Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

NOVEL IMAGE DESCRIPTORS AND LEARNING METHODS FOR IMAGE CLASSIFICATION APPLICATIONS

by
Ajit Puthenputhussery

Image classification is an active and rapidly expanding research area in computer vision and machine learning due to its broad applications. With the advent of big data, the need for robust image descriptors and learning methods to process a large number of images for different kinds of visual applications has greatly increased. Towards that end, this dissertation focuses on exploring new image descriptors and learning methods by incorporating important visual aspects and enhancing the feature representation in the discriminative space for advancing image classification.

First, an innovative sparse representation model using the complete marginal Fisher analysis (CMFA-SR) framework is proposed for improving the image classification performance. In particular, the complete marginal Fisher analysis method extracts the discriminatory features in both the column space of the local samples based within class scatter matrix and the null space of its transformed matrix. To further improve the classification capability, a discriminative sparse representation model is proposed by integrating a representation criterion such as the sparse representation and a discriminative criterion. Second, the discriminative dictionary distribution based sparse coding (DDSC) method is presented that utilizes both the discriminative and generative information to enhance the feature representation. Specifically, the dictionary distribution criterion reveals the class conditional probability of each dictionary item by using the dictionary distribution coefficients, and the discriminative criterion applies new within-class and between-class scatter matrices for discriminant analysis. Third, a fused color Fisher vector (FCFV) feature is developed by integrating the most expressive features of the DAISY Fisher vector (D-FV) feature, the WLD-SIFT Fisher vector (WS-FV) feature, and the SIFT-FV

feature in different color spaces to capture the local, color, spatial, relative intensity, as well as the gradient orientation information. Furthermore, a sparse kernel manifold learner (SKML) method is applied to the FCFV features for learning a discriminative sparse representation by considering the local manifold structure and the label information based on the marginal Fisher criterion. Finally, a novel multiple anthropological Fisher kernel framework (M-AFK) is presented to extract and enhance the facial genetic features for kinship verification. The proposed method is derived by applying a novel similarity enhancement approach based on SIFT flow and learning an inheritable transformation on the multiple Fisher vector features that uses the criterion of minimizing the distance among the kinship samples and maximizing the distance among the non-kinship samples.

The effectiveness of the proposed methods is assessed on numerous image classification tasks, such as face recognition, kinship verification, scene classification, object classification, and computational fine art painting categorization. The experimental results on popular image datasets show the feasibility of the proposed methods.

NOVEL IMAGE DESCRIPTORS AND LEARNING METHODS FOR IMAGE CLASSIFICATION APPLICATIONS

by
Ajit Puthenputhussery

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

Department of Computer Science

August 2018

APPROVAL PAGE

NOVEL IMAGE DESCRIPTORS AND LEARNING METHODS FOR IMAGE CLASSIFICATION APPLICATIONS

Ajit Puthenputhussery

Dr. Chengjun Liu, Dissertation Advisor — 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. Vincent Oria, Committee Member — Date
Professor of Computer Science, NJIT

Dr. Senjuti Basu Roy, Committee Member — Date
Assistant Professor of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Ajit Puthenputhussery

Degree: Doctor of Philosophy

Date: August 2018

Date of Birth: October 19, 1992

Place of Birth: Kerala, India

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science
  New Jersey Institute of Technology, Newark, NJ, 2018
- Bachelor of Engineering in Computer Engineering
  University of Mumbai, Mumbai, India, 2014

Major: Computer Science

Publications:

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***, vol. 19, no. 8, pp. 1757-1770, Aug. 2017.

A. Puthenputhussery, Q. Liu and C. Liu, "Multiple Anthropological Fisher Kernel Framework and Its Application to Kinship Verification", the IEEE Winter Conference on Applications of Computer Vision (**WACV** 2018), Lake Tahoe, NV, USA, 2018, pp. 57-65.

A. Puthenputhussery, Q. Liu, H. Liu and C. Liu, "Generative and Discriminative Sparse Coding for Image Classification Applications", the IEEE Winter Conference on Applications of Computer Vision (**WACV** 2018), Lake Tahoe, NV, USA, 2018, pp. 1824-1832.

A. Puthenputhussery, Q. Liu, and C. Liu, "Sparse Representation Based Complete Kernel Marginal Fisher Analysis Framework for Computational Art Painting Categorization", the 14th European Conference on Computer Vision (**ECCV** 2016), October 8-16, 2016, Amsterdam, the Netherlands.

A. Puthenputhussery, Q. Liu, and C. Liu, "SIFT Flow Based Genetic Fisher Vector Feature for Kinship Verification", the 23rd IEEE International Conference on Image Processing (**ICIP** 2016), September 25-28, 2016, Phoenix, Arizona, USA.

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 (**WACV** 2016), Lake Placid, NY, 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 (**WACV** 2016), Lake Placid, NY, 2016.

Q. Liu, A. Puthenputhussery, and C. Liu, "Inheritable Fisher Vector Feature for Kinship Verification", the IEEE Seventh International Conference on Biometrics: Theory, Applications and Systems (**BTAS** 2015), September 8-11, 2015, Arlington, VA.

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 (**ICIP** 2015), September 27-30, 2015, Quebec City, Canada.

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 (**ICIP** 2015), September 27-30, 2015, Quebec City, Canada.

A. Puthenputhussery, and C. Liu, "Novel Sparse Kernel Manifold Learner for Image Classification Applications", in Recent Advances in Intelligent Image Search and Video Retrieval, C. Liu, Ed., Springer, pp. 91-114, 2017.

A. Puthenputhussery, S. Chen, J. Lee, L. Spasovic, and C. Liu, "Learning and Recognition Methods for Image Search and Video Retrieval", in Recent Advances in Intelligent Image Search and Video Retrieval, C. Liu, Ed., Springer, pp. 21-43, 2017.

To My Beloved Parents and my Dearest Sister, Brother-in-law, Nephew and Niece.

ACKNOWLEDGMENT

Foremost, I would like to express my heartfelt appreciation to my dissertation advisor, Dr. Chengjun Liu, for his invaluable advice, technical guidance, patience and kindness to bring this dissertation proposal to fruition. Over the past few years, Dr. Liu has been a constant source of encouragement and support which has helped me to present our research achievements in top conferences and reputed journals.

Secondly, I am extremely grateful to Dr. Ali Mili, Dr. Taro Narahara, Dr. Vincent Oria and Dr. Senjuti Basu Roy for serving on my committee. I want to thank them for the time they have spent to provide me with their valuable feedback and suggestions on my research. In addition, I would like to extend special thanks to Dr. Amiya Tripathy from Don Bosco Institute of Technology, Mumbai, India. I would also like to thank all my fellow graduate students Qingfeng Liu, Xin Zhong, Hao Liu and Shaobo Liu for their support and assistance. Specially, I would always be indebted to Qingfeng Liu, who has been a valued mentor during my doctoral study. I must also thank Ms. Angel Butler and Dr. George Olsen in the Computer Science department for providing me with valuable academic advice and helping me with different academic issues.

Most importantly, I heartly appreciate my family for standing by my side for every decision that I have taken in my life. I thank my parents, Mr. Varghese Puthenputhussery, and Mrs. Daisy Puthenputhussery, for their faith in all my ambitions. I am thankful to my sister, Ashly Thomas, and brother-in-law Soby Thomas for their affection and encouragement during the tough years of PhD study. I would also like to thank my nephew Ethan Thomas, and niece Angel Thomas for providing me endless happiness and retaining the child in me.

Finally, I would also like to thank my friends Souvik Sinha, Pragya Sardana, Animesh Dwivedi, and the e-board members of DeepCS and GSA for enriching my graduate life, outside research.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter Page

TABLE OF CONTENTS
(Continued)

Chapter Page

LIST OF TABLES

LIST OF TABLES
(Continued)

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

Content-based image classification applications have expanded greatly due to the advent of an image era of big data resulting in the availability of large number of color images in the internet. With the wide spread availability of digital cameras, cheap data storage and better access of Internet services around the world, millions of color images are created, shared and stored over the Internet. These large number of digital images necessitate the development of automated learning systems that can classify these images into different categories with minimal or no human intervention. Image classification is a challenging topic in the computer vision and machine learning research areas due to the complexity of different visual elements in images and the difficulty to correctly understand the semantics of images. Figure 1.1 shows a general image classification framework which contains three major steps. The first step is feature extraction from the input images to efficiently represent interesting parts of images as a compact feature vector. In some cases, the images may be pre-processed using some image pre-processing techniques to reduce the background noise. The second step is the feature enhancement process so as to ensure that the feature vectors extracted are discriminative to improve the classification performance. Note that only the images in the training set are used to learn the model for feature enhancement. The final step is classification where the enhanced features are used to learn a classifier and the labels are predicted for the images of the test set.

Recently, several machine learning methods such as sparse coding, discrimination analysis have been broadly applied for different image classification applications such as scene and object recognition [3, 27, 71, 123, 39, 58, 87, 65, 66], face recognition [71, 118, 113, 119, 129, 87, 65, 66], human action recognition [32, 58], kinship verification [77, 18, 112, 70, 85, 67, 64], and fine art painting classification [78, 104, 44, 81, 108, 87,

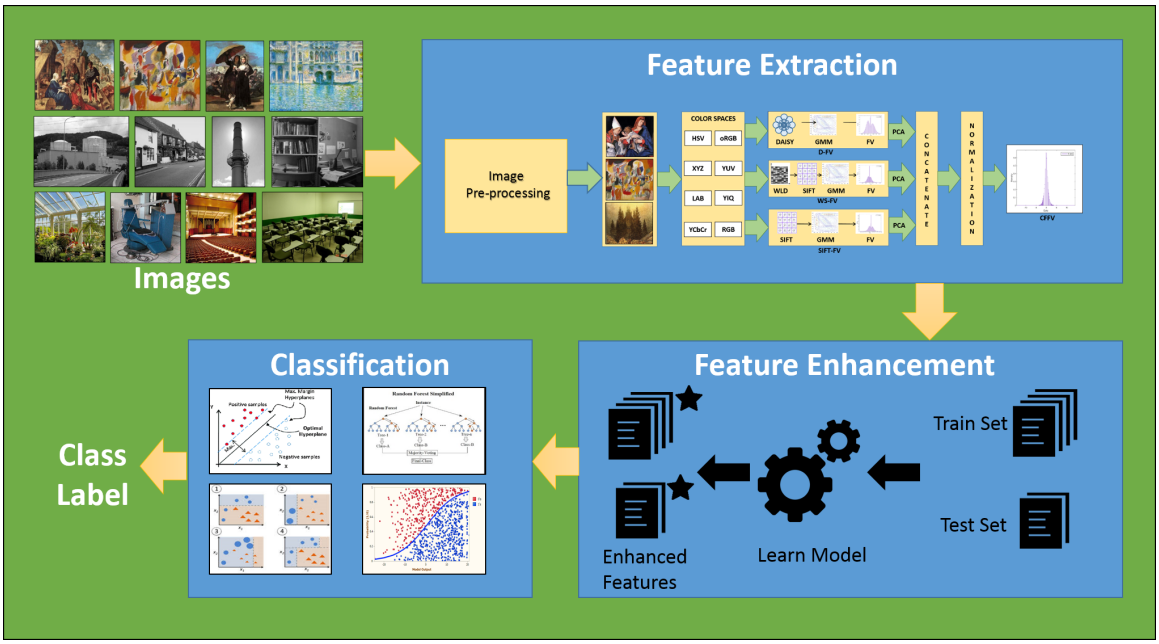

Figure 1.1 A general image classification framework with three major steps: feature extraction, feature enhancement and classification.

65, 66]. Studies in cognitive psychology [80, 107] show that the human visual system is more accurate and robust to find discriminative visual elements in images and a model based on the biological visual cortex is likely to achieve better performance.

This proposal, therefore, focuses on developing image descriptors and learning methods by incorporating cues from the human visual system. Specifically, first, a novel fused color Fisher vector (FCFV) feature is proposed in order to capture different visual information such as color, local, spatial, relative intensity and gradient orientation information. To handle the inconsistencies of different visual classes in images, the FCFV feature is computed by fusing the DAISY Fisher vector (D-FV) feature, Weber-SIFT Fisher vector (WS-FV) feature and the color SIFT Fisher vector features in different color spaces. Second, a sparse representation model using the complete marginal Fisher analysis (CMFA) framework is proposed to capitalize on both the representation aspect of sparse coding methods and the discrimination aspect of the enhanced marginal Fisher analysis method. A potential shortcoming of the MFA method [10] is the principal component

analysis (PCA) [25] step which may discard the null space of the local samples based within class scatter matrix containing important discriminatory information. Our proposed CMFA method extracts the discriminatory features in both the column space of the local samples based within class scatter matrix and the null space of its transformed matrix to enhance the discriminatory power. To further improve the classification capability, a discriminative sparse representation model is learned using the CMFA-SR features by integrating a representation criterion and a discriminative criterion. A variant of the above method is the sparse kernel manifold learner where the discriminative sparse representation model is learned on the FCFV features. Third, a novel discriminative dictionary distribution based sparse coding (DDSC) method is presented that provides new insights and leads to an effective representation and classification framework. Specifically, the proposed DDSC method integrates two new criteria, namely a discriminative criterion and a dictionary distribution criterion into the conventional sparse representation criterion. Finally, a new multiple anthropological Fisher kernel framework (M-AFK) is proposed for kinship verification applications. The genetic inheritable features in kinship relations are enhanced by matching densely sampled SIFT features using the SIFT flow algorithm [60]. An inheritable transformation is further applied to multiple Fisher vector features with the objective to increase the distance between the non-kinship samples and decrease the distance between the kinship samples.

The proposed methods are evaluated on several popular and publicly available image datasets associated with different image classification tasks such as scene and object classification, fine art painting classification, face recognition, kinship verification and fine grained image classification. Experimental results and analysis show the feasibility and effectiveness of the proposed methods.

This proposal is organized in the following manner. Chapter 2 discusses some related work by other researchers on image descriptors, manifold and deep learning methods, sparse coding algorithms, metric learning methods, kinship verification and painting

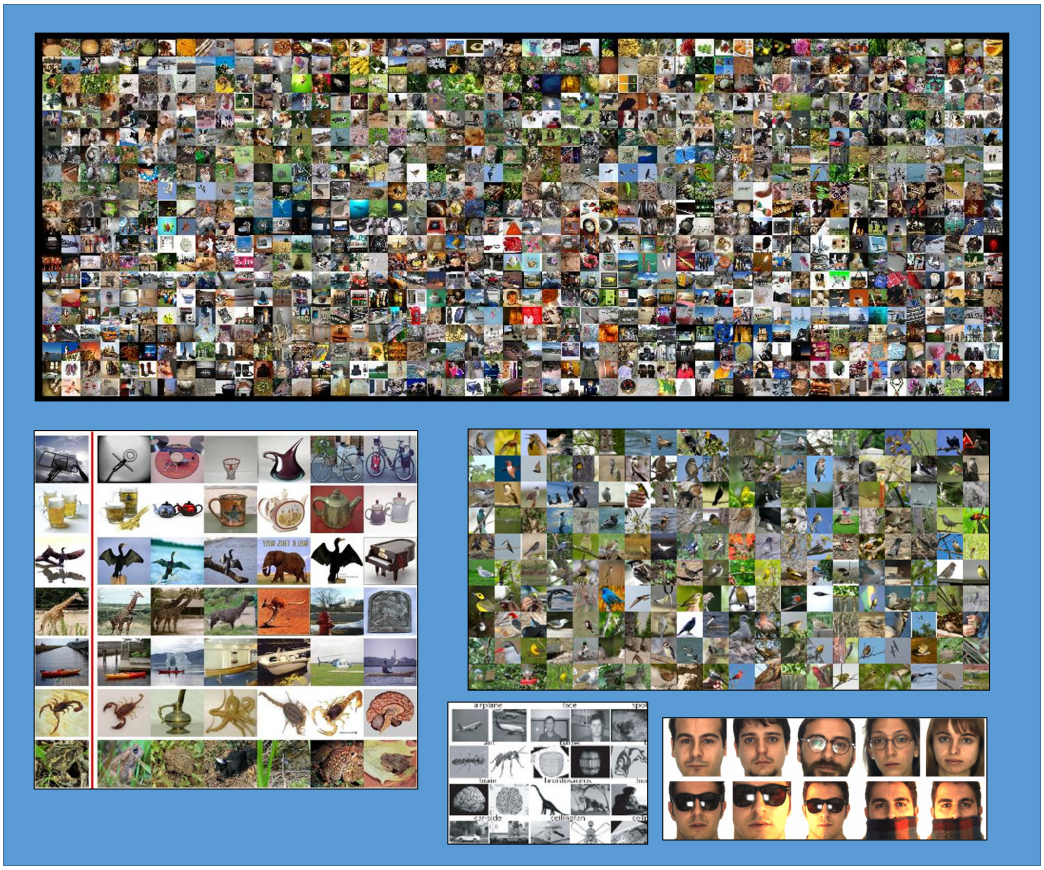

Figure 1.2 Some example images of different image classification datasets.

classification. Chapter 3 explains the FCFV feature and the SKML method for different image classification applications. Chapter 4 discusses CMFA-SR model and the derivation of the largest step size, optimization procedure and the screening rule. Chapter 6 introduces the SF-GFVF feature to enhance and encode the genetic features of parent and child image in kinship relations. Chapters 3, 4 and 6 also include detailed experimental results and analysis performed on various popular and publicly available image datasets. Finally, chapter 7 outlines some proposed research.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Image Descriptors

Local, color, spatial, intensity information, and gradient orientation information are the cues based on which human beings can distinguish between images, and hence they contribute significantly in image classification applications. Van de Weijer et al. [105] showed the effectiveness of color names learned from images for texture classification and action recognition. The work of van de Sande [104] showed that SIFT descriptor incorporated with color information result in a robust local descriptor for classification purposes. Guo et al. [33] proposed the complete LBP descriptor wherein a region in an image is represented by its center pixel and a local difference sign-magnitude transform. Shechtman et al. [95] proposed the self-similarity descriptor which measures similarity of visual entities based on matching internal layout of the image. Bosch et al. [9] introduced a PHOG descriptor that represents local image shape and its spatial layout, together with a spatial pyramid kernel so that the shape correspondence between two images can be measured by the distance between their descriptors using the kernel. The GIST descriptor developed by Oliva et al. [79] is based on a very low dimensional representation of the scene known as spatial envelope that generates a multidimensional space in which scenes sharing membership in semantic categories are projected as close as possible.

2.2 Manifold Learning and Deep Learning Methods

In image classification applications, different manifold learning methods, such as the locality sensitive discriminant analysis (LSDA) [10], the locality preserving projections [35], the marginal Fisher analysis (MFA) [118], have been widely used to preserve data locality in the embedding space. The MFA method based on the graph embedding framework was presented by Yan et al. [118] by designing two graphs that characterize the

intraclass compactness and the interclass separability. Cai et al. [10] proposed the LSDA method that maximizes the margins between data points of different classes by discovering the local manifold structure. A geometric l_p norm feature pooling (GLP) method was proposed by Feng et al. [23] to improve the discriminative power of pooled features by preserving their class-specific geometric information.

Popular deep learning methods, such as the convolutional neural networks (CNN), the deep autoencoders, and the recurrent neural networks, have received increasing attention in the multimedia community for challenging visual recognition tasks. Krizhevsky et al. [47] developed the AlexNet, which was the most notable deep CNN that contains 5 convolution layers followed by max-pooling layers, and 3 fully connected layers. The ZFNet proposed by Zeiler et al. [99] improved upon the AlexNet architecture by using smaller filter sizes, and developed a method to visualize the filters and weights correctly. He et al. [34] developed residual networks with a depth of upto 152 layers that contain skip connections and inter-block activation for better signal propagation between the layers.

2.3 Metric Learning

Metric learning methods have gained a lot of attention for computer vision and machine learning applications. In metric learning, an optimization objective function is developed from training images to learn the distance metric. Different metric learning methods have different objective functions designed for their specific purpose. Some representative metric learning methods include principal component analysis (PCA) [56], Laplacian eigen maps (LE) [7], linear discriminant analysis (LDA) [17], large margin nearest neighbor (LMNN) [112], information theoretic metric learning (ITML) [17] and cosine similarity metric learning (CSML) [77]. Some data samples in the training data provide more information towards learning the metric, therefore higher priority should be given to these data samples. But most existing metric learning methods do not differentiate important data samples and treat all the samples equally leading to reduction in accuracy. In [70],

Lu et al. proposed the 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. While NRML has achieved good performance in kinship verification, there are still some shortcomings. First, the NRML method derives the features and the metric learning independently, therefore theoretical relation cannot be established between them. Second, the objective function can face the issue of dominance of one term in the function over the other terms leading to inaccurate results.

2.4 Sparse Coding

Several sparse representation methods based on supervised learning methods have been developed for learning efficient sparse representations or incorporating discriminatory information by combining multiple class specific dictionary for different visual recognition applications. In particular, the sparse representation methods can be roughly categorized into three categories. The first category of sparse representation methods aims to learn a space efficient dictionary by fusing multiple atoms from the initial large dictionary. Fulkerson et al. [26] proposed an object localization framework that efficiently reduces the size of a large dictionary by constructiong small dictionaries based on the agglomerative information bottleneck. The work of Lazebnik et al. [49] present a technique for learning dictionaries by using the information-theoretic properties of sufficient statistics. Jiang et al. [41] presented an efficient greedy based optimization approach for modeling the discriminative dictionary learning by maximizing the monotonically increasing and submodular properties of a graph topology selection problem. Qiu et al. [89] developed an approach for dictionary learning of action attributes by integrating the mutual information for appearance information and class distributions between the learned dictionary and the rest of the dictionary space in the objective function.

The second category combines multiple class specific sub-dictionaries to improve the discriminatory power of the sparse representation method. Yang et al. [123] proposed a

Fisher discrimination discriminatory learning framework to learn a structured dictionary where each sub-dictionary has specific class labels. Mairal et al. [73] proposed a sparse representation based framework by jointly optimizing both the sparse reconstruction and class discrimination components for learning multiple dictionaries. Zhou et al. [132] presented a joint dictionary learning algorithm that jointly learns multiple class-specific dictionaries and a common shared dictionary by exploiting the visual correlation within a group of visually similar objects. A dictionary learning approach for positive definite matrices was proposed by Sivalingam et al. [101], where the dictionary is learned by alternating minimization of sparse coding and dictionary update stages.

The final category of sparse representation methods co-trains the sparse representation and discriminative dictionary by adding a discriminant term to the objective function. Yang et al. [121] proposed supervised hierarchical sparse coding models where the dictionary is learned via back-projection where implicit differentiation is used to relate the sparse codes to the dictionary. Jiang et al. [40] presented a label consistent K-SVD algorithm where a label consistency constraint and a classification performance criteria are integrated to the objective function to learn a reconstructive and discriminative dictionary. Zhang et al. [129] developed a discriminative K-SVD algorithm to learn an over-complete dictionary by directly incorporating labels in the dictionary learning stage.

To increase the computational efficiency of the sparse representation methods and to improve the scalability to large datasets, screening rules are receiving increasing attention by researchers. Wang et al. [111] proposed a sparse logistic regression screening rule to identify the zero components in the solution vector to effectively discard features for the l_1 regularized logistic regression. Xiang et al. [117, 116] presented a dictionary screening rule to select a subset of codewords to use in Lasso optimization and derived fast Lasso screening tests to find which data points and codewords are highly correlated. A new set of screening rules for the Lasso problem were developed by Wang et al. [109] that uses

non-expansiveness of the projection operator to effectively identify inactive predictors of the Lasso problem.

2.5 Kinship Verification

Facial images convey important characteristics such as identity information, kinship information, facial expressions, gender of a person, ethnicity, emotional information, mental state of a person and so on. Among these many characteristics, kinship is believed to be one of the most dominant one since children naturally inherit genetic features from their parents [115]. Subsequent studies in social sciences have confirmed that children resemble their parent more than other people and they may resemble a particular parent more at different ages [4]. Deghan et al. [18] proposed an algorithm that fuses the features and metrics using a gated autoencoders and a discriminative neural network. The hybrid framework learns the genetic features in parent-offspring relationships to improve the kinship verification performance. A neighborhood repulsed metric learning (NRML) method was proposed by Lu et al. [70] in which the distance of interclass samples are pushed as far as possible and the distance between intraclass samples within a kinship relation are pulled as close as possible for better verification accuracy. Lan et al. [48] proposed a quaternionic Weber local descriptor (QWLD) framework which uses quaternionic representation to handle all color channels of the image in a holistic way while preserving their relations, and applies Weber's law to ensure that the derived descriptors are robust and discriminative. A hierarchial learning representation was presented by Kohli et al. [46] that develops a compact feature representation by encoding relational information present in images using filters and contractive regularization penalty. The mixed bi-subject kinship verification problem is solved using a multi-view multi-task learning proposed by Qin et. al [88] where the transformation matrices for all the relations jointly learned as well as for a single relation is fused to improve the kinship verification performance. Zhou et al. [133] proposed a multiview scalable similarity learning (SSL) method by fusing

the diagonal similarity models from multiple feature representations in a coherent online process to leverage the interactions and correlations in multiview kin data.

2.6 Computational Fine Art Painting Categorization

Recently, several research efforts have been invested for painting classification using computer vision techniques. Shamir et al. [93] described a method for automated recognition of painters and schools of art based on their signature styles. Sablating et al. [92] examined the structural signature of a painting based on the brush strokes in potrait miniatures. The work of Zujovic et al. [134] described an approach to automatically classify digital pictures of paintings by using the salient aspects of a painting such as color, texture and edges. Shamir and Tarakhovsky [94] showed that automatic computer analysis can group artists by their artistic movements, and provide a map of similarities and influential links that is largely in agreement with the analysis of art historians. Siddique et al. [97] presented an efficient approach for learning a mixture of kernels by greedily selecting exemplar data instances corresponding to each kernel using AdaBoost for painting dataset classification. A multiple visual feature based framework was proposed by Shen [96] for automatic classification of western painting image collection. The work of Culjak et al. [16] offered an approach to automatically classify paintings into their genres by extracting features based on color and texture of the painting.

