criterion of Bayes error. Mathematically, the Bayes ideal feature is defined as a vector $\mathbf{b} = [p(1|\mathbf{y}), p(2|\mathbf{y}), ..., p(c{-}1|\mathbf{y})]^t \in \mathbb{R}^{c-1}$. This vector carries sufficient information to set up the Bayes classifier, which has the minimum Bayes error for classification. However, in practice, this vector is difficult to obtain. The well-known discriminant analysis applies another criterion different from the Bayes error and derives a $c - 1$ dimension feature, which is claimed as an approximation to the Bayes ideal feature. But the explicit relation is still unclear between the feature extracted by discriminant analysis and the Bayes ideal feature.

The following Proposition 4.4.1 demonstrates that our GDSRc method has a direct relation to the Bayes ideal feature, which guarantees the performance of our GDSRc method.

Proposition 4.4.1 *For a given test data* **y***, the GDSRc method computes the approximate Bayes ideal feature* ***b*** *and classify* **y** *to the* c*-th class, which is corresponding to the largest element in* ***b****.*

The proof of Proposition 4.4.1 is shown as follows.

Given the test data **y**, the ideal Bayes feature try to derive the posterior probability $p(c|\mathbf{y})$ for each class. The test data **y** can be approximated as a linear combination of the dictionary item as $\mathbf{y} \approx \sum_{j=1}^{k} w_j \mathbf{d}_j$. Besides, it is reasonable to assume that the posterior probability function $p(c|\mathbf{y})$ is a Lipschitz function. As a result, according to local coordinate coding [155], we have the following result:

$$\begin{aligned} p(c|\mathbf{y}) &\approx \sum_{j=1}^{k} w_j p(c|\mathbf{d}_j) \\ &\propto \sum_{j=1}^{k} w_j p(\mathbf{d}_j|c) p(c) \\ &\propto p(c) \sum_{j=1}^{k} w_j \sum_{\mathbf{x}_i \in \mathbf{X}_c} v_{ji} \end{aligned} \tag{4.20}$$

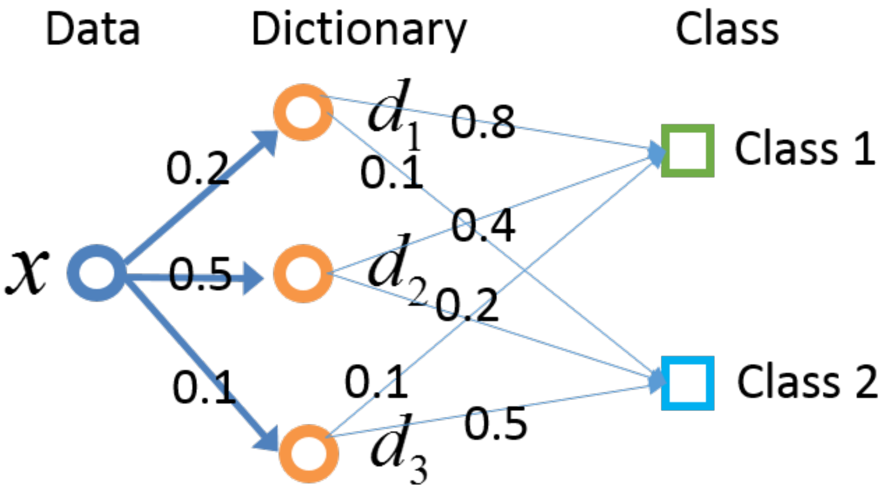
Note that we can normalize w_j so that $\sum_{j=1}^{k} w_j = 1$ to satisfy the condition of local coordinate coding.

Finally, based on this approximation to the ideal Bayes feature, our proposed GDSRc method chooses the largest elements for classification using the following equation given the same prior probability $p(c)$:

$$\begin{aligned} c^* &= \arg\max_c p(c|\mathbf{y}) \\ &\propto \arg\max_c \sum_{j=1}^{k} w_j \sum_{\mathbf{x}_i \in \mathbf{X}_c} v_{ji} \end{aligned} \tag{4.21}$$

As a result, the proposed GDSRc method takes advantage of the approximated Bayes ideal feature and establishes its connection to the Bayes decision rule for minimum error for classification.

Figure 4.3 shows an example of how the proposed GDSRc works. A test sample **x** has a linear combination of the dictionary items as $w_1 = 0.2$, $w_2 = 0.5$ and $w_3 = 0.1$. Then the score for each class using the learned dictionary distribution coefficients are calculated as follows: $w_1 \times 0.8 + w_2 \times 0.4 + w_3 \times 0.1 = 0.37$ for class 1 and $w_1 \times 0.1 + w_2 \times 0.2 + w_3 \times 0.5 =$



$$x = 0.2 \times d_1 + 0.5 \times d_2 + 0.1 \times d_3$$

$$\text{class } 1: 0.2 * 0.8 + 0.5 * 0.4 + 0.1 * 0.1 = 0.37$$
$$\text{class } 2: 0.2 * 0.1 + 0.5 * 0.2 + 0.1 * 0.5 = 0.17$$

Figure 4.3 Example of the GDSRc. The test sample **x** will be assigned to class 1 because 0.37 is larger than 0.17.

0.17 for class 2. Then the test sample **x** will be classified into class 1 since it has the largest score.

4.5 Experiments

To evaluate the effectiveness of the proposed GDSR method, it is tested on various visual recognition tasks, namely face recognition, computational fine art analysis, scene recognition and fine grained recognition. In particular, the datasets for evaluating the proposed GDSR method are listed in Table 4.1. Some sample images are also shown in Figure 4.4. Besides, additional comprehensive analysis on some critical issues concerning

Table 4.1 The Data Sets used and their Sizes

Task	Dataset	Size
Face recognition	AR face [94] Extended Yale face B [54]	4,000 2,414
Fine art analysis	Painting-91 dataset [45]	4266
Scene recognition	MIT-67 Indoor Scenes [109] 15 Scenes [52]	15,620 4,485
Fine grained recognition	CUB-200-2011 [128]	11788

about the performance is also presented for further investigating the properties of the proposed method.

All of our experiments are implemented using both C++ and Matlab in a desktop computer with 8 cores of Intel i7 CPU, 16 GB RAM and a GTX 745 GPU Card. The feature extraction process takes up to 1 hour. The training of our GDSR method takes up to 2.5 hour. The deep learning related fine tuning process is completed in a distributed computing system with one Tartan GPU up to 8 hours.

4.5.1 Face Recognition

Extended Yale face database B The Extended Yale Face Database B consists of 2414 frontal view face images from 38 individuals each with around 64 images taken under various lightening conditions. A cropped version of the database [54] is often applied, where all the images are manually aligned, cropped, and then re-sized to 168 $\times$ 192 .

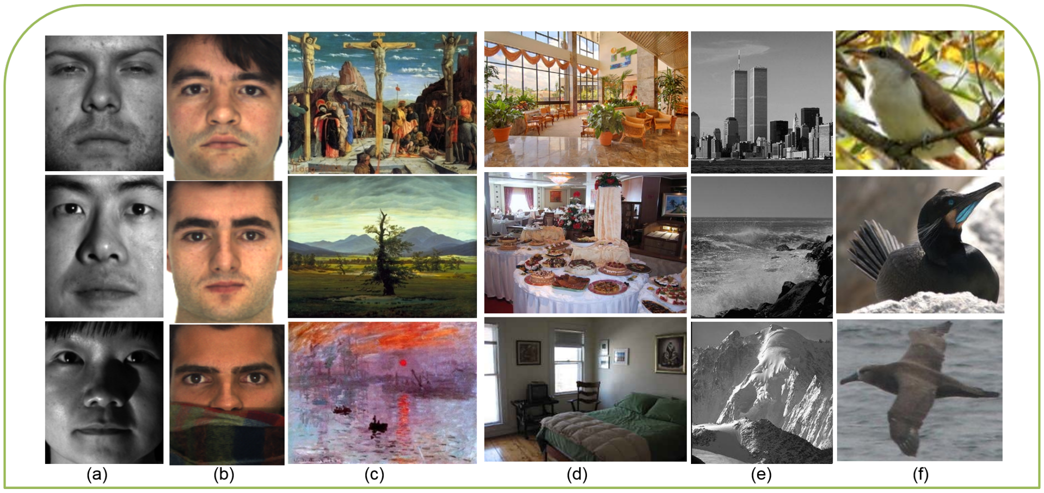


Figure 4.4 Sample images of the datasets: (a) the Extended Yale Face Database B, (b) the AR Face Database, (c) Painting-91 dataset, (d) the MIT-67 Indoor Scenes dataset, (e) the 15 Scenes dataset and (f) CUB-200-2011 dataset.

Two experimental settings are applied for fair comparison. First, we follow the experimental setting [151] that 20 images are randomly selected for training for each subject, and the remaining images (around 44 per subject) are for testing for 10 iterations. To show the robustness of our proposed method, we present results of our GDSR method under an extremely noisy condition, where the random faces [138] are used as the input. Specifically, the random faces [138] consists of the row vectors of a randomly generated transformation matrix from a zero-mean normal distribution, which is applied to project the face pattern vector into a dimension of 504 representation vector. Each row of the transformation matrix is normalized to unit length. Then the dimension is reduced to 350 and the dictionary size is 512. The model parameters are selected as follows: $\lambda = 0.1$ for the sparse representation criterion; $\gamma = 0.05$, $\sigma = 0.05$, $\eta = 0.1$ and $h = 0.1$ for the dictionary distribution criterion; $\alpha = 0.5$, and $\beta = 0.5$ for the discriminative criterion; and $k = 20$ for the GDSRc method.

Second, we follow the experimental setting [1], [42] that half images are randomly selected for training for each subject, and the remaining images are for testing for 10 iterations. The random faces are applied as well and the dimension of the representation

Table 4.2 Comparison with the State-of-the-Art Methods on the Extended Yale Face Database B under two Experimental Settings

Experimental setting 1	Accuracy $\%$
D-KSVD [161]	75.30
SRC [138]	90.00
FDDL [151]	91.90
The GDSR method	**95.19**
Experimental setting 2	Accuracy $\%$
LLC [133]	90.70
D-KSVD [161]	94.79 ± 0.49
LC-KSVD1 [42]	93.59 ± 0.54
LC-KSVD2 [42]	95.22 ± 0.61
FDDL [152]	96.07 ± 0.64
SRC [138]	96.32 ± 0.85
DBDL + SVM [1]	96.10 ± 0.25
DBDL [1]	97.31 ± 0.67
The GDSR method	**97.45** ± 0.40

vector is reduced from 504 to 350. the dictionary size is 512. The model parameters are selected as follows: $\lambda = 0.05$ for the sparse representation criterion; $\gamma = 0.05$, $\sigma = 0.05$, $\eta = 0.1$ and $h = 0.1$ for the dictionary distribution criterion; $\alpha = 0.1$, and $\beta = 0.5$ for the discriminative criterion; and k is the number of training samples per subject for the GDSRc method.

The final results shown in Table 4.2 demonstrate the effectiveness of the proposed methods under such a noisy condition.

AR Face Database The AR Face Database consists of over 4000 frontal view images for 126 individuals each with 26 pictures taken in two separate sessions. A subset of the data [94] that consists of 50 male subjects and 50 female subjects is chosen and are cropped to size 165*120. We follow three main widely adopted experimental settings to make fair comparisons.

The first experimental setting is defined in [42], [161], where the methods are evaluated by randomly selecting 20 images for training and the others for testing for each person for 10 iterations. In this experimental setting, the random faces [138], [42] with 540 dimensions are applied for fair comparison. Then the dimension is reduced from 540 to 400 and the size of the dictionary is 512. The model parameters are selected as follows: $\lambda = 0.1$ for the sparse representation criterion; $\gamma = 0.01$, $\sigma = 0.05$, $\eta = 0.1$ and $h = 0.1$ for the dictionary distribution criterion; $\alpha = 0.5$, and $\beta = 0.5$ for the discriminative criterion; and $k = 15$ for the GDSRc method.

The second experimental setting is defined in [138], [151], [152] where 14 images with only illumination change and expressions are selected for each person: the seven images from session 1 for training and the other seven from session 2 for testing. The pattern vector is formed as the concatenation of the column pixels. Then the dimension is reduced to 300 and the size of the dictionary is 512. The model parameters are selected as follows: $\lambda = 0.05$ for the sparse representation criterion; $\gamma = 0.01$, $\sigma = 0.05$, $\eta = 0.1$ and $h = 0.1$ for the dictionary distribution criterion; $\alpha = 0.5$, and $\beta = 0.5$ for the discriminative criterion; and $k = 7$ for the GDSRc method.

The third experimental setting is defined in [20], where the 26 images for each person are randomly permuted and the first half is taken for training, the rest for testing for total 10 iterations. The features are similar as those in the second experimental setting. Then the dimension is reduced to 500 and the size of the dictionary is 512. The model parameters are selected as follows: $\lambda = 0.1$ for the sparse representation criterion; $\gamma = 0.01$, $\sigma = 0.05$,

Table 4.3 Comparison with the other Popular Methods on AR Face Database under three Experimental Settings

Experimental setting 1	Accuracy %
D-KSVD [161]	95.00
SRC [138]	97.50
LC-KSVD2 [42]	97.80
FDDL [152]	96.22
DBDL + SVM [1]	95.69
DBDL [1]	97.47
The GDSR method	**98.50**
Experimental setting 2	Accuracy %
D-KSVD [161]	85.40
LC-KSVD [42]	89.7
JDL [164]	91.7
FDDL [152]	92.00
SRC [138]	94.99
The GDSR method	**96.29**
Experimental setting 3	Accuracy %
SRC [20]	93.75 ±1.01
ESRC [20]	97.36 ±0.59
SSRC [20]	98.58 ±0.40
The GDSR method	**99.02**±0.31%

$\eta = 0.1$ and $h = 0.1$ for the dictionary distribution criterion; $\alpha = 0.5$, and $\beta = 0.5$ for the discriminative criterion; and $k = 13$ for the GDSRc method.

