Color and deep learning features in face recognition

Lu, Ze

2018

Lu, Z. (2018). Color and deep learning features in face recognition. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/75855>

<https://doi.org/10.32657/10356/75855>

COLOR AND DEEP LEARNING FEATURES IN FACE RECOGNITION

NANYANG
TECHNOLOGICAL
UNIVERSITY

A Thesis Submitted to
the School of Electrical and Electronic Engineering of
the Nanyang Technological University

by

Ze Lu

in partial fulfillment of the requirement for
the Degree of Doctor of Philosophy

June 7, 2018

Abstract

Face recognition (FR) has been one of the most active research topics in computer vision for more than three decades. It has been widely applied in various practical scenarios such as access control system, massive surveillance, human computer interaction, etc. A conventional FR system consists of four stages, face detection, face alignment, face representation and face matching. Compared with the other three, face representation significantly affects the performance of a FR system because it determines whether the FR system is robust to real world variations, such as illumination, pose, occlusion, etc. Most of the early FR works have been limited to using grayscale images. Recently, research efforts have been dedicated to incorporate color information into the feature extraction process for improving FR performance. Specifically, different feature representations are extracted from face images in a certain color space and then fused together for classification. Main challenges of such color FR tasks include how to construct an effective color space to represent color images, and how to fuse different feature representations extracted from face images. To tackle the challenge of color space construction, we propose a framework to derive an effective color space LuC_1C_2 from the fundamental color space RGB . For the fusion of color features, we propose a Color Channel Fusion (CCF) method, a Covariance Matrix Regularization (CMR) method, and a color face descriptor, Ternary Color Local Binary Patterns (TCLBP). More recently, Convolutional Neural Networks (CNNs) have been proven effective for extracting high-level visual features from face images. However, there exist some problems in CNN feature representations. For example, the generalization ability of pre-trained CNN features is limited when training and testing data has

large differences, and the performance of CNNs drops dramatically when dealing with images of low resolution. Moreover, the feature fusion for different CNN architectures has not been thoroughly studied yet. To enhance the generalization ability of pre-trained CNN features, we investigate the combination of high-level CNN representations with low-level features, color pixel values, by score fusion. For different CNN architectures, we train a simplified ResNet model, ResNetShort, to fuse its features with those of VGG-Face by CMR. For low-resolution FR (LRFR), we propose a Deep Coupled ResNet model (DCR).

Color in the machine vision system is defined by a combination of 3 color components specified by a color space. Existing color spaces are based on different criteria and their performance is not consistent on different datasets. This motivates us to propose a framework for constructing effective color spaces. The proposed color space, LuC_1C_2 , consists of one luminance component and two chrominance components. The luminance component Lu is selected among four luminance candidates from existing color models by analysing their R,G,B coefficients and the color sensor properties. The chrominance components are derived by the discriminant analysis and the covariance analysis. Experiments show that both hand-crafted and CNN feature representations extracted from the LuC_1C_2 images perform consistently better than those extracted from images of other color spaces.

The fusion of multiple color features is important for achieving state-of-the-art FR performance. Existing color feature fusion methods either reduce the dimensionality of feature vectors in each color channel first and then concatenate all low-dimensional feature vectors, named as DR-Cat, or the vice versa, named as Cat-DR. In DR-Cat, existing methods simply reduce features in different color channels to the same number of dimensions and concatenate them. But the importance or reliability of features in different color channels is not the same. We propose a Color Channel Fusion (CCF) approach to select more features from more reliable and discriminative channels. Moreover, DR-Cat ignores the correlation information between different features while Cat-DR fully uses it. The correlation information estimated

from the training data may not be reliable. We propose a Covariance Matrix Regularization (CMR) technique to regularize the feature correlation estimated from training data before using it to train the feature fusion model. In addition to the fusion of different color features, we also jointly consider three color channels during color feature extraction by proposing the Ternary Color LBP (TCLBP) descriptor. Besides intra-channel LBP features, we extract the inter-channel LBP features by encoding the spectral structure of R,G,B component images at the same location.

CNNs have very large numbers of parameters that must be trained by millions of training examples. For every different application scenario, the common practice is to pre-train a CNN model on a very large dataset and then use it either as a fixed feature extractor or an initialization for fine-tuning on images from the application of interest. After successive convolutional layers in the CNN, high-level features are formed in the top layer. However, the low-level feature information might be lost in the top layer. The combination of high-level CNN features and low-level features can reduce the possible information loss in the top layer. Furthermore, low-level features depict basic characteristics of face images from the application of interest. We investigate the fusion of CNN features and the lowest-level features, color pixel values, to boost the generalization ability of pre-trained CNNs for different application scenarios. The two types of features are fused by the way of score fusion instead of feature-level fusion due to their big differences. To further improve the performance of CNNs, we train a simplified ResNet model, ResNetShort, and combine its features with those of VGG-Face by our proposed CMR technique. The two CNN models are trained from different face images by optimizing different loss functions through different architectures. This makes the learned discriminative information contained in ResNetShort features and VGG-Face features mutually complementary to each other thus better performance is achieved by the fusion of them. The FR performance of CNNs largely drops when CNNs are applied to face images of low-resolution. Existing CNN methods can not deal with different resolutions of probe images. We

propose a deep coupled network which extracts coupled mappings from face images to tackle the resolution degradation problem of probe images.

Acknowledgments

I would like to thank my supervisor, Professor Alex Kot, and my co-supervisor, Professor Xudong Jiang for their invaluable guidance, support and suggestions given to me in the past four years. I am lucky to have both of them as my supervisors. Professor Kot is a man of wise and broad knowledge. His suggestions to my research and to my career plan benefit me a lot. Professor Jiang spends much time discussing ideas, experiments, and paper writing with me. His devotion to research and approachable characteristics greatly inspire me. Their encouragements help me to overcome the difficulties encountered in my research.

I must express my gratitude to Yisi Chen, my wife, for her continued support and encouragement. Her love and understanding have accompanied me through my hardest times during PhD life. I am also grateful for my family members in China who experienced all of the ups and downs of my research.

I want to thank my friends and colleagues in the Rapid-Rich Object Search Lab, for their company and help. Jiong Yang, Tan Yu, Zhigang Tu, Zhenzhen Wang, Weixiang Hong, Junlin Hu, Jianfeng Ren, Chunluan Zhou, Renjie Huang, Peisong He, Haoliang Li, Yan Wang, Huijing Zhan, Renjie Wan, etc., without the friendship and happiness shared with them, I would never be able to go through so much difficulties in both research and life.

Finally, I would like to thank the Rapid-Rich Object Search Lab for providing the facilities and equipments which allow me to undertake this research.

Contents

Abstract	ii
Acknowledgments	vi
List of Figures	xii
List of Tables	xviii
List of Abbreviations	xxii
1 Introduction	1
1.1 Face recognition systems and an overview of face recognition techniques	1
1.1.1 Face recognition systems	1
1.1.2 An overview of techniques for face recognition	2
1.2 Motivations and objectives	4
1.2.1 Color face recognition	4
1.2.2 Deep learning face recognition	7
1.3 Main contributions	9
1.4 Organization	11
2 Related Work	13
2.1 Color face recognition techniques	13
2.1.1 Color face recognition problems	13
2.1.2 Color spaces	15
2.1.3 Fusion of color features	17

2.1.4	Summary	21
2.2	Deep learning and low-resolution face recognition techniques	23
2.2.1	Deep learning models in face recognition	23
2.2.2	Summary	29
2.2.3	Low-resolution face recognition approaches	31
2.2.4	Summary	34
2.3	Databases used for performance evaluation of face recognition in this thesis . .	35
2.3.1	AR	35
2.3.2	GT	35
2.3.3	FRGC	35
2.3.4	LFW	36
2.3.5	Multi-PIE	36
2.3.6	SCFace	36
3	Color Space LuC_1C_2	39
3.1	Introduction	39
3.2	The proposed color space LuC_1C_2	42
3.2.1	Overview of the proposed approach	42
3.2.2	Selection of the luminance component	44
3.2.3	Extraction of two chrominance components	47
3.3	Experiments	53
3.3.1	Databases	54
3.3.2	The dependence of face recognition performance on the correlation between two chrominance components C_1, C_2	55
3.3.3	Performance comparison of different color spaces under various con- ditions	56

3.3.4	Performance comparison of the proposed LuC_1C_2 color space with state of the arts on FRGC	65
3.3.5	Robustness improvement of the CNN model to low-resolution degradation using the proposed LuC_1C_2 color space	66
3.4	Summary	68
4	Color Feature Fusion	71
4.1	Color Channel Fusion	71
4.1.1	Introduction	71
4.1.2	Color channel fusion (CCF) approach	72
4.1.3	Experiments	77
4.1.4	Summary	80
4.2	Covariance Matrix Regularization	81
4.2.1	Introduction	81
4.2.2	Feature fusion in face recognition	83
4.2.3	Feature fusion with dimensionality reduction	84
4.2.4	Covariance Matrix Regularization for feature fusion	87
4.2.5	Experiments	90
4.2.6	Summary	97
4.3	Color face descriptor TCLBP	98
4.3.1	Introduction	98
4.3.2	The proposed color descriptor TCLBP	99
4.3.3	Experiments	104
4.3.4	Summary	106

5	Deep Learning Face Recognition	107
5.1	Enhance CNN performance using color pixel values	107
5.1.1	Introduction	107
5.1.2	Convolutional Neural Networks	108
5.1.3	Fine-tuning	110
5.1.4	Enhance the CNN performance by color pixel values	111
5.1.5	Experiments	115
5.1.6	Summary	118
5.2	Feature fusion of VGG-Face and ResNetShort	118
5.2.1	Introduction	118
5.2.2	VGG-Face and ResNetShort	119
5.2.3	Experiments	122
5.2.4	Summary	127
5.3	Deep Coupled ResNet Model	127
5.3.1	Introduction	127
5.3.2	Deep Coupled ResNet model	130
5.3.3	Experiments	134
5.3.4	Summary	137
6	Conclusions and Future Research	139
6.1	Color face recognition	139
6.2	Deep learning face recognition	141
6.3	Summary	143
6.4	Future research	144
	Publications	145
	References	146

List of Figures

1.1	The system flow for FR approaches.	3
1.2	The framework of color FR approaches.	5
2.1	Three LBP histograms obtained by applying the opponent LBP operation to two local regions from different color component images in the YIQ color space [1].	20
2.2	Color norm patterns and color angular patterns from local region images of different color component images from the YIQ color space [2].	20
2.3	Outline of the DeepFace architecture [3].	24
2.4	DeepID, DeepID2, and DeepID2+ structures [4–6].	25
2.5	DeepID3 structure [7].	26
2.6	Details of the VGG-Face architecture [8].	28
2.7	Residual learning: a building block [9].	29
2.8	Structure of the Resolution-Invariant Deep Network for resolution-robust feature extraction [10].	33
3.1	The framework of constructing an effective color space for the task of face recognition.	43
3.2	The normalized response of human cone cells for different wavelengths of light [11].	45
3.3	(a) The color filter array in cameras and (b) the camera spectral sensitivity of nikon-D70 [12].	46

3.4	Cropped images from AR.	54
3.5	Cropped images from GT.	55
3.6	Cropped images from Multi-PIE.	55
3.7	Cropped images of the FRGC database.	55
3.8	(a) The covariance between two chrominance components plotted against the angle between $\mathbf{u}_1$, $\mathbf{u}_2$ and (b) corresponding face recognition rates on the AR database.	57
3.9	(a) The covariance between two chrominance components plotted against the angle between $\mathbf{u}_1$, $\mathbf{u}_2$ and (b) corresponding face verification rates on the FRGC database.	57
3.10	The color face recognition framework used on AR, GT and Multi-PIE databases.	58
3.11	Face recognition rates against feature dimension on AR, each column specifies one type of feature (2 in total) and each row specifies one dimension reduction method (3 in total).	60
3.12	Face recognition rates against feature dimension on GT, each column specifies one type of feature (2 in total) and each row specifies one dimension reduction method (3 in total).	61
3.13	Face recognition rates against feature dimension on Multi-PIE, each column specifies one type of feature (2 in total) and each row specifies one dimension reduction method (3 in total).	62
3.14	The color face recognition framework used on FRGC databases.	63
3.15	Face verification rates against feature dimension on FRGC, each column specifies one type of feature (2 in total) and each row specifies one dimension reduction method (3 in total).	64
3.16	Face recognition rates of ResNet- LuC_1C_2 and ResNet- RGB against the resolution of face images on GT database.	69

3.17	Face verification accuracy of ResNet- LuC_1C_2 and ResNet- RGB against the resolution of face images on LFW database.	69
4.1	Color FR framework, C_i , f_i , and $f_i^{l_i}$ indicate color component images, channel-wise features, and low-dimensional features respectively.	73
4.2	(a), Channel-wise FR performance against PCA dimension on AR database; (b), within-class variations against PCA dimension on AR database.	74
4.3	(a), Channel-wise FR performance against ERE dimension on pose variation subset of MultiPIE; (b), discriminant value J against ERE dimension on pose variation subset of MultiPIE.	76
4.4	FR rate of 4 color spaces on AR using ERE-based CCF	78
4.5	FR rates of (a) PCA-based and (b) ERE-based methods using image-pixel values	80
4.6	FR rates of (a) PCA-based and (b) ERE-based methods using CLGWs	80
4.7	Exampling face images of the illumination variation subset from the Multi-PIE database.	91
4.8	Exampling face images from Georgia Tech face database.	91
4.9	Exampling face images of the AR database.	91
4.10	Exampling face images of the FRGC database.	92
4.11	Face recognition rates (%) of fusing features (pixel values or LBP) of 3 color channels (R,G,B) against the value of weights in CMR on Multi-PIE, GT and AR. Each column specifies one type of feature (pixel values or LBP) and each row specifies one dataset (Multi-PIE, GT and AR).	93
4.12	Face recognition rates (%) of fusing different types of features (pixel values and LBP of channel R) against the value of weights in CMR on Multi-PIE, GT and AR.	94
4.13	Face recognition rates against the value of weights of CMR for different numbers (s) of samples per subject on GT.	95

4.14	Face recognition rates against the value of weights of CMR for different numbers (s) of samples per subject on AR.	95
4.15	Feature extraction process of Intra-channel LBP.	100
4.16	Dimensionality of LBP-based color features.	101
4.17	Feature extraction process of Inter-channel LBP.	102
5.1	VGG-Face architecture, CONV indicates convolutional layers, POOL indicates pooling layers and FC indicates fully-connected layers.	111
5.2	Framework of the proposed method, DR indicates dimension reduction and RPs indicates raw pixels.	112
5.3	Sample face images from VGG Face database.	120
5.4	The ResNetShort architecture, where C, P, and F indicate convolutional, max pooling, and fully connected layers, respectively.	121
5.5	Normalized face images from CASIA-WebFace database.	121
5.6	Exampling face images from the illumination variation subset of Multi-PIE. . .	123
5.7	Exampling face images from Georgia Tech face database.	123
5.8	Exampling face images from AR database.	123
5.9	Sample face images from LFW database.	124
5.10	Face recognition rates (%) of fusing features extracted by different deep models (VGG-Face and ResNetShort) against the value of weights in CMR on Multi-PIE, GT and AR.	126

5.11	Architecture of the proposed Deep Coupled ResNet (DCR) model. The trunk network learns discriminant features (indicated by $\mathbf{v}$) shared by different resolutions of images, and the branch networks are trained as coupled mappings (indicated by $\mathbf{x}$ for HR features and $\mathbf{z}$ for LR features, respectively). C, P and F indicate convolutional layer, max-pooling layer and fully-connected layer, respectively. The number of output feature maps in convolutional layers and the number of outputs in fully connected layers are indicated by those on top of each layer. ‘ $\times h$ ’ represents a residual module that repeats for h times. k indicates the resolution of LR training images and β is a scaling parameter for center loss.	129
5.12	Example face images from CASIA-WebFace.	132

List of Tables

2.1	Face verification rates (%) for grayscale and color features using six different face resolutions of probe images on FRGC database, at the false accept rate of 0.1%. R from the RGB color space is used as the grayscale feature, while the $RQCr$ color space is employed as the color feature.	14
2.2	Comparisons of the face verification rates (FVR) (%) on FRGC database experiment 4 for various color spaces with the false accept rate (FAR) equivalent to 0.1%.	17
2.3	Comparisons of accuracy (%) on LFW and numbers of training images for different deep learning models.	29
2.4	Recognition rates (%) at distance d3/d2/d1 on SCface for different methods of LRFR.	34
2.5	Databases used for evaluation of color face recognition.	37
3.1	Performance comparison of LuC_1C_2 with state of the arts using raw pixels on FRGC	65
3.2	Performance comparison of LuC_1C_2 with state of the arts using complex features on FRGC	66
3.3	Averaged face recognition rate (AFRR) on GT database and averaged face verification accuracy (AFVR) on LFW database using ResNet- RGB and ResNet- LuC_1C_2 for images of resolution 4×4 , 6×6 , ..., 20×20	68