CHAPTER 3

SPARSE KERNEL MANIFOLD LEARNER FOR IMAGE CLASSIFICATION

3.1 Introduction

The human visual system is much more efficient and robust in classifying different visual elements in an image, therefore any image classification system based on the human visual system is likely to achieve good performance for classification tasks. Different visual aspects in an image such as color, edges, shape, intensity and orientation of objects help humans to identify and discriminate between images. Pioneer works in cognitive psychology believe that the human visual cortex represent images as sparse structures as it provides an efficient representation for later stages of visual processing [80, 107]. A sparse representation of a data-point can be represented as a linear combination of a small set of basis vectors allowing efficient storage and retrieval of data. Another advantage of sparse representation is that it adapts to varying level of information in the image since it provides a distributed representation of an image. Therefore, we introduce a hybrid feature extraction method to capture different kinds of information from the image and propose a discriminative sparse coding method based on manifold learning algorithm to learn an efficient and robust discriminative sparse representation of the image.

In this chapter, we first present novel DAISY Fisher vector (D-FV) and Weber-SIFT Fisher vector (WS-FV) features in order to handle the inconsistencies and variations of different visual classes in images. In particular, the D-FV feature enhances the Fisher vector feature by fitting dense DAISY descriptors [103] to a parametric generative model. We then develop the WS-FV by integrating Weber local descriptors [13] with SIFT descriptors and Fisher vectors are computed on the sampled WLD-SIFT features. An innovative fused Fisher vector (FFV) is proposed by fusing the principal components of D-FV, WS-FV and SIFT-FV (S-FV) features. We then assess our FFV feature in eight different color

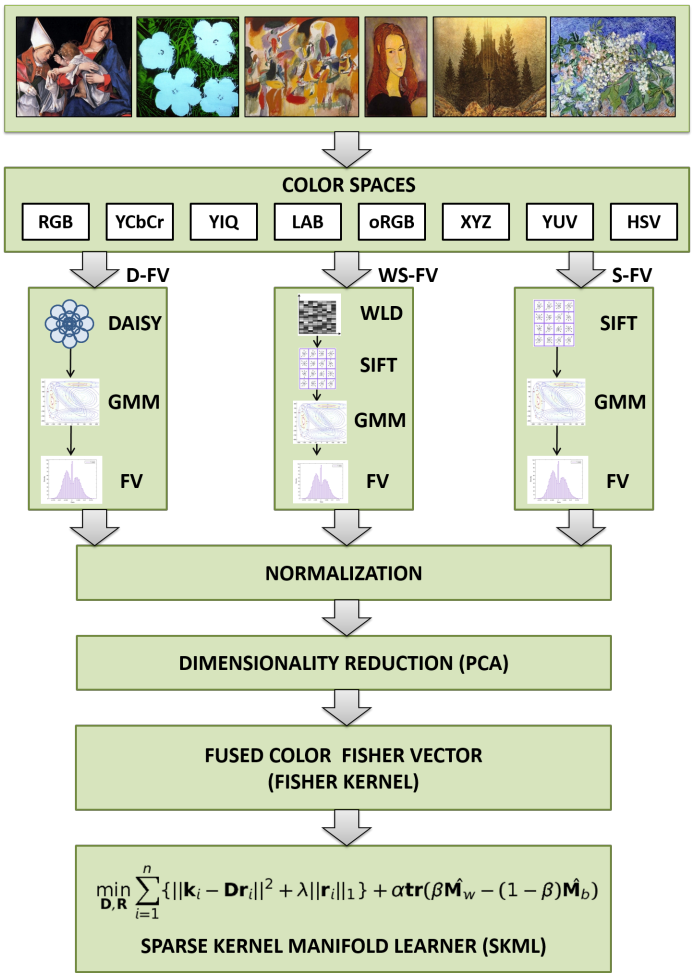

Figure 3.1 The framework of our proposed SKML method.

spaces and propose several color FFV features. The descriptors that are defined in different color spaces provide stability against image variations such as rotation, viewpoint, clutter and occlusions [100] which are essential for classification of images. We further extend this concept by integrating the FFV features in eight different color spaces to form a novel fused color Fisher vector (FCFV) feature. Finally, we use a sparse kernel manifold learner (SKML) method to learn a discriminative sparse representation by integrating the discriminative marginal Fisher analysis criterion to the sparse representation criterion. In particular, new intraclass compactness and interclass separability are define based on the sparse representation criterion under the manifold learning framework. The objective of the SKML method is to increase the interclass distance between data-points belonging to

different classes and decrease the intraclass distance between data-points of the same class. The SKML method can efficiently calculate a global shared dictionary without the need for computation of sub-dictionaries and hence is suitable for large datasets. The framework of our proposed SKML method is illustrated in Figure 3.1. Experimental results show that the proposed approach achieves better results compared to other popular image descriptors and state-of-the-art deep learning methods on different image classification datasets.

The rest of this chapter is organized in the following manner. Section 3.2 describes the details of the computation of different Fisher vector features and the SKML method. We present an extensive experimental evaluation and analysis of the proposed SKML method for different classification datasets in Sections 3.3 and 3.4 concludes the paper.

3.2 Novel Sparse Kernel Manifold Learner Framework

3.2.1 Fisher Vector

We briefly review the Fisher vector which is widely applied for visual recognition problems such as face detection and recognition [98], object recognition [38], etc. Fisher vector describes an image by what makes it different from other images [38] and focuses only on the image specific features. Particularly, let $\mathbf{X} = \{\mathbf{d}_t, t = 1, 2, ..., T\}$ be the set of T local descriptors extracted from the image. Let μ_λ be the probability density function of $\mathbf{X}$ with parameter λ, then the Fisher kernel [38] is defined as follows:

$$K(\mathbf{X}, \mathbf{Y}) = (\mathbf{G}_\lambda^X)^T \mathbf{F}_\lambda^{-1} \mathbf{G}_\lambda^Y \tag{3.1}$$

where $\mathbf{G}_\lambda^X = \frac{1}{T} \bigtriangledown_\lambda \log_{\mu_\lambda}(\mathbf{X})$, which is the gradient vector of the log-likelihood that describes the contribution of the parameters to the generation process. And $\mathbf{F}_\lambda$ is the Fisher information matrix of μ_λ.

Since $\mathbf{F}_\lambda^{-1}$ is symmetric and positive definite, it has a Cholesky decomposition as $\mathbf{F}_\lambda^{-1} = \mathbf{L}_\lambda^T \mathbf{L}_\lambda$. Therefore, the kernel $K(\mathbf{X}, \mathbf{Y})$ can be written as a dot product between normalized vectors $\mathbf{G}_\lambda$, obtained as $\mathbf{G}_\lambda^X = \mathbf{L}_\lambda \mathbf{G}_\lambda^X$ where $\mathbf{G}_\lambda^X$ is the Fisher vector of $\mathbf{X}$.

3.2.2 DAISY Fisher Vector (D-FV)

In this section, we present a new innovative DAISY Fisher vector (D-FV) feature where Fisher vectors are computed on densely sampled DAISY descriptors. DAISY descriptors are suitable for dense computation and offers precise localization and rotational robustness [103], therefore provides improved performance and better accuracy for classification. The DAISY descriptor [103] $\mathcal{D}(u_0, v_0)$ for location (u_0, v_0) is defined as follows:

$$\begin{aligned}\mathcal{D}(u_0, v_0) = [&\tilde{\mathbf{h}}^T_{\Sigma_1}(u_0, v_0),\\ &\tilde{\mathbf{h}}^T_{\Sigma_1}(\mathbf{I}_1(u_0, v_0, R_1)), ..., \tilde{\mathbf{h}}^T_{\Sigma_1}(\mathbf{I}_T(u_0, v_0, R_1)), ...,\\ &\tilde{\mathbf{h}}^T_{\Sigma_Q}(\mathbf{I}_1(u_0, v_0, R_Q)), ..., \tilde{\mathbf{h}}^T_{\Sigma_Q}(\mathbf{I}_T(u_0, v_0, R_Q))]^T\end{aligned} \tag{3.2}$$

where $\mathbf{I}_j(u, v, R)$ is the location with distance R from (u, v) in the direction given by j, Q represents the number of circular layers and $\tilde{\mathbf{h}}_\Sigma(u, v)$ is the unit norm of vector containing Σ-convolved orientation maps in different directions. The sampled descriptors are fitted to a Gaussian Mixture Model (GMM) with 256 parameters. The Fisher vectors are then encoded as derivatives of log-likelihood of the model.

3.2.3 Weber-SIFT Fisher Vector (WS-FV)

In this section, we propose a new Weber-SIFT Fisher vector (WS-FV) feature that computes the Fisher vector on Weber local descriptor (WLD) integrated with SIFT features so as to encode the color, local, relative intensity and gradient orientation information from an image. The WLD [13] is based on the Weber's law which states that the ratio of increment threshold to the background intensity is a constant. The descriptor contains two components differential excitation [13] and orientation [13] which are defined as follows.

$$\xi(x_c) = \arctan[\frac{\nu_s^{00}}{\nu_s^{01}}] \; and \; \theta(x_c) = \arctan(\frac{\nu_s^{11}}{\nu_s^{10}}) \tag{3.3}$$

where $\xi(x_c)$ is the differential excitation and $\theta(x_c)$ is the orientation of the current pixel x_c, $x_i(i = 0, 1, ...p-1)$ denotes the i-th neighbours of x_c and p is the number of neighbors,

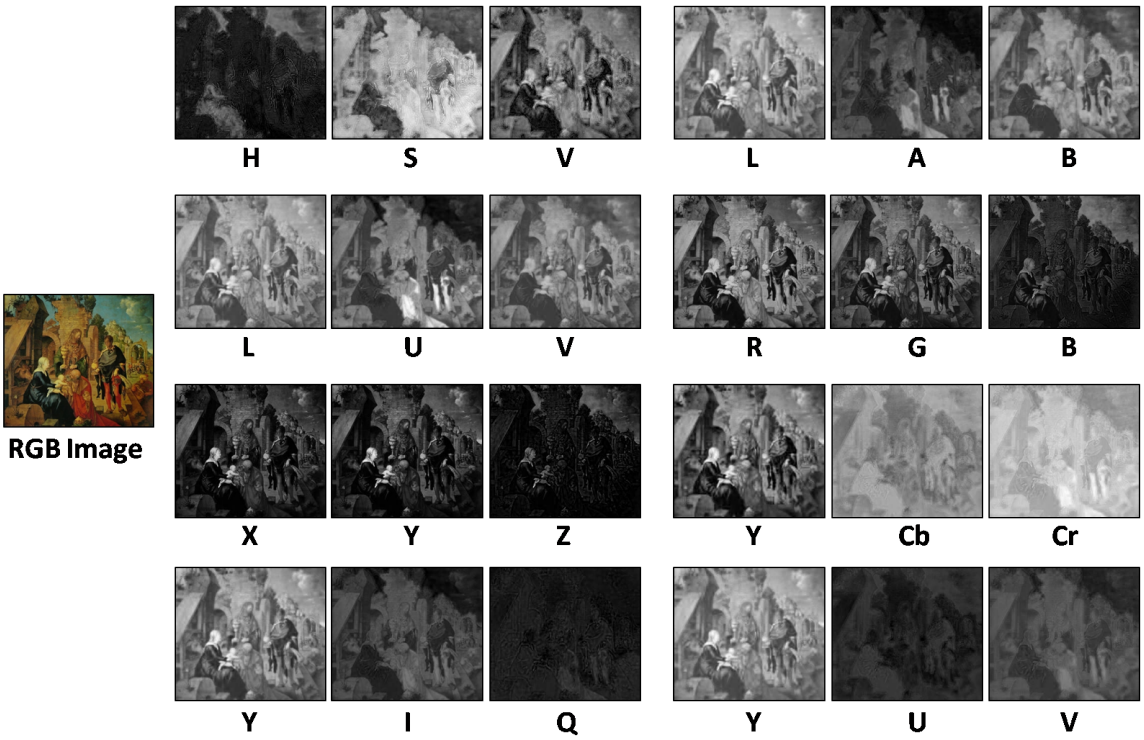

Figure 3.2 The color component images of a sample image from the Painting-91 dataset in different colorspaces.

ν_s^{00}, ν_s^{01}, ν_s^{10} and ν_s^{11} are the output of filters f_{00}, f_{01}, f_{10} and f_{11}, respectively. The WLD descriptor extracts the relative intensity and gradient information similar to humans perceiving the environment, therefore provides stability against noise and illumination changes. A parametric generative model is trained by fitting to the WLD-SIFT features and Fisher vectors are extracted by capturing the average first order and second order differences between the computed features and each of the GMM centers.

3.2.4 Fused Color Fisher Vector (FCFV)

In this section, we first present an innovative fused Fisher vector (FFV) feature that fuses the most expressive features of the D-FV, WS-FV and SIFT-FV features. The most expressive features are extracted by means of Principal Component Analysis (PCA) [25]. Particularly, let $\mathbf{X} \in \mathbb{R}^N$ be a feature vector with covariance matrix $\mathbf{\Sigma}$ given as follows:$\mathbf{\Sigma} = \mathbb{E}[(\mathbf{X} - \mathbb{E}(\mathbf{X}))][(\mathbf{X} - \mathbb{E}(\mathbf{X}))]^T$ where T represents transpose operation and $\mathbb{E}(.)$ represents

expectation. The covariance matrix can be factorized as follows [25]:$\mathbf{\Sigma} = \boldsymbol{\phi}\mathbf{\Lambda}\boldsymbol{\phi}^T$ where $\mathbf{\Lambda} = diag[\lambda_1, \lambda_2, \lambda_3, ..., \lambda_N]$ is the diagonal eigenvalue matrix and $\boldsymbol{\phi} = [\boldsymbol{\phi_1}\boldsymbol{\phi_2}\boldsymbol{\phi_3}....\boldsymbol{\phi_N}]$ is the orthogonal eigenvector matrix. The most expressive features of $\mathbf{X}$ is given by a new vector $\mathbf{Z} \in \mathbb{R}^K : \mathbf{Z} = \mathbf{P}^T\mathbf{X}$ where $\mathbf{P} = [\boldsymbol{\phi_1}\boldsymbol{\phi_2}\boldsymbol{\phi_3}....\boldsymbol{\phi_K}]$ and $K < N$.

We incorporate color information to our proposed feature as the color cue provides powerful discriminating information in pattern recognition and can be very effective for face, object, scene and texture classification [100, 57]. The descriptors defined in different color spaces provide stability against illumination, clutter, viewpoint and occlusions [100]. To derive the proposed FCFV feature, we first compute the D-FV, WS-FV and SIFT-FV in the eight different color spaces namely RGB, YCbCr, YIQ, LAB, oRGB, XYZ, YUV and HSV. Figure 3.2 shows the component images of a sample image from the Painting-91 dataset in different color spaces used in this paper. For each color space, we derive the FFV by fusing the most expressive features of D-FV, WS-FV and SIFT-FV for that color space. We then reduce the dimensionality of the eight FFV features using PCA, which derives the most expressive features with respect to the minimum square error. We finally concatenate the eight FFV features and normalize to zero mean and unit standard deviation to create the novel FCFV feature.

3.2.5 Sparse Kernel Manifold Learner (SKML)

In this section, we present a sparse kernel manifold learner (SKML) to learn a compact discriminative representation by considering the local manifold structure and the label information. In particular, new within class scatter and between class scatter matrices are defined constrained by the marginal Fisher criterion [118] and the sparse criterion so as to increase the interclass separability and reduce the intraclass compactness based on a manifold learning framework. A discriminative term is then integrated to the representation criterion of the sparse model so as to improve the pattern recognition performance.

The features used as input for the SKML method are the FCFV features extracted from the image. Given the Fisher kernel matrix $\mathbf{K} = [\mathbf{k}_1, \mathbf{k}_2, ..., \mathbf{k}_n] \in \mathbb{R}^{m \times n}$, which contains n samples in a m dimensional space, let $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_b] \in \mathbb{R}^{m \times b}$ denote the dictionary that represents b basis vectors and $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_n] \in \mathbb{R}^{b \times n}$ denote the sparse representation matrix which represents the sparse representation for m samples. The coefficient $\mathbf{r}_i$ in the sparse representation $\mathbf{R}$ correspond to the items in the dictionary $\mathbf{D}$.

In the proposed SKML method, we jointly optimize the sparse representation criterion and the marginal Fisher analysis criterion to derive the dictionary $\mathbf{D}$ and sparse representation $\mathbf{S}$ from the training samples. The objective of the marginal Fisher analysis criterion is to minimize the intraclass compactness and maximize the interclass separability. We define new discriminative intraclass compactness $\hat{\mathbf{M}}_w$ based on the sparse criterion as follows:

$$\hat{\mathbf{M}}_w = \sum_{i=1}^{n} \sum_{(i,j) \in N_k^w(i,j)} (\mathbf{r}_i - \mathbf{r}_j)(\mathbf{r}_i - \mathbf{r}_j)^T \tag{3.4}$$

where $(i, j) \in N_k^w(i, j)$ represents the (i, j) pairs where sample $\mathbf{k}_i$ is among the nearest neighbors of sample $\mathbf{k}_j$ of the same class or vice versa.

And the discriminative interclass separability $\hat{\mathbf{M}}_b$ is defined as:

$$\hat{\mathbf{M}}_b = \sum_{i=1}^{m} \sum_{(i,j) \in N_k^b(i,j)} (\mathbf{r}_i - \mathbf{r}_j)(\mathbf{r}_i - \mathbf{r}_j)^T \tag{3.5}$$

where $(i, j) \in N_k^b(i, j)$ represents nearest (i, j) pairs among all the (i, j) pairs between samples $\mathbf{k}_i$ and $\mathbf{k}_j$ of different classes.

Therefore, we define the modified optimization criterion as:

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{R}} \sum_{i=1}^{n} \{||\mathbf{k}_i - \mathbf{D}\mathbf{r}_i||^2 + \lambda ||\mathbf{r}_i||_1\} + \alpha \mathbf{tr}(\beta \hat{\mathbf{M}}_w - (1-\beta)\hat{\mathbf{M}}_b) \\ &s.t. ||\mathbf{d}_j|| \leq 1, (j = 1, 2, ..., b) \end{aligned} \tag{3.6}$$

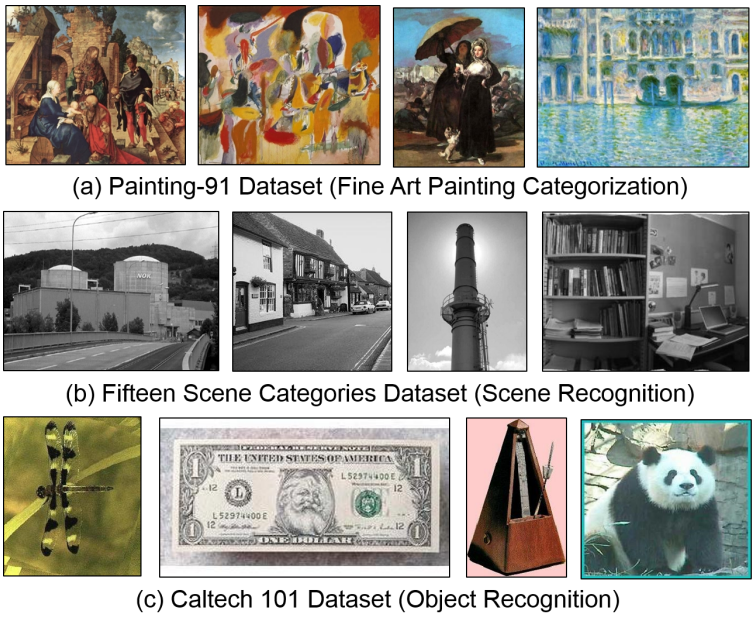

Figure 3.3 Some sample images from different visual recognition datasets used for evaluation of the proposed SKML method.

where **tr(.)** denotes the trace of a matrix, the parameter λ controls the sparsity term, the parameter α controls the discriminatory term, the parameter β balances the contributions of the discriminative intraclass compactness $\hat{\mathbf{M}}_w$ and interclass separability $\hat{\mathbf{M}}_b$.

Let $\mathbf{L} = \mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_t$ are the test data matrix and t be the number of test samples, then as the dictionary $\mathbf{D}$ is already learned, the discriminative sparse representation for the test data can be derived by optimizing the following criterion:

$$\min_{S} \sum_{i=1}^{t} \{||\mathbf{l}_i - \mathbf{D}\mathbf{s}_i||^2\} + \lambda||\mathbf{s}_i||_1 \tag{3.7}$$

The discriminative sparse representation for the test data is defined as $\mathbf{S} = [\mathbf{s}_1, ..., \mathbf{s}_t] \in \mathbb{R}^{b \times t}$ and has both the sparseness and discriminative information since we learn the dictionary from the the training process.

3.3 Experiments

We assess the performance of our proposed SKML method on three different image classifiction datasets namely the Painting-91 dataset [44], the CalTech 101 dataset [50] and the 15 Scenes dataset [53]. Figure 3.3 shows some sample images from different visual recognition datasets used for evaluation.

3.3.1 Painting-91 Dataset

This section assesses the effectiveness of our proposed features on the challenging Painting-91 dataset [44]. The dataset contains 4266 fine art painting images by 91 artists. The images are collected from the Internet and each artist has variable number of images ranging from 31 (Frida Kahlo) to 56 (Sandro Boticelli). The dataset classifies 50 painters to 13 styles with style labels as follows: abstract expressionism (1), baroque (2), constructivism (3), cubbism (4), impressionism (5), neoclassical (6), popart (7), post-impressionism (8), realism (9), renaissance (10), romanticism (11), surrealism (12) and symbolism (13).

Art painting categorization is a challenging task as the variations in subject matter, appearance, theme and styles are large in the art paintings of the same artists. Another issue is that the similarity gap between paintings of the same styles is very small due to common influence or origin. In order to effectively classify art paintings, key aspects such as texture form, brush stroke movement, color, sharpness of edges, color balance, contrast, proportion, pattern, etc. have to be captured [91]. Painting art images are different from photographic images due to the following reasons: (i) Texture, shape and color patterns of different visual classes in art images (say, a multicolored face or a disproportionate figure) are inconsistent with regular photographic images. (ii) Some artists have a very distinctive style of using specific colors (for ex: dark shades, light shades etc.) and brush strokes resulting in art images with diverse background and visual elements. The proposed SKML framework uses FCFV features which captures different kinds of information from the

painting image and the SKML method aims to improve the discrimination between classes essential for computational fine art painting categorization.

Table 3.1 Comparison of the Proposed SKML Feature with Popular Image Descriptors for Artist and Style Classification Task of the Painting-91 Dataset

No.	Feature	Artist CLs	Style CLs
1	LBP [78, 44]	28.5	42.2
2	Color-LBP [44]	35.0	47.0
3	PHOG [9, 44]	18.6	29.5
4	Color-PHOG [44]	22.8	33.2
5	GIST [44]	23.9	31.3
6	Color-GIST [44]	27.8	36.5
7	SIFT [68, 44]	42.6	53.2
8	CLBP [33, 44]	34.7	46.4
9	CN [105, 44]	18.1	33.3
10	SSIM [44]	23.7	37.5
11	OPPSIFT [104, 44]	39.5	52.2
12	RGBSIFT [104, 44]	40.3	47.4
13	CSIFT [104, 44]	36.4	48.6
14	CN-SIFT [44]	44.1	56.7
15	Combine(1 - 14) [44]	53.1	62.2
16	**SKML**	**63.09**	**71.67**

Performance in Different Color Spaces This section demonstrates the performance of our proposed SKML feature in eight different color spaces namely RGB, YCbCr, YIQ, LAB, oRGB, XYZ, YUV and HSV as shown in Table 3.3. Among the single color descriptors, the YIQ-FFV feature performs the best with classification accuracy of 59.22%

Table 3.2 Comparison of the Proposed SKML Feature with State-of-the-art Deep Learning Methods for Artist and Style Classification Task of the Painting-91 Dataset

No.	Feature	Artist CLs	Style CLs
1	MSCNN-0 [81]	55.15	67.37
2	MSCNN-1 [81]	58.11	69.69
3	MSCNN-2 [81]	57.91	70.96
4	MSCNN-3 [81]	-	67.74
5	CNN F_1 [108]	55.40	68.20
6	CNN F_2 [108]	56.25	68.29
7	CNN F_3 [108]	56.40	68.57
8	CNN F_4 [108]	56.35	69.21
9	CNN F_5 [108]	56.35	69.21
10	**SKML**	**63.09**	**71.67**

for the artist classification task whereas the RGB-FFV feature gives the best performance of 66.43% for the style classification task. The SKML feature is computed by using a sparse representation model on the fusion of the FFV features in eight different color spaces and it achieves the best performance in both artist and style classification re-emphasizing the fact that adding color information is particularly suitable for classification of art images.

Artist and Style Classification This section evaluates the performance of our proposed method on the task of artist and style classification. The artist classification is a task wherein a painting image has to be classified to its respective artist whereas the style classification task is to assign a style label to the painting image. The artist classification task contains 91 artists with 2275 train and 1991 test images. Similarly, the style classification task contains 13 style categories with 1250 train and 1088 test images. Table 3.1 shows the comparison of the proposed SKML feature with other state-of-the-art image

Table 3.3 Classification Performance of the FFV Feature in Different Color Spaces on the Painting-91 Dataset

Feature	Artist CLs	Style CLs
RGB-FFV	59.04	66.43
YCbCr-FFV	58.41	65.82
YIQ-FFV	59.22	66.26
LAB-FFV	49.30	59.98
oRGB-FFV	57.50	65.46
XYZ-FFV	56.41	64.32
YUV-FFV	57.70	64.25
HSV-FFV	51.43	60.58
SKML	**63.09**	**71.67**

descriptors. The color LBP descriptor [44] is calculated by fusing the LBP descriptors computed on the R,G and B channels of the image. Similar strategy is used to compute the color versions of PHOG and GIST descriptor. The opponent SIFT [104] for the painting image is computed by first converting the image to the opponent color space and then fusing the SIFT descriptors calculated for every color channel. The SSIM (self similarity) descriptor [44] is computed using a correlation map to estimate the image layout. The combination of all image descriptors listed in Table 3.1 gives a classification accuracy of 53.1% and 62.2% for the artist and style classification tasks, respectively. Experimental results show that our proposed SKML feature significantly outperforms popular image descriptors and their fusion, and achieves the classification performance of 63.09% and 71.67% for artist and style classification, respectively.