The experimental results that are presented in Table 4.3 show that the our GDSR method is able to improve upon the other popular methods under all the three experimental settings.

4.5.2 Computational Fine Art Analysis

The Painting-91 dataset [45] contains 4266 fine art painting images by 91 artists. The images are collected from the Internet and covers artists from different eras. There are variable number of images per artist ranging from 31 (Frida Kahlo) to 56 (Sandro Boticelli). The dataset classifies 50 painters to 13 style categories with style labels namely: (1) abstract expressionism, (2) baroque, (3) constructivism, (4) cubism, (5) impressionism, (6) neoclassical, (7) popart, (8) post-impressionism, (9) realism, (10) renaissance, (11) romanticism, (12) surrealism, and (13) symbolism. Following the experimental protocol [45], two tasks, namely artist classification and style classification, are assessed by using a predefined training data and test data containing 2275 training images and 1991 test images.

In order to represent the painting art images, we use a hybrid feature extraction step where Fisher vector features are extracted from SIFT descriptors, Weber local descriptors [14] and DAISY descriptors [122] so as to extract the local, spatial, relative intensity and gradient orientation information from the painting image. The color cue provides powerful discriminatory information in general [66], therefore we further incorporate the color information by computing the above Fisher vector features in different color spaces namely YCbCr, YIQ, oRGB, XYZ, YUV and HSV. To further improve the results, we combine our Fisher vector features with the feature extracted from a pre-trained CNN model, namely GoogleNet [120]. Note that the results of the single GoogleNet feature are also reported, which are only 46.71 and 55.79 for artist and style classification, respectively.

Artist classification involves classifying a painting to its respective artist among all the 91 artists. The dimension is reduced to 2000 and the size of the dictionary is 1024. The model parameters are selected as follows: $\lambda = 0.05$ for the sparse representation criterion; $\gamma = 0.05$, $\sigma = 0.05$, $\eta = 0.1$ and $h = 0.1$ for the dictionary distribution criterion; $\alpha = 0.1$, and $\beta = 0.5$ for the discriminative criterion; and $k = 25$ for GDSRc method.

The style classification task deals with the problem of categorizing a painting to the 13 style classes defined in the dataset. Then the dimension is reduced to 1200 and the size of the dictionary is 1024. The model parameters are selected as follows: $\lambda = 0.05$ for the sparse representation criterion; $\gamma = 0.05$, $\sigma = 0.05$, $\eta = 0.1$ and $h = 0.1$ for the dictionary distribution criterion; $\alpha = 0.1$, and $\beta = 0.5$ for the discriminative criterion; and $k = 40$ for GDSRc method.

Experimental results in Table 4.4 show that our proposed GDSR method achieves the state-of-the-art results in both the artist and style classification tasks, and significantly outperforms other methods. Figure 4.5 shows the confusion matrix for the 13 style categories of the Painting-91 dataset using our proposed GDSR method. It can be discovered that the first style (abstract expressionism) and the 13-th style (symbolism) have the best classification accuracy. While the 6-th style (neoclassical) is the most difficult categories to classify as there are large confusions in the Figure 4.5 between the style categories with the second style (baroque) and the 10-th style (renaissance). Similarly, the 9-th style (realism) and show large confusions in the confusion diagram with the 11-th style (romanticism). The reason is that the styles neoclassical, realism, baroque and romanticism belong to the same art movement period which is a period of time where a group of artists follow a common goal resulting in higher similarity between these styles. The style art movement results derived by our proposed GDSR method confirms the effectiveness of the proposed method.

Table 4.4 Comparison with other Popular Methods for Artist and Style Classification on the Painting-91 Dataset

Feature	Artist Classification	Style Classification
LBP [99, 45]	28.50	42.20
Color-LBP [45]	35.00	47.00
PHOG [7, 45]	18.60	29.50
Color-PHOG [45]	22.80	33.20
GIST [100, 45]	23.90	31.30
Color-GIST [45]	27.80	36.50
SIFT [87, 45]	42.60	53.20
CLBP [31, 45]	34.70	46.40
CN [125, 45]	18.10	33.30
SSIM [114, 45]	23.70	37.50
OPPSIFT [124, 45]	39.50	52.20
RGBSIFT [124, 45]	40.30	47.40
CSIFT [124, 45]	36.40	48.60
CN-SIFT [45]	44.10	56.70
Combine(1 - 14) [45]	53.10	62.20
MSCNN-1 [103]	58.11	69.67
MSCNN-2 [103]	57.91	70.96
CNN F_3 [102]	56.40	68.57
CNN F_4 [102]	56.35	69.21
GoogleNet [120]	46.71	55.79
GDSR	**67.06**	**77.09**

Table 4.5 Comparison with the State-of-the-Art Methods on the MIT-67 Indoor Scenes Dataset

Methods	Mean Accuracy $\%$
ROI + Gist [109]	26.1
DPM [101]	30.4
Object Bank [56]	37.6
miSVM [57]	46.4
D-Parts [119]	51.4
DP + IFV [43]	60.8
D3 [139]	78.13
VGG16-Place365 [163]	76.53
The GDSR method	**82.97**

4.5.3 Scene Recognition

The MIT-67 Indoor Scenes Dataset The MIT-67 Indoor Scenes dataset [109] is a very challenging indoor scene recognition dataset, which contains 67 indoor categories with 15620 images. Commonly used experimental setting [109] are followed, wherein 80*67 images are used for training and 20*67 images for testing.

The pretrained VGG16-Place365 CNN model [163] is applied to extract the feature, whose dimension is further reduced from 4096 to 3500. The dictionary size is selected as 2048. The model parameters are selected as follows: $\lambda = 0.05$ for the sparse representation criterion; $\gamma = 0.05$, $\sigma = 0.05$, $\eta = 0.1$ and $h = 0.01$ for the generative criterion; $\alpha = 0.1$, and $\beta = 0.5$ for the discriminative criterion; and $k = 75$ for GDSRc method. The results shown in Table 4.5 demonstrate that the proposed method is able to achieve the state-of-the-art results.

Table 4.6 Comparison with the State-of-the-Art Methods on the 15 Scenes Dataset

Methods	Accuracy %
LLC [133]	89.20
D-KSVD [161]	89.10
LC-KSVD1 [42]	90.40
LC-KSVD2 [42]	92.90
LaplacianSC [26]	89.7
DHVFC [27]	86.4
VGG16-Place365 [163]	92.15
DBDL [1]	98.73
The GDSR method	**98.75** ± 0.15

The 15 Scenes Dataset The 15 Scenes dataset [52] contains 4485 images from 15 scene categories, each with the number of images ranging from 200 to 400. Following the experimental protocol defined in [52], 100 images per class are randomly selected for training and the remaining for testing for 10 iterations. First, the spatial pyramid features provided by [42], which are obtained by using a four-level spatial pyramid and a codebook with a size of 200, are applied to represent the image as a vector with the dimension of 3000 for fair comparison. Then the dimension is reduced to 1000 and the size of the dictionary is 1024. The model parameters are selected as follows: $\lambda = 0.05$ for the sparse representation criterion; $\gamma = 0.05$, $\sigma = 0.05$, $\eta = 0.1$ and $h = 0.1$ for the dictionary distribution criterion; $\alpha = 0.1$, and $\beta = 0.5$ for the discriminative criterion; and $k = 100$ for the GDSRc method. The results shown in Table 4.6 demonstrate that the proposed method is able to achieve better results than the other state-of-the-art methods.

Table 4.7 Comparison with the other Popular Methods on the CUB-200-2011 Dataset

Methods	Accuracy %
PN-DCN [8]	75.70
CoSeg[49]	82.80
B-CNN [60]	84.80
TS-CNN [73]	76.90
PS-CNN [37]	76.60
ProCRC [10]	78.30
The GDSR method	**85.34**

4.5.4 Fine Grained Recognition

The Caltech-UCSD Birds-200-2011 (CUB-200-2011) dataset, which contains 5994 images for training and 5794 images for testing, is applied to evaluate the performance for fine grained recognition. In both training and testing phases, our experiment only use the bounding box without the part annotation. We combine the feature extracted from VGG-19 CNN [117] that is fine tuned on the CUB-200-2011 dataset and the feature derived from GoogleNet [120] that is fine tuned on the NABird dataset [34], which leads to a vector of 5120 in dimension. The dimensionality is then reduced to 1000 and the size of the dictionary is 512. The model parameters are selected as follows: $\lambda = 0.05$ for the sparse representation criterion; $\gamma = 0.05$, $\sigma = 0.05$, $\eta = 0.1$ and $h = 0.1$ for the generative criterion; $\alpha = 0.1$, and $\beta = 0.5$ for the discriminative criterion; and $k = 29$ for the GDSRc method. The results on Table 4.7 show the effectiveness of the proposed method.

4.6 Comprehensive Analysis

4.6.1 Explicit Modeling of the Generative Information

This section analyzes an explicit modeling of the generative information of the dictionary items. Specifically, the Figure 4.6 is presented, where the x axis represents each dictionary item and the y axis represents the normalized approximated value of $p(\mathbf{d}_j|c) \propto \sum_{\mathbf{x}_i \in \mathbf{X}_c} v_{ji}$

for each class. As illustrated in Figure 4.6, each dictionary item carries a soft class membership and captures information from different classes.

4.6.2 Evaluation of Dimensionality and Dictionary Size

This section presents an analysis of the performance under different sizes of dictionary and different values of dimensionality of the feature for scene recognition on the MIT-67 Indoor Scenes dataset. Specifically, the dictionary sizes of 128, 256, 512, 1024, 2048 are evaluated for dimensionality 500, 1000, 1500, 2000, 2500, 3000, 3500 and 4000, respectively. From Figure 4.7, we can conclude that (1) on the one hand, larger dictionary size usually contributes better performance, (2) on the other hand, the low dimension feature space is more sensitive to the over-completeness (dictionary size is large than the dimensionality) of the dictionary but higher dimension no longer requires that. For example, the dimensionality 500 requires at least dictionary size of 512 to achieve good performance (above 80%), while higher dimensionality, such as 3500, just requires dictionary size of 256.

4.6.3 Evaluation of the Size of the Training Data

We evaluate the performance of our GDSR method when the size of the training data varies for each category on the Extended Yale Face Database B. From Figure 4.8, it is able to conclude that on the one hand, our GDSR method achieves better performance with larger size of training data size available; on the other hand, the performance with large size dictionary is more sensitive to the size of the training data than small size dictionary.

4.6.4 Evaluation of GDSRc

We present the evaluation of GDSRc when different values of T (defined in Section 4.4) are applied for different sizes of dictionary. The experimental results in Figure 4.9 show that a large value of T is often preferred for better performance regardless of the size of

Table 4.8 Comparison with other Classifiers on the MIT-67 Indoor Scenes Dataset

Methods	MIT-67 Indoor Scenes %
KNN	76.70
Linear-SVM	79.60
RBF-SVM	80.07
GDSRc	**82.97**

the dictionary. As a matter of fact, the value of T is usually set as the size of the training samples of each class for best performance.

Besides, we also present the comparison with other classifiers when the same feature representation is applied. Specifically, the KNN classifier (K = 3), the linear kernel based SVM and the RBF kernel based SVM are applied for comparison. The experimental results in Table 4.8 show that our GDSRc can achieves better results.

4.6.5 The CNN Features

In order to harness the power of deep convolutional neural network (CNN) [50], we also applies current state-of-the-art deep CNN models for extracting the features. Specifically, two representative deep CNN models, namely the VGG net [117] and the GoogleNet [120] are applied for three data sets in our experiments: the Painting-91 dataset, the CUB-200-2011 dataset and the MIT-67 Indoor Scenes dataset.

The power of VGG net comes from a smaller 3×3 receptive fields and a deeper model layers (the model structure with the best performance has 19 layers). Besides, data augmentation techniques, such as multiple cropping and dense evaluation, also play an important role in the performance. While the power of the GoogleNet not only comes from its deeper model layers, but also a network in network [59] module called Inception [120], which harnesses the computing resources inside the network.

As for the Painting-91 dataset, a single GoogleNet and its combination with the CMFFV [107], [108], [105] feature are applied for comparison. As shown in 4.4, the combination of GoogleNet, the CMFFV [107], [108], [105] and the proposed GDSR method can achieve 67.06 and 77.09 for artist and style classification, respectively. In comparison, a single GoogleNet feature can only achieve 46.71 and 55.79 for artist and style classification, respectively. A single CMFFV feature can only achieve 65.78 and 73.16 for artist and style classification, respectively.

As for the CUB-200-2011 dataset, we combine the feature extracted from VGG-19 CNN [117] that is fine tuned on the CUB-200-2011 dataset and the feature derived from GoogleNet [120] that is fine tuned on the NABird dataset [34]. As shown in Table 4.7, we can achieve 85.34 for fine-grained recognition, which improves other state-of-the-art deep learning methods.

As for the MIT-67 Indoor Scenes dataset, the VGG Net that is pre-trained from the Place365 data set [163] is applied to extract the feature directly without fine tuning. As shown in Table 4.5, the deep CNN feature combined with the proposed GDSR method can achieve 82.97, which significantly improves upon other methods.

According to our analysis of the results with deep CNN features, we have the following empirical findings when the target data set is relatively small (less than the million level size), and different from the dataset (e.g. ImageNet data set) used for pre-training of the deep CNN model.

- Previous layers of the deep CNN model are more generic, while the last few layers are more data set specific.
- When the fine tuning is applied to the target data set, a small base learning rate is often selected for the whole model, but a higher learning rate is selected for the last few layers in order to learn more information from the target data set.