4.1	Best recognition rate (%) & its dimension using pixel value	79
4.2	Best recognition rate (%) & its dimension using CLGWs	79
4.3	Face recognition performances of the best single feature, DR-Cat, Cat-DR and CMR using pixel values of multiple color channels on Multi-PIE, GT, AR and FRGC.	96
4.4	Face recognition performances of the best single feature, DR-Cat, Cat-DR and CMR using LBP of multiple color channels on Multi-PIE, GT, AR and FRGC.	96
4.5	Face recognition performances of the best single feature, DR-Cat, Cat-DR and CMR using pixel values and LBP of channel R on Multi-PIE, GT, AR and FRGC.	97
4.6	Results on Georgia Tech	105
4.7	Results on FRGC	105
4.8	Results on LFW	106
5.1	Face verification accuracy (%) using CNN and pixel values on LFW	116
5.2	Face verification rate (%) at FAR = 0.1% using CNN and pixel values on FRGC	117
5.3	Comparison between the pre-trained VGG-Face model and our trained Resnet- Short model, CONV and FC indicate convolutional and fully connected layers, respectively.	122
5.4	Face verification accuracy (%) of ResNetShort, VGG-Face, DeepID and Canon- ical CNN on LFW.	125
5.5	Face recognition/verification performances of the best single feature and CMR using CNN features of VGG-Face and ResNetShort models on Multi-PIE, GT, AR and LFW.	127
5.6	Face verification accuracy of different approaches using different probe sizes on LFW	136
5.7	Face recognition rates of different approaches at different distances on SCface	137

6.1	Performance comparison of LuC_1C_2 with state of the arts using complex features on FRGC	140
6.2	Averaged face recognition rate (AFRR) on GT database and averaged face verification accuracy (AFVR) on LFW database using ResNet- RGB and ResNet- LuC_1C_2 for images of resolution $4 \times 4, 6 \times 6, \dots, 20 \times 20$	140
6.3	Face recognition performances of the best single feature, DR-Cat, Cat-DR and CMR using pixel values and LBP of channel R on Multi-PIE, GT, AR and FRGC.	141
6.4	Face verification rate (%) at FAR = 0.1% using CNN and pixel values on FRGC	142
6.5	Face recognition/verification performances of the best single feature and CMR using CNN features of VGG-Face and ResNetShort models on Multi-PIE, GT, AR and LFW.	142
6.6	Face recognition rates of different approaches at different distances on SCface	143

List of Abbreviations

FR	Face Recognition
CNN	Convolutional Neural Network
SIFT	Scale-Invariant Feature Transform
PCA	Principal Component Analysis
ERE	Eigenfeature Regularization and Extraction
LDA	Linear Discriminant Analysis
EFM	Enhanced Fisher linear discriminant Model
JB	Joint Bayesian
HR	High Resolution
LR	Low Resolution
SR	Super Resolution
CM	Coupled Mapping
MDS	Multidimensional Scaling
DMDS	Discriminative Multidimensional Scaling
LDMDS	Local-consistency-preserving Discriminant Multidimensional Scaling
LRFR	Low-Resolution Face Recognition
RICNN	Resolution-Invariant Convolutional Neural Network
DCR	Deep Coupled ResNet
DR-Cat	Dimensionality Reduction then Concatenation
Cat-DR	Concatenation then Dimensionality Reduction
B.S.	Best Single
CMR	Covariance Matrix Regularization
CLGW	Color Local Gabor Wavelets
LCVBP	Local Color Vector Binary Patterns
LBP	Local Binary Pattern
CLBP	Color Local Binary Pattern
TCLBP	Ternary-Color Local Binary Pattern
CCF	Color Channel Fusion

CCC	Color Channel Concatenation
WDF	Weighted Decision-Level Fusion
FRGC	Face Recognition Grand Challenge
GT	Georgia Tech.
FVR	Face Verification Rate
FAR	False Accept Rate
FRR	Face Recognition Rate
AFRR	Averaged Face Recognition Rate
FVA	Face Verification Accuracy
AFVA	Averaged Face Verification Accuracy
FLOPs	Floating-Point Operations per Second
MFLOPs	Mega Floating-Point Operations per Second
MB	Mega Byte
GPU	Graphic Processing Unit

Chapter 1

Introduction

The research work in this thesis focuses on feature representations in face recognition (FR). Specifically, we study 1) how to derive an effective color space from *RGB* to represent color images; 2) how to fuse different color features extracted from face images; 3) how to tackle problems in Convolutional Neural Network (CNN) features and enhance the FR performance. In this chapter, we first briefly introduce FR systems and give an overview of current FR techniques, then discuss motivations and objectives of our research works, finally present our major contributions and the organization of this thesis.

1.1 Face recognition systems and an overview of face recognition techniques

1.1.1 Face recognition systems

These days, video cameras have been widely used for various applications, and a great number of images are being produced at every second of our lives. Many of these images are selfies or group photos which contain faces. FR has thus become popular due to its potential use in a wide range of practical applications, such as automatic access control system, e-passport, criminal recognition, forensic sciences, driver's licenses, missing identification, surveillance systems, social networks, etc. A FR system is a computer application capable of identifying or verifying a person from a digital image or a video frame from a video source. It is increasingly

becoming dominant over the other biometric systems for several reasons. 1), Face is the only human readable biometric modality which turns out is incredibly important to humans who buy and use biometric systems. Our psychology tells us that we are far more comfortable with a system that can be audited and checked by people. 2), Ubiquity of visible light cameras. The visible light camera is arguably the most ubiquitous sensor in the world by which we can process a biometric signature from a human. Cameras exist on street corners, in banks, built into ATMs, laptops, and nearly every phone. 3), Perfect equilibrium between convenience and security. A FR system is adequately secure for most purposes, and face is the most convenient of any biometric modality.

1.1.2 An overview of techniques for face recognition

FR is a general topic that includes both face identification and face verification (also called authentication). Face identification is designed to identify a person based on the image of a face. This face image has to be compared with all the registered persons (one-to-many matching). Face verification is concerned with validating a claimed identity based on the image of a face, and either accepting or rejecting the identity claim (one-to-one matching). A conventional FR system usually consists of four stages: face detection, face alignment, face representation, and face matching as shown on Fig. 1.1. Face representation and face matching are two more important stages in a FR system [13]. Good face representations indicate features extracted from face images and contain discriminative information for separating different subjects. Good face matching represents classifiers which effectively distinguish various face patterns. In unconstrained environments, face images always display many variations including poses, expressions, illuminations, occlusions, resolutions, and backgrounds. These variations increase the distance of face samples from the same subject and decrease the distance of face samples from different subjects. Good face representations differentiate inter-personal differences while being robust to intra-personal variations. Face representation is thus more important than face

matching as it has bigger impact on the final performance of a FR system. Generally speaking, there exist two categories of face representation methods in recent research works, descriptor-based methods [14–17] and deep-learning-based methods [3, 6].

Figure 1.1: The system flow for FR approaches.

Descriptor-based methods can be further categorized into two categories: global features and local features. Representative global features are principal component analysis (PCA) [14] and linear discriminant analysis (LDA) [15]. Typical local features are local binary pattern (LBP) [16] and Gabor wavelets [17]. Global features transform a face image into a high-dimensional feature vector and learn a feature subspace to preserve the statistical information of face images. Different with global features, local features describe the structure pattern of each local patch and then combine the statistics of all patches into a concatenated feature vector to represent the whole face image. In [18], authors densely sample multi-scale local descriptors centered at dense facial landmarks and concatenate them. They empirically find that over-completed representation, with sufficient training data, is necessary to obtain state-of-the-art results. While there exist a huge body of works dealing with descriptor-based face representations, most of them are based on grayscale images [2]. Recently, some research efforts have been dedicated to incorporate color information into the process of feature extraction for improving FR performance [1, 2, 19], where different global and local features derived from an image of a certain color space are fused together. It is shown that color information contains complementary and discriminant information for FR tasks.

More recently, there has been increased demand for recognition of unconstrained face images, such as those collected from the internet or captured by mobile devices and surveillance cameras. The performance of existing descriptor-based methods is not good enough in unconstrained environments and they usually need some strong priors to engineer them by hand. Deep neural networks have shown much better face recognition performance than descriptor-based methods as in [8, 20]. Instead of using hand-crafted features, researchers train deep neural networks on large datasets of labelled faces to obtain robustness to pose, illumination, and other variational conditions. The impressive performance of deep neural networks relies on: (1) large scale of training data used to learn network parameters and (2) powerful computation devices such as thousands of CPU and/or GPU cores. Convolutional Neural Network (CNN) is an approach purely driven by data, which learns its deep features from the pixel values of the face images. The DeepFace in [3] is a deep neural network trained for the task of FR on over 4,000 subjects. Their best performance is achieved by combining three deep networks based on different alignment methods and color channels. DeepID [4] is a compact and therefore relatively cheap to compute network. The authors in [4] propose to use an ensemble of 25 such networks, each operating on a different face patch. Both PCA and a Joint Bayesian model [21] are employed to compute similarity scores between deep features. FaceNet [22] is a CNN trained to directly optimize the feature embedding itself, instead of an intermediate bottleneck layer. For training, authors select triplets of roughly aligned matching or non-matching face images produced by an online triplet mining approach.

1.2 Motivations and objectives

1.2.1 Color face recognition

Torres et al. [23] apply a modified PCA scheme to FR and their results show that the use of color information improves the recognition performance when compared with the same scheme using the luminance information. The improvement can be significant when large

facial expression and illumination variations are present or the resolution of face images is low [24, 25]. Since then, considerable research efforts have been devoted to the efficient utilization of facial color information to enhance the FR performance [1, 2, 19]. Color images consist of 3 component images while gray-scale images comprise only one component image, and this makes the framework of color FR as shown on Fig. 1.2 different with that of gray-scale FR.

Figure 1.2: The framework of color FR approaches.

Color in the machine vision system is defined by a combination of 3 color components specified by a color space. Many color spaces have been proposed to find the optimal way of representing color images for FR. In the early studies, color configurations are made through a combination of intuition and empirical comparisons without a systematic strategy. Thereafter, color spaces are proposed by seeking 3 sets of optimal coefficients to combine the fundamental R, G and B components based on a criterion. A consistent property of effective color spaces of FR is found out in [26] by analysing the transformation matrices of effective color spaces from the *RGB* color space. Based on this characteristic of effective color spaces, researchers propose color space normalization techniques, which are able to convert weak color spaces into effective ones. However, the performance of existing color spaces [1, 26, 27] is not consistent for different databases since they are proposed based on different criteria. By analysing color spaces that demonstrate good classification capabilities for FR tasks, we find that they are all composed of one luminance component and two chrominance components. This configuration

reduces the correlation between different color components and enhances the discriminating power of the color space. The luminance component can be chosen from existing color spaces by analysing their R, G, B sensor properties. A common characteristic of chrominance components is that the sum of R, G, B coefficients is zero. Under this condition, we can use the discriminant analysis and the covariance analysis to derive the optimal two chrominance components. In this way, a framework to construct effective color spaces is proposed for FR.

The next step in the color FR framework is feature extraction and feature fusion. The fusion of multiple color features is important for achieving state-of-the-art FR performance. Existing color feature fusion methods either reduce the dimensionality of feature vectors in each color channel first and then concatenate all low-dimensional feature vectors, named as DR-Cat, or the vice versa, named as Cat-DR. For DR-Cat, features in different color channels are usually processed separately first and then concatenated together into a feature vector for classification due to the high dimensionality of color component images or generated color features. Note that dimensionality reduction is applied on features in each color channel separately before the channel fusion. Specifically, the dimensionalities of different low-dimensional features are set to be equal as in [1, 28]. In fact, the reliability and importance of features in different color channels are not the same, which should be considered in determining dimensionalities of different low-dimensional features. We propose a Color Channel Fusion (CCF) method to select more features from more reliable and discriminative channels. Furthermore, DR-Cat ignores the correlation information between different features which is useful for classification. In Cat-DR, on the other hand, the correlation information estimated from the training data is fully used for training feature fusion models. But the correlation information may be unreliable and cause overfitting especially when the number of training samples is limited. We propose a Covariance Matrix Regularization (CMR) technique to solve problems of DR-Cat and Cat-DR. It works by assigning weights to cross-feature covariances in the covariance matrix of training data. Thus the feature correlation estimated from training data is regularized before being used to train feature fusion models.

In addition to fusing different color features, we also jointly consider three color channels during the process of feature extraction and propose a LBP-based color face descriptor, TCLBP. LBPs [16, 29–31] have gained reputation as powerful face descriptors as they have shown great robustness to variations such as facial pose, illumination, misalignment, etc. In color FR, a few research efforts have been proposed to incorporate color information into the extraction of LBP-based features [2, 25, 32]. However, it should be noted that existing color LBP features are restricted to extracting inter-channel features from each pair of color channels using the same spatial structure as that used for the intra-channel features. Also, they suffer from the curse of high dimensionality. What’s more, pixel values from different channels in certain color spaces are not quantitatively comparable. The TCLBP descriptor is proposed to tackle these mentioned problems.

1.2.2 Deep learning face recognition

Compared with descriptor-based FR methods, CNNs have shown much better face recognition performance in unconstrained environments. CNNs are high-capacity feature extractors with very large numbers of parameters that must be learned from millions of training examples [33]. It is impossible to collect millions of annotated images and implement the complex training process for every different FR scenario. In practice, it is common to pretrain a CNN on a very large dataset and then use the pre-trained CNN model either as a fixed feature extractor or an initialization for fine-tuning on images from the application of interest [34–38]. However, our experimental results show that the generalization ability of pre-trained CNNs is limited when training and testing datasets have large differences in viewpoints, image sizes, scene context, illumination, expressions or other factors, etc. Although high-level features are formed in the top layer, the low-level feature information might be lost at the same time. The combination of high-level CNN features and low-level features can reduce the possible information loss in the top layer. Moreover, low-level features contain raw information of face images from the

application of interest, which is complementary to feature representations in pre-trained CNN models. We investigate the fusion of CNN features and the lowest-level features, color pixel values, to boost the generalization ability and enhance the recognition performance of pre-trained and fine-tuned CNNs. The two types of features are fused by the way of score fusion instead of feature-level fusion due to their big differences.

The pre-trained VGG-Face model [8] has been widely used as a feature extractor for classifying face images as in [38–40]. Its network is characterized by using 3×3 convolutional layers stacked on top of each other in increasing depth. Different from the architecture of VGG-Face, ResNet in [20] consists of residual modules which conduct additive merging of signals. The authors in [20] argue that residual connections are inherently important for training very deep architectures. Accordingly, we train a simplified ResNet model, ResNetShort, of 29 layers with parameters optimized by the CASIA-WebFace dataset [41]. The two CNN models of VGG-Face and ResNetShort are trained from different face images by optimizing different loss functions through different deep architectures. This makes the learned discriminative information contained in VGG-Face features and ResNetShort features mutually complementary to each other. To investigate the effectiveness of fusing features from different CNN architectures, we combine the feature representations of VGG-Face and ResNetShort by CMR and compare the obtained FR performance with that of VGG-Face or ResNetShort.

Promising results have been achieved in FR tasks under challenging conditions such as occlusion [42], variations in pose and illumination [43]. While many CNN-based approaches have been developed for recognizing high resolution (HR) face images [22, 44, 45], there are few studies focused on FR in surveillance systems, where HR cameras are not available or there is a long distance between the camera and the subject. Under the condition of low-resolution (LR) images, CNN approaches developed for HR images usually decline [46, 47]. Also, existing CNN methods can not deal with different resolutions of probe images. Their performance can be largely boosted by using feature representations which are robust to the

resolution change. Motivated by the superior performance of ResNet and coupled mappings (CMs) [47], we propose to build a Deep Coupled ResNet model for LRFR tasks.

1.3 Main contributions

Based on the motivations discussed in the previous section, three branches of works are proposed for FR tasks in this thesis. Specifically, for the color space construction, we propose a framework to construct an effective color space LuC_1C_2 . For the color feature fusion, we propose a Color Channel Fusion (CCF) method, a Covariance Matrix Regularization (CMR) method and a Ternary Color LBP (TCLBP) descriptor. For CNN features, we investigate the combination of CNN representations with color pixel values by score fusion, the fusion of features from VGG-Face and ResNetShort by CMR, and propose a Deep Coupled ResNet (DCR) model for low-resolution FR tasks.