Table 3.2 shows the performance of the proposed SKML features compared with state-of-the-art deep learning methods. MSCNN [81] stands for multi-scale convolutional

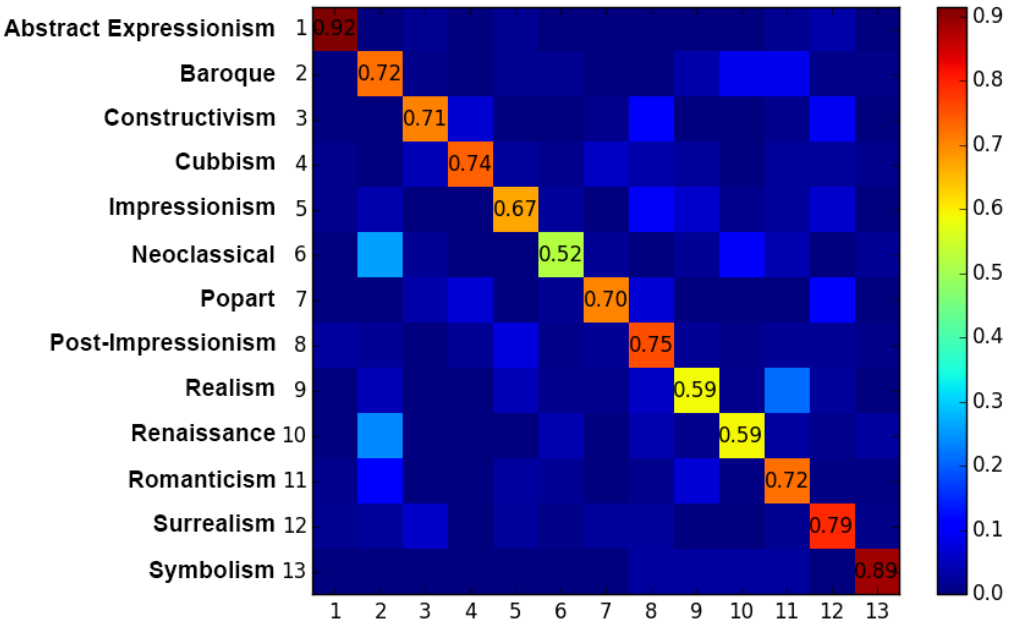

Figure 3.4 The confusion matrix for 13 style categories of the Painting-91 dataset using the SKML feature.

neural network which extracts features in different scales using multiple CNNs. The cross layer convolutional neural network (CNN F) [108] computes features from multiple layers of CNN to improve discriminative ability instead of only extracting from the top-most layer. The best performing CNN for the artist classification and style classification is MSCNN-1 [81] and MSCNN-2 [81], respectively. Our proposed SKML method achieves better result compared to state-of-the-art deep learning methods such as multi scale CNN and cross layer CNN.

Figure 3.4 shows the confusion matrix for the 13 style categories using the SKML feature where the rows denote the actual classes while the columns denote the assigned classes. It can be observed that the best classified categories are 1 (abstract expressionism) and 13 (symbolism) with classification rates of 92% and 89%, respectively. The most difficult category to classify is category 6 (neoclassical) as there are large confusions between the styles baroque and neoclassical. Similarly, the other categories that create confusion are the styles baroque and renaissance.

Table 3.4 Art Movement Associated with Different Art Styles

Art Movement	Art Style
Renaissance	renaissance
Post Renaissance	baroque, neoclassical, romanticism, realism
Modern Art	popart, impressionism, post impressionism, surrealism, cubbism, symbolism, constructivism, abstract expressionism

Comprehensive Analysis of Results Table 3.4 shows the art movements associated with different art styles. Interesting patterns can be observed from the confusion diagram in Figure 4.2. The art styles within an art movement show higher confusions compared to the art styles between the art movement periods. An art movement is a specific period of time wherein an artist or group of artists follow a specific common philosophy or goal. It can be seen that there are large confusions for the styles baroque and neoclassical. Similarly, the style categories romanticism and realism have confusions with style baroque. The style categories baroque, neoclassical, romanticism and realism belong to the same art movement period - post renaissance. Similarly, popart paintings have confusions with style category surrealism within the same art movement but none of the popart paintings are misclassified as baroque or neoclassical. The only exception to the above observation is the style categories renaissance and baroque as even though they belong to different art movement period, there are large confusions between them. The renaissance and baroque art paintings have high similarity as the baroque style evolved from the renaissance style resulting in few discriminating aspects between them [91].

Artist Influence In this section, we analyze the influence an artist can have over other artists. We find the influence among artists by looking at similar characteristics

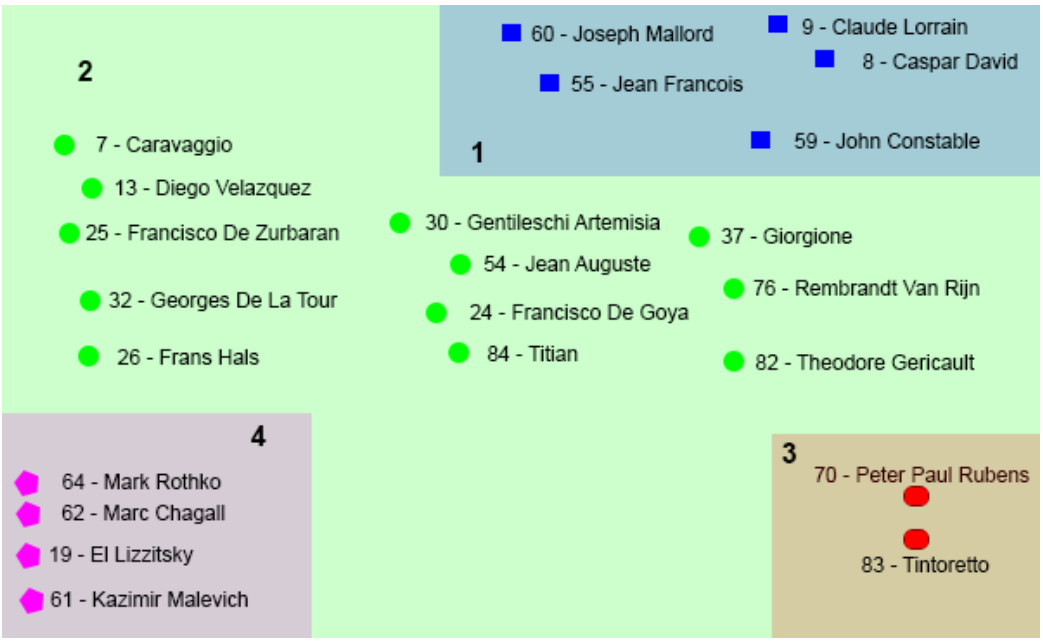

Figure 3.5 The artist influence cluster graph for the Painting-91 dataset.

between the artist paintings. Artist influence may help us to find new connections among artists during different art movement period and also understand the influence among different art movement periods. In order to calculate the artist influence, we calculate the correlation score between the paintings of different artists. Let $\mathbf{a}_{ik}$ denote the feature vector representing the painting by artist k where $i = 1, .., n_k$ and let n_k be the total number of paintings by artist k. We calculate $\mathbf{A}_k$ which is the average of the feature vector of all paintings by artist k. We then compute a correlation matrix by comparing the average feature vector of each artist with all other artists. Finally, clusters are defined for artists with high correlation score. Figure 3.5 show the artist influence cluster graph with correlation threshold of 0.70.

Interesting observations can be deduced from Figure 3.5. Every cluster can be associated with a particular style and time period. Cluster 1 shows artists with major contributions to the styles realism and romanticism and they belong to the post renaissance art movement period. Cluster 2 has the largest number of artists associated with the styles renaissance and baroque. Cluster 3 represents artists for the style Italian renaissance that

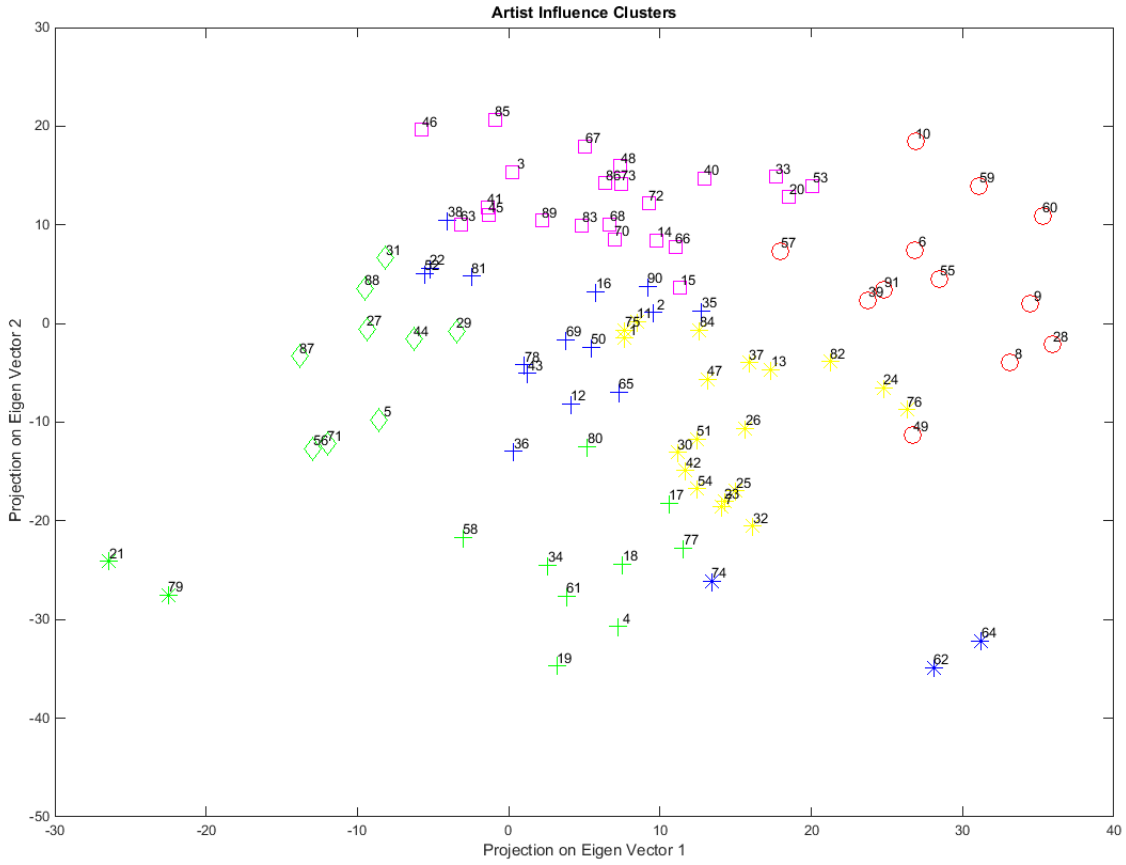

Figure 3.6 The artist influence cluster graph using k means clustering for the Painting-91 dataset.

took place in the 16^{th} century. And cluster 4 shows artists associated with style abstract expressionism in the modern art movement period (late 18^{th} - 19^{th} century).

We further show the k-means clustering graph with cosine distance to form clusters of similar artists. Figure 3.6 shows the artist influence graph clusters for paintings of all artists with k set as 8. First, the average of the feature vector of all paintings of an artist is calculated as described above. We then apply k-means clustering algorithm with k set as 8. The artist influence graph is plotted using the first two principal components of the average feature vector. The results of Figure 3.5 have high correlation with the results of the artist influence cluster graph in Figure 3.6.

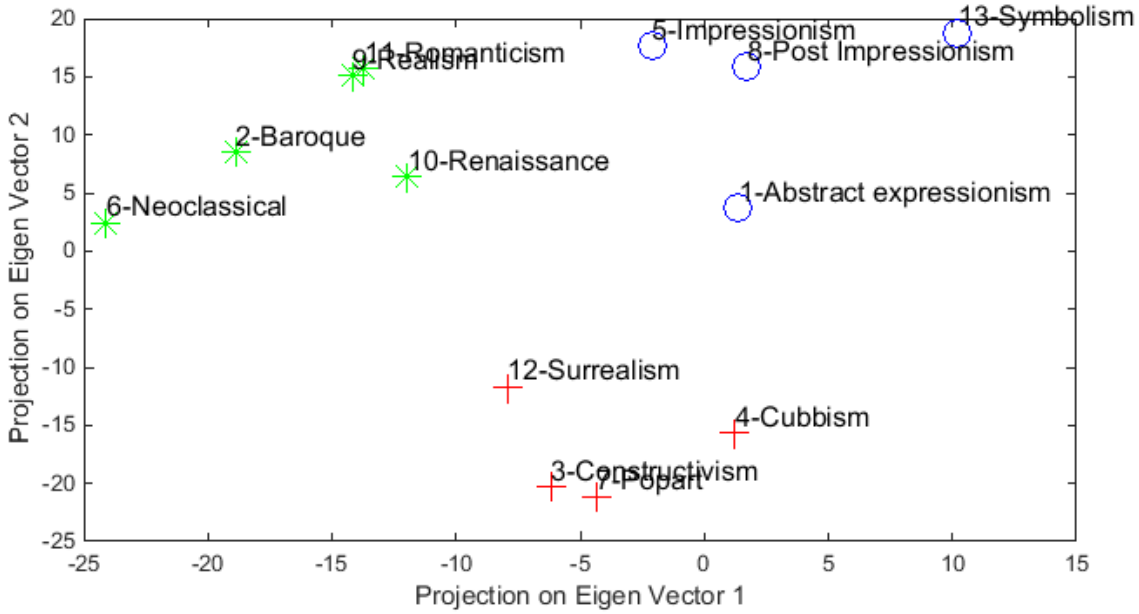

Figure 3.7 The style influence cluster graph for the Painting-91 dataset.

Style Influence In this section, we study the style influence so as to find similarities between different art styles and understand the evolution of art styles in different art movement periods. The style influence is calculated in a similar manner as the artist influence. First, we calculate the average of the feature vector of all paintings for a style. We then apply k-means clustering method with cosine distance to form clusters of similar styles. We set the number of clusters as 3 based on the different art movement periods. The style influence graph is plotted using the first two principal components of the average feature vector.

Figure 3.7 shows the style influence graph clusters with k set as 3. Cluster 1 contains the styles of the post renaissance art movement period with the only exception of style renaissance. The reason for this may be due the high similarity between styles baroque and renaissance as the style baroque evolved from the style renaissance [91]. The styles impressionism, post impressionism and symbolism in cluster 2 show that there are high similarities between these styles in the modern art movement period as the three styles

have a common french and belgian origin. Similarly, styles constructivism and popart in cluster 3 show high similarity in the style influence cluster graph.

We further show the results based on the correlation matrix computed by comparing the average feature vector of all paintings of each style with all other styles. We set the correlation threshold as 0.7.

$$Renaissance => Baroque, Neoclassical$$

$$Romanticism => Realism$$

$$Impressionism => Post\ impressionism$$

$$Constructivism => Popart$$

The results are in good agreement with the style influence cluster graph and support the observation that the art styles within an art movement show higher similarity compared to the art styles between the art movement periods. The styles baroque and neoclassical belong to the same art movement period and the style baroque has evolved from the style renaissance. Similarly, other styles belong to the modern art movement period. It can be observed from the style influence cluster graph that the style pairs romanticism:realism, impressionism:post impressionism and constructivism:popart are plotted close to each other in the graph indicating high similarity between these styles.

3.3.2 Fifteen Scene Categories Dataset

The fifteen scene categories dataset [50] contains 4485 images from fifteen scene categories namely, office, kitchen, living room, bedroom, store, industrial, tall building, inside cite, street, highway, coast, open country, mountain, forest, and suburb with 210 to 410 images per category. We follow the experimental protocol as described in [50] wherein 1500 images are used for training whereas the remaining 2985 images are used for testing. The train/test split is determined randomly with the criterion that 100 images are selected for every scene category as train images and the remaining images are used as test images.

Table 3.5 Comparison Between the Proposed Method and Other Popular Methods on the Fifteen Scene Categories Dataset

Method	Accuracy (%)
KSPM [50]	81.40
DHFVC [28]	86.40
LLC [110]	80.57
KSPM [50]	81.40
LaplacianSC [27]	89.70
DHFVC [28]	86.40
D-KSVD [129]	89.10
LC-KSVD [40]	90.40
Hybrid-CNN [131]	91.59
SKML	**96.25**

Table 3.5 shows the comparison of the proposed SKML features with popular learning methods. The LLC method [110] extracts a feature descriptor by using a locality constraint for projection to a local co-ordinate system. The DHFVC method [28] uses a hierarchical visual feature coding architecture based on restricted Boltzmann machines (RBM) for encoding of SIFT descriptors. A over-complete dictionary is learned by the D-KSVD algorithm [129] by integrating the classification error to the objective criterion whereas the LC-KSVD approach [40] adds a label consistency constraint combined with the classification and reconstruction error to form a single objective function. Another popular sparse coding method is LaplacianSC [27] which preserves the locality of features by using a similarity preserving criterion based on Laplacian framework. The sparse coding methods D-KSVD, LC-KSVD and LaplacianSC achieves an accuracy of 89.10%, 90.40% and 89.70%, respectively. The state-of-the-art deep learning method such as hybrid CNN

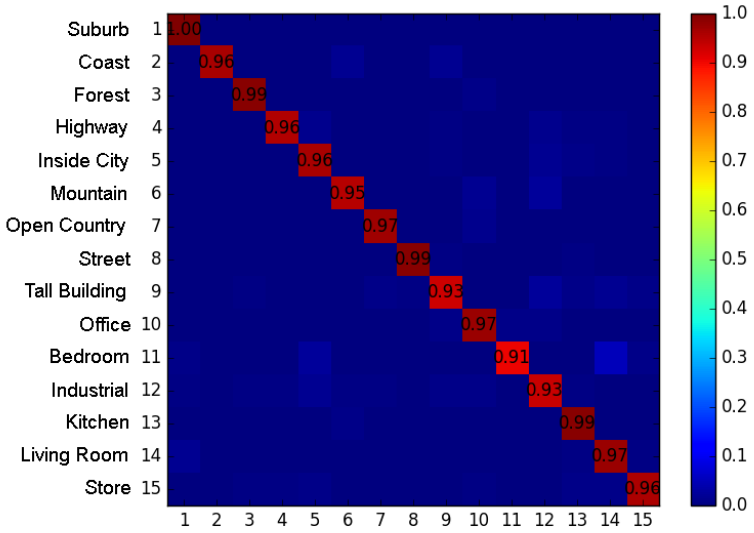

Figure 3.8 The confusion matrix diagram of the 15 scene categories dataset using the proposed SKML feature.

[131] which is trained on a combination of training set of ImageNet-CNN and Places-CNN achieves a performance of 91.59%. The experimental results in Table 3.5 show that our proposed SKML method achieves higher performance of 96.25% compared to popular sparse coding and deep learning methods.

The confusion diagram for the fifteen scene categories dataset is shown in Figure 3.8. The suburb category out of the fifteen scene categories achieves the best classification rate of 100%. The scene category with the lowest accuracy is the bedroom category with a classification rate of 91% as it has large confusions with the living room category. The living room scene category contains similar visual elements as the bedroom scene category resulting in high confusions between the two categories. The other scene categories that create confusion are tall building and industrial since both categories have some common visual semantics.

Figure 3.9 shows the t-SNE visualization for the fifteen scene categories dataset. The t-SNE method is a visualization technique used to fit high dimensional data to a plot using a non-linear dimensionality reduction technique to better understand the clusters of data of

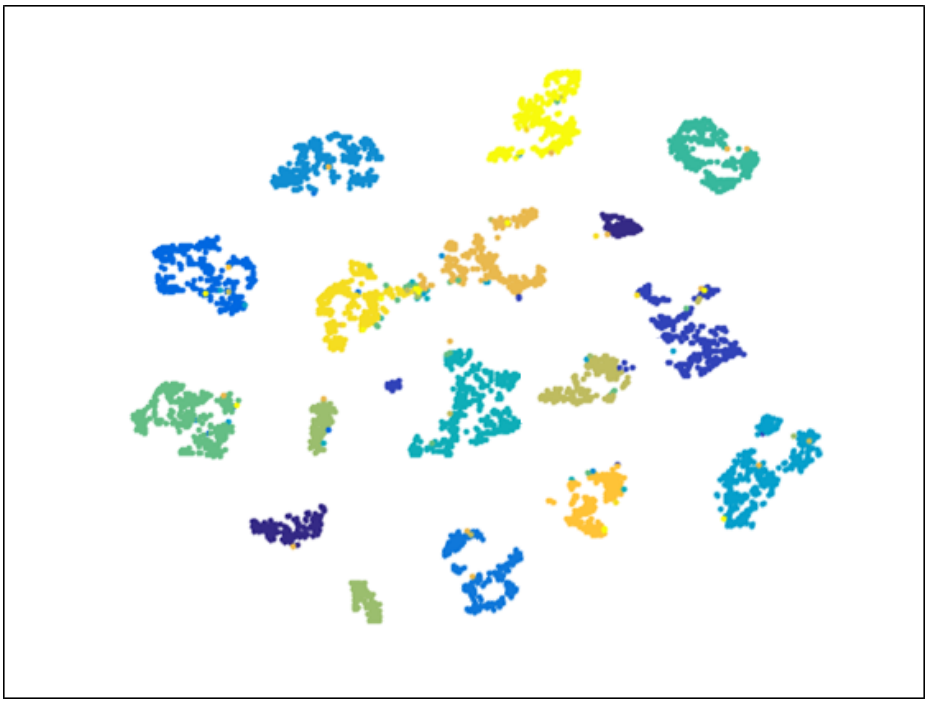

Figure 3.9 The t-SNE visualization of the 15 scene categories dataset using the proposed SKML feature.

different categories in a dataset. It can be seen from Figure 3.9 that our proposed SKML method improves the separability between clusters of different class. Another advantage of our proposed method is that it encourages better localization of data-points belonging to the same class resulting in better performance.

3.3.3 CalTech 101 Dataset

The Caltech 101 dataset [53] contains 9144 images of objects belonging to 101 categories. Every category has about 40 to 800 images and size of each image is roughly 300 X 200 pixels. The experimental protocol used for the CalTech 101 dataset is described in [110]. In particular, the training procedure involves five sets where each set contains 30, 25, 20, 15 and 10 train images per category, respectively and for every set, the test split contains the remaining images. In order to have a fair comparison with other methods, we report the performance as the average accuracy over all the categories.

Table 3.6 Comparison Between the Proposed SKML Method and Other Popular Methods on the Caltech 101 Dataset

Method	10	15	20	25	30
LLC [110]	59.77	65.43	67.74	70.16	73.44
SPM [50]	–	56.40	–	–	64.60
SVM-KNN [127]	55.80	59.10	62.00	–	66.20
SRC [113]	60.10	64.90	67.70	69.20	70.70
D-KSVD [129]	59.50	65.10	68.60	71.10	73.00
LC-KSVD [40]	63.10	67.70	70.50	72.30	73.60
CNN-M + Aug [12]	–	–	–	–	87.15
SKML	**82.47**	**84.46**	**85.35**	**86.61**	**87.95**

The experimental results in Table 3.6 shows the detailed classification performance of the proposed SKML mathod and other popular learning methods for the CalTech 101 dataset. The SPM (spatial pyramid matching) method [50] divides an image to sub-regions and computes histogram over these sub-regions to form a spatial pyramid. The SVM-KNN method [127] finds the nearest neighbors of the query image and trains a local SVM based on the distance matrix computed on the nearest neighbors. The sparse coding method SRC [113] uses a sparse representation method computed by l_1 minimization and achieves classification accuracy of 70.70% for the set with training size 30 images per category. The deep learning method CNN-M + Aug [12] is similar to the architecture of ZFNet [126] but also incorporates additional augmentation techniques such as flipping and cropping to increase the training size. It can be seen from Table 3.6 that our proposed method achieves better performance compared to other learning methods. Another advantage of the SKML method is that no additional data augmentation techniques are required to improve the performance.

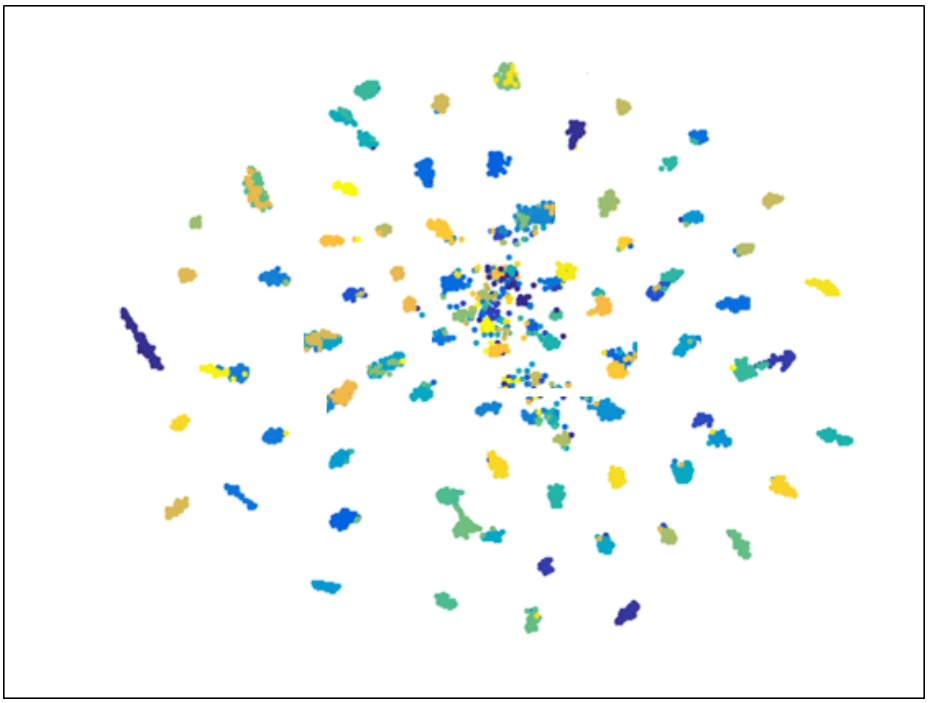

Figure 3.10 The t-SNE visualization for the CalTech 101 dataset using the proposed SKML feature.