- Fusing different deep CNN models, such as VGG net and GoogleNet, often leads to better performance.
- The combination of deep CNN feature and conventional computer vision features, such as the CMFFV [107] feature, can achieve better performance.
- A discriminatively learned representation and classification, such as our GDSR method, will boost the performance for the deep CNN feature.

4.6.6 Evaluation of the Effect of the Proposed GDSR Method

To evaluate the contribution of the individual steps to the overall recognition rate, we conduct experiments on the MIT-67 Indoor Scenes dataset using the initial input features as described in the above section. In order to have a fair comparison, we use the RBF-SVM classifier for classification instead of the GDSRc method since it depends on both the generative and discriminative criteria. It can be seen from Table 4.9 that the GDSR method (both discriminative and generative criteria) achieves the best performance of 80.67% since it incorporates both the discriminative and the generative information.

We further discuss the effects of our proposed method on the initial features and how it encourages better clustering and discrimination among different classes of a dataset. To visualize the effect of our proposed method, we use the popular t-SNE visualization technique [90] that produces visualization of high dimensional data in scatter plots by reducing the dimensionality to two-dimension. Figure 4.10 shows the t-SNE visualizations of the initial features used as input and the features extracted after applying the GDSR method for different datasets. It can be seen from Figure 4.10 that the proposed GDSR method helps to reduce the distance between data-points of the same class leading to formation of higher density clusters for data-points of the same class. Another advantage is that the GDSR method assists to increase the distance between clusters of different classes resulting in better discrimination among them. The GDSR method uses both the

Table 4.9 Evaluation of the Contribution of Generative and Discriminative Criterion in GDSR Method using the MIT-67 Indoor Scenes Dataset

Method	Accuracy (%)
GDSR with only discriminative criterion	77.24
GDSR with only generative criterion	78.51
Proposed GDSR (both criteria)	**80.67**

Table 4.10 Comparison between the Proposed Method and other Popular Methods on the Caltech 256 Dataset

Method	30	45	60
ScSPM [149]	34.02	37.46	40.14
IFK [104]	40.80	45.00	47.90
LLC [133]	41.19	45.31	47.68
M-HMP [6]	48.00	51.90	55.20
ZFNet CNN [156]	70.60	72.70	74.20
GDSR	**72.39**	**75.13**	**76.90**

generative and discriminative information, therefore, encourages better separation between data samples of different classes.

4.6.7 Evaluation of Different Training Size

The Caltech 256 dataset [30] is applied here to evaluate the influence of different training data size on the proposed GDSR method. The initial input features used are extracted from a pre-trained ZFNet [156]. For the GDSR method, we set the dictionary size to 1024, and the parameters as λ = 0.05, $h = 0.1$, α = 0.1, and β = 0.5. The experimental results in Table 4.10 show that our proposed method is able to achieve better results compared to other methods for different training data size, namely 30, 45, 60 for each class.

4.7 Conclusions

This chapter presents a new generative and discriminative sparse representation (GDSR) method, which leads to a new effective representation and classification schema. In particular, the generative criterion reveals the class conditional probability of each dictionary item. The discriminative criterion applies new within-class and between-class scatter matrices for discriminant analysis to enhance the discriminative capability. In addition, a new generative and discriminative sparse representation based classification (GDSRc) method is proposed by utilizing both discriminative and generative information. Experimental results on several visual recognition tasks show the effectiveness of the proposed methods.

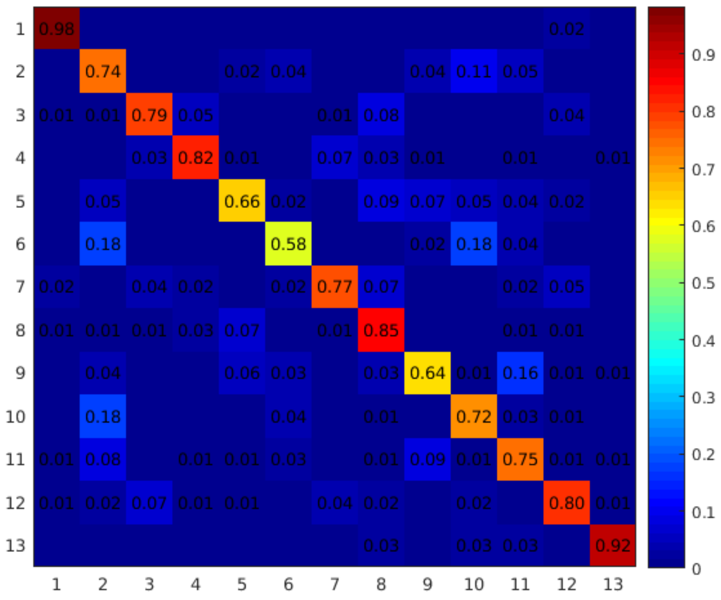


Figure 4.5 The confusion matrix for 13 style categories of the Painting-91 dataset

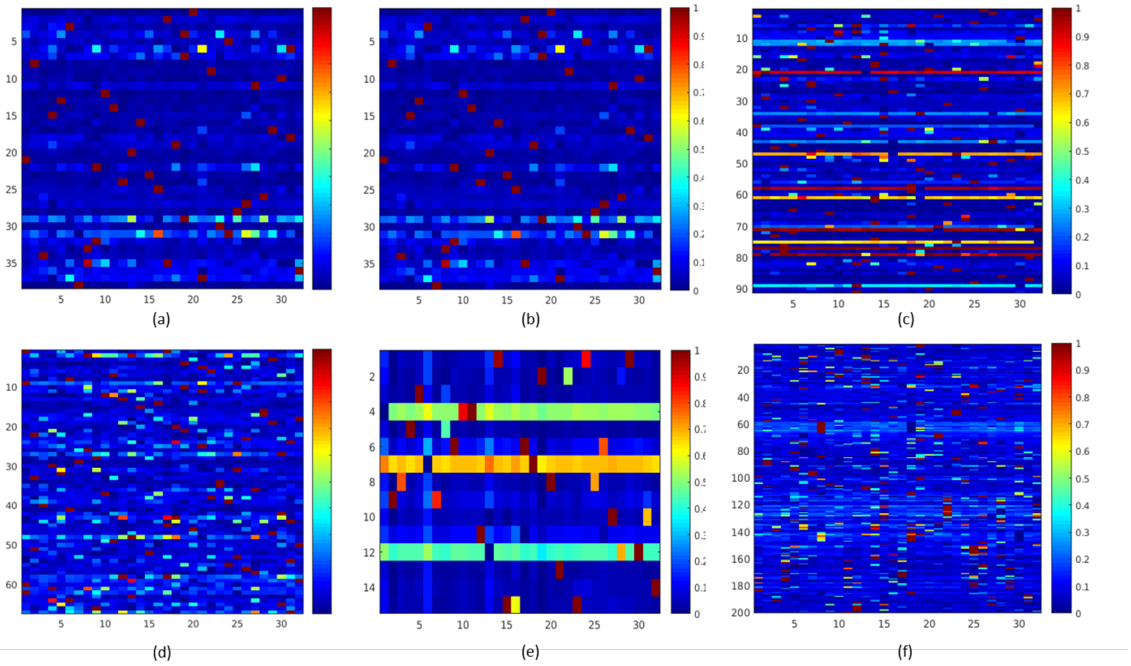


Figure 4.6 The explicit modeling of the generative information using different data sets for dictionary size 32: (a) Extended Yale face database B, (b) AR Face Database, (c) Painting-91 dataset, (d) MIT-67 Indoor Scenes, (e) 15 Scenes dataset and (f) CUB-200-2011 dataset.

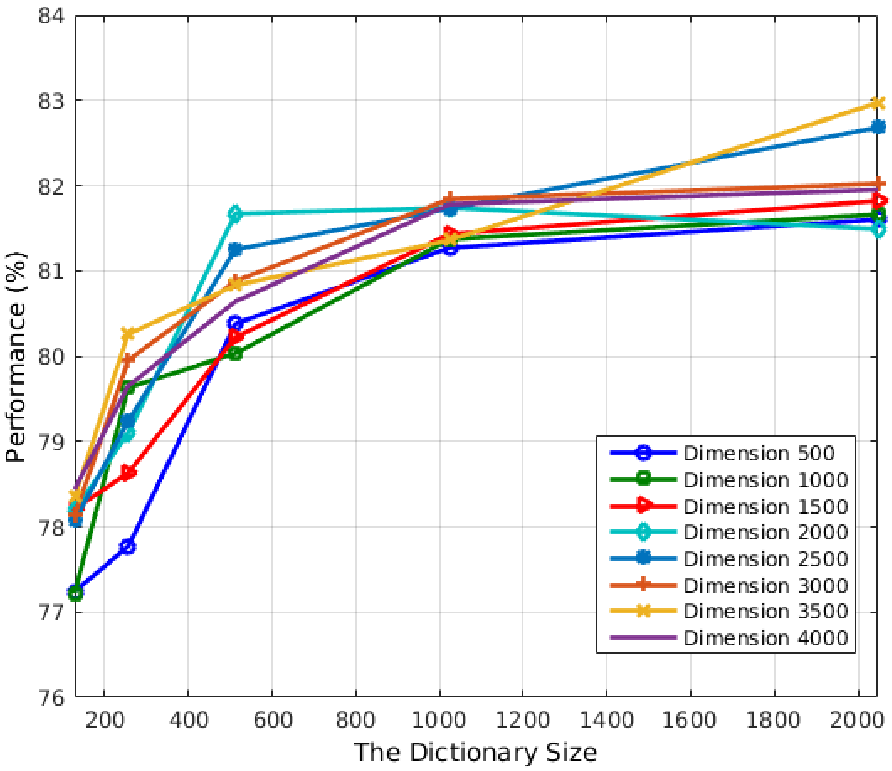


Figure 4.7 The performance of the proposed GDSR method under different sizes of dictionary and different values of dimensionality of the feature.

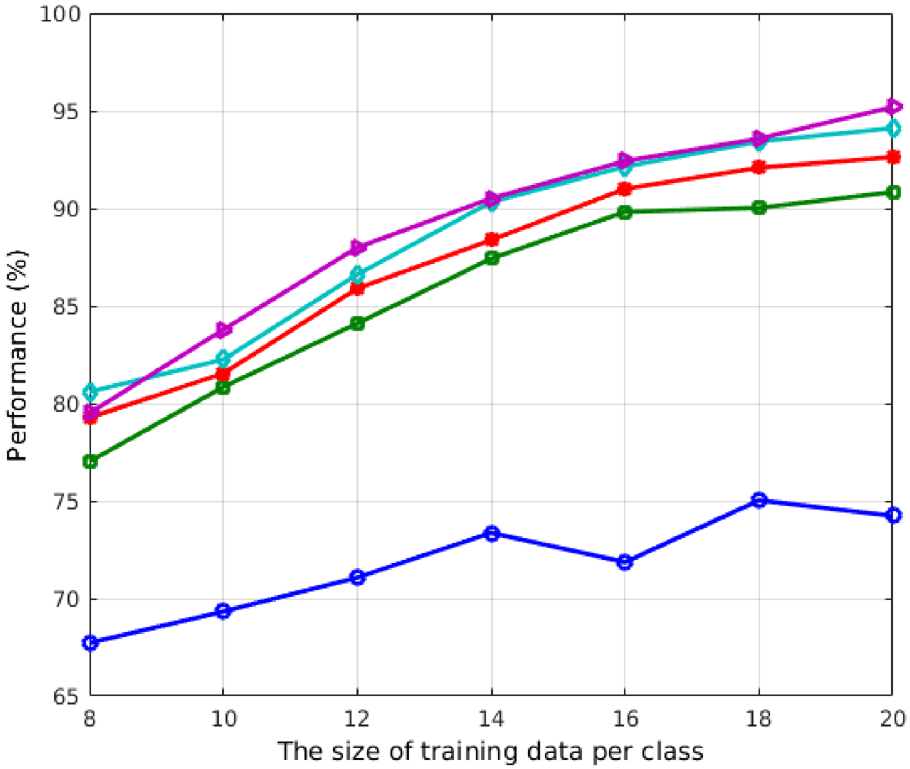


Figure 4.8 The performance of the proposed GDSR method when the size of training data in each class varies on the Extended Yale Face Database B.