For color space conversion, we propose an effective color space LuC_1C_2 . It consists of one luminance component and two chrominance components. The luminance component Lu is selected among four luminance candidates from existing color models by analysing their R,G,B coefficients and the color sensor properties. In order to generate the two effective chrominance components C_1, C_2 , the directions of their transform vectors are derived by the discriminant analysis and the covariance analysis in the chrominance subspace of the RGB color space. The magnitudes of their transform vectors are derived according to the discriminant values of Lu, C_1, C_2 . Experiments show that both descriptor-based and deep-learning-based color features extracted from the LuC_1C_2 images perform consistently better than those extracted from images of other color spaces such as RGB .

For color feature fusion methods using DR-Cat, we propose the Color Channel Fusion (CCF) method to select more features from more reliable and discriminant color channels instead of choosing the same number of low-dimensional features from different color channels. In CCF, the dimension reduction rule of a single color channel is integrated across all three

color channels. Extensive experiments show CCF outperforms existing methods consistently using different types of features. Furthermore, the Covariance Matrix Regularization (CMR) technique is proposed for color feature fusion methods using Cat-DR. Instead of modifying eigenvalues of covariance matrices as in conventional regularization techniques [48–52], CMR solves the overfitting problem by regularizing the off-diagonal cross-feature covariances in the covariance matrix of training data. Thus the trace of covariance matrices remains unchanged and the feature correlation estimated from the training data is suppressed before being used to train the feature fusion model. In this way, the obtained model does not adapt too much to the estimated correlation and hence the overfitting is reduced. Besides fusing different color features, we also jointly consider all three color channels during feature extraction by proposing the TCLBP face descriptor. It consists of intra-channel LBPs and inter-channel LBPs. The main contribution of TCLBP is the inter-channel LBP feature of extremely low dimensionality, which is generated by encoding the spectral structure of R,G,B component images at the same location.

For deep learning FR, we investigate the combination of high-level image representations learned by CNNs with low-level features, color pixel values, by score fusion to enhance the generalization ability of pre-trained CNN models. The color pixel values depict the low-level characteristics of face images from the application of interest. They provide information complementary to the high-level CNN features and reduce the possible information loss. Experiments show color pixel values can be used to boost the FR performance of pre-trained CNNs with and without fine-tuning. Furthermore, we train a simplified ResNet model, ResNet-Short, of 29 layers and combine its feature representations with those of VGG-Face by CMR to achieve improved FR performance. For low-resolution FR, we propose the Deep Coupled ResNet (DCR) model. It consists of one big trunk CNN and two small branch networks. We train the trunk CNN only once and fix its parameters to learn discriminant features shared by face images of different resolutions. Two branch networks are trained to learn resolution-specific coupled-mappings so that HR gallery images and LR probe images are projected to a

space where their distances are minimized. In reality, there can be various resolutions of probe images to be matched with HR gallery images, the proposed DCR model solves this problem by training different pairs of small branch networks while using the same big trunk network.

1.4 Organization

The organization of the thesis is as below:

In Chapter 2, we give brief introductions of existing color FR approaches, recent deep learning architectures for FR, and low-resolution FR methods.

In Chapter 3, our proposed color space LuC_1C_2 is presented, extensive experiments show the superior performance of LuC_1C_2 over the other color spaces for both descriptor-based and deep-learning-based FR methods.

In Chapter 4, we propose the Color Channel Fusion approach for DR-Cat methods to select more features from more reliable and discriminative channels. To solve overfitting problems in Cat-DR methods, we propose a Covariance Matrix Regularization technique. A novel LBP-based color face descriptor, Ternay-Color LBP, is also proposed to jointly consider information in three color channels during feature extraction.

In Chapter 5, we investigate the combination of image representations learned by CNNs with color pixel values to improve the generalization ability of pre-trained CNNs with and without fine-tuning. Furthermore, we achieve better FR performance by fusing feature representations of ResNetShort and VGG-Face models through CMR. For LRFR, we propose the Deep Coupled ResNet (DCR) model to extract coupled mappings from different resolutions of face images.

Chapter 2

Related Work

2.1 Color face recognition techniques

2.1.1 Color face recognition problems

In early FR approaches, the original color image in the *RGB* space is usually transformed into an image of grayscale by weightedly combining the R, G, and B components [53]. A presumption is that color information would provide little or no increase in system accuracy. Recent research efforts, however, reveal that color may provide useful information to face recognition. For color FR works reported so far, questions could be generally categorized as follows. 1) Is color information helpful in improving the FR accuracy compared with using grayscale images only? 2) Which color space performs the best for providing discriminate power needed to perform the reliable FR tasks? 3) How to fuse different color features to make use of the rich facial information for improving FR performance?

There exists discriminant information in color and it has been proven useful for tasks like object detection and identification. Furthermore, color features have been broadly used in pattern recognition because of its robust characteristics [54, 55]. In contrast to the intensity-driven features, color-based features are known to be less susceptible to resolution changes for objection recognition [56].

For the task of face recognition, results in [23] show that the use of the color information, embedded in a eigen approach, improves the recognition rate when compared to the same

Table 2.1: Face verification rates (%) for grayscale and color features using six different face resolutions of probe images on FRGC database, at the false accept rate of 0.1%. R from the RGB color space is used as the grayscale feature, while the $RQCr$ color space is employed as the color feature.

Image resolution	112×112	86×86	44×44	25×25	20×20	15×15
R	79.16	75.19	76.33	65.72	62.72	52.70
$RQCr$	82.06	78.32	77.18	76.56	74.10	68.32

scheme which uses only the luminance information. Moreover, the psychophysical results of the FR test in human visual systems show that the contribution of facial color becomes evident when the shapes of faces are getting degraded [57]. The experimental results in [58] on CVL [59] and CMU-PIE [60] color face databases show better performance of color 2DPCA features over those of the gray-level. It has been reported in [25] and [24] that the effectiveness of color information can become significant for improving FR performance when face images are taken under strong variations in illumination, as well as with low spatial resolutions. In [24], authors use non-negative matrix factorization to recognize color face images. The experiment is conducted on a subset of color images in the AR [61] database for robustness against facial expressions and illumination variations. Color face recognition results are compared with the results obtained from the gray scale images of the same dataset. Results show improved accuracy of color image recognition over gray level image recognition when large facial expressions and illumination variations are present. In [25], it is shown that color features can considerably boost the face recognition performance when compared with intensity features. Moreover, authors in [25] conduct experiments on the FRGC database [62] to show that facial color cue reduces the recognition error rate by at least an order of magnitude compared with intensity features under the condition of low resolution images. Their results are shown on Table 2.1, where the Bayesian algorithm is used for classification.

2.1.2 Color spaces

A color image contains three component images. Each pixel of a color image is specified in a color space, which serves as a color coordinate system. Among various color spaces, *RGB* is quite basic and commonly-used. The others are often computed from it by either linear or nonlinear transformations [63]. In general, the three components of a color can be defined in many different ways leading to a wide variety of color spaces [64]. It has been observed that different color spaces (or color models) possess distinct characteristics and effectiveness in terms of discriminating power for visual classification tasks [65]. When utilizing the color information in FR tasks, we should select a discriminative color space, and this selection plays an important role for obtaining the best FR performance.

Most of the early color FR methods are restricted to using a fixed color-component configuration comprising of three color components, which are mostly made through a combination of intuition and empirical comparisons [24–26, 63, 66, 67], without any systematic selection strategy. For example, the *YUV* color space has been shown to be able to increase the recognition rates of the *RGB* color space [23]. What’s more, compared with the other color component images, the *R* component image in the *RGB* color space and the *V* component image in the *HSV* color space have been demonstrated to perform better for FR tasks in [68]. Also, the *YQCr* color space consisting of *Y*, *Q* component images from the *YIQ* space, and the *Cr* color component image from the *YCbCr* space, has been shown to deliver better performance than *HSV* and $L^*a^*b^*$ color spaces [66, 68]. In [66] and [25], selecting components from different color spaces and combining them as a new color space is proven helpful for improving the FR performance. In [25], a color space named as *RQCr* is proposed. Their experimental results show that the proposed color representation achieves better FR performance than the others. In [65], the authors propose a boosting color-component feature selection framework to choose the optimal set of color components from a pool of various color components.

The above fixed color component configurations have limitation to attaining the best result for a given FR task. This is because fixed color components might be effective for a particular FR problem but could not work well for other FR problems under other FR operating conditions (e.g., illumination variations). Research findings in [26] show that, color spaces computed by linear conversions from the *RGB* space achieve higher FR accuracies than those computed by nonlinear conversions from the *RGB* space. The use of linear color space conversions is thus preferred to obtain an enhanced FR performance. For example, a general discriminant model is proposed in [69] for color face recognition. This model involves two sets of variables: a set of color component combination coefficients for color image representation and a set of projection basis vectors for image discrimination. An iterative whitening-maximization algorithm is designed to find the optimal solution of the model. For the experiment 4 in the FRGC dataset, their method achieves the face verification rate of 74.91% at the false accept rate of 0.1%. Liu in [63] proposes three conversions from the *RGB* color space, resulting three new color representations, i.e., the so-called uncorrelated color space (UCS), the independent color space (ICS), and the discriminating color space (DCS). Specifically speaking, the UCS uses PCA [70] to decorrelate the R, G, and B component images, the ICS is obtained by independent component analysis [71]. The DCS applies discriminant analysis [70] to derive three new component images. Decorrelation of component images in three new color spaces helps reduce redundancy and is an important criteria in pattern classifier designation [70]. Compared with UCS, the ICS and DCS boost the discriminating power of their component images by ICA and the discriminant analysis, respectively. Experiments on the FRGC database [62] show that the ICS, DCS, and UCS achieve the face verification rate of 73.69%, 71.42%, and 69.92%, respectively, at the false accept rate of 0.1%, compared to the RGB color space, the 2-D KL color space, and the FRGC baseline algorithm with the face verification rate of 67.13%, 59.16%, and 11.86%, respectively, at the same false accept rate. In [26], the authors find out a common characteristic of the powerful color space for FR by analysing the transformation matrices of

Table 2.2: Comparisons of the face verification rates (FVR) (%) on FRGC database experiment 4 for various color spaces with the false accept rate (FAR) equivalent to 0.1%.

Color space	FVR (32×32 images)	FVR (64×64 images)
Grayscale	36.94	-
<i>RGB</i>	45.20	67.13
<i>HSV</i>	58.42	-
<i>XYZ</i>	57.88	-
<i>YUV</i> [23]	67.29	-
<i>YIQ</i> [68]	69.75	-
<i>LSLM</i>	69.88	-
<i>YQCr</i> [68]	71.85	-
<i>RQCr</i> [25]	72.51	-
<i>ZRG – NII</i> [26]	72.86	-
Extended IWM [69]	74.91	-
ICS [63]	-	73.69
DCS [63]	-	71.42
UCS [63]	-	69.92

different color spaces from the *RGB* color space. The common characteristic is that the summations of the elements from the second or the third row of the conversion matrix are both zero. Based on the characteristic of powerful color spaces, they propose two color space normalization techniques, which are able to convert weak color spaces into powerful ones, so that better FR performance can be obtained by making use of these normalized color spaces. Experiments conducted on the FRGC database show that their color space normalization approaches can be applied to weak color spaces to boost the FR performance [26].

Table 2.2 gives a comparison of above reviewed color configurations or color spaces used for face recognition.

2.1.3 Fusion of color features

Local texture features [72–74] such as Gabor wavelets and local binary pattern (LBP) are considered as effective face descriptors as they shown great robustness to variations such as pose, expression, illumination and occlusion. The Gabor image representation [75], which captures

well the salient visual features corresponding to spatial localization, orientation selectivity, and spatial frequency, displays robust characteristics in dealing with image variabilities. The LBP, which are originally introduced in [76] for texture analysis, have been successfully extended to describe faces, due to the finding that faces can be seen as a composition of micropatterns that are well described by the LBP operators. Unlike local features, global features such as principal component analysis (PCA) [14] and linear discriminant analysis (LDA) [15] lexically convert each face image into a high-dimensional feature vector and learn a feature subspace to preserve the statistical information of face images. In color FR, different feature extraction processes are employed to take advantage of the rich facial information in three different color channels. After various global and local color features are extracted from an image of a certain color space, how to fuse them for improving face recognition performance is an important problem.

In [77], the authors propose a color-and-frequency-feature (CFF) approach for the task of FR. To begin with, the color space RIQ consisting of the R component image from the RGB color space and the I, Q components of the YIQ color space is constructed. The CFF method extracts different features from the real part, the imaginary part, and the magnitude of the R, I , and Q color-component images, respectively, in the discrete-cosine-transform frequency domain. Different feature vectors are fused by the weighted similarity score fusion. In [19], the authors propose to combine multiple global and local features extracted from the $RCrQ$ color space. Specifically speaking, three different image encoding schemes are proposed, i.e., a patch-based Gabor image representation for the component image R , a multiresolution LBP feature fusion scheme for the component image Cr , and a component-based DCT multiple encoding for the component image Q . This approach achieves the face verification rate of 92.43% at false accept rate of 0.1% on FRGC version 2 Experiment 4 [62], through combining three similarity matrices by a weighted summation rule. In [67, 69], authors perform fusion for three color components by concatenating them into an augmented vector. Due to the magnitude differences between different color component images, some components might dominate

the concatenated feature vector compared with the other component images. Thus the authors normalize each component image by a zero-mean-unit-variance approach before concatenating different vectors. Then the dimension reduction approach EFM [17] is applied on the concatenated vectors for classification.

Two color local texture features, i.e., color local Gabor wavelets (CLGWs) and color local binary pattern (CLBP), are proposed in [1] for FR tasks. Given a color image, the Gabor or LBP operator is applied on each separate color channel. In addition, the texture operator is extended to make use of opponent color channels as shown on Fig. 2.1. These two color local texture features exploit information from spatio-chromatic texture patterns of different spectral channels within a certain local face region. In order to combine different features, multiple color local texture features corresponding different color channels are fused by a feature-level fusion method. Authors claim that directly classifying concatenated feature vectors suffers from the degradation of FR performance due to the high dimensionality and the redundant information. To solve this problem, they extract low-dimensional features first and then independently normalize them prior to their concatenation and classification. Experimental results show that the color local texture features deliver excellent face recognition performance for face images taken under severe variation in illumination, and for low resolution face images, compared with grayscale texture features.

Authors in [2] propose a color face descriptor, i.e., local color vector binary patterns (LCVBPs), for FR tasks. It consists of two patterns: color norm patterns and color angular patterns as shown on Fig. 2.2. In particular, a method is designed to extract color angular patterns, which encode the texture patterns of the multiple inter-band angles (one per pair of different color bands). To perform the final classification, the LCVBP feature is produced by fusing different features extracted from both color norm patterns and color angular patterns. The combination method is the same as that used in [1], which reduces the dimensionality of features in different color channels first then concatenate all low-dimensional features together for classification.

Figure 2.1: Three LBP histograms obtained by applying the opponent LBP operation to two local regions from different color component images in the YIQ color space [1].

Figure 2.2: Color norm patterns and color angular patterns from local region images of different color component images from the YIQ color space [2].

2.1.4 Summary

This subsection gives brief introductions to three categories of color FR research works reported so far. 1) Whether color information is helpful in improving the recognition accuracy compared with using grayscale images only. 2) Which color space performs the best for providing discriminate power needed to perform the reliable classification tasks. 3) What is the optimal feature set and how to combine different color features for improving FR performance. For category 1), the effectiveness of color information for FR has been proved by numerous color FR works. Color FR has made tremendous progress during the past several years. Nevertheless, there still exist some problems needed to be solved for category 2) and category 3).

In recent color-related face recognition works, researchers try to select a color space from existing color models or learn a color space from the given training data as the optimal one. *RQCr* [25], *DCS* [63] and *ZRG – NII* [26] color spaces have gained reputation as effective color spaces. However, at the moment, how to construct effective color spaces has not been thoroughly studied. Previous color spaces are designed based on different criteria thus their performance is not consistent on different databases. In chapter 3, we propose a color space LuC_1C_2 based on a framework of constructing effective color spaces for face recognition tasks. It consists of one luminance component Lu and two chrominance components C_1, C_2 . The luminance component Lu is selected among 4 different luminance candidates by analysing their R, G, B coefficients and the color sensor properties. To find the two effective chrominance components C_1, C_2 , the directions of their transform vectors are determined by the discriminant analysis and the covariance analysis in the chrominance subspace of the RGB color space. The magnitudes of their transform vectors are determined according to the discriminant values of Lu, C_1, C_2 . The proposed color space achieves higher face verification rates than state-of-the-art color spaces for both descriptor-based and CNN-based approaches. In particular, the face

verification performance of CNN models trained by LuC_1C_2 images is consistently better than those trained by RGB images under the condition of low-resolution.