The t-SNE visualization for the CalTech 101 dataset is shown in Figure 3.10. It can be seen that our proposed SKML method helps to increase the interclass separability between clusters having data-points belonging to different class categories as our method integrates a discriminative criterion to the objective function encouraging better clustering of data-points. Another advantage is that our method reduces the intraclass distance between data-points belonging to the same class in a cluster resulting in improved pattern recognition performance.

3.4 Conclusion

This chapter presents a sparse kernel manifold learner framework for different image classification applications. First, a new hybrid feature extraction step is performed by introducing D-FV and WS-FV features to capture different aspects of image and encode important discriminatory information. We then derive an innovative FFV feature by integrating the

D-FV, WS-FV and SIFT-FV features. The FFV features are computed in eight different color spaces and fused to produce the novel FCFV feature. Finally, we propose a sparse kernel manifold learner (SKML) method by integrating a discriminative marginal Fisher criterion to the representation criterion to improve the classification performance. The SKML method aims is to minimize the intraclass compactness and maximize the interclass separability constrained on the discriminative sparse objective function. Experimental results on different image classification datasets show the effectiveness of the proposed method.

CHAPTER 4

SPARSE REPRESENTATION BASED COMPLETE MFA FRAMEWORK

4.1 Introduction

Image Classification, which aims to categorize different visual objects into several predefined classes, is a challenging topic in both computer vision and multimedia research areas. Recently, sparse coding algorithms have been broadly applied in multimedia research, for example, in face recognition [71, 39, 113, 119, 129], in disease recognition [24], in scene and object recognition [3, 40, 123, 27, 71, 39, 23], in hand written digit recognition [121], and in human action recognition [32]. Pioneer research in cognitive psychology [80, 107] reveals that the biological visual cortex adopts a sparse representation for visual perception in the early stages as it provides an efficient representation for later phases of processing. Besides, manifold learning methods, such as discriminant analysis [74, 45], marginal Fisher analysis [118], have been successfully applied to preserve data locality in the embeded space and learn discriminative feature representations [118, 58, 25].

The marginal Fisher analysis (MFA) method improves upon the traditional linear discriminant analysis or LDA by means of the graph embedding framework that defines an intrinsic graph and a penalty graph [118]. The intrinsic graph connects each data sample with its neighboring samples of the same class to define the intraclass compactness, while the penalty graph connects the marginal points of different classes to define the interclass separability. The MFA method, however, does not account for the null space of the local samples based within class scatter matrix, which contains important discriminatory imformation. We present a complete marginal Fisher analysis (CMFA) method that extracts the discriminatory features in both the column space of the local samples based within class scatter matrix and the null space of its transformed matrix. The rationale of extracting

features in both spaces is to enhance the discriminatory power by further utilizing the null space, which is not accounted for in the marginal Fisher analysis method.

To further improve the classification capability and to ensure an efficient representation, we propose a discriminative sparse representation model using the CMFA framework by integrating a representation criterion such as the sparse coding and a discriminant criterion. Sparse coding facilitates efficient retrieval of data in multimedia as it generates a sparse representation such that every data sample can be represented as a linear combination of a small set of basis vectors due to the fact that most of the coefficients are zero. Another advantage is that the sparse representation may be overcomplete, allowing more flexibility in matching data and yielding a better approximation of the statistical distribution of the data. Sparse coding, however, is not directly related to classification as it does not address discriminant analysis of the multimedia data. We present a discriminative sparse representation model by integrating a representation criterion, such as the sparse representation, and a discriminative criterion, which applies the new within-class and between-class scatter matrices based on the marginal information, for improving the classification capability. Furthermore, we propose the largest step size for learning the sparse representation to address the convergence issues of our proposed optimization procedure. Finally, we present a dictionary screening rule that discards the dictionary items with null coefficients to improve the computational efficiency of the optimization process without affecting the accuracy.

Our proposed CMFA-SR method is assessed on different image classification tasks using representative datasets, such as the Painting-91 dataset [44], the fifteen scene categories dataset [50], the MIT-67 indoor scenes dataset [90], the Caltech 101 dataset [53], the Caltech 256 object categories dataset [31], the AR face dataset [74], and the extended Yale B dataset [52]. The experimental results show the feasibility of our proposed method.

4.2 Sparse Representation Using the Complete Marginal Fisher Analysis

The motivation of this work is to derive a novel learning method by integrating the state-of-the-art feature extraction methods, such as the sparse representation [113] and the marginal Fisher analysis [118], as well as leveraging our research on enhancing discrimination analysis [62, 15]. Specificaly, the pioneer work on the marginal Fisher analysis [118] improves upon the traditional discriminant analysis by introducing K Nearest Neighbors, or KNN samples in the graph embedding framework. Our new complete MFA method further enhances the disriminatory power by introducing two processes that analyze both the column space and the null space of the local (KNN) samples based within-class scatter matrix. In addition, our novel discriminative sparse representation approach fuses both the sparse represetation criterion and the discrimination criterion to improve upon the conventional sparse representation that does not consider classification.

4.2.1 Complete Marginal Fisher Analysis

The marginal Fisher analysis or MFA method improves upon the traditional discriminant analysis method by introducing the K Nearest Neighbors or KNN for defining both the intraclass compactness and the interclass separability, respectively [118]. The motivation behind the MFA approach rests on the graph embedding framework that utilizes both the intrinsic graph and the penalty graph [118]. Our recent research also reveals the importance of local smaples, such as the KNN samples, for designing effective learning systems [63, 106]. The application of local samples has its theoretical roots in the statistical learning theory and the stuctrual risk minimization principle in general, and in the design of support vector machines in particular, such as the support vectors, which are local samples. We, therefore, leverage the ideas of the MFA method and local samples, coupled with the analysis of the column space and the null space of the local (KNN) samples based within-class scatter matrix, and propose our novel complete marginal Fisher analysis method.

Specifically, let the sample data matrix be $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m] \in \mathbb{R}^{h \times m}$, where m is the number of samples of dimension h. Let $\mathbf{W} \in \mathbb{R}^{h \times h}$ be a projection matrix, which will be derived through the following optimization process. The k_1 nearest neighbors based within-class scatter matrix is defined as follows:

$$\mathbf{S}_w = \mathbf{W}^T\mathbf{X}(\mathbf{D} - \mathbf{A})\mathbf{X}^T\mathbf{W} \tag{4.1}$$

where $\mathbf{A}$ is a binary matrix with nonzero elements $\mathbf{A}_{ij}$ corresponding to the k_1 nearest neighbors of the sample $\mathbf{x}_i$ or the sample $\mathbf{x}_j$ from the same class [86]. $\mathbf{D}$ is a diagonal matrix, whose diagonal elements are defined by the summation of the off-diagonal elements of $\mathbf{A}$ row-wise.

The k_2 nearest neighbors based between-class scatter matrix is defined as follows:

$$\mathbf{S}_b = \mathbf{W}^T\mathbf{X}(\mathbf{D}' - \mathbf{A}')\mathbf{X}^T\mathbf{W} \tag{4.2}$$

where $\mathbf{A}'$ is a binary matrix with nonzero elements $\mathbf{A}'_{ij}$ corresponding to the k_2 nearest neighbors of the sample $\mathbf{x}_i$ or the sample $\mathbf{x}_j$ from two different classes [86]. $\mathbf{D}'$ is a diagonal matrix, whose diagonal elements are defined by the summation of the off-diagonal elements of $\mathbf{A}'$ row-wise.

Applying the k_1 nearest neighbors based within-class scatter matrix $\mathbf{S}_w$ and the k_2 nearest neighbors based between-class scatter matrix $\mathbf{S}_b$, we are able to derive the optimal projection matrix $\mathbf{W}$ by maximizing the following critirion J_1 [25]:

$$\begin{aligned} J_1 &= \mathbf{tr}(\mathbf{S}_w^{-1}\mathbf{S}_b) \\ &= \mathbf{tr}((\mathbf{W}^T\mathbf{X}(\mathbf{D} - \mathbf{A})\mathbf{X}^T\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{X}(\mathbf{D}' - \mathbf{A}')\mathbf{X}^T\mathbf{W})) \end{aligned} \tag{4.3}$$

The MFA method first applies pricipal component analysis or PCA for dimensionality reduction [118]. A potential problem with this PCA step is that it may discard the null space of the k_1 nearest neighbors based within-class scatter matrix, which contains important discriminative information. Previous research on linear discriminant

analysis shows that the null space of the within-class scatter matrix contains important discriminative information whereas the null space of the between-class scatter matrix contains no useful discriminatory information [14, 125].

We, therefore, propose a new method, a complete marginal Fisher analysis method, which extracts features from two subspaces, namely, the column space of the k_1 nearest neighbors based within-class scatter matrix $\mathbf{S}_w$ and the null space of the transformed $\mathbf{S}_w$ by removing the null space of the mixture scatter matrix, i.e., $\mathbf{S}_m = \mathbf{S}_w + \mathbf{S}_b$. We then extract two types of discriminatory features in these two subspaces: the discriminatory features in the column space of $\mathbf{S}_w$, and the discriminatory features in the null space of the transformed $\mathbf{S}_w$.

4.2.2 Extraction of the Discriminatory Features in Two Subspaces

Let $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_h$ be the eigenvectors of $\mathbf{S}_w$, whose rank is p. The space $\mathbb{R}^h$ is thus divided into the column space, $span\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_p\}$, and its orthogonal complement, i.e., the null space of $\mathbf{S}_w$, $span\{\boldsymbol{\beta}_{p+1}, \boldsymbol{\beta}_{p+2}, ..., \boldsymbol{\beta}_h\}$. Let the transformation matrix $\mathbf{T}_p$ be defined as follows: $\mathbf{T}_p = [\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_p]$. The k_1 nearest neighbors based within-class scatter matrix $\mathbf{S}_w$ and the k_2 nearest neighbors based between-class scatter matrix $\mathbf{S}_b$ may be transformed into the column space as follows: $\mathbf{S}_w^{'} = \mathbf{T}_p^T \mathbf{S}_w \mathbf{T}_p$, $\mathbf{S}_b^{'} = \mathbf{T}_p^T \mathbf{S}_b \mathbf{T}_p$.

The optimal projection matrix $\boldsymbol{\xi} = [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, ..., \boldsymbol{\xi}_p]$ is derived by means of maximizing the following critirion $J_1^{'}$ [25]:

$$\begin{aligned} J_1^{'} &= \mathbf{tr}((\mathbf{S}_w^{'})^{-1}\mathbf{S}_b^{'}) \\ &= \mathbf{tr}((\mathbf{T}_p^T \mathbf{S}_w \mathbf{T}_p)^{-1}\mathbf{T}_p^T \mathbf{S}_b \mathbf{T}_p) \end{aligned} \tag{4.4}$$

The discriminatory features in the column space of $\mathbf{S}_w$ are derived as follows:

$$\mathbf{U}^c = \boldsymbol{\xi}^T \mathbf{T}_p^T \mathbf{X} \tag{4.5}$$

The computation of the discriminatory features in the null space of the transformed $\mathbf{S}_w$ consists of the following steps. First, we will discard the null space of the mixture scatter matrix, $\mathbf{S}_m = \mathbf{S}_w + \mathbf{S}_b$, by transforming both $\mathbf{S}_w$ and $\mathbf{S}_b$ into the column space of $\mathbf{S}_m$, respectively: $\mathbf{S}_w^{''}$ and $\mathbf{S}_b^{''}$. The rationale for discarding the null space of the mixture scatter matrix is due to the fact that both the within class scatter matrix and the between class scatter matrix are nullified in this null space. As a result, the null space of the mixture scatter matrix does not carry discriminatory information. Second, we compute the null space of $\mathbf{S}_w^{''}$, and then transform $\mathbf{S}_b^{''}$ into this null space in order to derive the discriminatory features $\mathbf{U}^n$.

Specifically, let $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_k]$ be the transformation matrix that is defined by the eigenvectors of $\mathbf{S}_m$ corresponding to the nonzero eigenvalues, where $k \leq h$. The scatter matrices $\mathbf{S}_w$ and $\mathbf{S}_b$ may be transformed into the column space of $\mathbf{S}_m$ as follows: $\mathbf{S}_w^{''} = \boldsymbol{\alpha}^T \mathbf{S}_w \boldsymbol{\alpha}$, $\mathbf{S}_b^{''} = \boldsymbol{\alpha}^T \mathbf{S}_b \boldsymbol{\alpha}$. Next, we compute the eigenvectors of $\mathbf{S}_w^{''}$, whose null space is spanned by the eigenvectors corresponding to the zero eigenvalues of $\mathbf{S}_w^{''}$. Let $\mathbf{N}$ be the transformation matrix defined by the eigenvectors that span the null space of $\mathbf{S}_w^{''}$. Then, we transform $\mathbf{S}_b^{''}$ into the null space of $\mathbf{S}_w^{''}$ as follows: $\mathbf{S}_b^{'''} = \mathbf{N}^T \mathbf{S}_b^{''} \mathbf{N}$. Finally, we diagonalize the real symmetric matrix $\mathbf{S}_b^{'''}$ and derive its eigenvectors. Let $\boldsymbol{\zeta}$ be the transformation matrix defined by the eigenvectors of $\mathbf{S}_b^{'''}$ corresponding to the non-zero eigenvalues. The discriminatory features in the null space of the transformed $\mathbf{S}_w$ are derived as follows:

$$\mathbf{U}^n = \boldsymbol{\zeta}^T \mathbf{N}^T \boldsymbol{\alpha}^T \mathbf{X} \tag{4.6}$$

In order to obtain the final set of features, the discriminatory features extracted in the column space and the null space are fused and normalized to zero mean and unit standard deviation.

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}^c \\ \mathbf{U}^n \end{bmatrix} \tag{4.7}$$

4.2.3 Discriminative Sparse Representation Model

In this section, we present a sparse representation model CMFA-SR that uses a discriminative sparse representation criterion with the rationale to integrate a representation criterion such as sparse coding and a discriminative criterion so as to improve the classification performance.

Given m training samples, our complete marginal Fisher analysis method derives the feature matrix: $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_m] \in \mathbb{R}^{l \times m}$. Let $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_r] \in \mathbb{R}^{l \times r}$ be the dictionary defined by the r basis vectors and $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_m] \in \mathbb{R}^{r \times m}$ be the sparse representation matrix denoting the sparse representation of the m samples. Note that the coefficients $\mathbf{a}_i$ correspond to the items in the dictionary $\mathbf{D}$.

In our proposed CMFA-SR model, we optimize a sparse representation criterion and a discriminative analysis criterion to derive the dictionary $\mathbf{D}$ and the sparse representation $\mathbf{S}$ from the training samples. We use the representation criterion of the sparse representation to define new discriminative within-class matrix $\hat{\mathbf{H}}_w$ and discriminative between-class matrix $\hat{\mathbf{H}}_b$ by considering only the k nearest neighbors. Specifically, using the sparse representation criterion the descriminative within class matrix is defined as $\hat{\mathbf{H}}_w = \sum_{i=1}^{m} \sum_{(i,j) \in N_k^w(i,j)} (\mathbf{s}_i - \mathbf{s}_j)(\mathbf{s}_i - \mathbf{s}_j)^T$, where $(i,j) \in N_k^w(i,j)$ represents the (i,j) pairs where sample $\mathbf{u}_i$ is among the k nearest neighbors of sample $\mathbf{u}_j$ of the same class or vice versa. The discriminative between class matrix is defined as $\hat{\mathbf{H}}_b = \sum_{i=1}^{m} \sum_{(i,j) \in N_k^b(i,j)} (\mathbf{s}_i - \mathbf{s}_j)(\mathbf{s}_i - \mathbf{s}_j)^T$, where $(i,j) \in N_k^b(i,j)$ represents k nearest (i,j) pairs among all the (i,j) pairs between samples $\mathbf{u}_i$ and $\mathbf{u}_j$ of different classes. As a result, the new optimization criterion is as follows:

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{S}} \sum_{i=1}^{m} \{||\mathbf{u}_i - \mathbf{D}\mathbf{s}_i||^2 + \lambda ||\mathbf{s}_i||_1\} + \alpha \mathbf{tr}(\beta \hat{\mathbf{H}}_w - (1-\beta)\hat{\mathbf{H}}_b) \\ &s.t. ||\mathbf{d}_j|| \leq 1, (j = 1, 2, ..., r) \end{aligned} \tag{4.8}$$

where the parameter λ controls the sparseness term, the parameter α controls the discriminatory term, the parameter β balances the contributions of the discriminative

within class matrix $\hat{\mathbf{H}}_w$ and between class matrix $\hat{\mathbf{H}}_b$, and **tr(.)** denotes the trace of a matrix. In order to derive the discriminative sparse representation for the test data, as the dictionary $\mathbf{D}$ is already learned, we only need to optimize the following criterion: $\min_B \sum_{i=1}^{t}\{||\mathbf{y}_i - \mathbf{D}\mathbf{b}_i||^2\} + \lambda||\mathbf{b}_i||_1$ where $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_t$ are the test samples and t is the number of test samples. The discriminative sparse representation for the test data is defined as $\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_t] \in \mathbb{R}^{r\times t}$. Since the dictionary $\mathbf{D}$ is learned from the training optimization process, it contains both the sparseness and the discriminative information, therefore the derived representation $\mathbf{B}$ is the discriminative sparse representation for the test set.

4.3 The Optimization Procedure

In this section, we provide a detailed analysis of the largest step size for learning the sparse representation to address the convergence issues of the algorithm. We also introduce a screening rule to safely remove the dictionary items with null coefficients without affecting the performance to improve the computational efficiency of the proposed model.

4.3.1 Largest Step Size for Learning the Sparse Representation

In this section, we present and prove the largest step size for learning the sparse representation using the FISTA algorithm [6]. In particular, after applying some linear algebra transformations, the scatter matrices $\hat{\mathbf{H}}_w$ and $\hat{\mathbf{H}}_b$ in Equation 4.8 can be defined as :

$$\begin{aligned}\hat{\mathbf{H}}_w &= 2\mathbf{S}(\mathbf{D}_{\hat{\mathbf{H}}_w} - \mathbf{W}_{\hat{\mathbf{H}}_w})\mathbf{S}^T \\ \hat{\mathbf{H}}_b &= 2\mathbf{S}(\mathbf{D}_{\hat{\mathbf{H}}_b} - \mathbf{W}_{\hat{\mathbf{H}}_b})\mathbf{S}^T\end{aligned} \tag{4.9}$$

where $\mathbf{W}_{\hat{\mathbf{H}}_w}$ and $\mathbf{W}_{\hat{\mathbf{H}}_b}$ are matrices whose values $\mathbf{W}_{\hat{\mathbf{H}}_w}(i,j) = 1$ if the pair (i,j) is among the k nearest pairs in the same class otherwise 0, $\mathbf{W}_{\hat{\mathbf{H}}_b}(i,j) = 1$ if the pair (i,j) is among the set $\{(i,j), i \in \pi_c, j \notin \pi_c\}$ otherwise 0, $\mathbf{D}_{\hat{\mathbf{H}}_w}$ and $\mathbf{D}_{\hat{\mathbf{H}}_b}$ are diagonal matrices whose values are $\mathbf{D}_{\hat{\mathbf{H}}_w}(i,i) = \sum_j \mathbf{W}_{\hat{\mathbf{H}}_w}(i,j)$ and $\mathbf{D}_{\hat{\mathbf{H}}_b}(i,i) = \sum_j \mathbf{W}_{\hat{\mathbf{H}}_b}(i,j)$.

Therefore, the objective function of the sparse representation in equation 4.8 can be converted to the following form:

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{S}} \sum_{i=1}^{m} \{||\mathbf{u}_i - \mathbf{D}\mathbf{s}_i||^2 + \lambda||\mathbf{s}_i||_1\} + \alpha \mathbf{tr}(\mathbf{S}\mathbf{M}\mathbf{S}^T) \\ &s.t. ||\mathbf{d}_j|| \leq 1, (j = 1, 2, ..., r) \end{aligned} \tag{4.10}$$

where $\mathbf{M} = 2(\beta(\mathbf{D}_{\hat{\mathbf{H}_w}} - \mathbf{W}_{\hat{\mathbf{H}_w}}) - (1-\beta)(\mathbf{D}_{\hat{\mathbf{H}_b}} - \mathbf{W}_{\hat{\mathbf{H}_b}}))$ for the proposed CMFA-SR method. We further optimize the objective function in Equation 4.10 by alternatively updating the sparse representation and the discriminative dictionary by decomposing into two separate objective functions for each training sample $\mathbf{u}_i$ given as follows:

$$\min_{\mathbf{s}_i} ||\mathbf{u}_i - \mathbf{D}\mathbf{s}_i||^2 + \alpha M_{ii}\mathbf{s}_i^t\mathbf{s}_i + \alpha \mathbf{s}_i^t \mathbf{g}_i + \lambda||\mathbf{s}_i||_1 \tag{4.11}$$

where $\mathbf{g}_i = \sum_{j \neq i} M_{ij}\mathbf{s}_j = [g_{i1}, g_{i2}, ..., g_{ik}]^t$ and $M_{ij}(i, j = 1, 2, .., m)$ is the value of the element in the i-th row and j-th column of the matrix $\mathbf{M}$. We optimize the above objective function by alternatively applying the FISTA algorithm [6] to learn the sparse representation and the Lagrange dual method [51] for updating the dictionary. In order to derive the largest step size for learning the sparse representation, we rewrite the objective function in Equation 4.11 in the form of $a(\mathbf{s}_i) + b(\mathbf{s}_i)$, where $a(\mathbf{s}_i) = ||\mathbf{u}_i - \mathbf{D}\mathbf{s}_i||^2 + \alpha M_{ii}\mathbf{s}_i^t\mathbf{s}_i + \alpha \mathbf{s}_i^t \mathbf{g}_i$ and $b(\mathbf{s}_i) = \lambda||\mathbf{s}_i||_1$.

To guarantee the convergence of the FISTA algorithm, an important quantity to be determined is the step size. Given the objective function $F(x) = f(x) + g(x)$, where $f(x)$ is a smooth convex function and $g(x)$ is a non-smooth convex function, the theoretical analysis [5] shows that

$$F(x_k) - F(x^*) \leq \frac{2||x_0 - x^*||^2}{s * (k+1)^2} \tag{4.12}$$

where x_k is the solution generated by the FISTA algorithm at the k-th iteration, x^* is the optimal solution, and s is the largest step size for convergence. This theoretical result means that the number of iterations of the FISTA algorithm required to obtain an

ϵ-optimal solution (x_t), such that $F(x_t) - F(x^*) \leq \epsilon$, is at most $\lceil C/\sqrt{\epsilon} - 1 \rceil$, where $C = \sqrt{2||x_0 - x^*||^2/s}$ Therefore, the step size plays an important role for the convergence of the algorithm and the largest step size can lead to less required iterations for the convergence of the FISTA algorithm.

We now, theoretically, derive the largest step size required for learning the sparse representation for each training sample.

Proposition 1. *The largest step size that guarantees convergence of the FISTA algorithm is* $\frac{1}{Lip(a)}$*, where* $Lip(a)$ *is the smallest Lipschitz constant of the gradient* ∇a *and* $Lip(a) = 2E_{\max}(\boldsymbol{D}^t\boldsymbol{D} + \alpha M_{ii}\boldsymbol{I})$ *which is twice the largest eigenvalue of the matrix* $(\boldsymbol{D}^t\boldsymbol{D} + \alpha M_{ii}\boldsymbol{I})$.

Proof. Function $a(\mathbf{s}_i)$ can be generalized as follows:

$$a(\mathbf{x}) = ||\mathbf{D}\mathbf{x} + \mathbf{b}||^2 + \alpha M_{ii}\mathbf{x}^t\mathbf{x} + \alpha\mathbf{x}^t\mathbf{c} \tag{4.13}$$

Taking the first derivative and finding the difference, we get

$$\nabla a(\mathbf{x}) - \nabla a(\mathbf{y}) = 2(\mathbf{D}^t\mathbf{D} + \alpha M_{ii}\mathbf{I})(\mathbf{x} - \mathbf{y}) \tag{4.14}$$

The Lipschitz constant of the gradient ∇a satisfies the following inequality

$$||\nabla a(\mathbf{x}) - \nabla a(\mathbf{y})|| \leq Lip(a)||\mathbf{x} - \mathbf{y}|| \tag{4.15}$$

Therefore, the smallest Lipschitz constant of the gradient ∇a is

$$Lip(a) = 2E_{\max}(\mathbf{D}^t\mathbf{D} + \alpha M_{ii}\mathbf{I}) \tag{4.16}$$

which is twice the largest eigenvalue of the matrix $(\mathbf{D}^t\mathbf{D} + \alpha M_{ii}\mathbf{I})$.