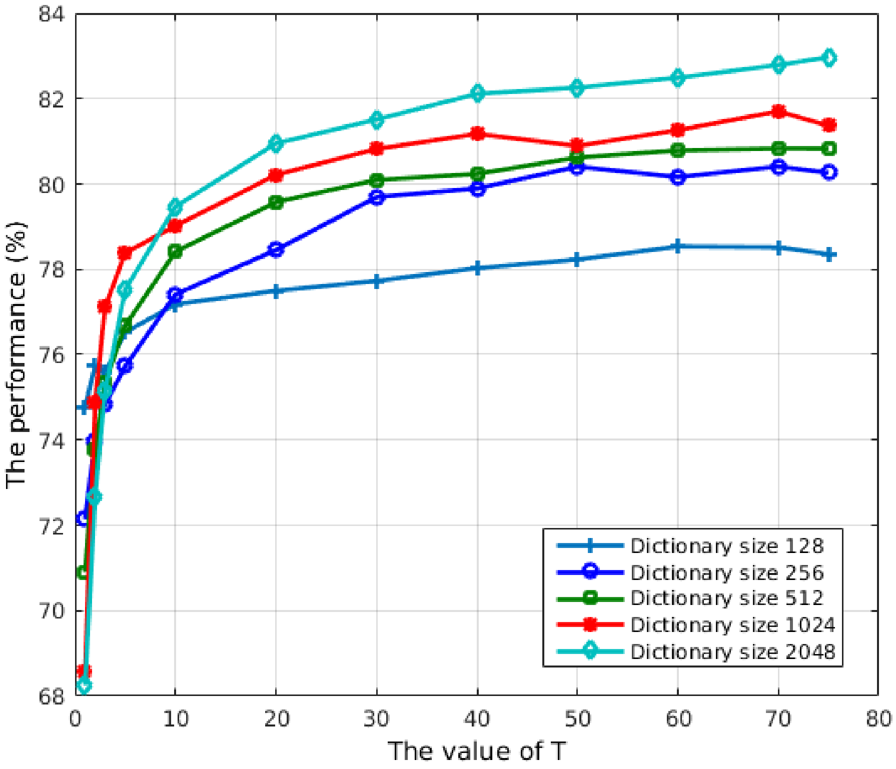
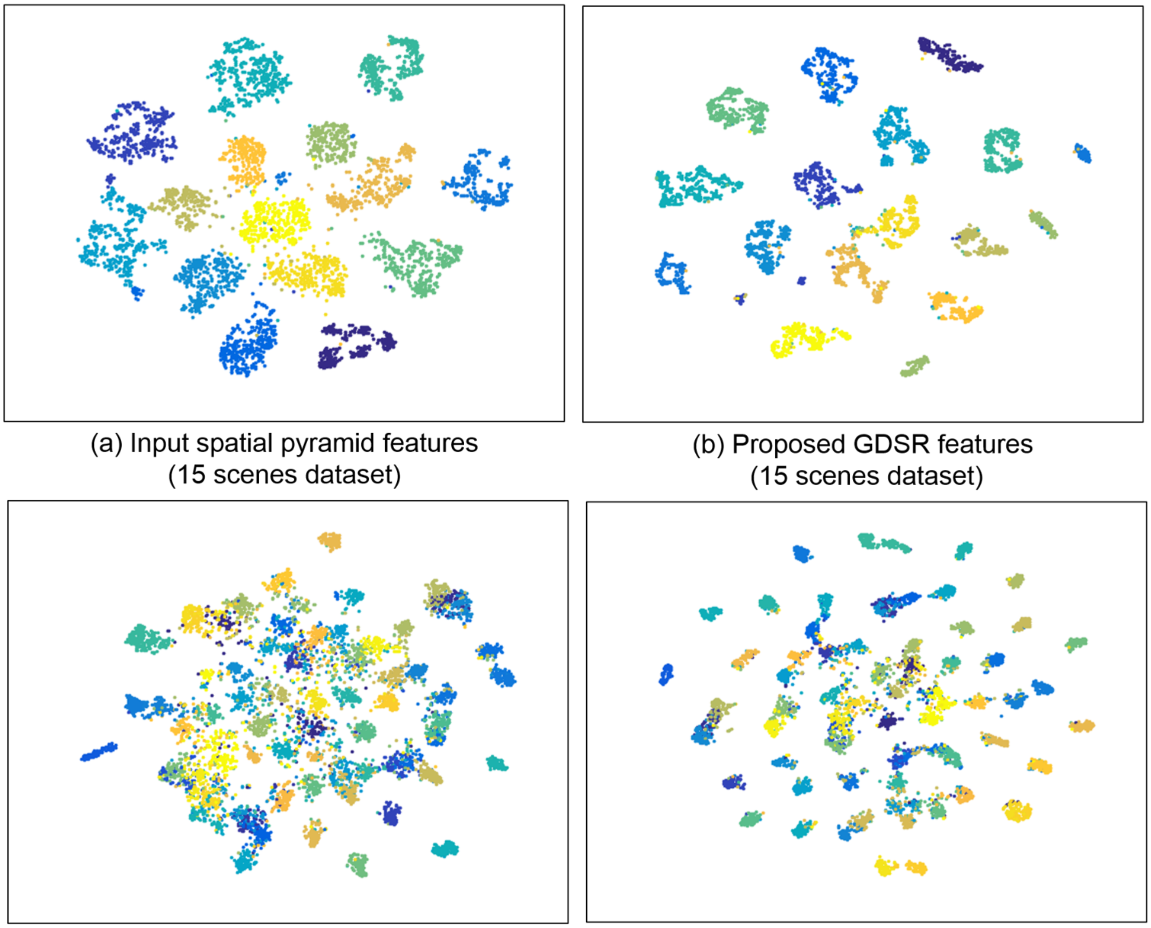


Figure 4.9 The performance of the proposed GDSRc method when the value of T varies on the MIT-67 Indoor Scenes dataset.



(a) Input spatial pyramid features
(15 scenes dataset)

(b) Proposed GDSR features
(15 scenes dataset)

(c) Input CNN features
(MIT-67 scenes dataset)

(d) Proposed GDSR features
(MIT-67 scenes dataset)

Figure 4.10 The t-SNE visualization of the initial input features and the features extracted after applying the proposed GDSR method.

CHAPTER 5

SCORE SPACE BASED MULTIPLE METRIC LEARNING FOR KINSHIP VERIFICATION

5.1 Introduction

Kinship verification has been an important topic in anthropology for many years. Pioneer work in anthropology [95], [2], [9] believes that there are some genetic related features which are inherited by children from their parents that can be used to determine the kinship relations. Recently, kinship verification from facial images is gaining increasing attention as an emerging research area in artificial intelligence [140], [24], [141], [89], [18], [146], [85], [106], [82], [80].

Many feature methods have been proposed for describing facial images[62], [70], [137], [11], [13], Fisher vector [116]. These features, which are designed specifically for distinguishing one image from others (the discriminative ability), cannot guarantee that a child image is more similar to its parent image than to other images (the inheritable ability). The major reason is that these features are designed or learned for recognition of face image thus cannot characterize the genetic relations between kinship images. Another reason is that the inherent similarity gap between kinship images is much larger than that in the face recognition problem, e.g., LFW [35], which means similarity between discriminative features is not sufficient for kinship verification. Lu et al. [89] proposed to apply a metric learning method on several features and proposed the MNRML method by combining different metrics on different features. The subsequent work [145], [146], [88] followed their idea by combining features and metric learning methods sequentially. In their methods, the features and the metric learning methods are developed in different paradigms independently, which may attenuate the effect when they are combined. Besides,

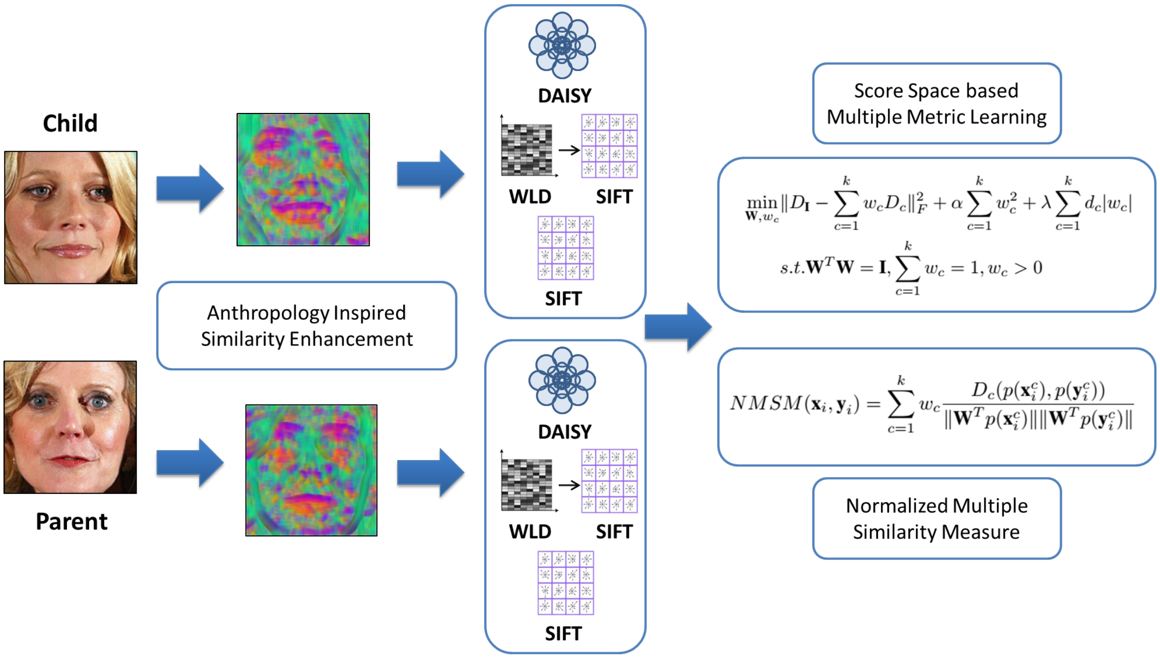

Child
Anthropology Inspired Similarity Enhancement
Parent
DAISY
WLD
SIFT
SIFT
DAISY
WLD
SIFT
SIFT
Score Space based Multiple Metric Learning
$\min_{\mathbf{W},w_c} \|D_{\mathbf{I}} - \sum_{c=1}^{k} w_c D_c\|_F^2 + \alpha \sum_{c=1}^{k} w_c^2 + \lambda \sum_{c=1}^{k} d_c |w_c|$
$s.t. \mathbf{W}^T\mathbf{W} = \mathbf{I}, \sum_{c=1}^{k} w_c = 1, w_c > 0$
$NMSM(\mathbf{x}_i, \mathbf{y}_i) = \sum_{c=1}^{k} w_c \frac{D_c(p(\mathbf{x}_i^c), p(\mathbf{y}_i^c))}{\|\mathbf{W}^T p(\mathbf{x}_i^c)\| \|\mathbf{W}^T p(\mathbf{y}_i^c)\|}$
Normalized Multiple Similarity Measure

Figure 5.1 The whole process of the proposed SML framework.

most metric learning methods are based on the Mahalanobis distance metric, which may not achieve the best performance in some scenarios.

To address these issues, this chapter proposes a novel score space based multiple metric learning (SML) method for kinship verification. The proposed SML method, which goes beyond the Mahalanobis distance metric, derives a semantically meaningful similarity between images by combining multiple anthropology inspired features and their metrics into a unified paradigm. Specifically, three novel anthropology inspired features (AIF) are first extracted, namely the AIF-SIFT, AIF-WLD and AIF-DAISY features. The process of deriving the anthropology inspired features consists of an anthropology inspired similarity enhancement method and the extraction of opponent color SIFT [44], color WLD-SIFT and DAISY [122] descriptors based on the enhanced image. In particular, the similarity enhancement method is applied to kinship image pairs by extending the SIFT flow method [61] and generating the enhanced images by reinforcing similar facial parts. The opponent SIFT descriptor, the color WLD-SIFT descriptor and the DAISY descriptor are then extracted on the enhanced images.

Second, a novel score space based multiple metric learning (SML) method is derived by learning a new metric and weights for multiple features in a unified paradigm. In particular, the new metric is learned while fixing the weights by balancing the behavior of pushing away the k-nearest non-kinship samples while pulling close the kinship ones for each training pairs. The weights are updated while fixing the transformation.

Finally, a novel normalized multiple similarity measure is proposed based on the observation that fractional power transformation is able to transform data into a near Gaussian shape with a stable variance which is well suited for dot product based similarity measure like cosine similarity measure from the Bayes decision rule for minimum error point of view [64]. The whole process of the proposed method is illustrated in Figure 5.1. The proposed method is then evaluated on two challenging kinship databases, KinFaceW-I and KinFaceW-II data set [89]. The experimental results show that the proposed method is able to achieve the state-of-the-art results.

5.2 Anthropology Inspired Feature Extraction

Naini et al. [95] analyzed the contributions of heredity and environment on external facial features. Their anthropological results [95] show that eyes, chin and parts of the forehead show higher visual resemblance between parents and their offspring and provide large feedback. From the computer vision point of view, these high resemblance in facial regions between kinship image pairs exhibit three important properties as follows given the notations that $\mathbf{p} = (x, y)$ are the grid coordinate of images, $\mathbf{d}(\mathbf{p}) = (u(\mathbf{p}), v(\mathbf{p}))$ is the displacement vector at $\mathbf{p}$, $u(\mathbf{p})$ and $v(\mathbf{p})$ are two integers that represent the displacements of x and y axes from the coordinates $\mathbf{p}$, respectively, s_1, s_2 are the two dense SIFT descriptors to be measured and ε represents the set of all the spatial neighborhoods.

- First, these facial regions between kinship image pairs have high visual resemblance (e.g., their eyes resemble each other), which means their local descriptors are similar, namely $\|s_1(\mathbf{p}) - s_2(\mathbf{p} + \mathbf{d}(\mathbf{p}))\|$ is small.

- Second, these facial regions should be at similar relative locations on two faces (e.g., their eyes appear at similar locations on two faces), which means there may be a small displacement between the centers of two local descriptors, namely $\|\mathbf{d}(\mathbf{p})\|$ is small.

- Third, the neighborhood regions of high resemblance facial regions tend to be similar (e.g., the neighborhood small regions around the center of eyes tend to be smoothly changed), which means $\|\mathbf{d}(\mathbf{p}) - \mathbf{d}(\mathbf{q})\|$ is small where $(\mathbf{p}, \mathbf{q}) \in \varepsilon$.

Inspired by these anthropological observations, we propose three novel anthropology inspired features to capture these high resemblance facial regions between parents and their children. First, we present a new anthropology inspired similarity enhancement (AISE) method by extending the SIFT flow [61] method from the scene alignment to kinship image pairs. The SIFT flow algorithm matches densely sampled SIFT features and finds correspondence estimated by SIFT flow. The objective function for SIFT flow [61] is defined as follows:

$$\begin{aligned} E(\mathbf{d}) = \sum_{\mathbf{p}} (\|s_1(\mathbf{p}) - s_2(\mathbf{p} + \mathbf{d}(\mathbf{p}))\|_1) + \\ \sum_{\mathbf{p}} \eta(\|\mathbf{d}(\mathbf{p})\|_1) + \sum_{\mathbf{p},\mathbf{q} \in \varepsilon} \theta(\|\mathbf{d}(\mathbf{p}) - \mathbf{d}(\mathbf{q})\|_1) \end{aligned} \tag{5.1}$$

As we have seen, the SIFT flow method, which satisfies three properties of high visual resemblance facial regions between kinship pairs, is very suitable to be extended to kinship image pairs for capturing the inheritable information between parents and children. Then the estimated SIFT flow can be applied to reinforce the high visual resemblance facial regions and generate similarity enhanced images.