For color feature fusion methods at the feature-level, existing methods either reduce the dimensionality of feature vectors in each color channel first and then concatenate all low-dimensional feature vectors, named as DR-Cat [1], or the vice versa, named as Cat-DR [18, 78]. For DR-Cat, due to the high dimensionality of images or generated color features, different color channels are usually processed separately and then concatenated together into a feature vector for classification [1, 2]. Specifically speaking, the number of low-dimensional feature dimensions in different color channels is set the same in existing methods. However, the importance or reliability of features in different color channels is not the same. In chapter 4, we propose a Color Channel Fusion (CCF) approach using jointly dimension reduction algorithms to select more features from more reliable and discriminative channels. Experiments using two different dimension reduction approaches, two different types of features on 3 image datasets show that CCF achieves consistently better performance than the existing Color Channel Concatenation (CCC) method which deals with different color channels equally. Moreover, DR-Cat ignores the correlation information between different features which is useful for classification. In Cat-DR, on the other hand, the correlation information estimated from the training data is fully used for feature fusion. But it may not be reliable especially when the number of training samples is limited. We propose a Covariance Matrix Regularization (CMR) technique in chapter 4 to solve problems of DR-Cat and Cat-DR. It works by assigning weights to cross-feature covariances in the covariance matrix of training data. Thus the feature correlation estimated from training data is regularized before being used to train the feature fusion model. The proposed CMR is applied to 3 feature fusion schemes: fusion of pixel values from 3 color channels, fusion of LBP features from 3 color channels, fusion of pixel values and LBP features from a single color channel. Extensive experiments of face recognition and verification are conducted on databases including MultiPIE [79], Georgia Tech [80], AR [61] and LFW

[81]. Results demonstrate that our proposed CMR technique considerably and consistently outperforms the best single feature, DR-Cat and Cat-DR.

In addition to the fusion of different color features, we also jointly consider information in three color channels during color feature extraction by proposing the Ternary Color LBP (TCLBP) descriptor. Color LBPs have shown excellent performance for color face recognition tasks, such as CLBP [1] and LCVBP [2]. However, these methods encode the inter-channel information on pairs of color channels by applying the same spatial structure as that used in the intra-channel encoding. This leads to a feature vector of very high dimensionality yet ineffective for encoding inter-channel information. Moreover, the difference of pixel values across color channels may not be a proper measure if they are not quantitatively comparable. We propose TCLBP in chapter 4, to encode the inter-channel information more effectively and efficiently. Extensive experiments on 4 public face databases, Color FERET [82], Georgia Tech, FRGC and LFW, are conducted to verify the effectiveness of the proposed TCLBP color feature for FR tasks. Experimental results show that our proposed TCLBP leads to visibly better FR performance than Color LBP, CLBP and LCVBP consistently.

2.2 Deep learning and low-resolution face recognition techniques

2.2.1 Deep learning models in face recognition

Among the many methods proposed in the literature, Convolutional Neural Networks (CNNs) [83, 84] have taken the community of FR by storm, significantly improving the state-of-the-art performance. This progress has been due to two factors: (i) end-to-end learning for the FR task using a CNN, and (ii) the availability of very large scale training datasets. CNN is usually used as a feature extractor, which is a learnable function composing several linear and non-linear operators [85]. In this subsection, we are going to briefly introduce recent progress of CNN architectures used in FR.

Taigman et al. [3] learn a nine-layer CNN model, DeepFace shown on Fig. 2.3, on the frontalized faces generated with a general 3D shape model from a dataset of 4 million examples spanning 4000 unique identities. This CNN model contains more than 120 million parameters by utilizing several locally connected layers without weight sharing, instead of the standard convolutional layers. In addition to using a very large amount of training data, DeepFace uses an ensemble of CNNs. Their best face verification performance of 97.35% on LFW [81] stems from an ensemble of three networks using different alignment methods and color channels.

Figure 2.3: Outline of the DeepFace architecture [3].

The DeepFace is extended by the DeepID series by Sun et al. in [4–7]. DeepID shown on Fig. 2.4 is learned as a classifier to differentiate around 10,000 face subjects in the training data and configured to keep reducing the number of neurons along the feature extraction hierarchy [4]. Compact identity-related features are formed in the top layers with only a small number of hidden neurons. DeepID extracts features from various face patches to form complementary and over-complete representations. Both PCA and a Joint Bayesian model [21] are used to train a metric for classification. DeepID2 shown on Fig. 2.4 uses both face identification and verification signals as supervision [5]. Specifically, softmax (identification) and contrastive (verification) cost are combined to construct the objective function. The DeepID2+ shown on Fig. 2.4 branches out a fully connected layer after each convolution layer [6]. Very deep neural networks are used in DeepID3 shown on Fig. 2.5 in [7], which is rebuilt from stacked convolution and inception layers. Compared to DeepFace, DeepID series do not use 3D face alignment, but a simpler 2D affine alignment. However, the final models involving a large number of CNNs are quite complicated.

Figure 2.4: DeepID, DeepID2, and DeepID2+ structures [4–6].

Figure 2.5: DeepID3 structure [7].

DeepFace and DeepID series are based on complex systems of several steps, which combine the outputs of several deep convolutional networks. Recently in [22], authors present a CNN model, named FaceNet, which learns a mapping from face images to an Euclidean space where distances correspond to the measure of face similarity. Two different deep network architectures are explored. The first architecture is based on the 22-layer Zeiler & Fergus [86] model. The second architecture is based on the GoogLeNet style Inception model [87], which is based on layers that use different sizes of convolutional filters and pooling layers in parallel. A point of difference of FaceNet is in their use of the triplet loss, which is computed from two matching face thumbnails and a non-matching face thumbnail and the minimizing the loss separates the positive pair from the negative by a distance margin. An online triplet mining method is used to ensure consistently increasing difficulty of triplets as the network trains. In training, this loss is applied to multiple layers, not just the final one. This method currently achieves the best performance on LFW [81] and YTF [88].

Authors in [8] propose a large face dataset of 2.6M images of 2622 subjects. A very deep CNN, VGG-Face, comprising a long sequence of convolutional layers, is also proposed as shown on Fig. 2.6. The filter size in convolutional layers is set to be 3×3 . A 2D similarity transformation is used for face alignment. The triplet loss [22] and Joint Bayesian model [21] are used for metric learning. This model demonstrates that stacking small filters to approximate large filters and building very deep convolution networks not only reduce the number of parameters but also increase the non-linearity of the network.

A CNN architecture similar to VGG-Face is proposed in [89] for unconstrained face verification. PReLU is used as an alternative to ReLU as the non-linear unit in their work. Moreover, two local normalization layers are added after first two convolutional layers to mitigate the effect of illumination variations. An average pooling layer is used instead of a fully-connected layer as the last layer to generate a compact and discriminative feature representation. Using the Joint Bayesian metric which has achieved good performances on face verification problem-

layer type name	0 input	1 conv	2 relu	3 conv	4 relu	5 mpool	6 conv	7 relu	8 conv	9 relu	10 mpool	11 conv	12 relu	13 conv	14 relu	15 conv	16 relu	17 mpool	18 conv
support	–	3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3
filt dim	–	3	–	64	–	–	64	–	128	–	–	128	–	256	–	256	–	–	256
num filts	–	64	–	64	–	–	128	–	128	–	–	256	–	256	–	256	–	–	512
stride	–	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
pad	–	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1

layer type name	19 relu	20 conv	21 relu	22 conv	23 relu	24 mpool	25 conv	26 relu	27 conv	28 relu	29 conv	30 relu	31 mpool	32 conv	33 relu	34 conv	35 relu	36 conv	37 softmax
support	1	3	1	3	1	2	3	1	3	1	3	1	2	7	1	1	1	1	1
filt dim	–	512	–	512	–	–	512	–	512	–	512	–	–	512	–	4096	–	4096	–
num filts	–	512	–	512	–	–	512	–	512	–	512	–	–	4096	–	4096	–	2622	–
stride	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1
pad	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0

Figure 2.6: Details of the VGG-Face architecture [8].

s, their model trained on the CASIA-WebFace dataset achieves the face verification accuracy of 97.45% on LFW.

Authors in [44] propose a complex deep learning approach to extract face representation by exploring the multimodal information. An ensemble of CNNs extract complementary facial features from the original face image, the face image frontalized by a 3-dimensional model, and various face patches. Then, all extracted features are concatenated together to form an augmented feature vector, whose dimension is further reduced by Stacked Auto-Encoder. Their ensemble system gains the face verification accuracy of 99.0% on the LFW dataset using the Joint Bayesian (JB) model to measure the similarity between features.

Recent evidence in [90] reveals that network depth of CNN is of crucial importance for good performance. Many visual recognition tasks have also greatly benefited from very deep models. However, an obstacle to training deep neural networks is the notorious problem of vanishing/exploding gradients. This makes deep learning systems difficult to optimize. In [9], authors address this problem by introducing a deep residual learning framework. Instead of hoping each few stacked layers directly fit a desired underlying mapping, they explicitly let these layers fit a residual mapping by skipping layers or using “shortcut connections” as shown on Fig. 2.7. As these residual networks are easier to optimize, they can gain accuracy from considerably increased depth. On the ImageNet dataset [91], the authors evaluate residual nets

Table 2.3: Comparisons of accuracy (%) on LFW and numbers of training images for different deep learning models.

Model	Training images	Accuracy (%) on LFW
DeepFace [3]	4.4M	97.35
VGG-Face [8]	2.6M	98.95
DeepID3 [7]	300K	99.47
MMDFR [44]	494K	99.02
FaceNet [22]	200M	99.63
UFV [89]	494K	97.45

with a depth of up to 152 layers. An ensemble of these residual nets won the 1st place on the ILSVRC 2015 classification task.

Figure 2.7: Residual learning: a building block [9].

Table 2.3 gives introductions to numbers of training images and face verification accuracies of above reviewed deep learning models on LFW database.

2.2.2 Summary

This subsection reviews the most relevant progress for deep-learning-based architectures used in FR. The success of CNNs is attributed to their ability to learn rich image representations. But training CNNs relies on estimating millions of parameters and requires a very large number of annotated training images. It is impossible to collect the training data and train a CNN model from scratch for every different application scenario. A widely-used alternative is to fine-tune the pre-trained CNN model that has been trained using a large set of labeled images. However, our

experiments show that pre-trained CNNs with or without fine-tuning cannot provide satisfactory FR performance when training and testing datasets have large differences. To address this problem, we investigate combining high-level CNN features with low-level color pixel values in chapter 5. Color pixel values are basic low-level features and they keep the most of the original appearance information of face images from the application of interest. This information is complementary to the CNN representations. These two features are combined by weighted similarity score fusion instead of feature-level fusion considering their big differences. Experiments are conducted on LFW and FRGC databases using the widely-used pre-trained CNN model, VGG-Face [8]. Results show that the low-level information contained in color pixels greatly improves the face verification rate/accuracy of VGG-Face with or without fine-tuning.

The fusion of multiple features is important for achieving state-of-the-art FR results. Different from the architecture of VGG-Face, ResNet in [9] consists of residual modules which conduct additive merging of signals. The authors in [9] argue that residual connections are inherently important for training very deep architectures. It is natural to study the combination of VGG-Face with ResNet, which would allow two models to reap the benefits of each other. In chapter 5, we train a Residual model by referring to the CNN architecture used in [45] and name it as ResNetShort. Training images are taken from the recently released CASIA-WebFace dataset and the supervising signals include both softmax loss and center loss [45]. Thus compared with VGG-Face, the obtained ResNetShort model uses different training data, a different network architecture, and different supervising loss functions. The features of ResNetShort are combined with those of the pre-trained VGG-Face model by a feature-level fusion method, CMR, proposed in chapter 4. Extensive experiments conducted on four popular face databases show that better FR performance is achieved by the combined features when compared with features of VGG-Face or ResNetShort alone.

2.2.3 Low-resolution face recognition approaches

With the growing installation of surveillance cameras in many areas, there is an increasing demand of face recognition technology for surveillance applications. Recognition performance will be dramatically degraded under this condition [92, 93] when face region in the scene is normally very small. In this subsection, we focus on the task of matching low-resolution (LR) probe images to high-resolution (HR) gallery images. Therefore, the main challenge of this LRFR problem is to handle the mismatch of resolution. Recently, quite a few LRFR approaches have been developed [10, 47, 94, 95], and these methods can be broadly classified into two categories: super-resolution-based (SR) methods and subspace-based methods. SR methods hallucinate image details, like image edges and textures from tremendous training natural images. Although SR methods can generate visually appealing HR images [92, 96–98], they often introduce some artificial parts, which usually decrease the recognition performance. Also, the target of SR methods is not absolutely consistent with that of FR thus their performance is unsatisfactory. Moreover, time-consuming sophisticated SR algorithms are not suitable for real-time FR applications.

The other category of subspace-based methods usually adopt coupled mappings to project LR and HR faces into a latent subspace. Li et al. propose coupled locality preserving mappings in [47]. Kernel trick is introduced in [95] to learn coupled mappings. Deep neural network is applied to addressing the LRFR problem in [10]. Both LR images and HR images are fed into a resolution-invariant deep network during training, which achieves state-of-the-art performance on SCface database [99]. Multidimensional scaling (MDS) is also employed to solve the LRFR problem in [100, 101].

The method of coupled mappings (CMs) is proposed in [47], it projects the face images of different resolutions into a unified feature space which favors the task of classification. These CMs are learned through minimizing the difference between the correspondences (i.e., low-resolution image and its high-resolution counterpart). This principle could be formulated by

the objective function in equation (2.1). N_t indicates the number of the training images. f_L, f_H are coupled mappings to be learned, the first one is for LR feature vectors $\mathbf{l}_i$ and the other one is for HR features $\mathbf{h}_i$. $\|\cdot\|$ indicates the L2 norm. An improved version with a penalty weighting matrix added to the objective function is also proposed in [47].

$$J(f_L, f_H) = \sum_{i=1}^{N_t} \|f_L(\mathbf{l}_i) - f_H(\mathbf{h}_i)\|. \quad (2.1)$$

Instead of using a linear projection to solve the problem of computing similarity metrics between HR and LR images, authors in [95] learn coupled kernel matrices to map the face images with different resolutions onto an infinite subspace and carry out the recognition step in the new space. Comparing multimodal data is difficult for conventional methods in practice due to the lack of efficient similarity computation. Their proposed method solves this problem by minimizing the dissimilarities captured by their kernel Gram matrices in the LR and HR spaces.

A CNN approach is proposed in [10] to address the problem of matching a LR face image against a gallery of relatively high resolution (HR) face images. The authors treat the discrimination information of HR and LR face images equally to boost the performance. They mix the real HR images with the upsampled LR ones to learn resolution invariant features in a supervised way with a deep convolutional network shown on Fig. 2.8. Finally, the cosine distance metric is employed to obtain recognition results. Three advantages are claimed for their approach: (i) Since resolution invariant features can be learned offline from training data, it is fast and suitable for databases of large size; (ii) any image is available for training; (iii) the obtained model has good generalization ability since no test data is used for training.

Biswas et al. [100] learn a mapping matrix via Multidimensional Scaling (MDS), which projects the HR and LR images into a common space where the distances between them approximate the distances had the probe images been captured in the same conditions as the

Figure 2.8: Structure of the Resolution-Invariant Deep Network for resolution-robust feature extraction [10].

gallery images. Images from the same class are coupled together while the distances of different classes approximate their distances in the HR space. SIFT-based descriptors at fiducial locations on the face image are employed as the feature representations.

In [101], a Discriminative Multidimensional Scaling (DMDS) method is proposed. It makes use of both inter-class distances and intra-class distances. DMDS projects the HR and LR images into a common space in which images of same subject are coupled together. Different from MDS in [100], in DMDS, the distances of different subjects can be larger than their distances in the HR space since the interclass constraint is added in the objective function. The new constraint ensures discriminability. Furthermore, the authors also propose the local-consistency-preserving DMDS (LDMDS) to preserve the local consistency, which considers not only the relationship of HR-LR images, but also the relationship of HR-HR images as well as that of LR-LR images.

A comparison of recognition results for recently proposed LRFR methods on the SCface database is concluded in Table 2.4.

Table 2.4: Recognition rates (%) at distance d3/d2/d1 on SCface for different methods of LRFR.