Hence, as shown in the FISTA algorithm [6], the largest step size that assures the convergence of the FISTA algorithm is the reciprocal of the smallest Lipschitz constant of the gradient ∇a. □

4.3.2 Updating the Dictionary

After the sparse representation **S** is learned using the FISTA algorithm, we have to learn the optimal dictionary **D**. The objective function in Equation 4.10 is a constrained optimization problem with inequality constraints, which may be solved using the Lagrange optimization method and the Kuhn-Tucker condition [51]. In order to solve the primal optimization, we take the first derivative with respect to **D** and set it to zero. The dual optimization problem can be formulated as follows:

$$\Lambda^* = \min_{\Lambda} \mathbf{tr}(\mathbf{U}\mathbf{S}^t(\mathbf{S}\mathbf{S}^t + \Lambda)^{-1}\mathbf{S}\mathbf{U}^t + \Lambda - \mathbf{U}^t\mathbf{U}) \tag{4.17}$$

where Λ is a diagonal matrix whose diagonal values are the dual parameters of the primal optimization problem. We solve the dual problem defined in Equation 4.17 using the gradient descent method and the dictionary **D** is updated using the following equation:

$$\mathbf{D} = \mathbf{U}\mathbf{S}^t(\mathbf{S}\mathbf{S}^t + \Lambda^*)^{-1} \tag{4.18}$$

4.3.3 The Dictionary Screening Rule

In this section, we present a dictionary screening rule to improve the computational efficiency during the optimization of the objective function defined in Equation 4.11. During the optimization procedure, the computational complexity is generally introduced due to an oversized dictionary. In our proposed dictionary screening rule, we first identify dictionary items with corresponding coefficient score set as zero by checking the sparse coefficient vectors. We then derive a trimmed dictionary by deleting the zero coefficient dictionary items to improve the computational efficiency. The trimmed dictionary is utilized by the FISTA algorithm [6] to obtain a compact sparse representation. We finally reintroduce the deleted zero coefficients back to compute the final sparse representation. Therefore, the dictionary screening rule improves the computational efficiency of the proposed sparse representation framework by computing a trimmed dictionary utilized by the FISTA algorithm.

The following proposition rule identifies the zero coefficients, so that the corresponding dictionary items may be deleted in order to compute the trimmed dictionary.

Proposition 2. *Given a training sample* $\boldsymbol{u}_i(i = 1,2,..,m)$ *and a dictionary item* $\boldsymbol{d}_j(j = 1,2,..,k)$*, the sparse coefficient* s_{ij} *is zero if* $|\boldsymbol{u}_i\boldsymbol{d}_j - \frac{\alpha}{2}\boldsymbol{g}_i^t\boldsymbol{I}_j| < (\lambda_{\max} - \sqrt{(||\boldsymbol{d}_j||^2 + \alpha M_{ii})(||\boldsymbol{u}_i||^2 + \frac{\alpha}{4M_{ii}}||\boldsymbol{g}_i||^2)}(\frac{\lambda_{\max}}{\lambda} - 1)$ *where* s_{ij} *is the* j*-th element of the sparse representation* $\boldsymbol{s}_i$*,* $\lambda_{\max} = \max_{1\leq j\leq k}|\boldsymbol{u}_i^t\boldsymbol{d}_j - \frac{\alpha}{2}\boldsymbol{g}_i^t\boldsymbol{I}_j|$ *and* $\boldsymbol{I}_j \in \mathbb{R}^{k\times 1}$ *is a vector with zero values for all elements except the* j*-th element which has a value 1.*

Proof. We first establish a relation between our proposed method and the traditional sparse representation lasso method. The objective function in Equation 4.11 is identical to the following equation:

$$\min_{\mathbf{s}_i} ||\mathbf{u}_i - \mathbf{D}\mathbf{s}_i||^2 + ||\sqrt{\alpha M_{ii}}\mathbf{s}_i + \sqrt{\frac{\alpha}{4M_{ii}}}\mathbf{g}_i||^2 + \lambda||\mathbf{s}_i||_1 \tag{4.19}$$

Therefore, the objective function in equation 4.11 can be rewritten as follows:

$$\min_{\mathbf{s}_i} ||\mathbf{u}_i^* - \mathbf{D}^*\mathbf{s}_i||^2 + \lambda||\mathbf{s}_i||_1 \tag{4.20}$$

where $\mathbf{u}_i^* = (\mathbf{u}_i^t - \sqrt{\frac{\alpha}{4M_{ii}}}\mathbf{g}_i^t)^t \in \mathbb{R}^{(n+k)\times 1}$ and $\mathbf{D}^* = (\mathbf{D}^t, \sqrt{\alpha M_{ii}}\mathbf{I})^t \in \mathbb{R}^{(n+k)\times k}$. Note that $||\mathbf{d}_j^*||^2 = ||\mathbf{d}_j||^2 + \alpha M_{ii} \leq 1 + \alpha M_{ii}$ and $||\mathbf{u}_i^*||^2 = ||\mathbf{u}_i||^2 + \frac{\alpha}{4M_{ii}}||\mathbf{g}_i||^2$.

According to the projection theorem in [111], we observe that $||\boldsymbol{\theta}_i(\lambda) - \boldsymbol{\theta}_i(\lambda_{\max})||_2 \leq ||\frac{\mathbf{u}_i^*}{\lambda} - \frac{\mathbf{u}_i^*}{\lambda_{\max}}||_2$, where $\boldsymbol{\theta}_i(\lambda)$ and $\boldsymbol{\theta}_i(\lambda_{\max})$ are the solutions of the dual problem associated with the values of λ. The condition given in proposition 3 for identifying dictionary items with zero coefficients is $|\mathbf{u}_i\mathbf{d}_j - \frac{\alpha}{2}\mathbf{g}_i^t\mathbf{I}_j| < (\lambda_{\max} - \sqrt{(||\mathbf{d}_j||^2 + \alpha M_{ii})(||\mathbf{u}_i||^2 + \frac{\alpha}{4M_{ii}}||\mathbf{g}_i||^2)}(\frac{\lambda_{\max}}{\lambda} - 1)$, which is equal to $|(\mathbf{d}_j^*)^t\boldsymbol{\theta}_i(\lambda_{\max})| < 1 - ||\mathbf{u}_i^*||_2||\mathbf{d}_j^*||_2|\frac{1}{\lambda} - \frac{1}{\lambda_{\max}}|$.

Thus, we have the following relations.

$$
\begin{aligned}
|\boldsymbol{\theta}_i^t(\lambda)\mathbf{d}_j^*| &= |(\mathbf{d}_j^*)^t\boldsymbol{\theta}_i(\lambda)| \\
&\leq |(\mathbf{d}_j^*)^t\boldsymbol{\theta}_i(\lambda) - (\mathbf{d}_j^*)^t\boldsymbol{\theta}_i(\lambda_{\max})| + |(\mathbf{d}_j^*)^t\boldsymbol{\theta}_i(\lambda_{\max})| \\
&\leq ||(\mathbf{d}_j^*)||_2||\boldsymbol{\theta}_i(\lambda) - \boldsymbol{\theta}_i(\lambda_{\max})||_2 \\
&+ 1 - ||(\mathbf{d}_j^*)||_2||\frac{\mathbf{u}_i^*}{\lambda} - \frac{\mathbf{u}_i^*}{\lambda_{\max}}||_2 \\
&\leq ||(\mathbf{d}_j^*)||_2||\frac{\mathbf{u}_i^*}{\lambda} - \frac{\mathbf{u}_i^*}{\lambda_{\max}}||_2 \\
&+ 1 - ||(\mathbf{d}_j^*)||_2||\frac{\mathbf{u}_i^*}{\lambda} - \frac{\mathbf{u}_i^*}{\lambda_{\max}}||_2 \\
&= 1
\end{aligned}
\tag{4.21}
$$

It is shown in [117] that the dual variable $\boldsymbol{\theta}_i$ in the Lagrange dual function of the lasso problem defined in Equation 4.20 satisfies

$$
|\boldsymbol{\theta}_i^t\mathbf{d}_j^*| \leq 1 \implies s_{ij} = 0 \tag{4.22}
$$

Hence, the proposition 3 is proved. □

4.4 Experiments

Our proposed CMFA-SR method has been evaluated on some challenging visual recognition tasks: (i) fine art painting categorization using the Painting-91 dataset [44], (ii) scene recognition using the fifteen scene categories [50] and the MIT-67 indoor scenes dataset [90], (iii) object recognition using the Caltech 101 dataset [53] and the Caltec 256 object categories [], and (iv) face recognition using the AR face database [74] and the extended Yale B dataset [52]. Specifically, the datasets used in our experiments are detailed in Table 4.1 and some sample images are shown in Figure 4.1.

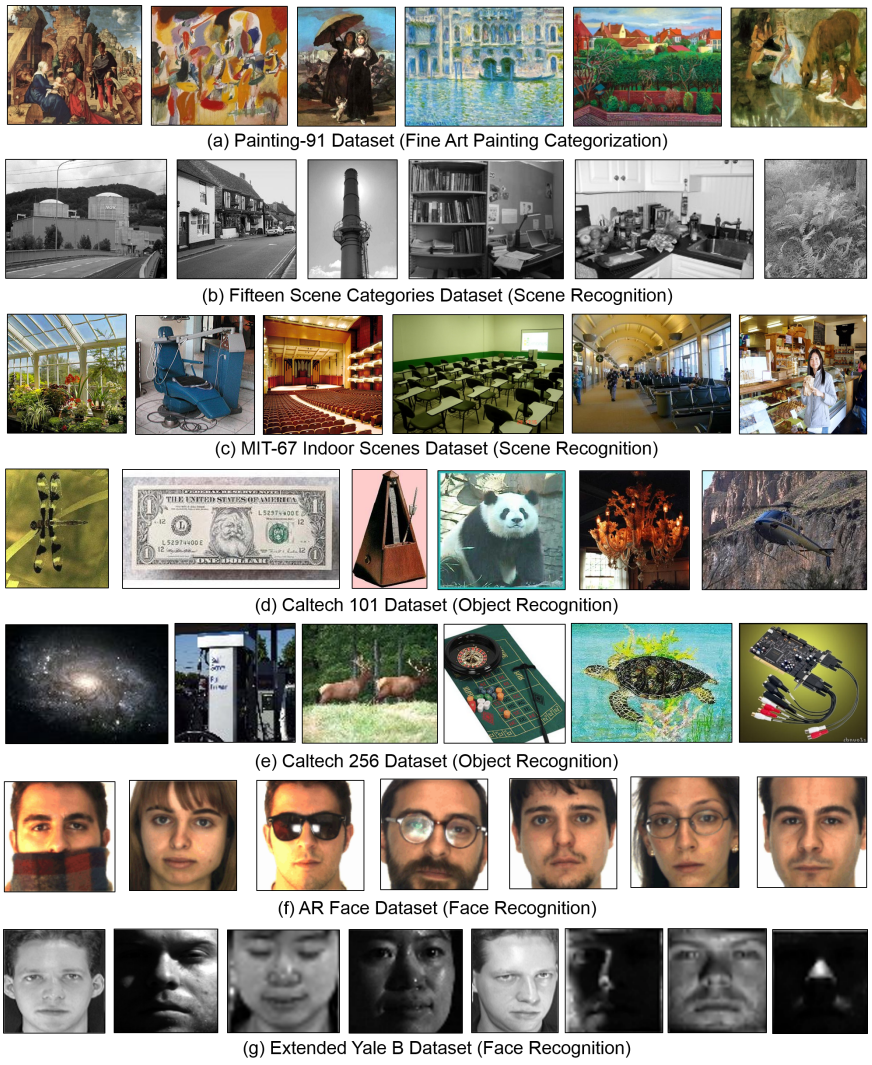

Figure 4.1 Some example images of the different datasets used for evaluation.

4.4.1 Painting-91 Dataset

The Painting-91 dataset [44] is a challenging dataset of fine art painting images collected from the Internet and contains two tasks: artist classification and style classification. We follow the experimental protocol in [44] which uses a fixed train and test split for both the tasks. The initial features used are fused Fisher vector (FFV) features [83] which are extracted using a hybrid feature extraction step as described in [84]. We further compute the FFV features in different color spaces namely RGB, XYZ, YUV, YCbCr,

Table 4.1 Different Tasks and their Associated Datasets Used for Evaluation of the Proposed CMFA-SR Method

Dataset	Task	# Classes	Total # Images	Reference
Painting-91 [44]	artist classification	91	4266	[44]
Painting-91 [44]	style classification	13	2338	[44]
15 Scenes [50]	scene recognition	15	4485	[50]
MIT-67 Scenes [90]	scene recognition	67	15620	[90]
Caltech 101 [53]	object recognition	101	9144	[110]
Caltech 256 [31]	object recognition	256	30607	[110]
AR Face [74]	face recognition	126	4000	[40]
Extended Yale B [52]	face recognition	38	2414	[129]

YIQ, LAB, HSV and oRGB to incorporate color information as the color cue provides powerful discriminatory information.

Artist Classification. The artist classification task classifies a painting image to its respective artist and is a challenging task as there are large variations in the appearance, styles and subject matter of the paintings of the same artist. The dictionary size is set as 512, and the parameters λ = 0.05, α = 0.2 and β = 0.4 are selected for the CMFA-SR method. The experimental results are summarized in column 3 of Table 4.2. MSCNN is the abbreviation for multi-scale convolutional neural networks. The classification is performed using RBF-SVM with parameters C = 20 and γ = 0.00007. Our proposed method consistently outperforms other popular image descriptors and state-of-the-art deep learning methods for the artist classification task.

Style Classification. The style classification task deals with the problem of categorizing a painting to the 13 style classes defined in the dataset. For the CMFA-SR method, the

Table 4.2 Comparison Between the Proposed Method and Other Popular Methods for Artist and Style Classification Task of the Painting-91 Dataset

No.	Method	Artist Cls.	Style Cls.
1	LBP [78, 44]	28.50	42.20
2	Color-LBP [44]	35.00	47.00
3	PHOG [9, 44]	18.60	29.50
4	Color-PHOG [44]	22.80	33.20
5	GIST [79, 44]	23.90	31.30
6	Color-GIST [44]	27.80	36.50
7	SIFT [68, 44]	42.60	53.20
8	CLBP [33, 44]	34.70	46.40
9	CN [105, 44]	18.10	33.30
10	SSIM [95, 44]	23.70	37.50
11	OPPSIFT [104, 44]	39.50	52.20
12	RGBSIFT [104, 44]	40.30	47.40
13	CSIFT [104, 44]	36.40	48.60
14	CN-SIFT [44]	44.10	56.70
15	Combine(1 - 14) [44]	53.10	62.20
16	MSCNN-1 [81]	58.11	69.67
17	MSCNN-2 [81]	57.91	70.96
18	CNN F_3 [108]	56.40	68.57
19	CNN F_4 [108]	56.35	69.21
20	**CMFA-SR**	**65.78**	**73.16**

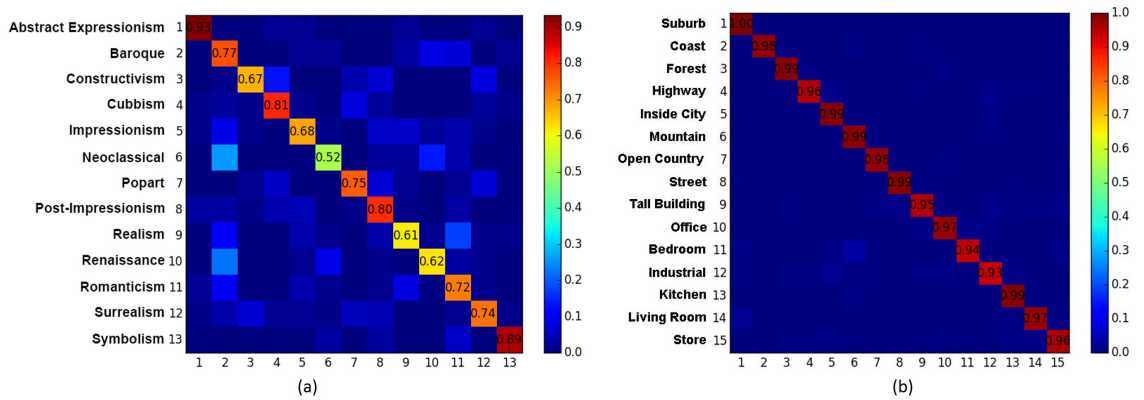

Figure 4.2 The confusion matrix for (a)13 style categories of the Painting-91 dataset (b) 15 scene categories dataset.

dictionary size is set as 256 and the same parameters are used as the artist classification task. The fourth column in Table 4.2 shows the recognition results. Experimental results demonstrate that our proposed CMFA-SR method achieves better performance compared to other popular image descriptors and deep learning methods for style classification.

Figure 4.2 (a) shows the confusion matrix for the 13 style categories of the Painting-91 dataset. It can be seen that the style categories with the best performance are 1 (abstract expressionism) and 13(symbolism) with classification rates of 93% and 89%, respectively. The most difficult style category to classify is category 6 (neoclassical) as there are large confusions between the style categories baroque and neoclassical. The other style category pairs that create confusion are the styles neoclassical: renaissance and the styles renaissance: baroque.

4.4.2 Fifteen Scene Categories Dataset

For the fifteen scene categories dataset [50], we follow the experimental protocol as in [50] where for 10 iterations, 100 images per class are randomly selected for each iteration from the dataset for training and the remaining images are used for testing. The initial input features used are the spatial pyramid features provided by [40] obtained by using a four-level spatial pyramid with a codebook of size 200. For the CMFA-SR method, the

Table 4.3 Comparison Between the Proposed Method and Other Popular Methods on the Fifteen Scene Categories Dataset

Method	Accuracy (%)
LLC [110]	80.57
KSPM [50]	81.40
DHFVC [28]	86.40
D-KSVD [129]	89.10
LaplacianSC [27]	89.70
LC-KSVD [40]	90.40
Places-CNN [131]	90.19
Hybrid-CNN [131]	91.59
DAG-CNN [124]	92.90
CMFA-SR	**98.45**

Table 4.4 Comparison Between the Proposed Method and Other Popular Methods on the MIT-67 Indoor Scenes Dataset

Method	Accuracy (%)
ROI + GIST [90]	26.10
Object Bank [55]	37.60
Discriminative parts [102]	51.40
VC + VQ [54]	52.30
DP + IFV [42]	60.80
Places-CNN [131]	68.24
Hybrid-CNN [131]	70.80
DAG-CNN [124]	77.50
CMFA-SR	**81.12**

dictionary size is set as 1024 and the parameters $\lambda = 0.05$, $\alpha = 0.2$, and $\beta = 0.4$ are selected. The RBF-SVM is used for classification with parameters set as $C = 7$ and $\gamma = 0.0001$. The experimental results in Table 4.3 show that the proposed method improves upon other popular sparse representation and deep learning methods by more than 5%. Figure 4.2 (b) shows the confusion matrix for the fifteen scene categories dataset.

4.4.3 MIT-67 Indoor Scenes Dataset

The MIT-67 indoor scenes dataset [90] is a challenging indoor scenes recognition dataset with a variable number of images per category where each category has atleast 100 images. We use experimental settings as in [90] where 80*67 images are used for training and 20*67 images are used for testing. The performance measure provided is the average classification accuracy over all the categories. We extract features for images of the MIT-67 indoor scenes dataset using a pre-trained convolution neural network Places-CNN [131]. For the proposed CMFA-SR method, the dictionary size is set as 512 and the parameters $\lambda = 0.05$, $\alpha = 0.1$, and $\beta = 0.5$ are selected, whereas for the RBF-SVM, parameters are set as $C = 2$ and $\gamma = 0.0001$. It can be seen from Table 4.4 that our method improves over the performance of Places-CNN by 13%. Our proposed CMFA-SR method helps to significantly improve the initial CNN features by encouraging better separation between the samples of different class and assist in the formation of compact clusters for the samples of same class (see Subsection 4.4.10). Experimental results in Table 4.4 show that the proposed method is able to achieve significantly better results and outperform other popular sparse representation and deep learning methods.

4.4.4 Caltech 101 Dataset

For the Caltech 101 dataset [53], we use the experimental settings as in [110], where we randomly split the dataset into 10, 15, 20, 25 and 30 training images per category and at the most 50 test images per category in order to have a fair comparison with other methods. The

Table 4.5 Comparison Between the Proposed Method and Other Popular Methods on the Caltech 101 Dataset

Method	10	15	20	25	30
SVM-KNN [127]	55.80	59.10	62.00	–	66.20
SPM [50]	–	56.40	–	–	64.60
LLC [110]	59.77	65.43	67.74	70.16	73.44
D-KSVD [129]	59.50	65.10	68.60	71.10	73.00
SRC [113]	60.10	64.90	67.70	69.20	70.70
LC-KSVD [40]	63.10	67.70	70.50	72.30	73.60
CNN-M + Aug [12]	–	–	–	–	87.15
CMFA-SR	**83.11**	**85.88**	**86.95**	**87.61**	**88.28**

performance measure provided is the average accuracy over all the classes. We evaluate our methods with features that are extracted using a pre-trained convolutional neural network CNN-M [12]. The dictionary size is selected as 512 and the parameters are set as λ = 0.05, α = 0.1, and β = 0.5 for the CMFA-SR method. The parameters of the RBF-SVM are C = 4 and γ = 0.00001. The experimental results shown in Table 4.5 show that even without using different fine tuning techniques as in [12], our proposed method is able to achieve comparable results to other state-of-the-art deep learning methods.

4.4.5 Caltech 256 Dataset

The Caltech 256 dataset [31] is an extended version of the Caltech 101 dataset and a more challenging object recognition dataset. We follow the experimental settings as specified in [110], where the dataset is randomly divided to 15, 30, 45 and 60 training images per category and at the most 25 test images for 3 iterations. The methods are evaluated using features extracted from a pre-trained ZFNet [99]. For the CMFA-SR method, we set the dictionary size to 1024, and the parameters as λ = 0.05, α = 0.1, and β = 0.5. The RBF-

Table 4.6 Comparison Between the Proposed Method and Other Popular Methods on the Caltech 256 Dataset

Method	15	30	45	60
ScSPM [120]	27.73	34.02	37.46	40.14
IFK [82]	34.70	40.80	45.00	47.90
LLC [110]	34.36	41.19	45.31	47.68
M-HMP [8]	40.50	48.00	51.90	55.20
ZFNet CNN [99]	65.70	70.60	72.70	74.20
CMFA-SR	**67.85**	**71.44**	**74.27**	**76.31**

SVM is used for classification with C = 2 and γ = 0.0001. The experimental results in Table 4.6 show that our proposed method is able to achieve better results compared to other learning methods.

4.4.6 AR Face Dataset

For the AR face dataset, a subset of the data [74] is selected containing 50 male and 50 female subjects and the images are cropped to 165*120 in order to follow the standard evaluation procedure. We evaluate our proposed method using two common experimental settings to have a fair comparison with other methods. We follow the first experimental setting as in [40] and [129] where we randomly select 20 training images and the remaining are selected for testing, for each person for 10 iterations. The model parameters are set as λ = 0.1, α = 0.2, and β = 0.6 and the dictionary size is selected as 512 for the CMFA-SR method. RBF-SVM is used for classification with parameters set as C= 4, γ = 0.0001.

The second experimental setting is defined in [20] where we randomly consider 26 images per person of which 13 images are used for training and the remaining 13 for testing for total of 10 iterations. The dictionary size is set to 512, and the parameters are set as λ = 0.1, α = 0.2, β = 0.5, and C= 1, γ = 0.0007 for the RBF-SVM classifier.

Table 4.7 Comparison Between the Proposed Method and Other Popular Methods on the AR Face Dataset

Method (Setting 1)	Accuracy (%)
D-KSVD [129]	95.00
LC-KSVD [40]	97.80
CMFA-SR	**98.95**
Method (Setting 2)	Accuracy (%)
SRC [113]	93.75 $\pm$ 1.01
ESRC [19]	97.36 $\pm$ 0.59
SSRC [20]	98.58 $\pm$ 0.40
CMFA-SR	**98.65** $\pm$ 0.42

The experimental results in Table 4.7 using our proposed CMFA-SR method for both the experimental settings show that our method is able to improve upon other popular methods.

4.4.7 Extended Yale B Dataset

As for the extended Yale B dataset, a common evaluation procedure is to use a cropped version of the dataset [52] where the images are manually aligned, cropped and resized to 192 x 168 pixels. The experimental setting as in [122] is followed wherein 20 images per subject are randomly selected for training and the remaining images are used for testing, for a total of 10 iterations. Note that this experimental setting is more difficult than that in [129]. We first scale the image to 42 X 48 and and we obtain the pattern vector using random faces [113]. The dictionary size is selected as 512. We set the parameters λ = 0.06, α = 0.2, and β = 0.5 for the CMFA-SR method. The classification is done using RBF-SVM with parameters C = 4 and γ = 0.001. Experimental results in Table 4.8 show that the proposed method achieves better results compared to other popular methods.

Table 4.8 Comparison Between the Proposed Method and Other Popular Methods on the Extended Yale B Dataset

Method	Accuracy (%)
D-KSVD [129]	75.30
SRC [113]	90.00
FDDL [122]	91.90
CMFA-SR	**94.94**

4.4.8 Evaluation of the Size of the Dictionary

In this section, we analyze the impact of different dictionary sizes on the performance of the CMFA-SR method. In particular, dictionary sizes of 1024, 512, and 256 are used for a comparative assessment of the performance. The results are presented in Figure 4.3 and we can deduce that the performance of the CMFA-SR method increases upto a certain dictionary size and then reaches a stable performance. We can also observe that for small datasets, a fairly good performance is achieved with a small dictionary size, whereas in case of large datasets such as the Caltech 101, a larger dictionary size is required. This indicates that a large dataset requires a larger dictionary as the dictionary captures the variability of the dataset.