To visualize the effectiveness of our method, the top three principal components of the SIFT descriptors of the image are mapped to the principal components of the RGB space, as shown in Figure 5.2. The purple and the orange regions in the visualization highlight the high visual resemblance regions in the kinship images. It can be discovered that these regions focus on eyes, mouth, chin and parts of the forehead. Therefore our proposed AISE method derives interesting phenomena that are consistent to the anthropology results in [95]. Other interesting patterns can also be deduced for different relations from Figure 5.2. It can be observed that the father-son and mother-daughter relation show large visual correspondence in different parts of facial regions leading to the deduction that individuals of the same gender in kinship relations share higher visual resemblance. It can also be seen that mother-daughter relation has higher genetic responses compared to father-daughter relation confirming the observation that mothers resemble their daughters more as in [2]. The experimental results in Section 5.4.2 also confirm such an observation.

Then the AIF-SIFT, AIF-WLD and AIF-DAISY descriptors are extracted from the similarity enhanced images derived by our anthropology inspired similarity enhancement method. Therefore we name these three anthropology inspired features as AIF-SIFT, AIF-WLD and AIF-DAISY. In particular, the AIF-SIFT feature is computed in the opponent color space [44] of the enhanced image. We then derive densely sampled SIFT features from the image encoded by the Weber local descriptors (WLD) and the process is repeated separately for the three components of the image resulting in color AIF-WLD feature. To improve the robustness against photometric and geometric transformations of the enhanced image, dense AIF-DAISY descriptors are computed with parameters radius of descriptor set as 15, number of rings as 3, number of histograms per ring as 8 and number of histogram bins as 8 resulting in a 200 dimension AIF-DAISY descriptor.

5.3 Score Space based Multiple Metric Learning Method

The complementary nature of discriminative and generative approach [96] leads to the generative score space. One example is the Fisher score [39], which has been widely applied for visual classification problems such as face recognition [116], object recognition [40]. In this section, we extend the Fisher score from classification problem to metric

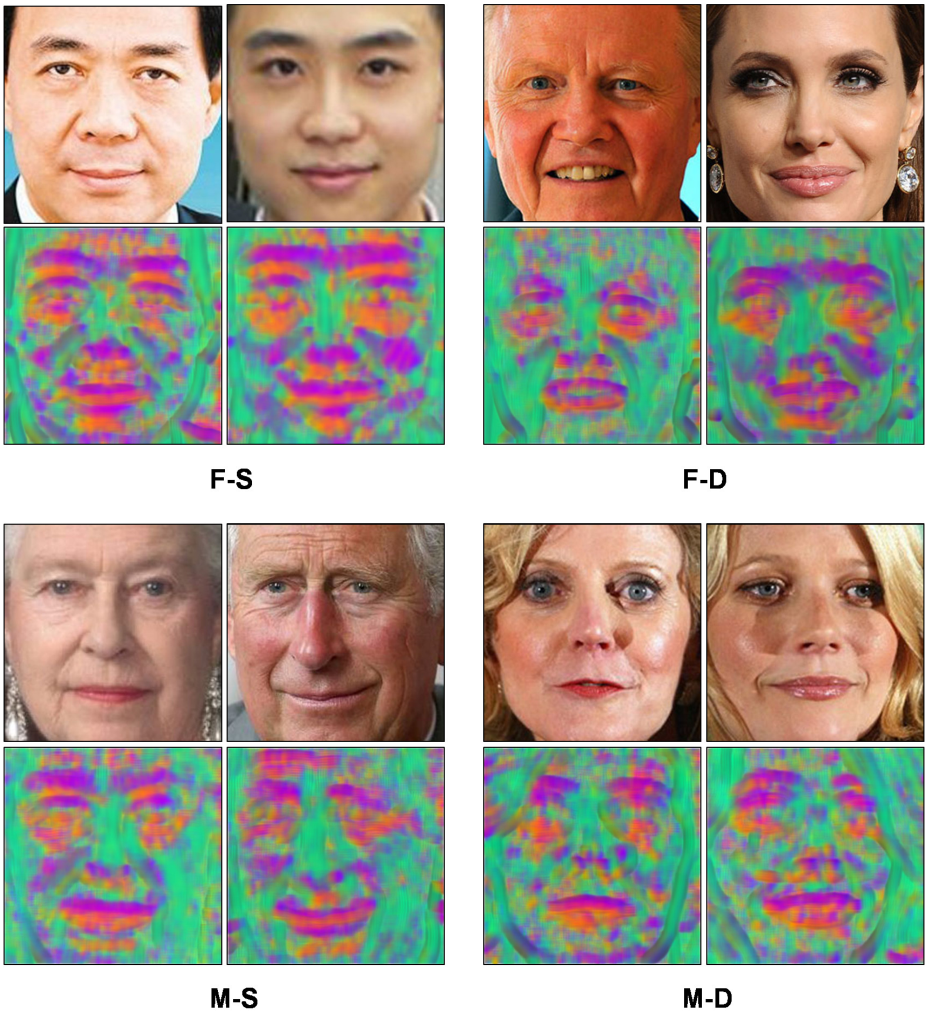


Figure 5.2 Visualization of SIFT images of different kinship relations using the top three principal components of SIFT descriptors extracted from the image. The purple and orange regions in the visualization highlight the inheritable genetic feature regions in the kinship images.

learning problem. Particularly, let $\mathbf{X}_i = \{\mathbf{d}_t, t = 1, 2, ..., T\}$ be the set of T local descriptors (e.g. AIF-SIFT, AIF-WLD or AIF-DAISY) extracted from an image of the $i-$th pair. Then $\mathbf{Y}_i$ is defined similarly for the other image of the $i-$th pair. Let $p(\mathbf{X}|\boldsymbol{\lambda})$ be the probability density function of generating $\mathbf{X}_i$ or $\mathbf{Y}_i$ with a set of parameters $\boldsymbol{\lambda}$, then the Fisher score is defined as follows:

$$\mathbf{F}(\mathbf{X}_i) = \frac{1}{T} \bigtriangledown_{\boldsymbol{\lambda}} \log[p(\mathbf{X}_i|\boldsymbol{\lambda})] \tag{5.2}$$

As a matter of fact, the Fisher score is the gradient vector of the log-likelihood that describes the contribution of the parameters to the generation process. It describes the generative perspective of features. Based on the Fisher score, a score space based similarity measure, namely Fisher kernel [39], is derived as $K_F(\mathbf{X}_i, \mathbf{Y}_i) = (\mathbf{F}(\mathbf{X}_i))^T \mathbf{I}^{-1} \mathbf{F}(\mathbf{Y}_i)$ using the Fisher information matrix $\mathbf{I}$. The conventional Fisher kernel provides a natural similarity measure between images by considering the underlying probability distribution. However, three major issues inherent of the conventional Fisher kernel are still waiting for solutions. First, the conventional Fisher kernel fails to take into account of the label information. Second, the Fisher information matrix $\mathbf{I}$ is difficult to obtain and approximation techniques are not sufficient to guarantee performance. Third, it only measures the similarity for a single aspect between images, which depends on the type of the local image descriptors.

Therefore, this chapter presents a novel score space based multiple metric learning method to address these three issues by learning a new distance metric that captures the pairwise information, and the weights of multiple distance metrics that exploits information from different features. Specifically, the score space based multiple distance metric is defined as follows with the weights $w_c (c = 1, 2, ..., k)$: $D(\mathbf{X}_i, \mathbf{Y}_i) = \sum_{c=1}^{k} w_c D_c(\mathbf{X}_i^c, \mathbf{Y}_i^c) = \sum_{c=1}^{k} w_c (\mathbf{p}_i^c)^T \mathbf{M} (\mathbf{c}_i^c) = \sum_{c=1}^{k} w_c (\mathbf{p}_i^c)^T \mathbf{W}\mathbf{W}^T (\mathbf{c}_i^c) = \sum_{c=1}^{k} w_c (\mathbf{x}_i^c)^T (\mathbf{y}_i^c)$, where $\mathbf{p}_i^c = \mathbf{F}(\mathbf{X}_i)$, $\mathbf{c}_i^c = \mathbf{F}(\mathbf{Y}_i)$, $\mathbf{x}_i^c = \mathbf{W}^T \mathbf{p}_i^c$ and $\mathbf{y}_i^c = \mathbf{W}^T \mathbf{c}_i^c$ $(i = 1, 2, ..., m)$. It is easy to see that matrix $\mathbf{M} = \mathbf{W}\mathbf{W}^T$ is symmetric and positive definite. To keep the notation simple, we use

$D(\mathbf{x}_i, \mathbf{y}_i)$ instead of $D(\mathbf{X}_i, \mathbf{Y}_i)$ in the remaining parts of the chapter. The introduction of **W** alleviates the assumptions on the Fisher information matrix since **W** can be learned from the training data and contains sufficient information for recognizing kinship relations.

The derivation of **W** and w_c consists of two iterative procedures. Let $\mathbf{D} = \{(\mathbf{x}_i^c, \mathbf{y}_i^c) | \mathbf{x}_i^c, \mathbf{y}_i^c \in \mathbb{R}^{n \times 1} (i = 1, 2, ..., m, c = 1, 2, ..., k)\}$. The main purpose of the transformation **W** and weights w_c is to push away the nearby non-kinship samples as far as possible while pulling the kinship relation samples as close as possible, and approximate the ideal similarity matrix. In other words, the distance between $\mathbf{x}_i^c$ and $\mathbf{y}_i^c$ should be as small as possible if $\mathbf{x}_i^c$ and $\mathbf{y}_i^c$ have kinship relations and otherwise the distance should be large. Therefore, the objective function for the SML method can be formulated as follows.

$$\begin{aligned} &\min_{\mathbf{W}, w_c} \| D_{\mathbf{I}} - \sum_{c=1}^{k} w_c D_c \|_F^2 + \alpha \sum_{c=1}^{k} w_c^2 + \lambda \sum_{c=1}^{k} d_c |w_c| \\ &s.t. \mathbf{W}^T \mathbf{W} = \mathbf{I}, \sum_{c=1}^{k} w_c = 1, w_c > 0 \end{aligned} \tag{5.3}$$

In this objective function, the third term of Equation 5.3 represents the criterion of pushing away the nearby non-kinship samples as far as possible while pulling the kinship samples as close as possible. While the first and second term show the reconstruction criterion and the regularization for the weights of different metrics The d_c is defined as $d_c = \sum_{i=1}^{m} 2 * D_c(\mathbf{x}_i^c, \mathbf{y}_i^c) - D_c(\mathbf{x}_i^c, (\mathbf{y}_i^c)^*) - D_c((\mathbf{x}_i^c)^*, \mathbf{y}_i^c) = \mathrm{Tr}\,(\mathbf{W}^T (2\mathbf{M}_1^c - \mathbf{M}_2^c - \mathbf{M}_3^c)\mathbf{W})$, where $\mathbf{M}_1^c = \sum_{i=1}^{m} \mathbf{p}_i^c (\mathbf{c}_i^c)^T$, $(\mathbf{x}_i^c)^*$ is the nearest neighbor of $\mathbf{x}_i^c$, $(\mathbf{y}_i^c)^*$ is the nearest neighbor of $\mathbf{y}_i^c$, $D_c \in \mathbb{R}^{m \times m}$ is the similarity matrix for the c-th feature ($c = 1, 2, ..., k$) and $D_{\mathbf{I}} \in \mathbb{R}^{m \times m}$ is the ideal similarity matrix which is derived by multiplying the scaled label vector with its transpose. Note that $\mathbf{M}_1^c$ is not symmetric, then we make it symmetric by using $\mathbf{M}_1^c = (\mathbf{M}_1^c + (\mathbf{M}_1^c)^T)/2$ without influencing the value of d_c. $\mathbf{M}_2^c$ and $\mathbf{M}_3^c$ can be computed in a similar way.

Now the problem becomes a constrained, non-negative, and weighted variant of the sparse representation problem and the term $\sum_{c=1}^{k} d_c |w_c|$, which corresponds to the criterion of pushing away the nearby non-kinship samples and pulling close the kinship samples, behaves as a regularization for the multiple metric learning problem.

The the objective function 5.3 then can be optimized using an iterative procedure. Specifically, given the fixed w_c, we can approximately update $\mathbf{W}$ by discarding the reconstruction criterion and optimizing the following objective function:

$$\begin{aligned}&\max_{\mathbf{W}} \operatorname{Tr}(\mathbf{W}^T \sum_{c=1}^{k} w_c(\mathbf{M}_2^c + \mathbf{M}_3^c - 2\mathbf{M}_1^c)\mathbf{W})\\ &s.t. \mathbf{W}^T\mathbf{W} = \mathbf{I}\end{aligned} \tag{5.4}$$

This can be done by deriving the eigenvectors of matrix $\sum_{c=1}^{k} w_c(\mathbf{M}_2^c + \mathbf{M}_3^c - 2\mathbf{M}_1^c)$.