Method	d3	d2	d1
RICNN [10]	74.0	67.0	23.0
MDS [100]	61.1	66.0	60.3
DMDS [101]	62.9	67.2	61.5
LDMS [101]	65.5	70.7	62.7

2.2.4 Summary

In this subsection, we review recently proposed approaches for the task of matching LR probe images to HR gallery images. Existing methods including super resolution, coupled mappings and multidimensional scaling yield only modest performance for the task of LRFR because they use pixel values or SIFT as the feature representations. RICNN in [10] learns resolution invariant features in a supervised way by mixing the real HR images with the upsampled LR ones. Although RICNN improves the performance of LRFR, it is sensitive to resolution change of probe images as indicated in [101]. We propose a novel CNN-based approach, the Deep Coupled ResNet (DCR) model in chapter 5. It consists of one trunk CNN and two branch networks. The trunk CNN, trained by face images of 3 significantly different resolutions, is used to extract discriminative features robust to the resolution change. Two branch networks, trained by HR images and images of the targeted LR, work as resolution-specific coupled mappings to transform HR and corresponding LR features to a space where their difference is minimized. Model parameters of branch networks are optimized using our proposed Coupled-Mapping (CM) loss function, which considers not only the discriminability of HR and LR features, but also the similarity between them. To deal with various possible resolutions of probe images in the face recognition task, different pairs of small branch networks are trained to be combined with the same trunk network. Thorough evaluation on LFW and SCface databases shows that the proposed DCR model achieves consistently and significantly better performance than the state-of-the-arts.

2.3 Databases used for performance evaluation of face recognition in this thesis

Here, we provide an introduction of databases used for evaluation of color face recognition in this thesis.

2.3.1 AR

The AR dataset [61] consists of around 4000 color face images of 126 people, which have different expressions, lighting conditions and occlusions. In the experiments of this thesis, 1400 frontal-face images captured across 2 sessions (separated by two weeks) of 100 subjects (50 males and 50 females) are selected. In each session, there are 7 undisguised images with different facial expressions and lighting conditions for each subject.

2.3.2 GT

The Georgia Tech (GT) [80] dataset consists of face images from 50 subjects. For each subject, there are 15 images. There exist variations of pose, expression, cluttered background and illumination in the images.

2.3.3 FRGC

Face images in the FRGC database [62] are partitioned into three datasets, i.e., training, target and query datasets. There are 12,776 controlled or uncontrolled images in the training set, 16,028 controlled images in the target set, and 8,014 uncontrolled images in the query set. The controlled images exhibit good quality, but the uncontrolled images show poor quality, such as large illumination variations, low resolution, and blurring. These uncontrolled factors pose grand challenges to face recognition performance. FRGC Experiment 4 has been reported to be the most challenging FRGC experiment [63], so it is chosen to assess the face recognition performance in the experiments.

2.3.4 LFW

The LFW dataset [81] has been broadly used as a benchmark database to assess face recognition approaches [4, 18]. It contains 13,233 face images from 5,749 persons. These images are collected from the internet and exhibit great variations of illumination, pose, expression, and occlusions.

2.3.5 Multi-PIE

The face images in the CMU Multi-PIE database are captured under variations of illumination, expression and pose across 4 sessions. The first 105 subjects which appear in all 4 sessions are used in the experiments.

2.3.6 SCFace

The SCface database contains images of 130 subjects taken in uncontrolled indoor environment using five video surveillance cameras of various qualities. The cameras are placed slightly above the subject's head, and the humans are not required to look at a fixed point during the recordings, thus the collected images are blurred. For each subject, there are 15 images taken at three distances (5 images at each distance), 4.20m (d1), 2.60m (d2) and 1.00m (d3), by surveillance cameras, and one frontal mugshot image taken by a digital camera. The surveillance-camera images suffer from low-resolution, uncontrolled illumination, different poses, and noises simultaneously. Face images captured at d1 are of the most poor quality compared with d2 and d3.

Databases used for evaluation of color face recognition in this thesis are summarized on Table 2.5.

Table 2.5: Databases used for evaluation of color face recognition.

Database (Year)	No. of subjects	Conditions	Image resolution	No. of images
AR (1998)	100	Illumination and expressions	165×120	1,400
GT (1999)	50	Illumination, expressions, and poses	150×150	750
FRGC (2006)	466	Illumination, expressions, and low resolution	130×150	36,807
LFW (2007)	5,749	Unconstrained environments	150×80	13,000
Multi-PIE (2008)	105	Illumination poses, and expressions	160×130	11,655
SCFace (2011)	130	Low resolution, illumination, different poses, and noises	100×70 , 60×40 , and 30×20	2,990

Chapter 3

Color Space LuC_1C_2

3.1 Introduction

Recent research efforts reveal that color provides useful information for machine learning and pattern recognition tasks, such as color constancy [54] and color histograms [102]. Color in the machine vision system is defined by a combination of 3 color components specified by a color space. For visual recognition tasks like object detection, indexing, retrieval and recognition [54, 102–108], different color spaces possess significantly different characteristics and effectiveness in the light of discriminating ability [64]. For example, the hue, saturation, value (HSV) color space and the luminance, chrominance-blue, chrominance-red (YCbCr) color space have been widely utilized in face detection tasks [105]. And the R component image in the RGB color space has been proven to be more powerful than the other component images for face recognition tasks in [68].

Face recognition has become a very active research area driven mainly by its broad applications in human-computer interaction, homeland security, and entertainment [109–111]. Color information plays a discriminative and complementary role in face recognition process. Torres et al. [23] applied a modified PCA scheme to face recognition and their results show that the use of color information improves the recognition rate when compared to the same scheme using the luminance information only. The improvement can be significant when large facial expression and illumination variations are present or the resolution of face images is

low [24, 25]. Since then, considerable research efforts have been devoted to the efficient utilization of facial color information to enhance the face recognition performance [1, 19, 26–28, 63, 65, 66, 69, 77, 112–114].

Recently, many color spaces have been proposed to find the optimal way of representing color images for face recognition. These color spaces are usually derived by linear or nonlinear transformations from the RGB color space. In the early studies, color configurations were made through a combination of intuition and empirical comparisons without a systematic strategy, such as YCbCr, YIQ [68], YUV [23] and the hybrid color space YQCr in [66]. In YIQ, the Y and Q components are taken from the YIQ space while the Cr component is taken from the YCbCr space. Among all kinds of configurations discussed in [25], RQCr, where R is taken from the RGB color space and Q, Cr are taken from the YIQ and YCbCr color spaces, respectively, shows the best face recognition performance on more than 3000 color facial images collected from three standard face databases [25]. Moreover, the R channel is known to be the best monochrome channel for face recognition [115] and QCr is the best chromaticity-component combination on the face recognition grand challenge database and evaluation framework [82]. Thus the RQCr color space was also used in [19] to extract multiple color features from face images.

Unlike the above heuristic selections of color components, the DCS color space was proposed by seeking 3 sets of coefficients to linearly combine the R, G and B component images based on a discriminant criterion in [27]. Its experimental results show that the DCS color space is effective for enhancing the face recognition performance obtained from RGB or $I_g(r-g)$ color spaces. Compared with other learning-based methods such as ICS and UCS [63], DCS retains the face spatial structure information. Later in [26], a common characteristic of effective color spaces for face recognition was found out by analysing the transformation matrices of different color spaces from the RGB color space. Based on this characteristic of effective color spaces, the authors proposed two color space normalization techniques, which are able

to convert weak color spaces into effective ones, so that better face recognition (FR) performance can be obtained by using these normalized color spaces. Among different normalized color spaces assessed in [26], the ZRG-NII color space, where Z describes the luminance information and N_1, N_2 are obtained by normalizing the ZRG color space, achieved the best FR performance. Therefore, the RQCr and ZRG-NII color spaces were considered to be the two most effective color representations devised for the purpose of face recognition in [1].

DCS, RQCr and ZRG-NII color spaces do achieve better face recognition performance than the others for some databases. However, their performance is not consistent for different databases. Besides, they were proposed based on different criteria. In this chapter, an effective color space is proposed based on a color-space-construction framework. By analysing color spaces that demonstrate good classification capabilities, we find that they are all composed of one luminance component and two chrominance components except the learning-based DCS color space. This configuration reduces the correlation between different color components and enhances the discriminating power of the color space. Based on the proposed framework, we construct an effective color space LuC_1C_2 . The luminance component Lu is selected among four luminance candidates from existing color models by analysing their R,G,B coefficients and the color sensor properties. In order to generate the two effective chrominance components C_1, C_2 , the directions of their transform vectors are derived by the discriminant analysis and the covariance analysis in the chrominance subspace of the RGB color space. The magnitudes of their transform vectors are derived according to the discriminant values of Lu, C_1, C_2 . In the experiments, the dependence of the face recognition performance on the correlation of two chrominance components is firstly validated. Then the FR performance of the proposed LuC_1C_2 color space is compared with those of DCS, RQCr, ZRG-NII and RGB color spaces on 4 benchmark databases (AR, GT, Multi-PIE and FRGC) using 2 distinct color features and 3 different face recognition methods. After that, we show that the fusion of multiple features extracted from the proposed color space achieves better face verification rate than published

state-of-the-art results on FRGC database. Finally on GT and LFW databases, we show that the CNN model trained using the proposed color space LuC_1C_2 performs consistently better than that trained using RGB color space when face images are in low-resolution.

The novelty of the proposed approach comes from 1) the framework of constructing an effective color space for face recognition; 2) R,G,B coefficients analysis and color sensor analysis for luminance components; 3) the generating process of two chrominance components; 4) an effective color space LuC_1C_2 which consistently performs better than state-of-the-art color spaces on four benchmark databases using two distinct features and three different face recognition methods; 5) achieving state-of-the-art FVR on FRGC database using our proposed color space; 6) improving the robustness of CNN to low-resolution degradation using the proposed color space LuC_1C_2 on GT and LFW databases.

3.2 The proposed color space LuC_1C_2

In this section, the framework of constructing an effective color space for face recognition tasks is presented first and then an effective color space LuC_1C_2 is constructed.

3.2.1 Overview of the proposed approach

In video, luminance (also called as gray, brightness, intensity, and lightness) is formed as a weighted sum of R,G,B components to represent the brightness of pixels in an image, and chrominance conveys the color information of the picture, separately from the accompanying luminance [116]. For color spaces widely-used in video compression standards (e.g., MPEG and JPEG) [105], luminance and chrominance are usually processed separately [117]. Motivated by this, we study the framework of constructing an effective color space for face recognition. To begin with, several existing effective color spaces are investigated. Color spaces including $I_1I_2I_3$ [118], YUV, YIQ, YCbCr, [23, 119], YQCr [66], RQCr [25], LSLM [120] and ZRG-NII [26] have shown good classification capabilities for face recognition. One

common characteristic of these color spaces is that they are all composed of one luminance component (I_1, Y, R, L, Z) and two chrominance components ($I_2I_3, UV, IQ, CbCr, QCr, SLM, N_1N_2$). Experimental results in [57] suggest that color cues do play a role in face recognition and their contribution becomes evident when shape cues are degraded. Under such conditions, face recognition performance with color images is significantly better than that with grayscale images. The separation of luminance and chrominance in [26] shows that when we reduce the correlation between different color components, the discriminative information possessed by different color components will become mutually complementary to each other. Based on these analyses, we propose the framework of constructing an effective color space for face recognition as shown on Fig. 3.1.

Figure 3.1: The framework of constructing an effective color space for the task of face recognition.

In the proposed framework, three color components of the proposed color space, Lu, C_1, C_2 , are computed through linear conversions from the RGB color space. The luminance component Lu is selected among 4 different luminance candidates by analysing their R,G,B coefficients and the color sensor properties. To obtain the two effective chrominance components C_1, C_2 , we first determine the directions a_1, a_2 of their transform vectors by the discriminant analysis and the covariance analysis in the chrominance subspace of the RGB color space, and then

determine the magnitudes τ_1, τ_2 of their transform vectors according to the discriminant values of Lu, C_1, C_2 .

3.2.2 Selection of the luminance component

In many face recognition algorithms, a color image of the RGB color space is converted into a monochrome image by linearly combining its three color components [69]. However, theoretical and experimental justifications are lacking for investigating which monochrome image is the best representation of the color image for the recognition purpose. In this section, 4 widely-used luminance components including I_1 from $I_1I_2I_3$ [121], R from RGB, Y from YUV [119] and L from LSLM [120] are compared. They are computed from the RGB color space by

$$\begin{pmatrix} I_1 \\ R \\ Y \\ L \end{pmatrix} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1 & 0 & 0 \\ 0.299 & 0.587 & 0.114 \\ 0.209 & 0.715 & 0.076 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \quad (3.1)$$

The $I_1I_2I_3$ color space proposed by Ohta et al. uses a K-L transformation to decorrelate the R,G,B components. The effectiveness of the luminance component I_1 for face verification was shown in [118]. But I_1 implicitly assumes a uniform distribution of light energy over the entire color space.

The color component R in the RGB color space has been shown to perform better than other intensity images including I_1 and Y or Gray for face retrieval [68]. In addition, the R channel of skin-tone color is known to be the best monochrome channel for face recognition [115]. Due to the extraordinary FR performance of the R component, it is usually used as a luminance component. However, the performance superiority of the R component over the others is reflected only on the FRGC database [62]. In general, it is not effective to select R as the luminance component as R discards the useful information by ignoring G and B components.

YUV, YIQ, YCbCr are three color spaces commonly used for efficient video transmission. In these 3 color spaces, the R,G,B components are separated into the luminance component Y and remaining chrominance components. Y performs the best to display pictures on monochrome (black and white) televisions [122]. And it is used in almost all research works doing traditional face recognition on monochrome images [16, 73, 123]. Nevertheless, there is no proof that Y is the optimal monochromatic form of color images for face recognition. The LSLM color space is obtained through a linear conversion from the RGB color space according to the opponent signals of the cones: blackwhite, redgreen, and yellowblue. And the L component describes the luminance information [120].

Figure 3.2: The normalized response of human cone cells for different wavelengths of light [11].

Fig. 3.2 [11] shows the normalized response of human cone cells against the wavelength of light, where the shapes of the curves are obtained by measurement of the light absorption by the cones, but the relative heights for the three types (L, S, M) are set equal. Among red, green and blue lights, we can see that the human eye is most sensitive to green light, followed by red light and then blue light. The blue light mainly stimulates S cones, the red light stimulates the two most common (M and L) of the three kinds of cones, and the green light stimulates M and

Figure 3.3: (a) The color filter array in cameras and (b) the camera spectral sensitivity of nikon-D70 [12].

L cones at a higher response level than the red light. Moreover, noise in the green channel is much less than that in the other two primary colors [124].

To produce a color image in the machine vision system, a digital device needs three different kinds of sensors to acquire the red, green, and blue parts of the spectrum [125]. 50% of the camera sensors in the color filter array are green sensors, which are twice as many as red or blue sensor as shown on Fig. 3.3(a). Fig. 3.3(b) shows the spectral sensitivity of the Nikon-D70 sensors for different wavelengths of the light. One observation from Fig. 3.3(b) is that the highest sensitivity of G sensor is almost twice as large as that of R or B sensor. Another observation from Fig. 3.3(b) is that, above the sensitivity level of 0.1, the passband width of G sensor (150nm) is much larger than that of R sensor (90nm) or B sensor (100nm). What's more, the spectrum of G sensor spreads from the central to the two sides of the visible spectrum rather than reside at one side of the visible spectrum like those of R and B sensors. To summarize, 50% of camera sensors in the color filter array are green sensors, each of them has higher sensitivity level and broader passband compared with those of red and blue sensors. Therefore, assigning the highest weight (significantly larger than 0.5) to G in the linear combination of R, G, B components is expected to produce a better luminance component Lu . The second highest weight should be assigned to R since it performs much better than B for face

recognition in [25, 68]. Following these analyses, we select L that has R,G,B weights of 0.209, 0.715 and 0.076 respectively as the luminance component Lu .

3.2.3 Extraction of two chrominance components

Chrominance is used in video systems to convey the color information of the picture, separately from the accompanying luminance signal [116]. Different chrominance components are computed through linear conversions from the RGB color space as shown below:

$$\begin{pmatrix} I_2 \\ I_3 \\ U \\ V \\ I \\ Q \\ S \\ LM \\ N1 \\ N2 \end{pmatrix} = \begin{pmatrix} 0.5 & 0 & -0.5 \\ -0.5 & 1 & -0.5 \\ -0.1471 & -0.2888 & 0.4359 \\ 0.6148 & -0.5148 & -0.1 \\ 0.596 & -0.274 & -0.322 \\ 0.211 & -0.523 & 0.312 \\ 0.209 & 0.715 & -0.924 \\ 3.148 & -2.799 & -0.349 \\ 0.821 & -0.215 & -0.606 \\ -0.179 & 0.785 & -0.606 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \quad (3.2)$$

These chrominance components are from the $I_1I_2I_3$, YUV, YIQ, LSLM, and ZRG-NII color spaces respectively.