4.4.9 Evaluation of the Size of the Training Data

We now evaluate the performance of our proposed CMFA-SR method when different sizes of training images per category are used. Figure 4.4 shows the performance of the CMFA-SR method for different training data sizes per category on the Caltech 101 dataset and 15 scenes dataset. The model parameters for both the datasets are set to values used in the corresponding experimental section. It can be observed from Figure 4.4 that the performance of the CMFA-SR method improves with the increase in the size of the

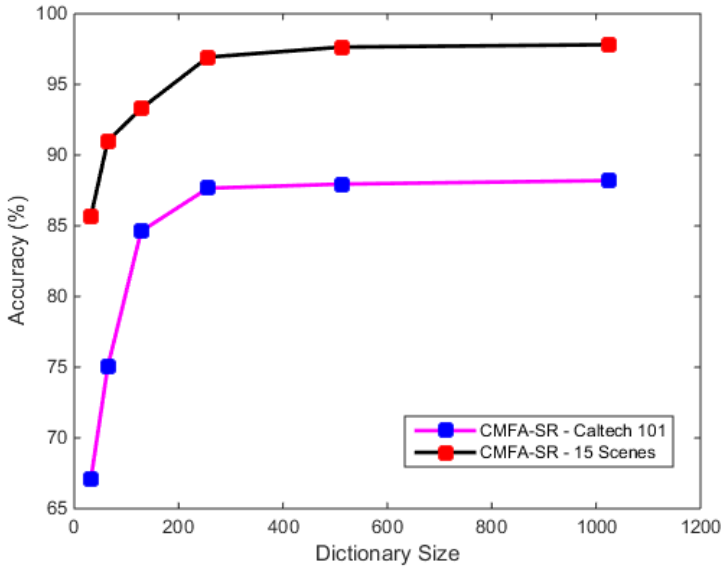

Figure 4.3 The performance of the proposed CMFA-SR method for different dictionary sizes on the Caltech 101 dataset and the 15 scenes dataset.

Table 4.9 Comparison of the Proposed CMFA-SR Features and the Deep Learning Features using the MIT-67 Indoor Scenes Dataset

Method	Accuracy (%)
Places-CNN [131]	68.24
CMFA-SR features	**81.12**

training data upto a certain value. After a certain training size, the performance only has minor variations indicating the robustness of the proposed method.

4.4.10 Evaluation of the Effect of the Proposed CMFA-SR Method

In order to understand the effectiveness of the proposed method, we first examine the effect of the CMFA-SR method using the deep learning features on the MIT-67 dataset. We extract the input CNN features extracted using the Places-CNN [131] on the MIT-67 dataset. The proposed method then processes these input CNN features to obtain the CMFA-SR features. Finally, the SVM classifier is used for classification. Table 4.9 shows

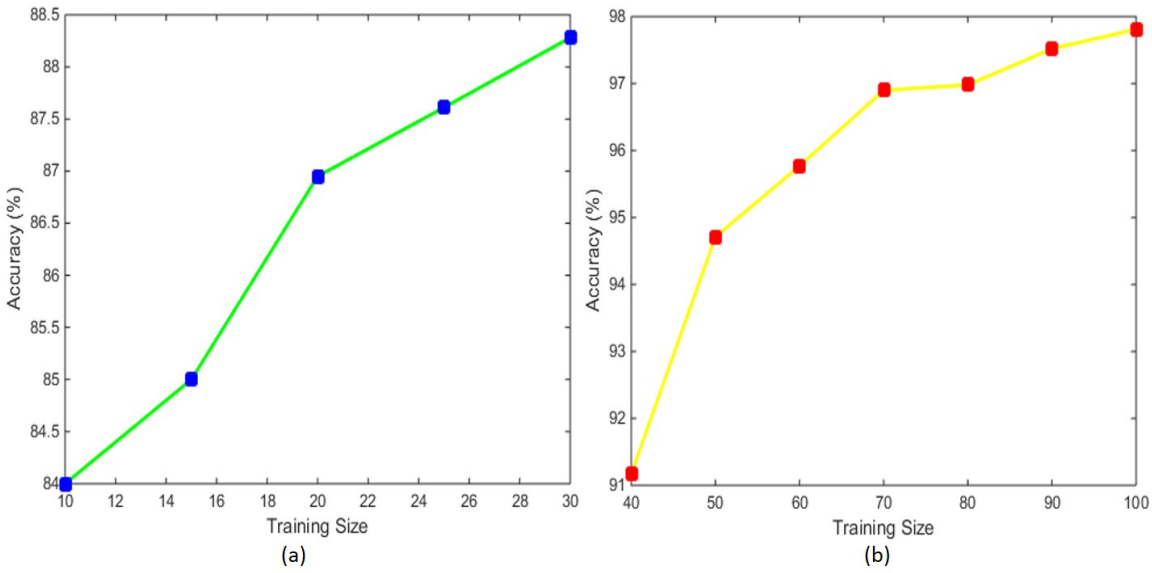

Figure 4.4 The performance of the proposed CMFA-SR method when the size of the training data varies on (a) Caltech 101 dataset (b) 15 scenes dataset.

Table 4.10 Comparative Evaluation of the Proposed CMFA-SR Features and the Hand Crafted Features Using the Painting-91 Dataset (artist classification task)

Method	Accuracy (%)
Fisher Vector features [84]	59.04
CMFA-SR features	**65.78**

the comparative evaluation of the proposed method and the deep learning method [131]. Specifically, our proposed method improves upon the performance of the deep learning method by a large margin.

To demonstrate the general importance of our proposed method, we conduct additional experiments on the Painting 91 dataset (artist classification task). The input features used are Fisher vector features computed as described in [84]. We then apply the proposed method to extract the CMFA-SR features and the final classification is performed by using the SVM classifier with the RBF kernel. Table 4.10 shows that our proposed

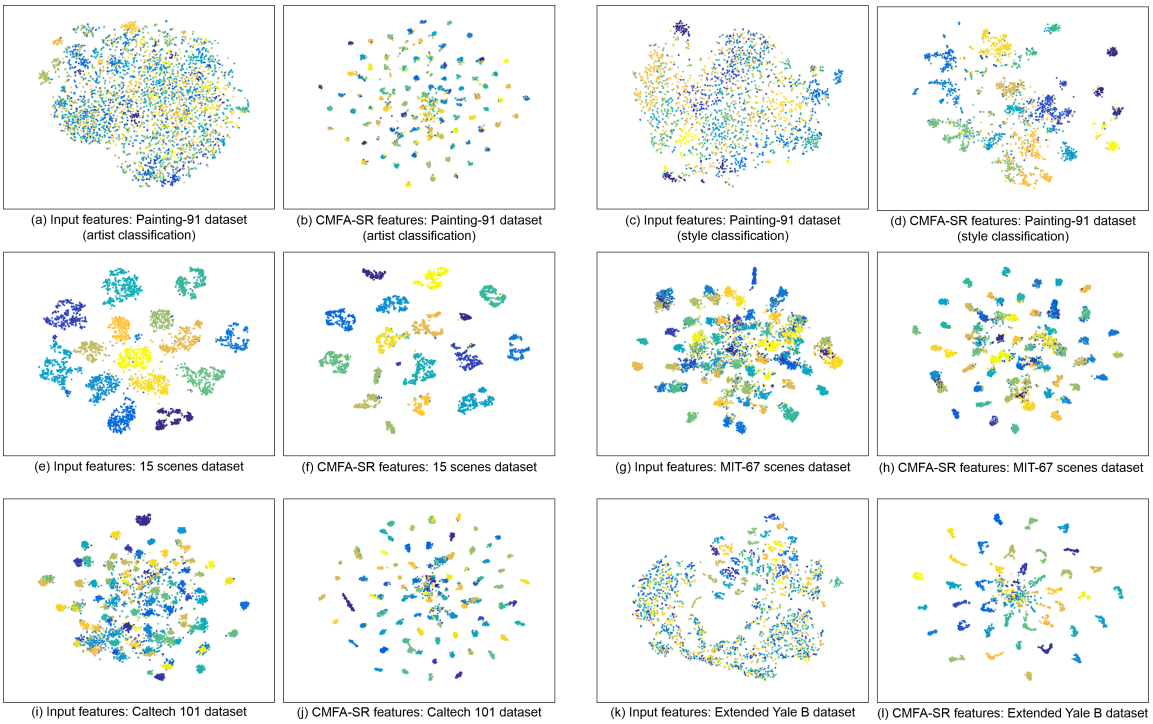

Figure 4.5 The t-SNE visualization of the initial input features and the features extracted after applying the proposed CMFA-SR method for different datasets.

method achieves the classification accuracy of 65.78%, compared to only 59.04% by the Fisher vector features method.

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 [72] that produces visualization of high dimensional data in scatter plots. Figure 4.5 shows the t-SNE visualizations of the initial features used as input and the features extracted after applying the CMFA-SR method for different datasets. It can be seen from Figure 4.5 that the proposed CMFA-SR method helps to reduce the distance among the data points of the same class, which leads to the formation of higher density clusters for these data points. Meanwhile the CMFA-SR method also helps increase the distance among the clusters of different classes resulting in better discrimination among them. Applying two types of discriminatory information, coupled with a discriminative sparse representation

Table 4.11 Evaluation of the Contribution of Individual Steps in the Proposed CMFA-SR Method Using the MIT-67 Indoor Scenes Dataset

Method	Accuracy (%)
Places-CNN (input features) [131]	68.24
CMFA features only (only subspace learning)	73.96
Dictionary learning features only	76.19
CMFA-SR features	**81.12**

Table 4.12 Evaluation of the Dictionary Screening Rule on the Caltech 101 Dataset with the Dictionary Size of 256, 512, and 1024

Method	256	512	1024
CMFA-SR without screening rule	0.45	2.62	5.78
CMFA-SR with screening rule	0.40	2.05	3.84

model, our proposed CMFA-SR method, which leads to better separation among the data samples from different classes, thus improves recognition performance.

To evaluate the contribution of the individual steps to the overall recognition rate, we conduct experiments on the MIT-67 dataset using the input CNN features extracted from the Places-CNN [131] as specified in [131]. Table 4.11 shows the performance evaluation of the individual steps in the proposed CMFA-SR method. Specifically, the CMFA-SR features (both CMFA and dictionary learning) achieves the best classification accuracy of 81.12% since it incorporates both the discriminatory features extracted using the CMFA method and the discriminative dictionary learning.

Table 4.13 Comparative Evaluation of the Proposed CMFA-SR Method with and without the Dictionary Screening Rule for the Dictionary Size 1024 Using the Caltech 101 Dataset

Method	Accuracy (%)
CMFA-SR without screening rule	88.49
CMFA-SR with screening rule	88.20

4.4.11 Evaluation of the Dictionary Screening Rule

We evaluate the performance of the proposed CMFA-SR method with and without the dictionary screening rule to understand the effectiveness of the screening rule. In particular, the performance is evaluated by calculating the average training time (s/per image), which is determined by dividing the total train time with the training sample size. The assessment is performed on the Caltech 101 dataset with the same settings as provided in the experiments section. Table 4.12 provides the average training time per image of the CMFA-SR method with and without dictionary screening rule for different dictionary sizes of 256, 512 and 1024 on the Caltech 101 dataset. It can be observed that the training time significantly reduces as the dictionary size of the CMFA-SR method increases. The training time efficiency is marginal for small dictionary sizes but for the dictionary size 1024, the screening rule improves the average training time per image by almost 33%. Table 4.13 shows the performance comparison of the proposed CMFA-SR method, with and without the screening rule for the dictionary size 1024 using the Caltech 101 dataset. It can be seen that there is a marginal loss of performance of less than 0.5% for the proposed method with the screening rule but it provides a significant improvement in the average training time by almost 33%.

4.4.12 Comparison with the L2 Norm Regularizer

We compare the proposed method with the L1 (sparsity regularizer) and L2 norm on the Painting-91 dataset and the 15 scenes dataset, respectively. The same input features

Table 4.14 Comparison of the Proposed Method with L1 and L2 Norm using the Painting-91 and 15 Scenes Dataset

dataset	Method	Accuracy (%)
Painting-91	Proposed method with L2 norm	59.82
Artist Cls. Task	**Proposed method with L1 norm**	**65.78**
Painting-91	Proposed method with L2 norm	64.32
Style Cls. Task	**Proposed method with L1 norm**	**73.16**
15 Scenes	Proposed method with L2 norm	92.26
	Proposed method with L1 norm	**98.45**

are used for the two datasets as described in Sections 4.4.1 and 4.4.2. The L2 norm based method is optimized using stochastic gradient decent algorithm and the RBF-SVM classifier is used for the final classification. Experimental results in Table 4.14 show that the L1 norm performs better than the L2 norm by a margin of between 5% and 8%. The L2 norm based method, even though possesses good analytical properties due to its differentiability, does not encourage model compression and removal of irrelevant features, which can be crucial for high-dimensional data. The L1 norm based method implicitly filters out a lot of noise from the model as well as stabilizes the estimates if there is high collinearity between the features resulting in a better generalized model. Another advantage of the L1 norm based method is that it is less sensitive to outliers, and therefore improves the pattern recognition performance.

4.5 Conclusion

We have presented in this chapter a complete marginal Fisher analysis (CMFA) method that extracts the discriminatory features in both the column space of the local samples based within class scatter matrix and the null space of its transformed matrix. We have also presented a discriminative sparse representation model by integrating a representation

criterion, such as the sparse representation, and a discriminative criterion, which applies the new within-class and between-class scatter matrices based on the marginal information, for improving the classification capability. We have finally proposed the largest step size for learning the sparse representation to address the convergence issues in optimization, and a dictionary screening rule to purge the dictionary items with null coefficients for improving the computational efficiency. Our experiments on different visual recognition tasks using representative datasets show the feasibility of our proposed method.

CHAPTER 5

DISCRIMINATIVE DICTIONARY DISTRIBUTION BASED SPARSE CODING

5.1 Introduction

Several machine learning and computer vision techniques have been broadly applied for different visual recognition tasks such as face recognition [110, 129, 40, 122, 113, 57, 119, 58], scene classification [106, 87, 27], and object classification [82, 110, 99]. However, in order to accurately classify images, a discriminative and robust representation is needed to capture the important aspects of the image. A major issue in computer vision applications is the high dimensionality of the image feature vector which can make the learning tasks more difficult and can have a dramatic impact on the performance. To solve this issue, sparse coding algorithms [61, 117, 113] have been widely used for data modeling by learning a dictionary that is adapted to the data to improve the feature representation. Sparse coding allows efficient retrieval of data as it generates sparse representations such that every data point can be represented as a linear combination of a small set of basis vectors. Another advantage is that the sparse representation can be overcomplete, allowing more flexibility in matching data and yielding a better approximation of the statistical distribution of the data.

Although the sparse representation method achieves impressive results in various challenging tasks, a potential limitation is the lack of dictionary distribution information since the dictionary is only derived from the representation criterion. The generative perspective remains ignored due to the intrinsic difficulty of estimating the class conditional probability accurately. The generative criterion models the data distribution and infers joint representations which may significantly affect the performance of the learning system. Another limitation in the conventional sparse representation criterion is the lack of discriminative criterion which helps to enhance the discrimination among data samples

of different categories. Previous works of research by [76, 37] show the complementary nature of discriminative and generative approaches and demonstrate the effectiveness of combining both the approaches.

To address these limitations, we present a discriminative dictionary distribution based sparse coding (DDSC) method in this chapter. Specifically, the dictionary distribution 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. To further improve the classification capability, we add a discriminative criterion that utilizes the underlying topology of the sparse representation by considering only the k nearest neighbors for defining a discriminant analysis criterion. In addition, we propose a new classification procedure that utilizes both the derived sparse representation and the dictionary distribution coefficients.

The proposed DDSC method iteratively updates the sparse representation, the dictionary and the dictionary distribution coefficients. In particular, the sparse representation is derived by using the FISTA algorithm [6], and the dictionary is constructed using a fast approximation and the Lagrange dual method. The effectiveness of the proposed DDSC method is evaluated on various visual recognition tasks, such as object recognition on the Caltech 256 dataset[31], computational fine art analysis on the Painting-91 dataset [44], scene recognition on the 15 scenes dataset [50] and the MIT-67 indoor scenes dataset [90], as well as face recognition on the AR face database [74] and the extended Yale face database B [52]. The experimental results show the feasibility of the proposed method.

5.2 Discriminative Dictionary Distribution based Sparse Coding (DDSC)

In this section, we derive a novel sparse representation model by exploiting both the discriminative and the dictionary distribution information to improve the classification performance. Dictionary learning plays a crucial role in the conventional sparse representation method.

Our proposed DDSC method explicitly models the class conditional probability of each dictionary item $p(\mathbf{d}_j|c)$, where $\mathbf{d}_j$ is j-th the dictionary item and c is the class label, and introduces a new discriminative criterion for enhancing the discriminative power of the dictionary. Given the training sample data matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ that contains m samples $[\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m]$, and each sample resides in the 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 DDSC method derives the sparse representation $\mathbf{w}_i \in \mathbb{R}^k (i = 1, 2, ..., m)$ for each training sample $\mathbf{x}_i$ ($\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m]$), and the dictionary distribution coefficients $\mathbf{v}_j \in \mathbb{R}^m (j = 1, 2, ..., k)$ for each dictionary item $\mathbf{d}_j$.

Specifically, the DDSC method is defined as follows:

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{W},\mathbf{V}} \{\sum_{i=1}^{m} ||\mathbf{x}_i - \mathbf{D}\mathbf{w}_i||^2 + \lambda||\mathbf{w}_i||_1\} + \gamma L(\mathbf{V}, \mathbf{D}) + \alpha H(\mathbf{W}) \\ &s.t. \quad ||\mathbf{d}_j|| \leq 1, (j = 1, 2, ..., k) \end{aligned} \tag{5.1}$$

The first term in Equation 5.1 is the conventional sparse representation criterion, where the parameter λ controls the L_1 normalization.

The second term $L(\mathbf{V}, \mathbf{D})$ is the dictionary distribution criterion, which is defined as follows:

$$L(\mathbf{V}, \mathbf{D}) = \sum_{j=1}^{k} ||\mathbf{d}_j - \mathbf{X}\mathbf{v}_j||^2 + \sigma||\mathbf{v}_j - \eta\mathbf{p}_j||^2 \tag{5.2}$$

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 $\mathbf{x}_i$, which is computed as follows:

$$p_{ji} = exp\{-\frac{1}{2h^2}||\mathbf{d}_j - \mathbf{x}_i||^2\} \tag{5.3}$$

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

The traditional view of the dictionary learning is to represent the training sample as a linear combination of the dictionary items. The dictionary items and the training samples consist of a bipartite graph and they influence each other mutually. In addition, the generative criterion also adds a constraint on the dictionary distribution coefficients vector $\mathbf{v}_j$ such that 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 $p(\mathbf{d}_j|c)$.

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

$$H(\mathbf{W}) = \mathbf{tr}(\beta \mathbf{S}_w^{'} - (1-\beta)\mathbf{S}_b^{'}) \tag{5.4}$$

where the new within-class scatter matrix is defined as $\mathbf{S}_w^{'} = \sum_{i=1}^{m} \sum_{(\mathbf{w}_i, \mathbf{w}_j) \in T_k^w} (\mathbf{w}_i - \mathbf{w}_j)(\mathbf{w}_i - \mathbf{w}_j)^t$, and 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 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 different classes.

This discriminative criterion utilizes the underlying topology of the sparse representation of 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 to $H(\mathbf{W}) = \mathbf{tr}(\mathbf{WLW}^t)$, 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 different classes, and $W_b(i,j) = 0$

otherwise. And, 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.

5.3 Optimization Procedure

In this section, we discuss the optimization procedure of the proposed DDSC method. The objective function in Equation 5.1 is optimized using 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 [51], while the dictionary distribution coefficients $\mathbf{v}_j$ are initialized using the value of $\eta \mathbf{p}_j$.

First, given the dictionary $\mathbf{D}$ and the dictionary distribution coefficients $\mathbf{V}$, the sparse representation $\mathbf{W}$ for each training sample $\mathbf{x}_i$ can be obtained by rewriting the objective function defined in Equation 5.1 as follows.

$$\min_{\mathbf{w}_i} ||\mathbf{x}_i - \mathbf{D}\mathbf{w}_i||^2 + \alpha L_{ii} \mathbf{w}_i^t \mathbf{w}_i + \alpha \mathbf{w}_i^t \mathbf{h}_i + \lambda ||\mathbf{w}_i||_1; \tag{5.5}$$

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 $\mathbf{L}$. We then apply the FISTA algorithm [6] to learn the sparse representation $\mathbf{w}_i$ for each training sample $\mathbf{x}_i$.

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.

$$\mathbf{v}_j = (\mathbf{X}^t \mathbf{X} + \sigma \mathbf{I})^{-1} (\mathbf{X}^t \mathbf{d}_j + \sigma \eta \mathbf{p}_j) \tag{5.6}$$

where $\mathbf{X}^t \mathbf{d}_j$ 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 $\mathbf{v}_j$ significantly helps to derive the dictionary as shown in the following sub-section.

Third, after learning 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.

$$\begin{aligned} &\min_{\mathbf{D}} ||\mathbf{X} - \mathbf{DW}||^2 + \gamma(||\mathbf{D} - \mathbf{XV}||^2 + \sigma||\mathbf{V} - \eta\mathbf{P}||^2) \\ &s.t. \quad ||\mathbf{d}_j|| \leq 1, (j = 1, 2, ..., k) \end{aligned} \tag{5.7}$$

where $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_k]$. The optimization of Equation 5.7 is not a trivial problem due to the exponential form of the vector $\mathbf{p}_j$ with respect to $\mathbf{d}_j$. We seek a more efficient approximation to derive the dictionary instead of using some generic solvers. It is based on the observation from Equation 5.6 that the coefficients of the nearest neighbors of the dictionary items are sufficient for an efficient approximation since the dictionary distribution coefficient vector $\mathbf{v}_j$ represents a similarity measure between the training samples and the dictionary items. Specifically, the approximation method consists of the following steps. (i) The influence of distant training samples are diminished by setting the elements whose absolute value is less than a threshold in $\mathbf{v}_j$ to zero. The resulting new vector is denoted as $\bar{\mathbf{v}}_j$. (ii) The dictionary is then derived by solving the following new optimization problem.

$$\begin{aligned} &\min_{\mathbf{D}} ||\mathbf{X} - \mathbf{DW}||^2 + \gamma||\mathbf{D} - \mathbf{X}\bar{\mathbf{V}}||^2 \\ &s.t. \quad ||\mathbf{d}_j|| \leq 1, (j = 1, 2, ..., k) \end{aligned} \tag{5.8}$$

Table 5.1 Description of the Datasets Used for Evaluation of the Proposed Method

Task	Dataset	#Samples
Object recognition	CalTech 256 [31]	30,607
Scene recognition	MIT-67 indoor scenes [90]	15,620
Scene recognition	15 scenes [50]	4,485
Fine art analysis	Painting-91 dataset [44]	4266
Face recognition	AR face [74]	4,000
Face recognition	Extended Yale face B [52]	2,414

where $\bar{\mathbf{V}}$ is a matrix containing $\bar{\mathbf{v}}_j$. This problem is a constrained optimization problem with inequality constraints, which is solved using the Lagrange optimization and the Karush-Kuhn-Tucker conditions [51].

5.4 Classification Procedure

After the dictionary **D** and the dictionary distribution coefficients **V** are derived, we present a new discriminative dictionary distribution based sparse coding classification (DDSCc) method. In particular, for the test data **y**, we derive sparse representation by optimizing the following criterion:

$$\min_{\mathbf{w}} \left\{ ||\mathbf{y} - \mathbf{D}\mathbf{w}||^2 + \lambda ||\mathbf{w}||_1 \right\} \tag{5.9}$$

where the representation $\mathbf{w} = [w_1, w_2, ..., w_k]^t$ contains both the generative and the discriminative information, as the dictionary **D** is learned during the training optimization process.

The DDSCc method is then applied based on the derived discriminative dictionary distribution based sparse coding **w** and the dictionary distribution coefficients **v**. Specifically,

the DDSCc method is defined as follows.

$$c^* = \arg\max_c \sum_{j=1}^{k} w_j \sum_{\mathbf{x}_i \in \mathbf{X}_c} v_{ji} \tag{5.10}$$

Note that we only select the top T largest values of v_{ji} for the DDSCc method.

5.5 Experiments

To evaluate the effectiveness of the proposed DDSC method, we conduct experiments on different visual recognition tasks, namely object recognition, scene recognition, face recognition, and computational fine art analysis. In particular, the datasets used for evaluating the proposed DDSC method are listed in Table 5.1. The parameters for the dictionary distribution criterion are selected as $\gamma = 0.05$, $\sigma = 0.05$ and $\eta = 0.1$ for all the datasets. We also present additional comprehensive analysis to further investigate the properties of the proposed method.

5.5.1 Scene Recognition

The 15 Scenes Dataset The 15 scenes dataset [50] contains 4485 images from 15 scene categories, each with the number of images ranging from 200 to 400. Following the experimental protocol defined in [50], 100 images per class are randomly selected for training and the remaining for testing for 10 iterations. First, the spatial pyramid features provided by [40], 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. The dimension is then reduced to 1000 and the size of the dictionary is 1024. The model parameters are selected as follows: $\lambda = 0.05$, $h = 0.1$, $\alpha = 0.1$, $\beta = 0.5$, and $k = 100$ for the DDSCc method. The results shown in Table 5.2 demonstrate that the proposed method is able to achieve better results compared to other learning methods.

Table 5.2 Comparison with Other State-of-the-art Methods on the 15 Scenes Dataset

Methods	Accuracy %
LLC [110]	89.20
D-KSVD [129]	89.10
LC-KSVD1 [40]	90.40
LC-KSVD2 [40]	92.90
LaplacianSC [27]	89.70
DHVFC [29]	86.40
VGG16-Place365 [130]	92.15
DDSC	**98.75** $\pm$ 0.15

The MIT-67 Indoor Scenes Dataset The MIT-67 indoor scenes dataset [90] is a very challenging indoor scene recognition dataset containing 15620 images with 67 classes. We use the experimental settings defined in [90], wherein 80 images per class are used for training and 20 images per class are used for testing. The initial input features are selected from a pretrained VGG16 CNN model [130] and the feature dimension is reduced from 4096 to 3500. The dictionary size is selected as 2048. The model parameters are selected as $\lambda = 0.05$, $h = 0.01$, $\alpha = 0.1$, $\beta = 0.5$ and $k = 75$ for DDSCc method. The results shown in Table 5.3 demonstrate that the proposed method achieves better results compared to other popular learning methods.