Then given the $\mathbf{W}$, we can optimize the following problem to derive w_c:

$$\begin{aligned}&\min_{w_c} ||D_{\mathbf{I}} - \sum_{c=1}^{k} w_c D_c||_F^2 + \alpha \sum_{c=1}^{k} w_c^2 + \lambda \sum_{c=1}^{k} d_c|w_c|\\ &s.t. \sum_{c=1}^{k} w_c = 1, w_c > 0\end{aligned} \tag{5.5}$$

We can apply the FISTA algorithm [3] to optimize the objective function defined in Equation 5.5. The structure of the FISTA algorithm remains the same but the proximal operator is different as our method is a constrained, non-negative, and weighted variation. We thus replace the original soft thresholding operator with an efficient projection operator [21] considering the non-negative constraint. We can also transform the objective function defined in Equation 5.5 into a quadratic programming problem by using the fact $\lambda \sum_{c=1}^{k} d_c|w_c| = \lambda \sum_{c=1}^{k} d_c w_c$ since $w_c > 0$. Then the objective function can be optimized efficiently.

After the SML is derived, a novel normalized multiple similarity measure (NMSM) is further proposed, where the SML is normalized as follows with the power transformation $p(\mathbf{x})$ defined as $p(\mathbf{x}) = sign(\mathbf{x})|\mathbf{x}|^{\beta}$, where β ($0 < \beta < 1$) is the power parameter, and both the power and the sign operations are element-wise.

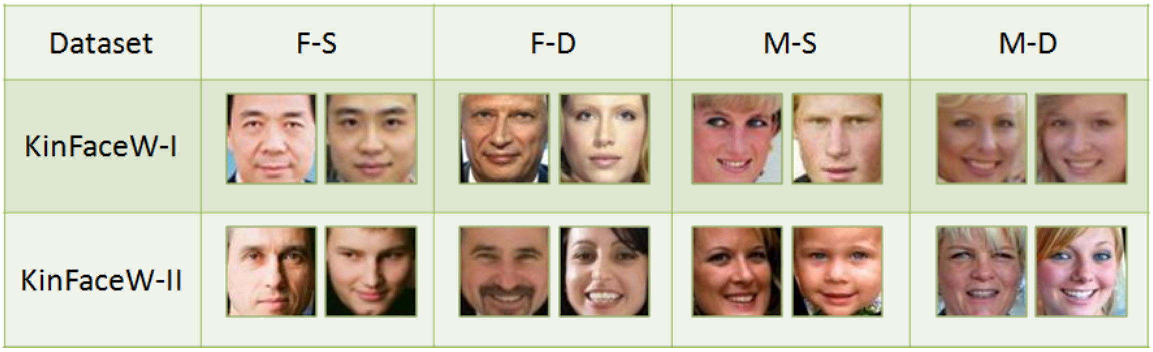


Figure 5.3 Example images from the KinFaceW-I and KinFaceW-II data set

$$NMSM(\mathbf{x}_i, \mathbf{y}_i) = \sum_{c=1}^{k} w_c \frac{D_c(p(\mathbf{x}_i^c), p(\mathbf{y}_i^c))}{\|\mathbf{W}^T p(\mathbf{x}_i^c)\| \|\mathbf{W}^T p(\mathbf{y}_i^c)\|} \tag{5.6}$$

The proposed NMSM takes advantage of normalization through fractional power transformation and the L_2 normalization. The fractional power transformation is able to transform from the data into a near Gaussian shape with a stable variance [40], [129]. With the help of the L_2 normalization, it can be proved that the NMSM is proportional to a weighted linear combination of the whitened cosine similarity measure for each feature, which shows its theoretical roots in the Bayes decision rule for minimum error [64] under some conditions such as multivariate Gaussian distribution assumption. Then the proposed NMSM establishes a relation between the score space based multiple metric learning and the Bayes rule induced similarity measure under the assumption of multivariate Gaussian distribution. Thus it is theoretically guaranteed to achieve good performance.

5.4 Experiments

This section evaluates the effectiveness of our proposed method on two challenging kinship databases: the KinFaceW-I data set and the KinFaceW-II data set [89], [88]. These two data sets contain images for four kinship relations: father-son (F-S), father-daughter (F-D), mother-son (M-S), and mother-daughter (M-D). In the KinFaceW-I data set, there are 156,

Table 5.1 Comparison between the SML and other Methods on the KinFaceW-I Data Set

Methods	F-S	F-D	M-S	M-D	Mean
CSML [97]	61.10	58.10	60.90	70.00	62.50
NCA [28]	62.10	57.10	61.90	69.00	62.30
LMNN [136]	63.10	58.10	62.90	70.00	63.30
NRML [89]	64.10	59.10	63.90	71.00	64.30
MNRML [89]	72.50	66.50	66.20	72.00	69.90
ITML [17]	75.30	64.30	69.30	76.00	71.20
DMML [145]	74.50	69.50	69.50	75.50	72.25
MPDFL [146]	73.50	67.50	66.10	73.10	70.10
GGA [18]	70.50	70.00	67.20	74.30	70.50
ANTH [18]	72.50	71.50	70.80	75.60	72.60
DGA [18]	76.40	72.50	71.90	77.30	74.50
Polito [88]	85.30	85.80	87.50	86.70	86.30
LIRIS [88]	83.04	80.63	82.30	84.98	82.74
NUAA [88]	86.25	80.64	81.03	83.93	82.96
CNN-Basic [159]	70.80	75.70	79.40	73.40	74.80
CNN-Points [159]	71.80	76.10	84.10	78.00	77.50
SML AIF-SIFT	75.61	72.75	75.04	85.87	77.32
SML AIF-WLD	85.27	78.40	81.01	84.18	82.22
SML AIF-DAISY	80.75	81.40	77.60	84.18	80.98
SML	**88.15**	**82.49**	**80.62**	**90.95**	**85.55**

134, 116, and 127 image pairs for each relation respectively. In the KinFaceW-II data set, there are 250 pairs of images for each kinship relation. Example images are shown in Figure 5.3. In our experiments, we conduct 5-fold cross validation where both data sets are divided into five folds having the same number of image pairs [89], [88].

5.4.1 Implementation Details

The AISE method is first applied to derive the similarity enhanced images. Second, we derive the AIF-DAISY feature and the color AIF-WLD feature on the similarity enhanced images. The dense color SIFT feature is derived with a step size of 1 and five scale patch sizes as 2, 4, 6, 8, 10. Then, the dimensionality of the opponent color SIFT feature is further reduced to 64 by PCA. The spatial information [116] is added to the SIFT feature with 2 more dimensions. The color AIF-WLD feature is computed similarly. The dimensionality of the AIF-DAISY feature is directly reduced to 66 by PCA. Afterwards, a Gaussian mixture model with 256 components is estimated for the Fisher score computation. Then the score space based multiple metric learning is learned from the data with the parameters $\alpha = 1$ and $\lambda = 0.1$ for both the KinFaceW-I data set and the KinFaceW-II data set. The normalized multiple similarity measure with $\beta = 0.5$ is applied. Finally a two class support vector machine is used to determine the kinship relations between images.

5.4.2 Comparison with the State-of-the-Art

This section presents the comparison between the proposed method and the state-of-the-art methods. Experimental results on Table 5.1 and Table 5.2 show that our method is able to achieve competitive and even better results than the state-of-the-art methods.

In particular, the Polito team [88] achieves the state-of-the-art 86.30% mean verification rate on the KinFaceW-I data set, which is slightly better than our method that achieves 85.55% mean verification rate. However, our method achieves 89.80% mean

Table 5.2 Comparison between the SML and other Methods on the KinFaceW-II Data Set

Methods	F-S	F-D	M-S	M-D	Mean
CSML [97]	71.80	68.10	73.80	74.00	71.90
NCA [28]	73.80	70.10	74.80	75.00	73.50
LMNN [136]	74.80	71.10	75.80	76.00	74.50
NRML [89]	76.80	73.10	76.80	77.00	75.70
MNRML [89]	76.90	74.30	77.40	77.60	76.50
ITML [17]	69.10	67.00	65.60	68.30	67.50
DMML [145]	78.50	76.50	78.50	79.50	78.25
MPDFL [146]	77.30	74.70	77.80	78.00	77.00
GGA [18]	81.80	74.30	80.50	80.80	79.40
DGA [18]	83.90	76.70	83.40	84.80	82.20
Polito [88]	84.00	82.20	84.80	81.20	83.10
LIRIS [88]	89.40	83.60	86.20	85.00	86.05
NUAA [88]	84.40	81.60	82.80	81.60	82.50
CNN-Basic [159]	79.60	84.90	88.50	88.30	85.30
CNN-Points [159]	81.90	89.40	92.40	89.90	88.40
SML AIF-SIFT	88.20	82.00	87.80	85.20	85.80
SML AIF-WLD	75.40	71.60	73.00	77.00	74.25
SML AIF-DAISY	87.80	85.00	89.20	86.00	87.00
SML	**91.40**	**87.20**	**90.80**	**89.80**	**89.80**

verification rate on the KinFaceW-II data set, which is significantly better than Polito, who only achieves 83.10% mean verification rate on the KinFaceW-II data set.

Meanwhile, our method achieves the state-of-the-art results: 89.80% mean verification rate on the KinFaceW-II data set, which is better than the convolutional neural network based method CNN-Points [159], which achieves 88.40% mean verification rate. Besides, our method obtains 85.55% mean verification rate on the KinFaceW-I data set, which improves upon CNN-Points [159] by a large margin around 8%. In summary, our method averagely obtains the state-of-the-art results on both data sets.

Note that our method can significantly improve upon other metric learning methods that use multiple features, such as MNRML [89], DMML [145]. The second observation is that our method often achieves better results on F-S and M-D kinship relations than F-D and M-S kinship relations, which is consistent to the anthropological results [2]. The reason is that the similarity variation between images of different gender is larger than that of the same gender and our proposed SML method is able to capture such a variation by learning the new transformation and the weights of multiple features. The third observation is that our proposed SML method achieves more improvement on the KinFaceW-II data set due to the availability of more training samples.

Please also see Figure 5.4 for comparison with other state-of-the-art results.

5.4.3 Evaluation of the Anthropology Inspired Features

This section assesses the effectiveness of the anthropology inspired features (AIF). Note that three anthropology inspired features (AIF-SIFT, AIF-WLD and AIF-DAISY features) are evaluated separately first (simply assign the weight of the feature set to 1, and others 0). Similarly, without applying the AISE method for deriving similarity enhanced images, we do the same for SIFT, WLD and DAISY features separately to obtain results. The experimental results on Table 5.3 show that the performance of the anthropology inspired features (AIF-SIFT, AIF-WLD and AIF-DAISY features) derived from the enhanced

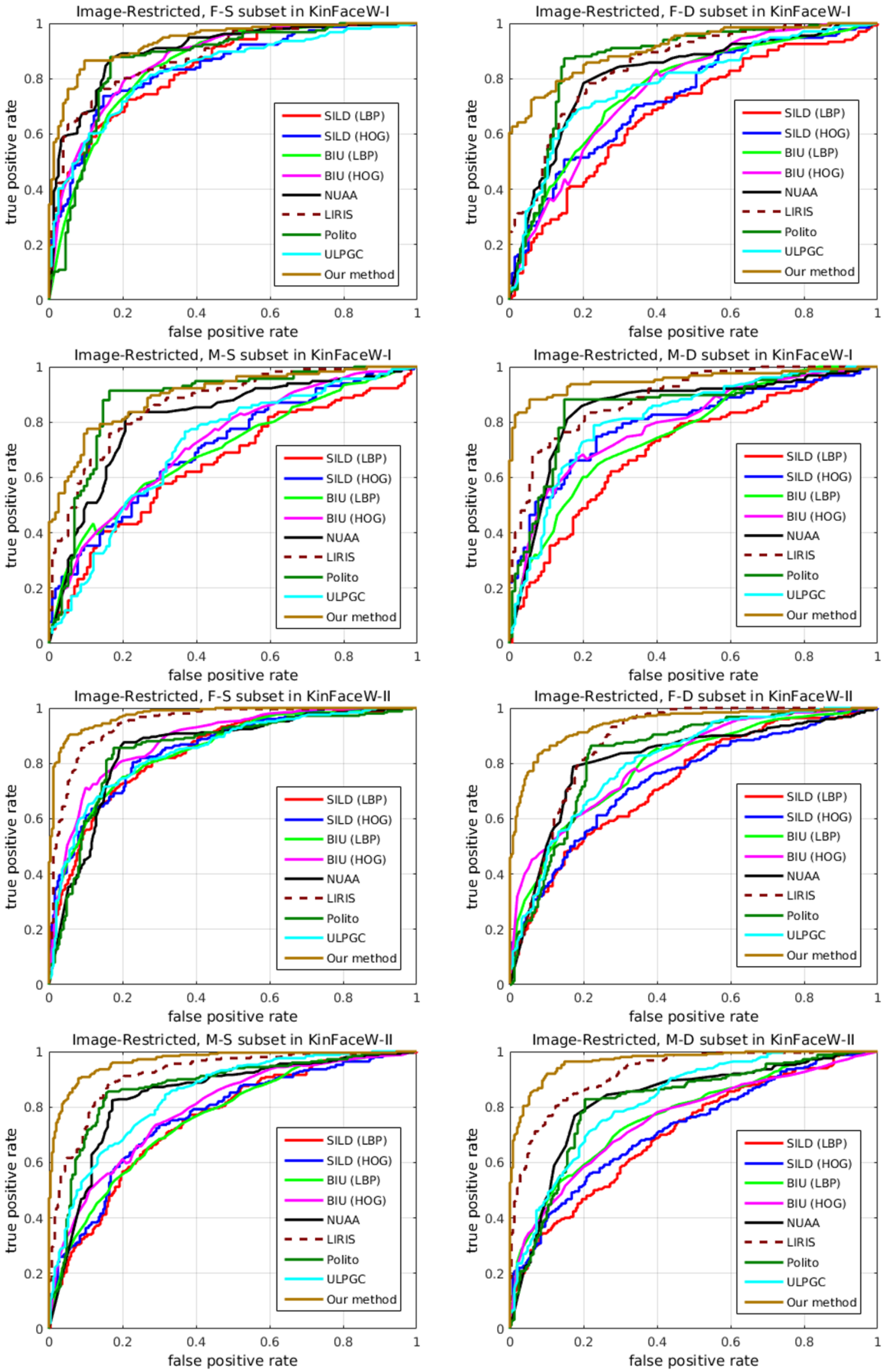


Image-Restricted, F-S subset in KinFaceW-I

Image-Restricted, F-D subset in KinFaceW-I

Image-Restricted, M-S subset in KinFaceW-I

Image-Restricted, M-D subset in KinFaceW-I

Image-Restricted, F-S subset in KinFaceW-II

Image-Restricted, F-D subset in KinFaceW-II

Image-Restricted, M-S subset in KinFaceW-II

Image-Restricted, M-D subset in KinFaceW-II

Figure 5.4 The ROC curve for comparison with other state-of-the-art results.