It is easy to find a common characteristic of these chrominance components that the sum of R,G,B coefficients $\mathbf{u}^T = (u_1, u_2, u_3)$ is zero for every chrominance component C as shown in (3.3) and (3.4).

$$C = \mathbf{u}^T \begin{pmatrix} R \\ G \\ B \end{pmatrix} = (u_1 \quad u_2 \quad u_3) \begin{pmatrix} R \\ G \\ B \end{pmatrix} \quad (3.3)$$

$$\sum_{j=1}^3 u_j = 0. \quad (3.4)$$

Our explanation for this characteristic is that, the luminance information contained in R, G and B components gets cancelled and the remaining information forms a chrominance component C. In order to generate a chrominance component, the sum of u_1, u_2, u_3 needs to be zero. Therefore, $\mathbf{u}$ is orthogonal to $\mathbf{b} = (1, 1, 1)^T$ in the 3-dimensional RGB color space as follows:

$$\sum_{j=1}^3 u_j = \mathbf{u}^T \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \mathbf{u}^T \mathbf{b} = 0. \quad (3.5)$$

Now the question is: how to find two appropriate 3×1 vectors $\mathbf{u}_1, \mathbf{u}_2$ to linearly combine R,G,B in (3.3) and produce two chrominance components C_1, C_2 respectively. According to the condition in (3.5), C_1, C_2 will be the two chrominance components as long as $\mathbf{u}_1$ and $\mathbf{u}_2$ lie in the 2-dimensional subspace orthogonal to $\mathbf{b}$, which is referred to as the chrominance subspace in this chapter.

But which pair of $\mathbf{u}_1, \mathbf{u}_2$ lying in the 2-dimensional subspace orthogonal to $\mathbf{b}$ is the optimum choice for enhancing the face recognition performance is still a question. It has been shown in recent studies [54, 126, 127] that the face recognition performance is related with both the discriminating power and the correlation of color components. To find $\mathbf{u}_1, \mathbf{u}_2$ that maximize the discriminative power and minimize the correlation of C_1, C_2 , we apply the discriminant analysis and the covariance analysis to the chrominance subspace of the RGB color space. As the eigenvector obtained from discriminant analysis and the vector obtained from covariance analysis are of arbitrary lengths, they provide only the direction information. Therefore, we represent $\mathbf{u}_1, \mathbf{u}_2$ by the magnitude and the direction as follows:

$$\mathbf{u}_1 = \tau_1 \mathbf{a}_1, \mathbf{u}_2 = \tau_2 \mathbf{a}_2, \quad (3.6)$$

where $\tau_i, \mathbf{a}_i (i = 1, 2)$ indicate the magnitude and the direction of vector $\mathbf{u}_i$ respectively.

For classification purposes, the Fisher criterion [128] defined in (3.7)[27] is an effective criterion because it maximizes the between-class variation and minimizes the within-class variation of the projected data simultaneously.

$$\max_{\mathbf{x}} \mathbf{J}(\mathbf{x}) = \max_{\mathbf{x}} \frac{\mathbf{x}^T \mathbf{S}_b \mathbf{x}}{\mathbf{x}^T \mathbf{S}_w \mathbf{x}} \quad (3.7)$$

In (3.7), $\mathbf{J}$ indicates the ratio of the inter-person variation to the intra-person variation of the data projected to $\mathbf{x}$. $\mathbf{S}_b, \mathbf{S}_w$ are the between-class scatter matrix and the within-class scatter matrix of the input data, respectively.

In our case, C_1 will be a chrominance component as long as $\mathbf{u}_1$ lies in the 2-dimensional chrominance subspace (3.5). Under this condition, the utilization of the Fisher criterion helps us find the optimal $\mathbf{u}_1$ which maximizes the discriminative power of C_1 . Thus pixel values of color images are transformed from the 3-dimensional RGB color space to the 2-dimensional chrominance subspace as below:

$$\mathbf{B} = \begin{pmatrix} \mathbf{n}_1^T \\ \mathbf{n}_2^T \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}, \quad (3.8)$$

where $\mathbf{n}_1, \mathbf{n}_2$ are two basis vectors that span the 2-dimensional chrominance subspace $\mathbf{B}$ orthogonal to $\mathbf{b}$. Although $\mathbf{n}_1, \mathbf{n}_2$ are not unique, our final results are independent of the choice of $\mathbf{n}_1, \mathbf{n}_2$, because any pair of orthogonal vectors of unit length orthogonal to $\mathbf{b}$ determines the same subspace $\mathbf{B}$. Suppose each image in the training dataset contains k pixels. Let $\mathbf{B}_{ij} \in \mathbf{R}^{2 \times k}$ denote the projection of the j -th RGB image of class i into the 2-dimensional chrominance subspace $\mathbf{B}$, where $i = 1, 2, \dots, p, j = 1, 2, \dots, q_i$. p indicates the number of classes and q_i indicates the number of images for class i .

To compute the value of $\mathbf{J}$ in (3.7), we define the within-class scatter matrix $\mathbf{S}_w \in \mathbf{R}^{2 \times 2}$ and the between-class scatter matrix $\mathbf{S}_b \in \mathbf{R}^{2 \times 2}$ as follows:

$$\mathbf{S}_w = \sum_{i=1}^p \sum_{j=1}^{q_i} (\mathbf{B}_{ij} - \bar{\mathbf{B}}_i)(\mathbf{B}_{ij} - \bar{\mathbf{B}}_i)^T, \quad (3.9)$$

$$\mathbf{S}_b = \sum_{i=1}^p q_i (\bar{\mathbf{B}}_i - \bar{\mathbf{B}})(\bar{\mathbf{B}}_i - \bar{\mathbf{B}})^T, \quad (3.10)$$

where

$$\bar{\mathbf{B}}_i = \frac{1}{q_i} \sum_{j=1}^{q_i} \mathbf{B}_{ij}, \quad (3.11)$$

$$\overline{\mathbf{B}} = \frac{1}{q} \sum_{i=1}^p \sum_{j=1}^{q_i} \mathbf{B}_{ij}. \quad (3.12)$$

$\overline{\mathbf{B}}_i$ indicates the mean matrix of training images in class i and $\overline{\mathbf{B}}$ indicates the mean matrix of all training images. q is the number of all training samples and $q = \sum_{i=1}^p q_i$.

As both $\mathbf{S}_b$ and $\mathbf{S}_w$ are positive definite matrices, $\mathbf{J}(\mathbf{x})$ in (3.7) is actually a generalized Rayleigh quotient [129]. Thus maximizing $\mathbf{J}(\mathbf{x})$ is equivalent to solving a generalized eigenvalue problem

$$\mathbf{S}_b \mathbf{x} = \lambda \mathbf{S}_w \mathbf{x}, \quad (3.13)$$

where $\mathbf{x} \in \mathbf{R}^{2 \times 1}$ is an eigenvector and λ is the corresponding eigenvalue. After $\mathbf{S}_b, \mathbf{S}_w$ are substituted into (3.13), the generalized eigenvector $\mathbf{x}$ corresponding to the largest eigenvalue is chosen as the optimal projection vector since the eigenvalue indicates the discriminative power of data projected to the corresponding eigenvector. Note that the eigenvector $\mathbf{x}$ obtained from (3.13) is of arbitrary length, which means it provides only the direction information. Here, we take $\mathbf{x}$ of the unit length. Then $\mathbf{a}_1$ is computed by

$$\mathbf{a}_1 = (\mathbf{n}_1, \mathbf{n}_2) \mathbf{x}. \quad (3.14)$$

And C_1 is given by

$$C_1 = \mathbf{u}_1^T \begin{pmatrix} R \\ G \\ B \end{pmatrix} = \tau_1 \mathbf{a}_1^T \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \quad (3.15)$$

The key of color face recognition techniques is how to effectively utilize the complementary information between color components and remove their redundancy according to [54, 130, 131]. The dependence of face recognition performance on the correlation between two chrominance components C_1, C_2 will be further validated through experimentation. The

discriminative power of C_1 has been maximized in the previous subsection, we then derive $\mathbf{a}_2$ by minimizing the correlation between C_1, C_2 . Similarly to C_1 , C_2 is given by

$$C_2 = \mathbf{u}_2^T \begin{pmatrix} R \\ G \\ B \end{pmatrix} = \tau_2 \mathbf{a}_2^T \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \quad (3.16)$$

The covariance provides a measure of the strength of the correlation between two or more sets of random variates [132]. Hence the covariance between C_1 and C_2 is defined as below to measure the correlation between them.

$$\begin{aligned} cov(C_1, C_2) &= E((C_1 - E(C_1))(C_2 - E(C_2))^T) \\ &= \mathbf{u}_1^T E\left(\begin{pmatrix} R - \bar{R} \\ G - \bar{G} \\ B - \bar{B} \end{pmatrix} \begin{pmatrix} R - \bar{R} \\ G - \bar{G} \\ B - \bar{B} \end{pmatrix}^T\right) \mathbf{u}_2 \\ &= \tau_1 \mathbf{a}_1^T \mathbf{M} \tau_2 \mathbf{a}_2, \end{aligned} \quad (3.17)$$

where E indicates the expectation and $\mathbf{M}$ is

$$\mathbf{M} = E\left(\begin{pmatrix} R - \bar{R} \\ G - \bar{G} \\ B - \bar{B} \end{pmatrix} \begin{pmatrix} R - \bar{R} \\ G - \bar{G} \\ B - \bar{B} \end{pmatrix}^T\right), \quad (3.18)$$

which can be estimated by the training data similarly to the computations of $\mathbf{S}_w$ (4.7) and $\mathbf{S}_b$ (4.6).

When $cov(C_1, C_2)$ becomes zero, C_1, C_2 will be uncorrelated and the information contained in them will become most complementary to each other. So we make $cov(C_1, C_2)$ equal to zero to minimize the correlation of C_1, C_2 :

$$cov(C_1, C_2) = \tau_1 \mathbf{a}_1^T \mathbf{M} \tau_2 \mathbf{a}_2 = 0. \quad (3.19)$$

$\mathbf{a}_2$ can be represented by the two basis vectors $\mathbf{n}_1, \mathbf{n}_2$ and its angle θ in the 2-dimensional chrominance subspace $\mathbf{B}$ as below:

$$\mathbf{a}_2 = \cos(\theta) \mathbf{n}_1 + \sin(\theta) \mathbf{n}_2, \theta \in [0, 2\pi). \quad (3.20)$$

Substitute (3.20) into (3.19), we have

$$\theta = \tan^{-1}\left(-\frac{\mathbf{a}_1^T \mathbf{M} \mathbf{n}_1}{\mathbf{a}_1^T \mathbf{M} \mathbf{n}_2}\right). \quad (3.21)$$

Together with $\mathbf{a}_1$ determined in (3.14), $\mathbf{a}_2$ can be computed by (3.20) and (3.21).

Lu, C_1, C_2 possess different discriminating power for face recognition. The discriminating power of a color component can be represented by the discriminant value J . Suppose the column vector $\mathbf{s}_{ij}$ contains all pixel values of Lu of the j th image in class i , where $i = 1, 2, \dots, p$ and $j = 1, 2, \dots, q_i$. p indicates the number of classes and q_i indicates the number of images for class i . The discriminant value of Lu , J_{Lu} , is defined as follows:

$$v_w = \sum_{i=1}^p \sum_{j=1}^{q_i} (\mathbf{s}_{ij} - \bar{\mathbf{s}}_i)^T (\mathbf{s}_{ij} - \bar{\mathbf{s}}_i), \quad (3.22)$$

$$v_b = q_i \sum_{i=1}^p (\bar{\mathbf{s}}_i - \bar{\mathbf{s}})^T (\bar{\mathbf{s}}_i - \bar{\mathbf{s}}), \quad (3.23)$$

$$J_{Lu} = \sqrt{\frac{v_b}{v_w}}, \quad (3.24)$$

where $\bar{\mathbf{s}}_i$ is the mean vector of class i and $\bar{\mathbf{s}}$ is the mean vector of all training data. The discriminant values of data projected onto vectors $\mathbf{a}_1$ and $\mathbf{a}_2$, $J_{\mathbf{a}_1}, J_{\mathbf{a}_2}$, can be computed similarly to (4.5)-(3.24).

To represent C_1, C_2 according to their discriminant values, we assign $\mathbf{u}_1, \mathbf{u}_2$ magnitudes τ_1, τ_2 proportional to the discriminant values of C_1, C_2 by the equation below:

$$\tau_0 : \tau_1 : \tau_2 = J_{Lu} : J_{\mathbf{a}_1} : J_{\mathbf{a}_2}, \quad (3.25)$$

where τ_0 is the magnitude of the transform vector of Lu . Thus

$$\tau_1 = \frac{J_{\mathbf{a}_1}}{J_{Lu}} \tau_0, \tau_2 = \frac{J_{\mathbf{a}_2}}{J_{Lu}} \tau_0. \quad (3.26)$$

Two chrominance components C_1, C_2 are computed by

$$C_1 = \mathbf{u}_1^T \begin{pmatrix} R \\ G \\ B \end{pmatrix} = \frac{J_{\mathbf{a}_1}}{J_{Lu}} \tau_0 \mathbf{a}_1^T \begin{pmatrix} R \\ G \\ B \end{pmatrix}, \quad (3.27)$$

$$C_2 = \mathbf{u}_2^T \begin{pmatrix} R \\ G \\ B \end{pmatrix} = \frac{J_{\mathbf{a}_2}}{J_{Lu}} \tau_0 \mathbf{a}_2^T \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \quad (3.28)$$

The generating process of two chrominance components C_1, C_2 is summarized in following algorithm.

- 1: Calculate two basis vectors $\mathbf{n}_1, \mathbf{n}_2$ of the 2-dimensional chrominance subspace orthogonal to $\mathbf{b}$ in (3.5).
- 2: Transform pixel values of color images from the 3-dimensional RGB color space to the 2-dimensional chrominance subspace $\mathbf{B}$ in (3.8) and compute the within-class scatter matrix $\mathbf{S}_w$ and the between-class scatter matrix $\mathbf{S}_b$ in (4.7)-(4.19).
- 3: Solve the eigenvalue problem in (3.13) to get $\mathbf{x}$ and calculate $\mathbf{a}_1$ in (3.14).
- 4: Compute $\mathbf{M}$ in (3.18) by training images and compute $\mathbf{a}_2$ by (3.20), (3.21).
- 5: Compute $J_{Lu}, J_{\mathbf{a}_1}, J_{\mathbf{a}_2}$ using (4.5)-(3.24) and substitute them into (3.27),(3.28) to compute C_1, C_2 .

Algorithm 1: Generating two chrominance components.

3.3 Experiments

The effectiveness of our proposed LuC_1C_2 color space for face recognition is extensively evaluated on five benchmark databases, AR [61], Georgia Tech (GT) [80], Multi-PIE, FRGC [62] and the LFW [81].

To begin with, we validate the dependence of FR performance on the correlation of C_1, C_2 through experimentation. The evaluation process of the proposed color space consists of 3 parts. In the first part, the face recognition performance of the proposed LuC_1C_2 color space is compared against that of three state-of-the-art color spaces, DCS [27], RQCr [25] and ZRG-NII [26], and the fundamental color space, RGB. Their face recognition/verification rates are

compared under various conditions: using 2 different features and 3 different dimension reduction methods on 4 databases. In the second part, the face verification rate (FVR) of LuC_1C_2 is compared with those of the state of the arts on the FRGC database. Multiple features are extracted from LuC_1C_2 images and fused at the decision level. Finally, the performances of the CNN model trained using the proposed color space LuC_1C_2 and that trained using RGB images are compared under the condition of low-resolution on GT and LFW databases.

3.3.1 Databases

3.3.1.1 AR

The face region is cropped from the original images and resized to 56×40 , which is similar to the image size used in [133]. Fig. 3.4 shows some cropped images of the AR database. We randomly select 10 from 100 subjects in the first session to do color space training and images from the second session are used for testing.

Figure 3.4: Cropped images from AR.

3.3.1.2 GT

The face region is cropped from the original images and rescaled to the size of 30×30 similar to that used in [112]. The first eight images of randomly selected 10 from 50 subjects are used to train the color space and the rest seven images of all subjects serve as testing images. Fig. 3.5 shows some cropped images of the GT database.

3.3.1.3 Multi-PIE

The Multi-PIE database contains face images captured under variations of illumination, poses and expressions in four recording sessions. We use the pose variation subset, which consists

Figure 3.5: Cropped images from GT.

of 105 subjects with 20 face images per subject across 4 sessions. All images in session 1 are used to do training and images in the remaining 3 sessions are used for testing. Face regions are cropped from original images and resized to the resolution of 50×40 for extraction of pixel values, and GW features. We show the exemplar images on Fig. 3.6.

Figure 3.6: Cropped images from Multi-PIE.