5.5.2 Computational Fine Art Analysis

The Painting-91 dataset [44] contains 4266 fine art painting images by 91 artists. 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) cubbism, (5) impressionism,

Table 5.3 Comparison with Other State-of-the-art Methods on the MIT-67 Indoor Scenes Dataset

Methods	Mean Accuracy %
ROI + Gist [90]	26.10
Object Bank [55]	37.60
miSVM [54]	46.40
D-Parts [102]	51.40
DP + IFV [42]	60.80
D3 [114]	78.13
VGG16-Place365 [130]	76.53
DDSC	**82.97**

(6) neoclassical, (7) popart, (8) post-impressionism, (9) realism, (10) renaissance, (11) romanticism, (12) surrealism, and (13) symbolism.

The initial input features used are Fisher vector features extracted as described in [86]. We follow the experimental protocol in [44] having two tasks, namely artist classification and style classification. 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: $\lambda = 0.05$, $h = 0.1$, $\alpha = 0.1$, and $\beta = 0.5$. k is set as 25 for the DDSCc 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: $\lambda = 0.05$, $h = 0.1$, $\alpha = 0.1$, $\beta = 0.5$, and $k = 40$ for DDSCc method. Experimental results in Table 5.4 show that our proposed DDSC method outperforms other popular methods in both the artist and style classification tasks.

Table 5.4 Comparison with Other Popular Methods on the Painting-91 Dataset

Feature	Artist Cls.	Style Cls.
RGBSIFT [104, 44]	40.30	47.40
CSIFT [104, 44]	36.40	48.60
CN-SIFT [44]	44.10	56.70
Combine(1 - 14) [44]	53.10	62.20
CNN F_3 [108]	56.40	68.57
CNN F_4 [108]	56.35	69.21
MSCNN-1 [81]	58.11	69.67
MSCNN-2 [81]	57.91	70.96
DDSC	**66.59**	**75.09**

5.5.3 Object Recognition

The Caltech 256 dataset [31] is an extended version of the Caltech 101 dataset and a more challenging object classification dataset containing 30607 images from 256 categories. We follow the experimental protocol defined in [110] where the entire dataset is partitioned randomly into 30, 45 and 60 training data samples per category and at the most 25 test data samples per category for 3 iterations. The initial input features used are extracted from a pre-trained ZFNet [99] resulting in feature vector with dimension 4096. We further reduce the dimension to 2000 using PCA. The performance is evaluated by calculating the average classification accuracy over all the categories. For the DDSC 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. k is set as 60 for the DDSCc method. Experimental results in Table 5.5 show that our proposed method achieves better results compared to other methods.

Table 5.5 Comparison Between the Proposed Method and Other Popular Methods on the Caltech 256 Dataset

Methods	30	45	60
IFK [82]	40.80	45.00	47.90
LLC [110]	41.19	45.31	47.68
M-HMP [8]	48.00	51.90	55.20
ZFNet CNN [99]	70.60	72.70	74.20
DDSC	**72.39**	**75.13**	**76.90**

5.5.4 Face Recognition

Extended Yale face database B The extended Yale face database B consists of 2414 face images from 38 individuals each with around 64 images taken under various lightening conditions. A cropped version of the database [52] is often applied, where all the images are manually aligned, cropped, and then re-sized to 168×192 .

Two experimental settings are applied for fair comparison. First, we follow the experimental setting [122] that 20 images are randomly selected for training for each subject, and the remaining images (around 44 per subject) are used for testing for 10 iterations. To show the robustness of our proposed method, we present results of our DDSC method under an extremely noisy condition, where the random faces [113] are used as the input. Specifically, the random faces [113] 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 with a dimension of 504. Each row of the transformation matrix is normalized to unit length. Then the dimension is reduced to 350 and the dictionary size is set as 512. The model parameters are selected as follows: $\lambda = 0.1$ for the sparse representation criterion, $h = 0.1$ for the dictionary distribution criterion, $\alpha = 0.5$, and $\beta = 0.5$ for the discriminative criterion. k is set as 20 for the DDSCc method.

Table 5.6 Comparison with Other Popular Learning Methods on the Extended Yale Face Database B

Experimental setting 1	Accuracy %
D-KSVD [129]	75.30
SRC [113]	90.00
FDDL [122]	91.90
DDSC	**95.19**
Experimental setting 2	Accuracy %
LLC [110]	90.70
D-KSVD [129]	94.79 $\pm$ 0.49
LC-KSVD1 [40]	93.59 $\pm$ 0.54
LC-KSVD2 [40]	95.22 $\pm$ 0.61
FDDL [122]	96.07 $\pm$ 0.64
SRC [113]	96.32 $\pm$ 0.85
DDSC	**97.45** $\pm$ 0.40

Second, we follow the experimental setting described in [2], [40] where half images are randomly selected for training for each subject, and the remaining images are used for testing for 10 iterations. The input features used are random faces and the dimension of the representation vector is reduced from 504 to 350. The dictionary size is set as 512. The model parameters are selected as follows: $\lambda = 0.05$, $h = 0.1$, $\alpha = 0.1$, and $\beta = 0.5$. k is set to 20 for the DDSCc method. The final results shown in Table 5.6 demonstrate the effectiveness of the proposed method under such a noisy condition.

AR face database The AR face database contains 4000 frontal view images for 126 individuals with 26 images per person. We follow the experimental protocol as described

Table 5.7 Comparison with Other Popular Methods on the AR Face Database

Experimental setting 1	Accuracy %
D-KSVD [129]	95.00
SRC [113]	97.50
LC-KSVD2 [40]	97.80
FDDL [122]	96.22
DDSC	**98.50**
Experimental setting 2	Accuracy %
D-KSVD [129]	85.40
LC-KSVD [40]	89.70
JDL [132]	91.70
FDDL [122]	92.00
SRC [113]	94.99
DDSC	**96.29**

in [74] where 50 male subjects and 50 female subjects are chosen. The images are cropped to size 165*120.

The first experimental setting is defined in [40], [129], 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 [113], [40] with 540 dimensions are applied for fair comparison. Then the dimension is reduced from 540 to 400 and the size of the dictionary is set as 512. The model parameters are selected as follows: $\lambda = 0.1$, $h = 0.1$, $\alpha = 0.5$, and $\beta = 0.5$. k is set as 15 for the DDSCc method.

The second experimental setting is defined in [113], [122] 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

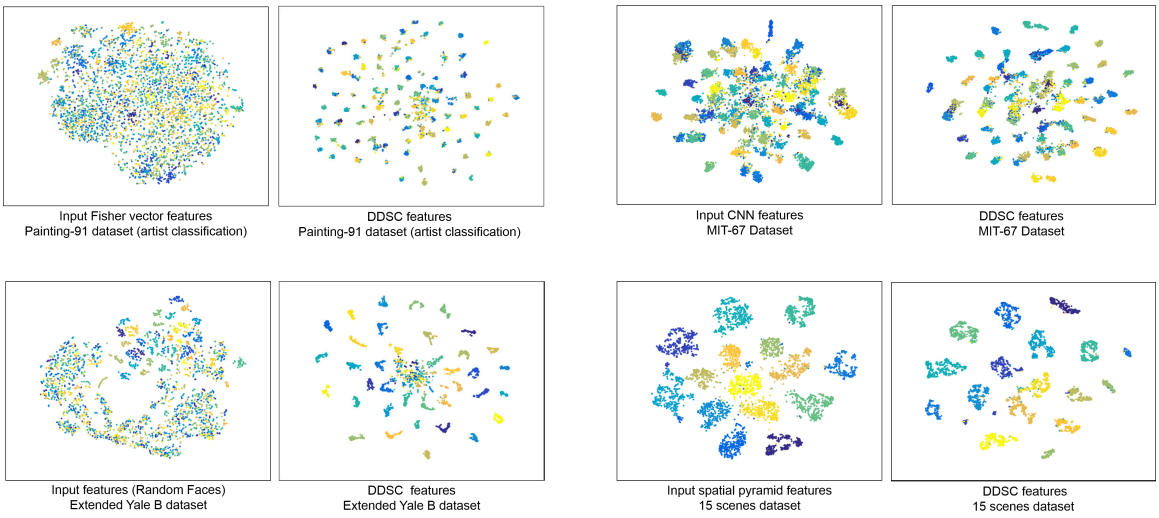

Figure 5.1 The t-SNE visualization of the initial input features and the features extracted after applying the proposed DDSC method.

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: $\lambda = 0.05$, $h = 0.1$, $\alpha = 0.5$, $\beta = 0.5$, and $k = 7$ for the DDSCc method. The experimental results presented in Table 5.7 show that the our DDSC method is able to improve upon the other popular methods under all the three experimental settings.

5.5.5 Evaluation of the Effect of the Proposed DDSC Method

To evaluate the contribution of the individual criterion to the overall classification accuracy, we conduct experiments on the MIT-67 dataset using the initial input features as described in the Experiments Section 5.5.1. In order to have a fair comparison, we use the RBF-SVM classifier for classification instead of the DDSCc method since it depends on both the dictionary distribution and discriminative criteria. It can be seen from Table 5.8 that the DDSC method (both discriminative and dictionary distribution criteria) achieves the best performance of 80.67% since it incorporates both the discriminative and the dictionary distribution information.

Table 5.8 Evaluation of the Contribution of Generative and Discriminative Criterion in DDSC Method Using the MIT-67 Scenes Dataset

Method	Accuracy (%)
DDSC with discriminative criterion	77.24
DDSC with dictionary distribution criterion	78.51
Proposed DDSC (both criteria)	**80.67**

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 [72] that produces visualization of high dimensional data in scatter plots. Figure 5.1 shows the t-SNE visualizations of the initial features used as input and the features extracted after applying the DDSC method for different datasets. It can be seen from Figure 5.1 that the proposed DDSC method helps to reduce the distance between images of the same class leading to formation of higher density clusters for images of the same class. Another advantage is that the DDSC method assists to increase the distance between clusters of different classes resulting in better discrimination among them. The DDSC method uses both the dictionary distribution and discriminative information, therefore, encourages better separation between data samples of different classes.

CHAPTER 6

MULTIPLE ANTHROPOLOGICAL FISHER KERNEL LEARNING

6.1 Introduction

Kinship verification is a challenging task as the correlated visual resemblance between parents and their offspring have to be captured. In order to effectively classify kinship relations, the genetic features between parent and child have to be enhanced and encoded in the feature representation. Many feature representation methods such as LBP [1], Gabor features [59], Fisher vector [98], learning-based (LE) descriptor [11], etc. have been proposed for representing face images. But these methods are not explicitly designed in order to capture and enhance the similarities and genetic relations between parent and child images. Another issue is that unlike traditional face recognition problem, the similarity gap between kinship images is much larger specifying the need for more powerful visual features.

To address these issues, this paper proposes a novel SIFT flow based genetic Fisher vector feature with applications to kinship verification. We enhance the genetic inheritable features of parent and child image in kinship relations by matching densely sampled SIFT features and visual correspondence between them using the SIFT flow algorithm [60]. We analyze and correlate the enhanced genetic features to the anthropological results and find interesting patterns in different kinship relations. We then apply an inheritable transformation with the objective of pushing the non-kinship samples as far as possible and pulling the kinship samples as close as possible. The experimental results on the two challenging kinship databases, the KinFace W-I and the Kinship W-II dataset [70] show the effectiveness of the proposed method. The framework of our proposed method is illustrated in Figure 6.1.

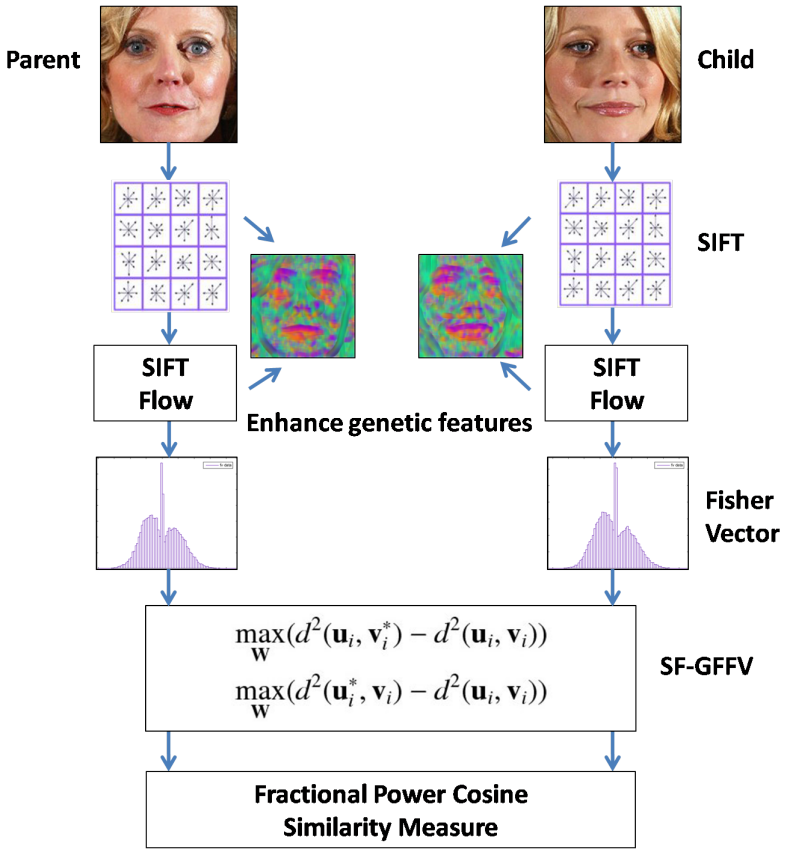

Figure 6.1 The framework of our proposed SF-GFVF feature.

6.2 SIFT Flow based GFVF Framework

6.2.1 SIFT Flow based Similarity Enhancement Method

We present a novel similarity enhancement method by extending the SIFT flow algorithm [60] for kinship images so as to find inheritable feature relations between the kinship images and enhance the similarities between them. The SIFT flow algorithm matches the densely sampled SIFT features and finds the correspondence estimated by SIFT flow. It can be formulated similarly as the optical flow wherein SIFT descriptors are matched instead of the pixel to pixel correspondences between two images. The SIFT flow is based on the criteria that the SIFT descriptors are matched along the flow vectors and the flow field is

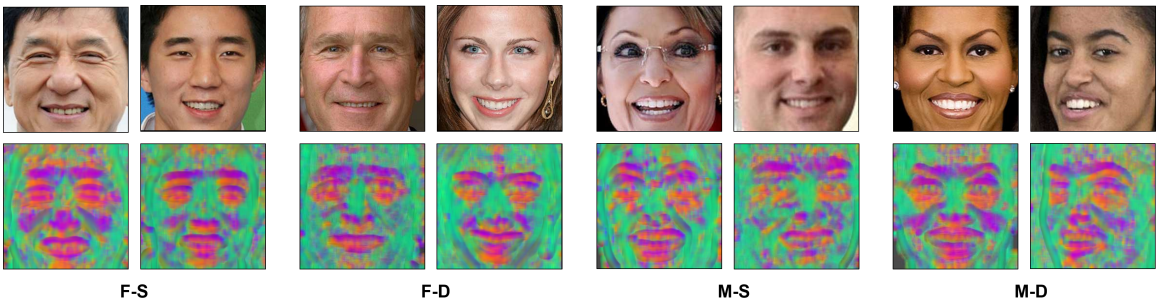

Figure 6.2 Visualization of SIFT images of different kinship relations using the top three principal components of SIFT descriptors.

smooth [60]. The energy function for SIFT flow [60] is defined as follows:

$$E(\mathbf{w}) = \sum_{\mathbf{p}} \min(\|s_1(\mathbf{p}) + s_2(\mathbf{p}+\mathbf{w}(\mathbf{p}))\|_1, t) + \sum_{\mathbf{p}} \eta(|u(\mathbf{p}) + v(\mathbf{p})|) + \sum_{\mathbf{p},\mathbf{q}\epsilon\varepsilon} \min(\alpha|u(\mathbf{p}) + u(\mathbf{q})|, d) + \min(\alpha|v(\mathbf{p}) + v(\mathbf{q})|, d) \tag{6.1}$$

where $\mathbf{p} = (x, y)$ are the grid coordinate of images, $\mathbf{w}(\mathbf{p}) = (u(\mathbf{p}), v(\mathbf{p}))$ is the flow vector at $\mathbf{p}$, s_1, s_2 are the two SIFT images to be matched and ε contains all the spatial neighborhoods.

To visualize the SIFT images, the top three principal components of the SIFT image are mapped to the principal components of the RGB space, as shown in Figure 6.2. The purple and the orange regions in the visualization highlight the inheritable genetic feature regions in the kinship images. Our objective is to enhance these genetic regions in the kinship images. For a query parent-child image pair, the SIFT flow is applied to match dense correspondences between the parent and the child SIFT descriptors. If the image pair is in kinship relation, the genetic facial regions are enhanced by adding weights to those specific facial regions.

Our proposed similarity enhancement method results in interesting phenomena that correlate the enhanced genetic features to the anthropological features. Naini et al. [75] analyzed the contributions of heredity and environment on external facial features. The

relative strength of genetic influence on different facial parameters is assessed using optical surface scanning and twin method. The anthropological results [75] show that eyes, chin and parts of the forehead show higher visual resemblance between parent and their offspring and provide large feedback. The results shown in Figure 6.2 show high correlation to the anthropological results with high feedback in parts of forehead and eye regions. Interesting patterns can be deduced for different relations from Figure 6.2. 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 [4].

6.2.2 Inheritable Genetic Transformation

We first briefly review the Fisher vector method. Fisher vector is widely used for visual recognition problems such as face recognition [98], object recognition [38]. Particularly, let $\mathbf{X} = \{\mathbf{d}_t, t = 1, 2, ..., T\}$ be the set of T local descriptors extracted from the image. Let μ_λ be the probability density function of $\mathbf{X}$ with a set of parameters λ, then the Fisher kernel [38] is defined as follows: $K(\mathbf{X}, \mathbf{Y}) = (\mathbf{G}_\lambda^X)^T \mathbf{F}_\lambda^{-1} \mathbf{G}_\lambda^Y$ where $\mathbf{G}_\lambda^X = \frac{1}{T} \bigtriangledown_\lambda \log[\mu_\lambda(\mathbf{X})]$, which is the gradient vector of the log-likelihood that describes the contribution of the parameters to the generation process. And $\mathbf{F}_\lambda$ is the Fisher information matrix of μ_λ. Essentially, the Fisher vector is derived from the explicit decomposition of the Fisher kernel as the symmetric and positive definite Fisher information matrix $\mathbf{F}_{\boldsymbol{\lambda}}$ has a Cholesky decomposition as $\mathbf{F}_{\boldsymbol{\lambda}}^{-1} = \mathbf{L}_{\boldsymbol{\lambda}}^T \mathbf{L}_{\boldsymbol{\lambda}}$. Therefore, the Fisher kernel $K(\mathbf{X}, \mathbf{Y})$ can be written as a dot product between two vectors $\mathbf{L}_{\boldsymbol{\lambda}} \mathbf{G}_{\boldsymbol{\lambda}}^{\mathbf{X}}$ and $\mathbf{L}_{\boldsymbol{\lambda}} \mathbf{G}_{\boldsymbol{\lambda}}^{\mathbf{Y}}$ which are defined as the **Fisher vectors** of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Fisher vector focuses on the image specific features and discards the image independent features but this does not guarantee enhancement of genetic features in parent and child images.

We therefore learn an inheritable genetic transformation $\mathbf{W}$ on the SIFT flow based genetic Fisher vector $\mathbf{p}_i(i = 1, 2, ..., m)$ and $\mathbf{c}_i(i = 1, 2, ..., m)$ for each training pairs ($\mathbf{p}_i$, $\mathbf{c}_i$) where $\mathbf{p}_i$ denotes the parent image and $\mathbf{c}_i$ denotes the child image. The learned SF-GFVF for the parent and child image are as follows: $\mathbf{u}_i = \mathbf{W}^T\mathbf{p}_i$ and $\mathbf{v}_i = \mathbf{W}^T\mathbf{c}_i$. The objective of learning the inheritable transformation is to minimize the distance between $\mathbf{u}_i$ and $\mathbf{v}_i$ if $\mathbf{u}_i$ and $\mathbf{v}_i$ have kinship relations and maximize the distance otherwise.

Let $\mathbf{D} = \{(\mathbf{u}_i, \mathbf{v}_i)|\mathbf{u}_i, \mathbf{v}_i \in \mathbb{R}^{n\times 1}(i = 1, 2, ..., m)\}$ be the training data that consists of m pairs of SIFT flow based genetic Fisher vector features derived from the kinship images. Therefore, multiple objectives for the SF-GFVF method can be formulated as:

$$\begin{aligned} &\max_{\mathbf{W}}(d^2(\mathbf{u}_i, \mathbf{v}_i^*) - d^2(\mathbf{u}_i, \mathbf{v}_i)) \\ &\max_{\mathbf{W}}(d^2(\mathbf{u}_i^*, \mathbf{v}_i) - d^2(\mathbf{u}_i, \mathbf{v}_i)) \end{aligned} \tag{6.2}$$

where $d^2(\mathbf{u}_i, \mathbf{v}_i) = (\mathbf{p}_i - \mathbf{c}_i)^T\mathbf{W}\mathbf{W}^T(\mathbf{p}_i - \mathbf{c}_i)$, $\mathbf{u}_i^*$ is the nearest neighbor of $\mathbf{u}_i$ and $\mathbf{v}_i^*$ is the nearest neighbor of $\mathbf{v}_i$. Note that there are 2*m objective functions in Equation 6.2 since $i = 1, 2, ..., m$.

In practice, it is difficult to solve a multiple objective problem for high dimensions since it is computationally expensive and a single solution may not exist. Therefore, linear scalarization [36] is applied in order to convert the multi-objective problem into a single objective function with a weighted sum of the individual objective functions. Assuming the same weight λ_i^2 for the objective functions of each training pair $(\mathbf{u}_i, \mathbf{v}_i)$, we want to maximize the following objective function:

$$\begin{aligned} &\max_{\mathbf{W}} \sum_{i=1}^{m} \lambda_i^2(d^2(\mathbf{u}_i, \mathbf{v}_i^*) + d^2(\mathbf{u}_i^*, \mathbf{v}_i) - 2 * d^2(\mathbf{u}_i, \mathbf{v}_i)) \\ &s.t. \sum_{i=1}^{m} \lambda_i = 1, \mathbf{W}^T\mathbf{W} = \mathbf{I} \end{aligned} \tag{6.3}$$

Then objective function in Equation 6.3 can be further simplified as $\mathrm{Tr}\,(\mathbf{W}^T(\mathbf{Q}_1 + \mathbf{Q}_2 - 2\mathbf{Q}_3)\mathbf{W})$ where $\mathbf{Q}_1 = \sum_{i=1}^{m} \lambda_i^2(\mathbf{p}_i - \mathbf{c}_i^*)(\mathbf{p}_i - \mathbf{c}_i^*)^T$. $\mathbf{Q}_2$ and $\mathbf{Q}_3$ can be computed in a similar way.

Then the algorithm of optimizing the objective function in Equation 6.3 is summarized as follows.

Algorithm 1 SF-GFVF Learning Algorithm

Input: Training Images: $\mathbf{D} = \{(\mathbf{u}_i, \mathbf{v}_i)|\mathbf{u}_i, \mathbf{v}_i \in \mathbb{R}^{n\times 1}(i = 1, 2, ..., m)\}$

Output: Inheritable tranformation $\mathbf{W}$

1: **Step 1 (Initialization)**

Initialize $\lambda_i = 1/m$ and $\mathbf{W} = \mathbf{I}$

2: **Step 2 W is fixed, optimize on** λ_i

$$\lambda_i = \frac{f^{-1}(\mathbf{u}_i, \mathbf{v}_i)}{\sum_{i=1}^{m} f^{-1}(\mathbf{u}_i, \mathbf{v}_i)} \tag{6.4}$$

where $f(\mathbf{u}_i, \mathbf{v}_i) = d^2(\mathbf{u}_i, \mathbf{v}_i^*) + d^2(\mathbf{u}_i^*, \mathbf{v}_i) - 2 * d^2(\mathbf{u}_i, \mathbf{v}_i)$

3: **Step 3** λ_i **is fixed, update W**

$$\begin{aligned} &\max_{\mathbf{W}} \mathrm{Tr}(\mathbf{W}^T(\mathbf{Q}_1 + \mathbf{Q}_2 - 2\mathbf{Q}_3)\mathbf{W}) \\ &s.t. \mathbf{W}^T\mathbf{W} = \mathbf{I} \end{aligned} \tag{6.5}$$

4: **Step 4 Continue to Step 2 if not converged**

After the SF-GFVF is derived, principal component analysis with whitening transformation is applied in order to extract the most expressive features. A fractional power cosine similarity measure (FPCSM) is then applied as follows to compute the similarity between two images.

$$FPCSM(\mathbf{u}_i, \mathbf{v}_i) = CS(sign(\mathbf{u}_i)|\mathbf{u}_i|^\alpha, sign(\mathbf{v}_i)|\mathbf{v}_i|^\alpha) \tag{6.6}$$

where $CS(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^T\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$ is the traditional cosine similarity measure and α $(0 < \alpha < 1)$ is the power parameter.

Table 6.1 Comparison Between the SF-GFVF and Other Popular Methods on the KinFaceW-I Dataset

Methods	F-S	F-D	M-S	M-D	Mean
CSML [77]	61.10	58.10	60.90	70.00	62.50
NCA [30]	62.10	57.10	61.90	69.00	62.30
LMNN [112]	63.10	58.10	62.90	70.00	63.30
NRML [70]	64.10	59.10	63.90	71.00	64.30
MNRML [70]	72.50	66.50	66.20	72.00	69.90
ITML [17]	75.30	64.30	69.30	76.00	71.20
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
SF-GFVF	76.27	**74.64**	**75.48**	**79.98**	**76.09**

The linear scalarization optimization procedure may be similar to metric learning methods such as NRML [70] in terms of mathematical formulas but the differences are as follows. (i) Our method uses multiple objective function instead of a common global objective function which helps to prevent dominance of one term in the function over other terms. (ii) Our method enhances the genetic features in kinship images and is proposed from the feature learning point of view and not the metric learning point of view.