Table 5.3 Evaluation of the Effectiveness of the Anthropology Inspired Features (AIF-SIFT, AIF-WLD and AIF-DAISY Features) on the KinFaceW-I and KinFaceW-II Data Set

KinFaceW-I	F-S	F-D	M-S	M-D	Mean
SIFT	73.41	69.02	66.40	79.56	72.09
WLD	73.35	65.69	70.69	71.70	70.36
DAISY	71.79	65.68	66.34	75.96	69.94
AIF-SIFT	75.61	72.75	75.04	85.87	77.32
AIF-WLD	85.27	78.40	81.01	84.18	82.22
AIF-DAISY	80.75	81.40	77.60	84.18	80.98

KinFaceW-II	F-S	F-D	M-S	M-D	Mean
SIFT	80.40	70.20	79.80	80.00	77.60
WLD	68.80	62.00	63.20	65.00	64.75
DAISY	76.40	69.80	71.00	70.60	71.95
AIF-SIFT	88.20	82.00	87.80	85.20	85.80
AIF-WLD	75.40	71.60	73.00	77.00	74.25
AIF-DAISY	87.80	85.00	89.20	86.00	87.00

images using the AISE method significantly improve the performance of SIFT, WLD and DAISY features without applying the AISE method. Such a significant improvement demonstrates the effectiveness of our AISE method.

5.4.4 Comparison of SML, SSML and FK

This section presents the comparison of our proposed SML method, single SML (SSML) method which uses a single feature, as well as Fisher kernel (FK) method [40] with a single feature when other experimental settings are fixed. Experimental results in Table 5.4 show that our proposed SML method improves upon the SSML and FK method on both datasets. We thus can make the following conclusions: (1) multiple features improves upon a single feature since SML method improves upon the SSML for all the features; (2) the learning of the new transformation improves the performance since SSML method improves upon FK method for all the three features.

5.4.5 Computational Complexity

This section presents the analysis of the computational complexity of our method. Empirically our method costs 165 and 527 seconds on two datasets respectively. Theoretically the cost is $O((m+n)*n^2)$ when updating **W** and $O(k^3)$ when updating w_c for each iteration. The total cost is $O(t((m+n)*n^2+k^3))$ for t iterations. In practice, $k = 3, m < n, t <= 10$, thus the total cost is $O(n^3)$.

5.5 Conclusion

This chapter presents a novel score space based multiple metric learning (SML) method for Kinship Verification. First, three new anthropology inspired features are extracted, namely the AIF-SIFT, AIF-WLD and AIF-DAISY features. Second, a novel score space based multiple metric learning method is proposed to combine multiple features and their metrics between images in a unified paradigm by iteratively learning a new transformation

Table 5.4 Comparison of SML, SSML and FK Methods on the KinFaceW-I and KinFaceW-II Data Set

KinFaceW-I	F-S	F-D	M-S	M-D	Mean
FK AIF-SIFT	73.43	72.05	71.16	79.87	74.13
FK AIF-WLD	80.81	76.54	76.70	79.41	78.36
FK AIF-DAISY	75.65	77.32	68.59	79.44	75.24
SSML AIF-SIFT	75.61	72.75	75.04	85.87	77.32
SSML AIF-WLD	85.27	78.40	81.01	84.18	82.22
SSML AIF-DAISY	80.75	81.40	77.60	84.18	80.98
SML	**88.15**	**82.49**	**80.62**	**90.95**	**85.55**

KinFaceW-II	F-S	F-D	M-S	M-D	Mean
FK AIF-SIFT	82.40	77.40	79.80	78.80	79.60
FK AIF-WLD	73.80	71.20	76.20	73.80	73.75
FK AIF-DAISY	85.60	81.80	85.00	82.60	83.75
SSML AIF-SIFT	88.20	82.00	87.80	85.20	85.80
SSML AIF-WLD	75.40	71.60	73.00	77.00	74.25
SSML AIF-DAISY	87.80	85.00	89.20	86.00	87.00
SML	**91.40**	**87.20**	**90.80**	**89.80**	**89.80**

and the weights. Third, a novel normalized multiple similarity measure is presented based on the SML. Experimental results show that the proposed method is able to achieve the state-of-the-art results for kinship verification.

CHAPTER 6

FUTURE WORK

This dissertation has presented three learning methods for visual recognition, namely a new locally linear K nearest neighbor method, or LLK method, a new generative and discriminative sparse representation (GDSR) method and a new Score space based multiple Metric Learning (SML) method. The key part of the LLK method and the GDSR method lies in the connection between a novel sparse representation method and a classification method. The input feature of these two methods also plays an important role in the performance. Therefore, in the future work, I will focus more on exploring the powerful deep learning method, such as convolutional neural network (CNN), for feature extraction for different visual recognition tasks and data sets. In the meantime, different model structures, parameter tuning, dataset augmentation about CNN are also important topics that I will work on.

I will also try to make a large scale data set for the fine art painting analysis problem. The new large scale paintings data set aims at providing both the computer vision community and the artist community a better understanding of the fine art paintings by using the cutting edge deep learning methods, especially the state-of-the-art deep CNN models, such as VGG Net, GoogleNet and ResNet.

Besides, I will also try to make a large data set for kinship verification with the goal of assisting the anthropology studies by using the advanced computer vision technologies.

Furthermore, I am also involved a project of intelligent incident detection and vehicle counting in videos. I will explore new video analysis techniques using both deep convolutional neural network and deep recurrent neural network.

Finally, I will continue exploring the computer vision problems in the real world, such as satellite image analysis and medical image analysis. I will apply the state of the art techniques to solve these problems for a better world.

BIBLIOGRAPHY

[1] Naveed Akhtar, Faisal Shafait, and Ajmal Mian. Discriminative bayesian dictionary learning for classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[2] Alexandra Alvergne, Charlotte Faurie, and Michel Raymond. Differential facial resemblance of young children to their parents: who do children look like more? *Evolution and Human Behavior*, 28(2):135 – 144, 2007.

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

[4] Peter Belhumeur, João Hespanha, and David Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[5] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Hierarchical Matching Pursuit for Image Classification: Architecture and Fast Algorithms. In *Advances in Neural Information Processing Systems*, pages 2115–2123, December 2011.

[6] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Multipath sparse coding using hierarchical matching pursuit. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 660–667, 2013.

[7] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *ACM International Conference on Image and Video Retrieval*, pages 401–408, 2007.

[8] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. In *British Machine Vision Conference*, 2014.

[9] Paola Bressan and Maria F Dal Martello. Talis pater, talis filius: Perceived resemblance and the belief in genetic relatedness. *Psychological Science*, 13(3):213–218, 2002.

[10] Sijia Cai, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. A probabilistic collaborative representation based approach for pattern classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2016.

[11] Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. Face recognition with learning-based descriptor. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2707–2714, 2010.

[12] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[13] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3025–3032, 2013.

[14] Jie Chen, Shiguang Shan, Chu He, Guoying Zhao, Matti Pietikainen, Xilin Chen, and Wen Gao. WLD: A robust local image descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1705–1720, Sept 2010.

[15] Yuejie Chi and Fatih Porikli. Classification and boosting with multiple collaborative representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1519–1531, 2014.

[16] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[17] Jason Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit Dhillon. Information-theoretic metric learning. In *International Conference on Machine Learning*, pages 209–216, 2007.

[18] Afshin Dehghan, Enrique G Ortiz, Ruben Villegas, and Mubarak Shah. Who do I look like? determining parent-offspring resemblance via gated autoencoders. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1757–1764, 2014.

[19] Weihong Deng, Jiani Hu, and Jun Guo. Extended src: Undersampled face recognition via intraclass variant dictionary. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1864–1870, 2012.

[20] Weihong Deng, Jiani Hu, and Jun Guo. In defense of sparsity based face recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 399–406, 2013.

[21] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l1-ball for learning in high dimensions. In *International Conference on Machine Learning*, pages 272–279, 2008.

[22] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[23] Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination in sparse supervised learning. Technical Report UC/EECS-2010-126, EECS Dept., University of California at Berkeley, September 2010.

[24] Ruogu Fang, Kevin D Tang, Noah Snavely, and Tsuhan Chen. Towards computational models of kinship verification. In *International Conference on Image Processing*, pages 1577–1580, 2010.

[25] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition (2Nd Ed.)*. Academic Press Professional, Inc., San Diego, CA, USA, 1990.

[26] Shenghua Gao, Ivor Wai-Hung Tsang, and Liang-Tien Chia. Laplacian sparse coding, hypergraph Laplacian sparse coding, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):92–104, 2013.

[27] Hanlin Goh, Nicolas Thome, Matthieu Cord, and Joo-Hwee Lim. Learning deep hierarchical visual feature coding. *IEEE Transactions on Neural Networks and Learning Systems*, 2014.

[28] Jacob Goldberger, Sam T. Roweis, Geoffrey E. Hinton, and Ruslan Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems*, 2004.

[29] Kristen Grauman and Bastian Leibe. Visual object recognition. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(2):1–181, 2011.

[30] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[31] Zhenhua Guo, Lei Zhang, and David Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663, June 2010.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

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

[34] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge J. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2015.

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

[36] Ke Huang and Selin Aviyente. Sparse representation for signal classification. In *Advances in Neural Information Processing Systems*, pages 609–616, 2006.

[37] Shaoli Huang, Zhe Xue, Dacheng Tao, and Ya Zhang. Part-stacked CNN for fine-grained visual categorization. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2016.

[38] Yongzhen Huang, Zifeng Wu, Liang Wang, and Tieniu Tan. Feature coding in image classification: A comprehensive study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 493–506, 2014.

[39] Tommi Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems*, pages 487–493, 1998.

[40] Herve Jegou, Florent Perronnin, Matthijs Douze, Jorge Sanchez, Patrick Perez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2012.

[41] Li jia Li and Fei fei Li. What, where and who? classifying event by scene and object recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[42] Zhuolin Jiang, Zhe Lin, and Larry S. Davis. Label consistent K-SVD: Learning a discriminative dictionary for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2651–2664, 2013.

[43] Mayank Juneja, Andrea Vedaldi, C. V. Jawahar, and Andrew Zisserman. Blocks that shout: Distinctive parts for scene classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 923–930, 2013.

[44] Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost van de Weijer, Andrew D Bagdanov, Antonio M Lopez, and Michael Felsberg. Coloring action recognition in still images. *International journal of computer vision*, 105(3):205–221, 2013.

[45] Fahad Shahbaz Khan, Shida Beigpour, Joost van de Weijer, and Michael Felsberg. Painting-91: a large scale database for computational painting categorization. *Machine Vision and Applications*, 25(6):1385–1397, 2014.

[46] Fahad Shahbaz Khan, Muhammad Anwer Rao, Joost van de Weijer, Andrew D. Bagdanov, Maria Vanrell, and Antonio Lopez. Color attributes for object detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012.

[47] Fahad Shahbaz Khan, Joost Van De Weijer, and Maria Vanrell. Top-down color attention for object recognition. In *International Conference on Computer Vision*, pages 979–986, Sept 2009.

[48] Rahat Khan, Joost Van de Weijer, Fahad Shahbaz Khan, Damien Muselet, Christophe Ducottet, and Cecile Barat. Discriminative color descriptors. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2866–2873, June 2013.

[49] Jonathan Krause, Hailin Jin, Jianchao Yang, and Fei-Fei Li. Fine-grained recognition without part annotations. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2015.

[50] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[51] Y lan Boureau, Jean Ponce, and Yann Lecun. A theoretical analysis of feature pooling in visual recognition. In *International Conference on Machine Learning*, pages 111–118, 2010.

[52] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.

[53] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2007.

[54] Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005.

[55] Fei-Fei Li, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples. In *Workshop on Generative-Model Based Vision, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[56] Li-Jia Li, Hao Su, Eric Xing, and Fei-Fei Li. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in Neural Information Processing Systems*, pages 1378–1386, 2010.

[57] Quannan Li, Jiajun Wu, and Zhuowen Tu. Harvesting mid-level visual concepts from large-scale internet images. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 851–858, 2013.

[58] Yifeng Li and Alioune Ngom. Fast kernel sparse representation approaches for classification. *International Conference on Data Mining*, pages 966–971, 2012.

[59] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv*, 2013.

[60] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *International Conference on Computer Vision*, 2015.

[61] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2011.

[62] Chengjun Liu. Gabor-based kernel PCA with fractional power polynomial models for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):572–581, May 2004.

[63] Chengjun Liu. Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):725–737, 2006.

[64] Chengjun Liu. The Bayes decision rule induced similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1086–1090, 2007.

[65] Chengjun Liu. Learning the uncorrelated, independent, and discriminating color spaces for face recognition. *IEEE Transactions on Information Forensics and Security*, 3(2):213–222, 2008.