3.3.1.4 FRGC

The face is cropped from the original images with backgrounds and rescaled to the size of 32×32 , which is the same as that used in [26, 27]. We use the training set for color space training and the other two sets for testing. Fig. 3.7 shows some cropped images of the FRGC database.

Figure 3.7: Cropped images of the FRGC database.

3.3.2 The dependence of face recognition performance on the correlation between two chrominance components C_1, C_2

Here, we conduct experiments to investigate the dependence of the face recognition performance on the correlation of two chrominance components C_1, C_2 . With $\mathbf{u}_1$ determined in

(3.27), we rotate $\mathbf{u}_2$ (3.28) in the 2-dimensional chrominance subspace so that different covariances (3.17) of C_1, C_2 and different LuC_1C_2 color spaces can be obtained, corresponding to different angles between $\mathbf{u}_1$ and $\mathbf{u}_2$.

We apply ERE [52] to raw pixels of LuC_1C_2 images and calculate the face recognition or verification rate. For AR database, the covariance between two chrominance components is plotted against δ (the angle between $\mathbf{u}_1$ and $\mathbf{u}_2$) on Fig. 3.8(a) and the face recognition rates obtained from different LuC_1C_2 color spaces are plotted against δ on Fig. 3.8(b). Similarly for the FRGC database, the covariance between C_1, C_2 is plotted against δ on Fig. 3.9(a) and face verification rates obtained from different LuC_1C_2 color spaces are plotted against δ on Fig. 3.9(b).

As we can observe on Fig. 3.8 and Fig. 3.9, the curve of covariance is very similar to that of the face recognition performance on both FRGC and AR databases. This observation provides evidence of the close dependence of the face recognition performance on the correlation of two chrominance components C_1, C_2 . Moreover, the covariance will be maximized and the face recognition or verification rate will reach its minimum when δ is around 0° ($\mathbf{u}_2$ is identical to $\mathbf{u}_1$) on both AR and FRGC datasets. The covariance will be minimized to 0 and the face recognition or verification rate will reach its maximum when δ is 100° on AR dataset and 60° on FRGC dataset rather than 90° ($\mathbf{u}_2$ is orthogonal to $\mathbf{u}_1$). Therefore, in order to achieve the best face recognition performance, we should minimize the correlation between C_1, C_2 rather than simply choose $\mathbf{u}_2$ orthogonal to $\mathbf{u}_1$ [26].

3.3.3 Performance comparison of different color spaces under various conditions

As shown in the color face recognition framework on Fig. 4.1, a translated, rotated, cropped and resized color face image is firstly transformed to a color space such as LuC_1C_2 , from the original RGB color space. Many recent color face recognition works conduct experiments on

Figure 3.8: (a) The covariance between two chrominance components plotted against the angle between u_1, u_2 and (b) corresponding face recognition rates on the AR database.

Figure 3.9: (a) The covariance between two chrominance components plotted against the angle between u_1, u_2 and (b) corresponding face verification rates on the FRGC database.

raw pixels [25, 26, 63, 68]. Also, Gabor wavelet has proven to be highly effective for FR [17]. Thus we extract these two representative features from face images of the new color space.

Figure 3.10: The color face recognition framework used on AR, GT and Multi-PIE databases.

In order to compare the face recognition rates of LuC_1C_2 against those of the other color spaces, three popular dimensionality reduction methods, i.e., PCA, Enhanced Fisher linear discriminant Model (EFM) [134] and eigenfeature regularization and extraction (ERE) [52], are applied to the extracted color features. Also note that PCA is commonly used as a benchmark for the evaluation of the performance in FR algorithms [82] and it may significantly enhance the recognition accuracy [135, 136]. Plenty of color face recognition methods employ EFM method for low-dimensional feature extraction [25–27]. ERE outperforms all other FR methods discussed in [1, 52, 137].

According to [1, 28, 65, 114], different color channels are processed separately and then concatenated together into a pattern vector for classification. So dimensionality reduction is implemented separately in 3 color channels on AR, GT and Multi-PIE. The low-dimensional color features from three color channels are concatenated into one vector to combine the information in three color channels. Note that all low-dimensional features would be normalized by removing the mean and dividing the standard deviation of the feature in each color channel before the final feature concatenation to avoid the effect of magnitude dominance of one

component over the others [26]. After that, the nearest-neighbor classifier is used to classify all query images, where the Mahalanobis distance is used for PCA and ERE, and the Cosine distance is used for EFM.

3.3.3.1 Results on AR

The face recognition rate (FRR), which is the ratio of the number of correctly classified query images to the total number of query images, is plotted against the feature dimension to present the overall face recognition performances of different color spaces on AR and GT databases. Note that, in this chapter, the PCA dimension indicates the number of principle components of PCA, the EFM dimension indicates the number of principal components used in the PCA step of EFM, and the ERE dimension indicates the parameter m , which is the dimension of significant face components in [52]. As the face recognition rate of the RGB color space is far below those of the other color spaces, it is not plotted in the figures. Instead, we show the best FRR of the RGB color space among all feature dimensions in the legend area such as “RGB(FRR(%))”. Using two color features and three face recognition methods, the face recognition rates in different color spaces are shown on Fig. 3.11. The results show that our proposed LuC_1C_2 color space performs better than three state-of-the-art color spaces consistently over all tested feature dimensions of two different features and three different FR methods.

3.3.3.2 Results on GT

Fig. 3.12 shows the face recognition rates of 5 color spaces using two color features and three face recognition methods on the GT database. Again, our proposed LuC_1C_2 color space achieves consistently better face recognition performance than state-of-the-art color spaces over all tested feature dimensions of two different features and three different FR methods.

3.3.3.3 Results on Multi-PIE

Fig. 3.13 shows the face recognition rates of 5 color spaces using two color features and three face recognition methods on the Multi-PIE database. As we can see, our proposed LuC_1C_2

Figure 3.11: Face recognition rates against feature dimension on AR, each column specifies one type of feature (2 in total) and each row specifies one dimension reduction method (3 in total).

Figure 3.12: Face recognition rates against feature dimension on GT, each column specifies one type of feature (2 in total) and each row specifies one dimension reduction method (3 in total).

color space achieves consistently better face recognition performance than state-of-the-art color spaces over all tested feature dimensions of two different features and three different FR methods.

Figure 3.13: Face recognition rates against feature dimension on Multi-PIE, each column specifies one type of feature (2 in total) and each row specifies one dimension reduction method (3 in total).

3.3.3.4 Results on FRGC

FRGC is the most frequently-used dataset to test the face recognition performances of different color spaces. To achieve fair comparisons with state-of-the-art results reported by other researchers, the commonly used evaluation framework on FRGC in [26, 27] is adopted, as shown on Fig. 3.14. This makes comparisons of our proposed LuC_1C_2 color space with other color spaces fair and reliable.

Figure 3.14: The color face recognition framework used on FRGC databases.

The face recognition performance on FRGC is reported by plotting the face verification rate (FVR) at FAR=0.1% against the feature dimension so as to be consistent with recently published studies [1, 26]. The best FVR of the RGB color space among all feature dimensions is shown in the legend area such as “RGB(FVR(%))”. As we can see from Fig. 3.15, for both of the two color features and the three face recognition methods, the face verification rates of the proposed LuC_1C_2 color space are consistently higher than those of DCS, ZRG-NII, RQCr and RGB color spaces over all tested feature dimensions. Note that the GW feature is extracted from face images of 64×64 rather than 32×32 , which shows quite poor face recognition performance. The decision-level feature fusion approach in [138] is used to fuse GW features extracted from 3 color channels.

Figure 3.15: Face verification rates against feature dimension on FRGC, each column specifies one type of feature (2 in total) and each row specifies one dimension reduction method (3 in total).

Table 3.1: Performance comparison of LuC_1C_2 with state of the arts using raw pixels on FRGC

Color space	FVR (%) at FAR = 0.1%
Gray [26]	36.94
RGB [26]	45.20
HSV	65.16
YUV [26]	67.29
YIQ [26]	69.75
RQCr [26]	72.51
ZRG-NII [26]	72.86
Extended IWM [69]	74.91
DCS [27]	76.26
LuC_1C_2	77.95

3.3.4 Performance comparison of the proposed LuC_1C_2 color space with state of the arts on FRGC

Many recent color face recognition works evaluate the effectiveness of different color spaces on raw pixels [25, 26, 63, 68]. Table 3.1 shows results cited directly from published papers [26, 27] on the FRGC database. Experimental settings are all set as below: raw pixels are used as features and EFM method is used for dimension reduction, where the feature dimension was set to 1000. Table 3.1 shows that (1) the use of color information greatly improves the FVR compared to the same scheme using the gray information only; (2) different color spaces possess significantly different classification capabilities for face recognition; (3) our proposed LuC_1C_2 color space performs significantly better than recently proposed color spaces on the FRGC database.

In recent years, many complex features and different feature fusion methods have been utilized to achieve state-of-the-art results on FRGC. In [74], authors fuse local patterns of Gabor magnitude and phase using block-based Fisher's linear discriminant. Gabor and LBP features are processed by Kernel LDA and then fused by decision-level fusion in [138]. In [139], multiscale Local Phase Quantization and multiscale LBP are fused using kernel fusion. Multi-directional multi-level dual-cross patterns are fused by PLDA and score averaging in

Table 3.2: Performance comparison of LuC_1C_2 with state of the arts using complex features on FRGC

Color space	FVR (%) at FAR = 0.1%
Gray [74]	85.2
Gray [138]	88.1
Gray [139]	91.59
Gray [140]	92.89
RIQ [77]	80.03
RQCr [19]	92.43
ZRG-NII [2]	93.17
LuC_1C_2	93.38

[140]. Authors in [77] use EFM to extract features from the real part, the imaginary part, and the magnitude of RIQ color images in the frequency domain. The features are then fused by means of concatenation. In [19], patch-based Gabor, multi-resolution LBP and component-based DCT features are extracted from R, Q, Cr images, respectively. EFM and the decision-level fusion approach are used to fuse different features. Authors in [2] combine different color features of color norm patterns and color angular patterns from the ZRG-NII color space by feature concatenation. We can see that the above state-of-the-arts use multiple different features and different feature fusion schemes for different color spaces.

To compare the FVR of our color space LuC_1C_2 with various state of the arts, we extract raw-pixel, GW and LBP features from LuC_1C_2 images and simply fuse them at the decision level with equal weights. Table 3.2 tabulates the face verification performance of various approaches. All the state-of-the-art results are directly cited from recently published papers. Table 3.2 shows that the proposed LuC_1C_2 color space achieves better FVR than all reported results on FRGC database.

3.3.5 Robustness improvement of the CNN model to low-resolution degradation using the proposed LuC_1C_2 color space

Recently, many deep learning approaches are developed for face recognition based on Convolutional Neural Networks (CNN) [141]. Promising results have been achieved under challeng-

ing conditions such as occlusion [42], variations in pose and illumination [43]. While many face recognition approaches have been developed for recognizing high resolution face images [22, 44, 45], there are few studies focused on face recognition of low-resolution images.

Here, we validate that our proposed color space can be used to improve the robustness of CNN models to low-resolution degradation. Two CNN models are trained to compare their face recognition performance under the condition of low-resolution. One of the models is trained by face images in the proposed color space LuC_1C_2 , and the other is trained by RGB face images. Different from former CNN architectures such as VGG, ResNet proposed in [9] consists of residual modules which conduct additive merging of signals. The authors in [9] argue that residual connections are inherently important for training very deep architectures. Therefore, a short version of ResNet model proposed in [142], which consists of 12 residue modules, is adopted as the architecture of CNN models in our experiments. The recently released CASIA-WebFace [41] database is used to train two ResNet models. The 434,793 images of 9,067 subjects, which contain at least 14 images per subject, compose the training set of CNN models. From the 434,793 images, 25,000 images are randomly selected to train the proposed LuC_1C_2 color space, which is fixed and used in this experiment. To normalize face images, we apply the affine transformation to them based on the positions of five facial points including two eye centers, the nose tip, and two mouth corners. All images are normalized to the size of 112×96 . An off-the-shelf face alignment approach in [143] is adopted for facial point detection. Once the training is finished, low-resolution face images can be fed into the two obtained CNN models, ResNet- LuC_1C_2 and ResNet- RGB , to get their feature representations. For the feature representation, we take the 512-dimensional feature vector after the first fully-connected layer in ResNet.

Experiments are conducted on two databases, GT and LFW [81], to compare the face recognition performance of ResNet- LuC_1C_2 and ResNet- RGB using low-resolution images. To make the face recognition task more challenging on GT database for CNN models, 2 images

Table 3.3: Averaged face recognition rate (AFRR) on GT database and averaged face verification accuracy (AFVR) on LFW database using ResNet- RGB and ResNet- LuC_1C_2 for images of resolution $4 \times 4, 6 \times 6, \dots, 20 \times 20$.

Database	ResNet- RGB	ResNet- LuC_1C_2
AFRR (%) on GT	76.5000	80.7885
AFVA (%) on LFW	84.7259	86.1630

are randomly selected from 15 images per subject as gallery data and training data for PCA, and remaining 13 images per subject serve as probe data. On the LFW dataset, we follow the “Unrestricted, Labeled Outside Data Results” protocol in [81] and compute the mean verification accuracy according to the 10-fold cross-validation scheme. PCA and cosine distance are used to compute the similarity scores between two CNN features. The total scatter matrix of PCA is computed using the 9 training folds of LFW data in the 10-fold cross validation.

On both GT and LFW, face images are normalized and aligned using same methods as on CASIA-WebFace images. Images are down-sampled to different lower sizes of $4 \times 4, 6 \times 6, \dots, 20 \times 20$ to obtain low-resolution images. Face recognition rates or face verification accuracy of ResNet- LuC_1C_2 and ResNet- RGB are plotted against the size of face images on Fig. 3.16 for GT database and on Fig. 3.17 for LFW database. Furthermore, the averaged face recognition rate or face verification accuracy across different resolutions of images for two CNN models are listed on Table 3.3. We can observe from Fig. 3.16, Fig. 3.17 and Table 3.3 that the CNN model trained using LuC_1C_2 color space achieves consistently higher face recognition rates or face verification accuracy than that trained using RGB color space under the condition of low-resolution images.

3.4 Summary

In this chapter, we propose a color space LuC_1C_2 based on a framework of constructing an effective color space for face recognition tasks. It consists of one luminance component Lu and two chrominance components C_1, C_2 . The luminance component Lu is chosen among

Figure 3.16: Face recognition rates of ResNet- LuC_1C_2 and ResNet-RGB against the resolution of face images on GT database.

Figure 3.17: Face verification accuracy of ResNet- LuC_1C_2 and ResNet-RGB against the resolution of face images on LFW database.

4 different luminance candidates by analysing their R,G,B coefficients and the color sensor properties. In order to produce the two effective chrominance components C_1, C_2 , the directions of their transform vectors are determined by the discriminant analysis and the covariance analysis in the chrominance subspace of the RGB color space. The magnitudes of their transform vectors are determined according to the discriminant values of Lu, C_1, C_2 . Extensive experiments are conducted on 5 benchmark databases. Experimental results using 2 distinct features and 3 different face recognition methods on 4 databases show that our proposed color space LuC_1C_2 achieves consistently better face recognition performance than state-of-the-art color spaces. The experiments also show that features extracted from the proposed color space achieve higher face verification rate than all published results on FRGC database. Furthermore, experiments on GT and LFW show that the ResNet model trained using LuC_1C_2 images performs consistently better than that trained using RGB images when testing images are of low-resolution.

Chapter 4

Color Feature Fusion

4.1 Color Channel Fusion

4.1.1 Introduction

Color possesses discriminative information for face recognition (FR). Torres et al. [23] applied a modified PCA scheme to three color components separately and combined the results. Their results show that the use of color information improves the recognition performance when compared with the same scheme using the luminance information. The improvement can be significant when large facial expression and illumination variations are present or the resolution of face images is low [24, 25]. In [144], a pose invariant face recognition system based on the probability distribution functions of pixels in different color channels was proposed. This method achieved much better performance than that obtained from gray-level face images. Since then, more and more researchers have turned to color information for better FR performance.

Recently, research effort in color FR has been dedicated to applying a multiple-feature encoding scheme to multiple and different color-component images [1]. For example, a discriminative color feature method is proposed in [28]. The dimensionality of each color component image is reduced independently and then all low-dimensional color features are concatenated to form an augmented pattern vector. Other than that work, good FR performance is achieved

in [1]. The authors propose color local Gabor wavelets (CLGWs) and color LBP (CLBP) features and apply five popular low-dimensional feature extraction techniques including PCA [14] and ERE [52] on each color channel separately. Similar to [2, 65], low-dimensional features are separately extracted from each color component and then combined.