6.3 Anthropology Inspired Feature Extraction

Naini et al. [75] analyzed the contributions of heredity and environment on external facial features. Their anthropological results [75] 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 [60] 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 [60] is defined as follows:

$$\begin{aligned} E(\mathbf{d}) = \sum_{\mathbf{p}} (\|s_1(\mathbf{p}) - s_2(\mathbf{p} + \mathbf{d}(\mathbf{p}))\|_1) + \\ \sum_{\mathbf{p}} \eta(\|\mathbf{d}(\mathbf{p})\|_1) + \sum_{\mathbf{p},\mathbf{q} \in \varepsilon} \theta(\|\mathbf{d}(\mathbf{p}) - \mathbf{d}(\mathbf{q})\|_1) \end{aligned} \tag{6.7}$$

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 6.2. 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 [75]. Other interesting patterns can also be deduced for different relations from Figure 6.2. 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 [4].

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 [43] 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.

6.4 Multiple Anthropological Fisher Kernel Framework

The complementary nature of discriminative and generative approach leads to the generative score space. One example is the Fisher score [37], which has been widely applied for visual classification problems such as face recognition [98], object recognition [38]. In this section, we extend the Fisher score from classification problem to metric 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. And $\mathbf{Y}_i$ 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:

$$\mathbf{F}(\mathbf{X}_i) = \frac{1}{T} \bigtriangledown_{\boldsymbol{\lambda}} \log[p(\mathbf{X}_i|\boldsymbol{\lambda})] \tag{6.8}$$

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 [37], 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 paper presents a novel multiple anthropological Fisher kernel framework 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 paper. The introduction of $\mathbf{W}$ alleviates the assumptions on the Fisher information matrix since $\mathbf{W}$ can be learned from the training data and contains sufficient information for recognizing kinship relations.

The derivation of $\mathbf{W}$ 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 $\mathbf{W}$ 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 M-AFK method can be formulated as follows.

$$\begin{aligned} &\min_{\mathbf{W}, w_c} \|D_{\mathbf{I}} - \sum_{c=1}^{k} w_c D_c\|_F^2 + \alpha \sum_{c=1}^{k} w_c^2 + \lambda \sum_{c=1}^{k} d_c |w_c| \\ &s.t. \mathbf{W}^T\mathbf{W} = \mathbf{I}, \sum_{c=1}^{k} w_c = 1, w_c > 0 \end{aligned} \tag{6.9}$$

In this objective function, the third term of Equation 6.9 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 (0.5 for

scaling in our experiment) 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 6.9 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:

$$\begin{aligned} &\max_{\mathbf{W}} \mathrm{Tr}(\mathbf{W}^T \sum_{c=1}^{k} w_c(\mathbf{M}_2^c + \mathbf{M}_3^c - 2\mathbf{M}_1^c)\mathbf{W}) \\ &s.t. \mathbf{W}^T\mathbf{W} = \mathbf{I} \end{aligned} \tag{6.10}$$

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:

$$\begin{aligned} &\min_{w_c} \|D_{\mathbf{I}} - \sum_{c=1}^{k} w_c D_c\|_F^2 + \alpha \sum_{c=1}^{k} w_c^2 + \lambda \sum_{c=1}^{k} d_c|w_c| \\ &s.t. \sum_{c=1}^{k} w_c = 1, w_c > 0 \end{aligned} \tag{6.11}$$

We can apply the FISTA algorithm [6] to optimize the objective function defined in Equation 6.11. 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 [22] considering the non-negative constraint. We can also transform the objective function defined in Equation 6.11 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 M-AFK is derived, a novel normalized multiple similarity measure (NMSM) is further proposed, where the M-AFK 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.

$$NMSM(\mathbf{x}_i, \mathbf{y}_i) = \sum_{c=1}^{k} w_c \frac{D_c(p(\mathbf{x}_i^c), p(\mathbf{y}_i^c))}{\|\mathbf{W}^T p(\mathbf{x}_i^c)\| \|\mathbf{W}^T p(\mathbf{y}_i^c)\|} \tag{6.12}$$

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 [38]. 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. This shows its theoretical roots to the Bayes decision rule for minimum error under some conditions such as the multivariate Gaussian distribution assumption, therefore, provides theoretical guarantee to achieve better performance.

6.5 Experiments

This section demonstrates the performance of our proposed method on two challenging kinship databases: the KinFaceW-I dataset and the KinFaceW-II dataset [70]. There are four kinship relations in both the datasets: father-son (F-S), father-daughter (F-D), mother-son (M-S), and mother-daughter (M-D). In KinFaceW-I dataset, each image pair in the kinship relation was acquired from different photos whereas in KinfaceW-II, they were obtained from the same photo. In the KinFaceW-I dataset, there are 156, 134, 116, and 127 image pairs for each of the relations defined above. In the KinFaceW-II dataset, there are 250 pairs of the images for each relation. In our experiments, we conduct 5-fold cross

Table 6.2 Comparison Between the SF-GFVF and Other Popular Methods on the KinFaceW-II Dataset

Methods	F-S	F-D	M-S	M-D	Mean
CSML [77]	71.80	68.10	73.80	74.00	71.90
NCA [30]	73.80	70.10	74.80	75.00	73.50
LMNN [112]	74.80	71.10	75.80	76.00	74.50
NRML [70]	76.80	73.10	76.80	77.00	75.70
MNRML [70]	76.90	74.30	77.40	77.60	76.50
ITML [17]	69.10	67.00	65.60	68.30	67.50
GGA [18]	81.80	74.30	80.50	80.80	79.40
DGA [18]	83.90	76.70	83.40	84.80	82.20
SF-GFVF	**87.20**	**79.60**	**88.00**	**87.80**	**85.65**

validation where both datasets are divided into five folds having the same number of image pairs [70].

6.5.1 Comparison Between the SF-GFVF and Other Popular Methods

This section presents the comparison between our proposed SF-GFVF method and other state-of-the-art deep learning and metric learning methods. In Tables 6.1 and 6.2, ANTH denotes anthropological results, GGA denotes gated autoencoders and DGA denotes discriminative autoencoders. It can be observed that the result on the KinFace W-II dataset is better than the KinFace W-I dataset due to the availability of more training samples. Another reason is that the KinFace W-II dataset contains kinship images from the same photo therefore helps to reduce the illumination and background noise compared to the KinFace W-I dataset which contains kinship images from the different photos. Experimental results in Tables 6.1 and 6.2 show that our method outperforms deep learning methods [18] and other metric learning based methods.

6.5.2 Comparison Between the SF-GFVF and FV

This section presents the comparison between our proposed SF-GFVF method and the original Fisher vector (FV) [38] method. Experimental results in Table 6.3 show that our proposed SF-GFVF method improves upon the original FV method by approximately 4 percent and 9 percent in the KinFace W-I and KinFace W-II datasets, respectively. The reason is that the original Fisher vector method focuses on image specific features but it does not enhance the genetic features in kinship images. Our method uses the SIFT flow algorithm and inheritable transformation to encode and enhance the facial genetic features in kinship relations.

Table 6.3 Comparison Between the SF-GFVF and Fisher Vector on the KinFaceW-I and KinFaceW-II Dataset

KinFaceW-I	F-S	F-D	M-S	M-D	Mean
FV	75.02	70.56	65.49	78.39	72.37
SF-GFVF	**76.27**	**74.64**	**75.48**	**79.98**	**76.09**
KinFaceW-II	F-S	F-D	M-S	M-D	Mean
FV	80.00	68.60	79.40	78.20	76.55
SF-GFVF	**87.20**	**79.60**	**88.00**	**87.80**	**85.65**

6.5.3 Comparison Between the M-AFK and Other Popular Methods

The experimental results in Tables 6.4 and 6.5 show that our method is able to achieve better performance compared to other multiple feature learning methods. 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 [4]. The reason is that the similarity variation between images of different gender is larger than that of the same gender and our proposed M-AFK method captures such a variation by learning the new transformation and the weights of multiple features.

Table 6.4 Comparison Between the M-AFK and Other Methods on the KinFaceW-I Dataset

Methods	F-S	F-D	M-S	M-D	Mean
LMNN [112]	63.10	58.10	62.90	70.00	63.30
NRML [70]	64.10	59.10	63.90	71.00	64.30
MNRML [70]	72.50	66.50	66.20	72.00	69.90
DGA [18]	76.40	72.50	71.90	77.30	74.50
Polito [69]	85.30	85.80	87.50	86.70	86.30
LIRIS [69]	83.04	80.63	82.30	84.98	82.74
NUAA [69]	86.25	80.64	81.03	83.93	82.96
CNN-Basic [128]	70.80	75.70	79.40	73.40	74.80
CNN-Points [128]	71.80	76.10	84.10	78.00	77.50
M-AFK	**88.15**	**82.49**	**80.62**	**90.95**	**85.55**

Table 6.5 Comparison Between the M-AFK and Other Methods on the KinFaceW-II Dataset

Methods	F-S	F-D	M-S	M-D	Mean
NRML [70]	76.80	73.10	76.80	77.00	75.70
MNRML [70]	76.90	74.30	77.40	77.60	76.50
DGA [18]	83.90	76.70	83.40	84.80	82.20
Polito [69]	84.00	82.20	84.80	81.20	83.10
LIRIS [69]	89.40	83.60	86.20	85.00	86.05
NUAA [69]	84.40	81.60	82.80	81.60	82.50
CNN-Basic [128]	79.60	84.90	88.50	88.30	85.30
CNN-Points [128]	81.90	89.40	92.40	89.90	88.40
M-AFK	**91.40**	**87.20**	**90.80**	**89.80**	**89.80**

6.6 Conclusion

This paper presents a SIFT flow based inheritable Fisher vector feature (SF-GFVF) and a multiple anthropological Fisher kernel framework (M-AFK) for kinship verification. The proposed SF-GFVF feature uses SIFT flow algorithm to enhance the genetic features in kinship images. An inheritable transformation is then applied to the enhanced Fisher vector by optimizing multiple objective functions. For the MAFK method, three new anthropology inspired features are extracted followed by the M-AFK framework. A normalized multiple similarity measure is then applied for effective normalization. Experimental results show that the proposed methods are able to outperform other popular methods for kinship verification.

CHAPTER 7

PLANNED WORK

This dissertation has presented four learning methods for image classification namely, a sparse representation model based on complete marginal Fisher analysis framework (CMFA-SR), a sparse kernel manifold learner (SKML), a discriminative dictionary distribution based sparse coding (DDSC) method and a multiple anthropological Fisher kernel framework (M-AFK). Sparse coding methods allow efficient retrieval of data by learning a dictionary that is adapted to data. The proposed CMFA-SR and DDSC methods uses a discriminative L1-norm regularizer which performs model compression by retaining useful discriminative features and setting null coefficients to irrelevant features which can be crucial in a high dimensional dataspace. Another advantage is that the L1 norm regularizer is less sensitive to outliers leading to a better generalized model. One of the issues of hand-crafted features as well as deep learning features in the high dimensional space is the existence of highly correlated features which may affect the classification performance. The deep learning methods such as convolutional neural networks (CNNs) have a huge network structure which can result in large redundancies in the network [21]. This may lead to the formation of highly similar and redundant features resulting in overfitting that may affect the classification performance. One future direction of work would be to integrate sparse coding method to CNNs to improve the recognition performance.

Our proposed CMFA-SR method currently uses an enhanced MFA method followed by discriminative sparse representation model and RBF-SVM classifier for classification. A single classifier training may be sensitive to the shape of the training data. In order to reduce such sensitivity due to a single model, the author would like to explore ensemble learning methods. Ensemble learning combine predictions from multiple classifiers which may help to reduce overfitting. Another advantage is that it improves the expressibility of

different classifiers in the ensemble resulting in a better approximation of the test label. A set of classifiers can be learned on the derived discriminative sparse coding features using feature selection and data sub-sampling techniques. A majority voting scheme among the different classifiers in the ensemble can then be used in order to predict the label for the test data.

BIBLIOGRAPHY

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

[2] N. Akhtar, F. Shafait, and A. Mian. Discriminative Bayesian dictionary learning for classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[3] U. L. Altintakan and A. Yazici. Towards effective image classification using class-specific codebooks and distinctive local features. *IEEE Transactions on Multimedia*, 17(3):323–332, March 2015.

[4] A. Alvergne, C. Faurie, and M. 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.

[5] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring. In *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 693–696, 2009.

[7] M Belkin and P Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003.

[8] L. Bo, X. Ren, and D. Fox. Multipath sparse coding using hierarchical matching pursuit. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[9] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *International Conference on Image and Video Retrieval*, (CIVR), pages 401–408, 2007.

[10] D. Cai, X. He, K. Zhou, J. Han, and H. Bao. Locality sensitive discriminant analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 708–713, 2007.

[11] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2707–2714. IEEE, 2010.

[12] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2014.

[13] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao. Wld: A robust local image descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1705–1720, Sept 2010.

[14] L. Chen, H. M. Liao, M. Ko, J. Lin, and G. Yu. A new lda-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33(10):1713 – 1726, 2000.

[15] S. Chen and C. Liu. Clustering-based discriminant analysis for eye detection. *IEEE Transactions on Image Processing*, 23(4):1629–1638, 2014.

[16] M. Culjak, B. Mikus, K. Jez, and S. Hadjic. Classification of art paintings by genre. In *International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1634–1639, May 2011.

[17] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 209–216, 2007.

[18] A. Dehghan, E. G. Ortiz, R. Villegas, and M. Shah. Who do i look like? determining parent-offspring resemblance via gated autoencoders. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1757–1764. IEEE, 2014.

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

[20] W. Deng, J. Hu, and J. Guo. In defense of sparsity based face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 399–406, 2013.

[21] M. Denil, B. Shakibi, L. Dinh, and N. de Freitas. Predicting parameters in deep learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2148–2156, 2013.

[22] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l1-ball for learning in high dimensions. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 272–279, 2008.

[23] J. Feng, B. Ni, Q. Tian, and S. Yan. Geometric lp-norm feature pooling for image classification. In *CVPR 2011*, pages 2609–2704, June 2011.

[24] Q. Feng and Y. Zhou. Kernel combined sparse representation for disease recognition. *IEEE Transactions on Multimedia*, 18(10):1956–1968, 2016.

[25] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press Professional, Inc., San Diego, CA, USA, 1990.

[26] B. Fulkerson, A. Vedaldi, and S. Soatto. *Localizing Objects with Smart Dictionaries*, pages 179–192. Berlin, Heidelberg, Springer, 2008.

[27] S. Gao, I. W. H. Tsang, and L. T. Chia. Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):92–104, 2013.

[28] H. Goh, N. Thome, M. Cord, and J. H. Lim. Learning deep hierarchical visual feature coding. *IEEE Transactions on Neural Networks and Learning Systems*, 25(12):2212–2225, 2014.

[29] H. Goh, N. Thome, M. Cord, and J. H. Lim. Learning deep hierarchical visual feature coding. *IEEE Transactions on Neural Networks and Learning Systems*, 25(12):2212–2225, Dec 2014.

[30] J. Goldberger, S. T. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *NIPS*, 2004.

[31] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. CNS-TR-2007-001, EECS, 2007.

[32] T. Guha and R. K. Ward. Learning sparse representations for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1576–1588, 2012.

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

[34] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[35] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[36] C-L Hwang and Abu Syed Md Masud. *Multiple objective decision makingmethods and applications: a state-of-the-art survey*, volume 164. Springer Science and Business Media, 2012.

[37] T. Jaakkola, M. Diekhans, and D. Haussler. Using the fisher kernel method to detect remote protein homologies. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 149–158. AAAI Press, 1999.

[38] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, Sept 2012.

[39] M. Jian and C. Jung. Semi-supervised bi-dictionary learning for image classification with smooth representation-based label propagation. *IEEE Transactions on Multimedia*, 18(3):458–473, March 2016.

[40] Z. Jiang, Z. Lin, and L. 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.

[41] Z. Jiang, G. Zhang, and L. S. Davis. Submodular dictionary learning for sparse coding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3418–3425, 2012.

[42] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 923–930, 2013.

[43] F. Khan, R. Anwer, J. van de Weijer, A. Bagdanov, A. Lopez, and M. Felsberg. Coloring action recognition in still images. *International journal of computer vision*, 105(3):205–221, 2013.

[44] F. Khan, S. Beigpour, J. van de Weijer, and M. Felsberg. Painting-91: a large scale database for computational painting categorization. *Machine Vision and Applications*, 25(6):1385–1397, 2014.

[45] T. Kim and J. Kittler. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):318–327, 2005.

[46] N. Kohli, M. Vatsa, R. Singh, A. Noore, and A. Majumdar. Hierarchical representation learning for kinship verification. *IEEE Transactions on Image Processing*, 26(1):289–302, 2017.

[47] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[48] R. Lan, Y. Zhou, and Y. Y. Tang. Quaternionic weber local descriptor of color images. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(2):261–274, 2017.

[49] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1294–1309, 2009.

[50] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178, 2006.

[51] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 801–808. MIT Press, 2007.

[52] K. Lee, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005.

[53] F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 178–178, 2004.

[54] Q. Li, J. Wu, and Z. Tu. Harvesting mid-level visual concepts from large-scale internet images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 851–858, 2013.

[55] L. Li-jia, S. Li-jia, F. Li, and P. Eric. Object bank: A high-level image representation for scene classification semantic feature sparsification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1378–1386. 2010.

[56] C. 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.

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

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

[59] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. In *IEEE Transactions on Image Processing*, pages 467–476, 2002.

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

[61] L. Liu, C. Shen, L. Wang, A. Hengel, and C. Wang. Encoding high dimensional local features by sparse coding based fisher vectors. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1143–1151, 2014.

[62] Q. Liu and C. Liu. A new locally linear knn method with an improved marginal fisher analysis for image classification. In *The IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–6. IEEE, 2014.

[63] Q. Liu and C. Liu. A novel locally linear knn method with applications to visual recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 28(9):2010–2021, Sept 2017.

[64] Q. Liu, A. Puthenputhussery, and C. Liu. Inheritable fisher vector feature for kinship verification. In *The IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6, Sept 2015.

[65] Q. Liu, A. Puthenputhussery, and C. Liu. Learning the discriminative dictionary for sparse representation by a general fisher regularized model. In *The IEEE International Conference on Image Processing (ICIP)*, pages 4347–4351, Sept 2015.

[66] Q. Liu, A. Puthenputhussery, and C. Liu. Novel general knn classifier and general nearest mean classifier for visual classification. In *The IEEE International Conference on Image Processing (ICIP)*, pages 1810–1814, Sept 2015.

[67] Q. Liu, A. Puthenputhussery, and C. Liu. A novel inheritable color space with application to kinship verification. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, March 2016.

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

[69] J. Lu, J. Hu, V. Liong, X. Zhou, A. Bottino, I. Islam, T. Vieira, X. Qin, X. Tan, S. Chen, Y. Keller, S. Mahpod, L. Zheng, K. Idrissi, C. Garcia, S. Duffner, A. Baskurt, M. Castrillon-Santana, and J. Lorenzo-Navarro. The FG 2015 Kinship Verification in the Wild Evaluation. In *FG 2015*, pages 1–7, May 2015.

[70] J. Lu, X. Zhou, Y. Tan, Y. Shang, and J. Zhou. Neighborhood repulsed metric learning for kinship verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):331–345, 2014.

[71] C. Luo, B. Ni, S. Yan, and M. Wang. Image classification by selective regularized subspace learning. *IEEE Transactions on Multimedia*, 18(1):40–50, Jan 2016.

[72] L. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[73] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[74] A. M. Martinez and A. C. Kak. Pca versus lda. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001.

[75] F. B. Naini and J. P. 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.

[76] A. Y. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems (NIPS)*, pages 841–848. 2002.

[77] H.V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In *Asian Conference on Computer Vision (ACCV)*, volume 6493, pages 709–720, 2011.

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

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

[80] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[81] K. Peng and T. Chen. A framework of extracting multi-scale features using multiple convolutional neural networks. In *The IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, June 2015.

[82] F. Perronnin, J. Sanchez, and T. Mensink. *Improving the Fisher Kernel for Large-Scale Image Classification*, pages 143–156. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[83] F. Perronnin, J. Snchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 143–156. 2010.

[84] A. Puthenputhussery, Q. Liu, and C. Liu. Color multi-fusion fisher vector feature for fine art painting categorization and influence analysis. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, March 2016.

[85] A. Puthenputhussery, Q. Liu, and C. Liu. Sift flow based genetic fisher vector feature for kinship verification. In *The IEEE International Conference on Image Processing (ICIP)*, pages 2921–2925, Sept 2016.

[86] A. Puthenputhussery, Q. Liu, and C. Liu. Sparse representation based complete kernel marginal fisher analysis framework for computational art painting categorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 612–627. 2016.

[87] 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*, 19(8):1757–1770, Aug 2017.

[88] X. Qin, X. Tan, and S. Chen. Mixed bi-subject kinship verification via multi-view multi-task learning. *Neurocomputing*, 214:350 – 357, 2016.

[89] Q. Qiu, Z. Jiang, and R. Chellappa. Sparse dictionary-based representation and recognition of action attributes. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 707–714, 2011.

[90] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 413–420, 2009.

[91] L. Rathus. *Foundations of art and design*. Wadsworth Cengage Learning, Boston, MA, 2008.

[92] R. Sablatnig, P. Kammerer, and E. Zolda. Hierarchical classification of paintings using face- and brush stroke models. In *The International Conference on Pattern Recognition (ICPR)*, 1998.

[93] L. Shamir, T. Macura, N. Orlov, D. M. Eckley, and I. G. Goldberg. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Transactions on Applied Perception*, 2010.

[94] L. Shamir and J. A. Tarakhovsky. Computer analysis of art. *J. Comput. Cult. Herit.*, 2012.

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

[96] J. Shen. Stochastic modeling western paintings for effective classification. *Pattern Recognition*, pages 293 – 301, 2009. Learning Semantics from Multimedia Content.

[97] B. Siddiquie, S.N. Vitaladevuni, and L.S. Davis. Combining multiple kernels for efficient image classification. In *The IEEE Winter Conference on Applications of Computer Vision (WACV) Workshop*, pages 1–8, 2009.

[98] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher Vector Faces in the Wild. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2013.

[99] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[100] A. Sinha, S. Banerji, and C. Liu. New color gphog descriptors for object and scene image classification. *Machine Vision Applications*, 25(2):361–375, 2014.

[101] R. Sivalingam, D. Boley, V. Morellas, and N. Papanikolopoulos. Positive definite dictionary learning for region covariances. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1013–1019, 2011.

[102] J. Sun and J. Ponce. Learning discriminative part detectors for image classification and cosegmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 3400–3407, 2013.

[103] Engin Tola, V. Lepetit, and P. 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.

[104] K.E.A. van de Sande, T. Gevers, and C.G.M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, Sept 2010.

[105] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 18(7):1512–1523, July 2009.

[106] Q. Liu and C. Liu. A novel locally linear knn model for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1329–1337, 2015.

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

[108] K. Peng and T. Chen. Cross-layer features in convolutional neural networks for generic classification tasks. In *The IEEE International Conference on Image Processing (ICIP)*, pages 3057–3061, Sept 2015.

[109] J. Wang, P. Wonka, and J. Ye. Lasso screening rules via dual polytope projection. *Journal of Machine Learning Research*, 16:1063–1101, 2015.

[110] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3360–3367, June 2010.

[111] J. Wang, J. Zhou, J. Liu, P. Wonka, and J. Ye. A safe screening rule for sparse logistic regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1053–1061. 2014.

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

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

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

[115] S. Xia, M. Shao, J. Luo, and Y. Fu. Understanding kin relationships in a photo. *Multimedia, IEEE Transactions on*, 14(4):1046–1056, Aug 2012.

[116] Z. J. Xiang and P. J. Ramadge. Fast lasso screening tests based on correlations. In *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2137–2140, 2012.

[117] Z. J. Xiang, H. Xu, and P. J. Ramadge. Learning sparse representations of high dimensional data on large scale dictionaries. In *Advances in Neural Information Processing Systems (NIPS)*, pages 900–908. 2011.

[118] S. Yan, D. Xu, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, Jan 2007.

[119] J. Yang and C. Liu. A general discriminant model for color face recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1–6, 2007.

[120] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1794–1801, 2009.

[121] J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse coding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3517–3524, 2010.

[122] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 543–550, 2011.

[123] M. Yang, L. Zhang, X. Feng, and D. Zhang. Sparse representation based fisher discrimination dictionary learning for image classification. *International Journal of Computer Vision*, 109(3):209–232, 2014.

[124] S. Yang and D. Ramanan. Multi-scale recognition with dag-cnns. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[125] H. Yu and J. Yang. A direct lda algorithm for high-dimensional datawith application to face recognition. *Pattern recognition*, 34(10):2067–2070, 2001.

[126] M. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[127] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2126–2136, 2006.

[128] K. Zhang, Y. Huang, C. Song, H. Wu, and L. Wang. Kinship verification with deep convolutional neural networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 148.1–148.12, September 2015.

[129] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2691–2698, 2010.

[130] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[131] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems (NIPS)*, pages 487–495, 2014.

[132] N. Zhou, Y. Shen, J. Peng, and J. Fan. Learning inter-related visual dictionary for object recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3490–3497, 2012.

[133] X. Zhou, H. Yan, and Y. Shang. Kinship verification from facial images by scalable similarity fusion. *Neurocomputing*, 197:136 – 142, 2016.

[134] J. Zujovic, L. Gandy, S. Friedman, B. Pardo, and T.N. Pappas. Classifying paintings by artistic genre: An analysis of features and classifiers. In *The IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5, Oct 2009.