[66] Chengjun Liu. Extracting discriminative color features for face recognition. *Pattern Recognition Letters*, 32(14):1796 – 1804, 2011.

[67] Chengjun Liu. Effective use of color information for large scale face verification. *Neurocomputing*, 101:43–51, 2013.

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

[69] Chengjun Liu and Harry Wechsler. Robust coding schemes for indexing and retrieval from large face databases. *IEEE Transactions on Image Processing*, 9(1):132–137, 2000.

[70] Chengjun Liu and Harry Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image processing*, 11(4):467–476, 2002.

[71] Chengjun Liu and Harry Wechsler. Independent component analysis of Gabor features for face recognition. *IEEE Transactions on Neural Networks*, 14(4):919–928, 2003.

[72] Chengjun Liu and Jian Yang. ICA color space for pattern recognition. *IEEE Transactions on Neural Networks*, 20(2):248–257, 2009.

[73] Jiang Liu, Chenqiang Gao, Deyu Meng, and Wangmeng Zuo. Two-stream contextualized CNN for fine-grained image classification. In Dale Schuurmans and Michael P. Wellman, editors, *AAAI Conference on Artificial Intelligence*, 2016.

[74] Lingqiao Liu, Lei Wang, and Xinwang Liu. In defense of soft-assignment coding. In *International Conference on Computer Vision*, pages 2486–2493, 2011.

[75] Qingfeng Liu, Yukhe Lavinia, Abhishek Verma, Joyoung Lee, Lazar Spasovic, and Chengjun Liu. Feature representation and extraction for image search and video retrieval. In *Recent Advances in Intelligent Image Search and Video Retrieval*, chapter 1. Springer, 2017.

[76] Qingfeng Liu and Chengjun Liu. A new locally linear KNN method with an improved marginal Fisher analysis for image classification. In *IEEE International Joint Conference on Biometrics*, 2014.

[77] Qingfeng Liu and Chengjun Liu. A novel hierarchical interaction model and hits map for action recognition in static images. In *IEEE International Conference on Systems, Man, and Cybernetics*, 2014.

[78] Qingfeng Liu and Chengjun Liu. A novel locally linear KNN model for visual recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.

[79] Qingfeng Liu and Chengjun Liu. Improved soft assignment coding for image classification. In *Recent Advances in Intelligent Image Search and Video Retrieval*, chapter 3. Springer, 2017.

[80] Qingfeng Liu and Chengjun Liu. Inheritable color space (incs) and generalized incs framework with applications to kinship verification. In *Recent Advances in Intelligent Image Search and Video Retrieval*, chapter 4. Springer, 2017.

[81] Qingfeng Liu and Chengjun Liu. A novel locally linear KNN method with application to visual recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2017.

[82] Qingfeng Liu, Ajit Puthenputhussery, and Chengjun Liu. Inheritable Fisher vector feature for kinship verification. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2015.

[83] Qingfeng Liu, Ajit Puthenputhussery, and Chengjun Liu. Learning the discriminative dictionary for sparse representation by a general Fisher regularized model. In *IEEE International Conference on Image Processing*, 2015.

[84] Qingfeng Liu, Ajit Puthenputhussery, and Chengjun Liu. Novel general knn classifier and general nearest mean classifier for visual classification. In *IEEE International Conference on Image Processing*, 2015.

[85] Qingfeng Liu, Ajit Puthenputhussery, and Chengjun Liu. A novel inheritable color space with application to kinship verification. In *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[86] Zhiming Liu and Chengjun Liu. Fusion of color, local spatial and global frequency information for face recognition. *Pattern Recognition*, 43(8):2882–2890, 2010.

[87] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[88] Jiwen Lu, Junlin Hu, Venice Erin Liong, Xiuzhuang Zhou, Andrea Bottino, Ihtesham Ul Islam, Tiago Figueiredo Vieira, Xiaoqian Qin, Xiaoyang Tan, Songcan Chen, Yosi Keller, Shahar Mahpod, Lilei Zheng, Khalid Idrissi, Christophe Garcia, Stefan Duffner, Atilla Baskurt, Modesto Castrillon-Santana, and Javier Lorenzo-Navarro. The FG 2015 Kinship Verification in the Wild Evaluation. In *International Conference on Automatic Face and Gesture Recognition*, pages 1–7, May 2015.

[89] Jiwen Lu, Xiuzhuang Zhou, Yap-Pen Tan, Yuanyuan Shang, and Jie Zhou. Neighborhood repulsed metric learning for kinship verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):331–345, 2014.

[90] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[91] Julien Mairal, Francis Bach, and Jean Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012.

[92] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *International Conference on Machine Learning*, page 87, 2009.

[93] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Supervised dictionary learning. In *Advances in Neural Information Processing Systems*, pages 1033–1040, 2008.

[94] Aleix Martínez and Avinash Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001.

[95] Farhad Naini and James Moss. Three-dimensional assessment of the relative contribution of genetics and environment to various facial parameters with the twin method. *American Journal of Orthodontics and Dentofacial Orthopedics*, 126(6):655 – 665, 2004.

[96] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems*, pages 841–848, 2002.

[97] HieuV. Nguyen and Li Bai. Cosine similarity metric learning for face verification. In *Asian Conference on Computer Vision*, volume 6493, pages 709–720, 2011.

[98] Zhenxing Niu, Gang Hua, Xinbo Gao, and Qi Tian. Context aware topic model for scene recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2743–2750, 2012.

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

[100] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[101] Megha Pandey and Svetlana Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *International Conference on Computer Vision*, pages 1307–1314, 2011.

[102] Kuan-Chuan Peng and Tsuhan Chen. Cross-layer features in convolutional neural networks for generic classification tasks. In *International Conference on Image Processing*, pages 3057–3061, Sept 2015.

[103] Kuan-Chuan Peng and Tsuhan Chen. A framework of extracting multi-scale features using multiple convolutional neural networks. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, June 2015.

[104] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the Fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, pages 143–156, 2010.

[105] Ajit Puthenputhussery, Qingfeng Liu, and Chengjun Liu. Color multi-fusion Fisher vector feature for computational painting categorization. In *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[106] Ajit Puthenputhussery, Qingfeng Liu, and Chengjun Liu. Sift flow based genetic Fisher vector feature for kinship verification. In *IEEE International Conference on Image Processing*, 2016.

[107] Ajit Puthenputhussery, Qingfeng Liu, and Chengjun Liu. Sparse representation based complete kernel marginal Fisher analysis framework for computational art painting categorization. In *European Conference on Computer Vision*, 2016.

[108] Ajit Puthenputhussery, Qingfeng Liu, and Chengjun Liu. A sparse representation model using the complete marginal Fisher analysis framework and its applications to visual recognition. *IEEE Transactions on Multimedia*, 2017.

[109] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 413–420, 2009.

[110] Kishore K. Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2013.

[111] Sreemananath Sadan and Jason J. Corso. Action bank: A high-level representation of activity in video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012.

[112] Jorge Sanchez, Florent Perronnin, Thomas Mensink, and Jakob J. Verbeek. Image classification with the Fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[113] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2014.

[114] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[115] Peichung Shih and Chengjun Liu. Comparative assessment of content-based face image retrieval in different color spaces. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(7):873–893, 2005.

[116] Karen Simonyan, Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Fisher vector faces in the wild. In *British Machine Vision Conference*, 2013.

[117] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[118] Vikas Sindhwani and Amol Ghoting. Large-scale distributed non-negative sparse coding and sparse dictionary learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 489–497, 2012.

[119] Jian Sun and Jean Ponce. Learning discriminative part detectors for image classification and cosegmentation. In *International Conference on Computer Vision*, pages 3400–3407, 2013.

[120] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.

[121] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag New York, Inc., New York, NY, USA, 2010.

[122] Engin Tola, Vincent Lepetit, and Pascal Fua. DAISY: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, May 2010.

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

[124] Koen Van De Sande, Theo Gevers, and Cees Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, Sept 2010.

[125] Joost Van De Weijer, Cordelia Schmid, Jakob Verbeek, and Diane Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 18(7):1512–1523, July 2009.

[126] Jan Van Gemert, Jan-Mark Geusebroek, Cor Veenman, and Arnold Smeulders. Kernel codebooks for scene categorization. In *European Conference on Computer Vision*, pages 696–709, 2008.

[127] Jan C Van Gemert, Cor J Veenman, Arnold WM Smeulders, and Jan-Mark Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010.

[128] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[129] Matthew P Wand, James Stephen Marron, and David Ruppert. Transformations in density estimation. *Journal of the American Statistical Association*, 86(414):343–353, 1991.

[130] Heng Wang, Muhammad Muneeb Ullah, Alexander Klser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, 2009.

[131] Jie Wang, Jiayu Zhou, Jun Liu, Peter Wonka, and Jieping Ye. A safe screening rule for sparse logistic regression. In *Advances in Neural Information Processing Systems*, pages 1053–1061, 2014.

[132] Jie Wang, Jiayu Zhou, Peter Wonka, and Jieping Ye. Lasso screening rules via dual polytope projection. In *Advances in Neural Information Processing Systems*, pages 1070–1078, 2013.

[133] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3360–3367, 2010.

[134] Jun Wang, Huyen T Do, Adam Woznica, and Alexandros Kalousis. Metric learning with multiple kernels. In *Advances in Neural Information Processing Systems*, pages 1170–1178, 2011.

[135] Zhaowen Wang, Jianchao Yang, Nasser Nasrabadi, and Thomas Huang. A max-margin perspective on sparse representation-based classification. In *International Conference on Computer Vision*, pages 1217–1224, 2013.

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

[137] Lior Wolf, Tal Hassner, and Yaniv Taigman. Descriptor based methods in the wild. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition, European Conference on Computer Vision*, 2008.

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

[139] Jianxin Wu, Bin-Bin Gao, and Guoqing Liu. Representing sets of instances for visual recognition. In *AAAI Conference on Artificial Intelligence*, pages 2237–2243, 2016.

[140] Siyu Xia, Ming Shao, and Yun Fu. Kinship verification through transfer learning. In *International Joint Conference on Artificial Intelligence*, pages 2539–2544, 2011.

[141] Siyu Xia, Ming Shao, Jiebo Luo, and Yun Fu. Understanding kin relationships in a photo. *IEEE Transactions on Multimedia*, 14(4):1046–1056, Aug 2012.

[142] Zhen James Xiang and Peter Ramadge. Fast lasso screening tests based on correlations. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2137–2140, 2012.

[143] Zhen James Xiang, Hao Xu, and Peter Ramadge. Learning sparse representations of high dimensional data on large scale dictionaries. In *Advances in Neural Information Processing Systems*, volume 24, pages 900–908, 2011.

[144] Eric Xing, Michael Jordan, Stuart Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, pages 505–512, 2002.

[145] Haibin Yan, Jiwen Lu, Weihong Deng, and Xiuzhuang Zhou. Discriminative multimetric learning for kinship verification. *IEEE Transactions on Information Forensics and Security*, 9(7):1169–1178, 2014.

[146] Haibin Yan, Jiwen Lu, and Xiuzhuang Zhou. Prototype-based discriminative feature learning for kinship verification. *IEEE Transactions on Cybernetics*, 2015.

[147] Shuicheng Yan, Dong Xu, Benyu Zhang, HongJiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007.

[148] Jian Yang and Chengjun Liu. Color image discriminant models and algorithms for face recognition. *IEEE Transactions on Neural Networks*, 19(12):2088–2098, 2008.

[149] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1794–1801, 2009.

[150] Meng Yang, Dengxin Dai, Lilin Shen, and Luc Van Gool. Latent dictionary learning for sparse representation based classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2014.

[151] Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang. Fisher discrimination dictionary learning for sparse representation. In *International Conference on Computer Vision*, pages 543–550, 2011.

[152] Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang. Sparse representation based Fisher discrimination dictionary learning for image classification. *International Journal of Computer Vision*, pages 1–24, 2014.

[153] Yang Yang, Shengcai Liao, Zhen Lei, Dong Yi, and Stan Z Li. Color models and weighted covariance estimation for person re-identification. In *International Conference on Pattern Recognition*, pages 1874–1879, Aug 2014.

[154] Di You, Onur Hamsici, and Aleix Martínez. Kernel optimization in discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):631–638, 2011.

[155] Kai Yu, Tong Zhang, and Yihong Gong. Nonlinear learning using local coordinate coding. In *Advances in Neural Information Processing Systems*, pages 2223–2231, 2009.

[156] Matthew Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833, 2014.

[157] Hao Zhang, Alexander Berg, Michael Maire, and Jitendra Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2126–2136, 2006.

[158] Jun Zhang, Youssef Barhomi, and Thomas Serre. A new biologically inspired color image descriptor. In *European Conference on Computer Vision*, pages 312–324. Springer Berlin Heidelberg, 2012.

[159] Kaihao Zhang, Yongzhen Huang, Chunfeng Song, Hong Wu, and Liang Wang. Kinship verification with deep convolutional neural networks. In *British Machine Vision Conference*, pages 148.1–148.12, September 2015.

[160] Lei Zhang, Meng Yang, and Xiangchu Feng. Sparse representation or collaborative representation: Which helps face recognition? In *International Conference on Computer Vision*, pages 471–478, 2011.

[161] Qiang Zhang and Baoxin Li. Discriminative K-SVD for dictionary learning in face recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2691–2698, 2010.

[162] Miao Zheng, Jiajun Bu, Chun Chen, Can Wang, Lijun Zhang, Guang Qiu, and Deng Cai. Graph regularized sparse coding for image representation. *IEEE Transactions on Image Processing*, 20(5):1327–1336, 2011.

[163] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *Arxiv*, 2016.

[164] Ning Zhou, Yi Shen, Jinye Peng, and Jianping Fan. Learning inter-related visual dictionary for object recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3490–3497, 2012.

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