Current color FR works focus on how to extract effective features from multiple color channels following the framework in Fig. 4.1. Due to the high dimensionality of color component images $C_i (i = 1, 2, 3)$ or generated color features f_i , different color channels are usually processed separately first and then concatenated together into a feature vector for classification. According to [135], dimensionality reduction is a critical module of face recognition. However, in current color FR methods, dimensionality reduction is applied on each color channel separately before the channel fusion. Specifically, the dimensionalities of different low-dimensional features $(f_1^{l_1}, f_2^{l_2}, f_3^{l_3})$ are set to be equal ($l_1 = l_2 = l_3$) in [1, 2, 28, 65]. In fact, the reliability and importance of features in different color channels are not the same, which should be considered in determining l_1, l_2, l_3 . Thus, the rules or ideas of dimension reduction in a single color channel should be integrated across all three color channels to achieve more effective feature extraction and channel fusion. This chapter is targeted at filling in this gap by proposing a color channel fusion method where jointly dimension reduction algorithms are used to select more features from reliable and discriminative channels. The main new contribution of this work is selecting more effective features over different color channels for better color channel fusion.

4.1.2 Color channel fusion (CCF) approach

The effectiveness of proposed color channel fusion approach is validated by applying it to two quite different dimensionality reduction methods (PCA and ERE) using two different types of features (image-pixel values and color local Gabor Wavelets (CLGWs) [1]). PCA is commonly used as a benchmark for the evaluation of the performance in FR algorithms [82]. ERE outperforms all other FR methods discussed in [1, 2, 52, 137]. CLGWs is a color local texture feature proposed in [1].

Figure 4.1: Color FR framework, C_i , f_i , and $f_i^{l_i}$ indicate color component images, channel-wise features, and low-dimensional features respectively.

Let I represent the color feature for a face image in one of the color channels H, S, V , then I_{ij} is the feature of j th face image in class i , where $i = 1, 2, \dots, p$, and $j = 1, 2, \dots, q$, so p is the number of subjects, q denotes the number of image samples in class i and there are in total $N = p \times q$ images. Also, $\bar{I}_i$ is the mean feature of samples in class i and $\bar{I}$ is the mean feature across all samples.

4.1.2.1 CCF with PCA

Let's take the dimension reduction process of I_{ij} as an example to explain PCA algorithm, the total scatter matrix of $\{I_{ij}\}$ is defined by

$$\mathbf{S}^t = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^q (I_{ij} - \bar{I})(I_{ij} - \bar{I})^T. \quad (4.1)$$

By solving the eigenvalue problem below, eigenvector matrix Φ of $\mathbf{S}^t$ is calculated and Λ is a diagonal matrix consisting of eigenvalues

$$\Lambda = \Phi^T \mathbf{S}^t \Phi. \quad (4.2)$$

Transformation matrix Φ_l is formed by eigenvectors in Φ corresponding to the l ($l \leq \text{rank of } \mathbf{S}^t$) largest eigenvalues, and I_{ij}^l is the resulting l -dimensional feature vector

$$I_{ij}^l = \Phi_l^T I_{ij}. \quad (4.3)$$

Figure 4.2: (a), Channel-wise FR performance against PCA dimension on AR database; (b), within-class variations against PCA dimension on AR database.

As the experimental results shown in Fig. 4.2(a), different color channels possess significantly different classification abilities for face recognition when PCA is used for dimensionality reduction. In other words, the reliability of features in different color channels is not the same.

The question is how to identify the reliability of features in different color channels. As analysed in [135, 136, 145], PCA improves the generalization capability by removing unreliable dimensions caused by the biased estimates of the within-class variations. The bias is most pronounced when eigenvalues tend toward equality [135, 136, 145]. Fig. 4.2(b) shows that the channel with larger within-class variations has flatter within-class variation spectrum, so this channel tends to be more unreliable. To apply this principle to the color channel fusion, more dimensions should be selected from the color channel whose within-class variations are smaller.

To derive within-class variation spectrum $V^w \in \mathbb{R}^{1 \times r}$ (r is the rank of S^t in (4.1)), within-class scatter matrix is calculated first

$$S^w = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^q (I_{ij} - \bar{I}_i)(I_{ij} - \bar{I}_i)^T. \quad (4.4)$$

Within-class variation along l -th ($l = 1, 2, \dots, r$) eigenvector $\Phi(:, l)$ of S^t in (4.2) is

$$V^w(l) = \Phi(:, l)^T S^w \Phi(:, l). \quad (4.5)$$

Within-class variation spectrums $V_H^w, V_S^w, V_V^w \in \mathbb{R}^{1 \times r}$ of H, S, V channels can be calculated in the same way. Then the following algorithm is used to determine l_H, l_S, l_V , which indicate the numbers of eigenvectors corresponding to largest eigenvalues used in (4.3) on H, S, V color channels respectively. Suppose $D = l_H + l_S + l_V$ is the number of dimensions required for the fused feature vector.

$$l_H = 1, l_S = 1, l_V = 1$$

for $n = 4$ to D **do**

$$U_H = \sum_{i=1}^{l_H} V_H^w(i)$$

$$U_S = \sum_{i=1}^{l_S} V_S^w(i)$$

$$U_V = \sum_{i=1}^{l_V} V_V^w(i)$$

$$U_{min} = \min\{U_H, U_S, U_V\}$$

if $U_{min} = U_H$ **then** $l_H = l_H + 1$

else if $U_{min} = U_S$ **then** $l_S = l_S + 1$

else $l_V = l_V + 1$

end if

end for

Initially, l_H, l_S, l_V are set to be 1. In each round of the loop from 4 to D , the channel with smallest sum of within-class variations in selected dimensions is chosen and in this channel, the eigenvector of $\mathbf{S}^t$ corresponding to the largest eigenvalue among the unselected ones is chosen. In this way, more dimensions from more reliable channels across H, S, V are chosen. It is easy to see that this algorithm also tends to make the sum of within-class variations in each color space (U_H, U_S, U_V) equal.

4.1.2.2 CCF with ERE

Besides the dimensionality reduction, regularization is another effective technique to improve the robustness of face recognition [52, 146, 147]. Eigenfeature regularization and extraction

Figure 4.3: (a), Channel-wise FR performance against ERE dimension on pose variation subset of MultiPIE; (b), discriminant value J against ERE dimension on pose variation subset of MultiPIE.

(ERE) [52] first regularizes eigenvalues of within-class scatter matrix based on an eigenspectrum model, and then extracts the most discriminant features for fast recognition.

From channel-wise FR performance of different color channels shown in Fig. 4.3(a), it can be observed that the classification ability of different color channels is not the same when ERE is used for dimensionality reduction. In order to apply ERE's discriminative rule of a single color channel to all three color channels, the criteria based on maximization of discriminant value J used in [52] is adopted. In this way, more features from more discriminant channels across H, S, V are selected.

J spectrums of different color channels are shown in Fig. 4.3(b). Suppose $\mathbf{A}_H, \mathbf{A}_S, \mathbf{A}_V$ are eigenvector matrices derived by ERE on channel H, S, V respectively. The following algorithm is used to determine l_H, l_S, l_V , which indicate the numbers of eigenvectors in $\mathbf{A}_H, \mathbf{A}_S, \mathbf{A}_V$ corresponding to largest J values used in ERE on H, S, V color channels. Suppose $D = l_H + l_S + l_V$ is the number of dimensions required for the fused feature vector.

$$l_H = 1, l_S = 1, l_V = 1$$

$$J = \{J_H(2 : \text{end}), J_S(2 : \text{end}), J_V(2 : \text{end})\} \text{ arranged in descending order}$$

for $n = 1$ to $D - 3$ **do**

if $J(n) \in J_H$ **then** $l_H \leftarrow l_H + 1$

```

else if  $J(n) \in J_S$  then  $l_S \leftarrow l_S + 1$ 
    else  $l_V \leftarrow l_V + 1$ 
end if
end for

```

To begin with, l_H, l_S, l_V are set to be 1. Discriminant value spectrums $J_H(2 : end), J_S(2 : end), J_V(2 : end)$ from H, S, V color channels are put together and arranged in descending order. l_H, l_S, l_V are determined by counting how many J values are from corresponding channel among $\{J(n)\}, n = (1, 2, 3, \dots, D - 3)$. Thus, more dimensions from more discriminant channels across H, S, V are selected.

4.1.2.3 Face recognition

With l_H, l_S, l_V eigenvectors corresponding to largest eigenvalues in PCA or discriminant values in ERE, three transformation matrices are formed to extract low-dimensional features from H, S, V color channels respectively. Then three low-dimensional feature vectors are normalized and concatenated. A minimum-Mahalanobis-distance classifier is subsequently performed between probe and gallery images to determine the identity of probe images.

4.1.3 Experiments

The proposed color channel fusion (CCF) algorithm is verified by applying it to two quite different dimensionality reduction approaches (PCA and ERE) using two different types of features (image-pixel values and CLGWs) on three publicly available datasets: AR [61] and two datasets from CMU Multi-PIE [79]. It is compared with the color channel concatenation (CCC) algorithm used in [1, 2, 28, 65, 144] and the decision-level fusion method using a weighted sum rule (WDF) in [19].

Figure 4.4: FR rate of 4 color spaces on AR using ERE-based CCF

4.1.3.1 Color space selection

The face recognition performance is not the same in different color spaces. It is well-known that HSV [23, 148, 149], RQCr [1, 2, 25], ZRG [1, 2, 26] and HSI [144] are better than other color spaces for image recognition. However, there is no common opinion about which of the four is consistently the best choice for color face recognition. An experiment on AR database is conducted to study these 4 popular color spaces using ERE-based CCF method. As shown in Fig. 4.4, HSV color space outperforms the others. Thus, our experiments make use of HSV color space. Using the other three color spaces, the relative performances of different color fusion approaches are similar to those using HSV space.

4.1.3.2 Face recognition under different variations

To validate the effectiveness of the proposed CCF approach for face recognition tasks under illumination variation, pose variation and mixed variations, we conduct 3 experiments on MultiPIE and AR databases.

Illumination: 18 flash-only (illumination 1-18) frontal images with neutral expression per subject in the Multi-PIE database are used in this experiment, which produces $18 \times 4 \times 105 = 7560$ images for training and testing. According to [133], if a class does not have sufficient

Table 4.1: Best recognition rate (%) & its dimension using pixel value

Database	illumination		pose		AR	
DR	PCA	ERE	PCA	ERE	PCA	ERE
CCF (dimensionality)	81.85 (180)	84.43 (165)	83.64 (300)	87.31 (75)	82.97 (240)	80.47 (150)
CCC (dimensionality)	77.71 (360)	78.81 (300)	79.35 (360)	81.85 (105)	79.31 (390)	77.31 (270)
WDF (dimensionality)	69.22 (120)	75.49 (300)	76.37 (210)	83.69 (45)	77.19 (240)	68.50 (180)

Table 4.2: Best recognition rate (%) & its dimension using CLGWs

Database	illumination		pose		AR	
DR	PCA	ERE	PCA	ERE	PCA	ERE
CCF (dimensionality)	87.61 (420)	90.16 (195)	80.90 (360)	84.03 (105)	86.63 (270)	89.39 (285)
CCC (dimensionality)	86.05 (600)	88.24 (300)	79.14 (420)	81.60 (105)	86.37 (360)	89.11 (285)
WDF (dimensionality)	80.40 (360)	86.03 (300)	75.03 (330)	78.27 (300)	78.09 (240)	83.20 (285)

training samples to represent some variations of its query image, they are represented by the non-class-specific component of other classes. To fully represent all variations for all classes in the training step, 4 different illuminations are randomly selected from 18 illuminations per subject in session 1. Images in the remaining 3 sessions are used for testing.

Pose: To do training, 3 different poses are randomly selected from 5 poses per subject in session 1. Images in the remaining 3 sessions are used for testing.

Mixed variations: 3 different variations are randomly selected from 7 mixed variations per subject in session 1 for training. Images in session 2 are used for testing.

Results and Analysis: The face recognition rates are averaged over 10 rounds of random selection of the training samples. The best recognition rate among all dimensions and its dimensionality at which the best recognition rate is achieved are shown in Table 4.1 for image-pixel value and in Table 4.2 for CLGWs feature. They show that the proposed CCF approach outperforms CCC and WDF methods consistently for all image variations, dimension reduction

Figure 4.5: FR rates of (a) PCA-based and (b) ERE-based methods using image-pixel values

Figure 4.6: FR rates of (a) PCA-based and (b) ERE-based methods using CLGWs

techniques and features. What's more, the dimensionality of the best recognition rate obtained by CCF is much lower than that of CCC. Although WDF gets its peak recognition rate at lower dimensionality in some cases, its recognition rate is always the lowest among the three methods. To provide more details, we plot the recognition rates against the dimensionality in Fig. 4.5 for color feature fusion methods using image-pixel values and in Fig. 4.6 for methods using CLGWs. The proposed CCF method outperforms the CCC method consistently over all dimensionality. The performance gains are significant for small number of features.

4.1.4 Summary

In this chapter, a color channel fusion method is proposed to make use of reliability and importance of features in different color channels. By integrating the dimension reduction rule of

a single color channel across all three color channels, a more effective channel fusion method is achieved. Extensive experiments on three color face datasets are conducted to validate the effectiveness and robustness of the proposed CCF method. It outperforms the CCC method and the WDF method consistently for two different dimension reduction approaches, two different types of features and 3 image variations: illumination, pose and mixed variations.

4.2 Covariance Matrix Regularization

4.2.1 Introduction

Face recognition has been a very active research area due to its increasing security demands, commercial applications and law enforcement applications [111, 150–154]. It is often the case in face recognition that no single feature is rich enough to capture all of the available information [155]. The robust face recognition requires multiple feature sets to be taken into account [156], which can be features of different color channels [1, 2, 19, 157], and different types of features [73, 138, 156].

Feature fusion often results in very high dimensionality. For example, multi-scale descriptors in [18] are densely extracted from dense landmarks and concatenated together to form a 100K-dimensional feature vector. The high dimensionality of feature vectors imposes great burdens on the robust face recognition task. Therefore, dimensionality reduction is a critical module of feature fusion. Existing feature fusion methods can be generally classified into two categories: DR-Cat and Cat-DR. DR-Cat applies dimensionality reduction to each feature before the concatenation of multiple features and Cat-DR does vice versa. Choi and et.al [1] use DR-Cat to reduce the dimension of each feature separately before concatenating all low-dimensional feature vectors in the column order. Tan in [138] uses PCA to reduce the dimensionality of Gabor wavelets and LBP prior to fusing them by averaging their similarity scores (same as DR-Cat). DR-Cat is also used in [2, 65, 77, 114]. By reducing the dimensionality of each feature separately before concatenating them together, DR-Cat ignores the

correlation information between different features. But the correlation information plays an important role in the process of feature fusion. In order to utilize the correlation information, Yang and et.al [67] employ Cat-DR to concatenate three color components into an augmented feature vector first and then apply PCA or EFM to the concatenated feature vector. Cat-DR is also used in [18] to fuse multi-scale descriptors centered at dense facial landmarks. The dimensionality of the augmented feature vector is reduced by PCA and LDA. In the case of perfect training data, Cat-DR utilizing the correlation information usually achieves better performance than DR-Cat. However, in practice, the limited training data may result in unreliable estimates of cross-feature correlations. This often leads to overfitting and performance degradation in Cat-DR.

To solve problems in feature fusion methods of DR-Cat and Cat-DR, we propose a covariance matrix regularization (CMR) technique. Instead of modifying eigenvalues of covariance matrices as in conventional regularization techniques [48–52], CMR works by regularizing the off-diagonal cross-feature covariances in the covariance matrix of training data. Thus the trace of covariance matrices remains unchanged and the feature correlation estimated from the training data is suppressed before being used to train the feature fusion model. In this way, the obtained model does not adapt too much to the estimated correlation and hence the overfitting is reduced. In the experimental part conducted on four public face databases including MultiPIE, GT, AR and FRGC, we vary the value of weights in CMR to show how it solves the problem of overfitting and improves the face recognition performance. Then, we study the relationship between the optimal value of weights in CMR and the number of training images per subject. Finally, we compare the performance of CMR against the best single feature, DR-Cat and Cat-DR by fusing features of multiple color channels, and multiple types of features.

4.2.2 Feature fusion in face recognition

4.2.2.1 Feature fusion schemes

Face recognition is an area that is well-suited for the fusion of multiple descriptors due to its inherent complexity and need for fine distinctions [156]. Multiple descriptors can be features extracted from different color channels. Y, I, Q components possess the property of decorrelation, which helps in reducing redundancy and is a key characteristic in pattern classifier designation. Thus features extracted from Y, I, Q color channels are fused in [157]. Similarly, R, Q, Cr features are fused in [1, 19] and Z, R, G features are fused in [2]. Furthermore, multiple descriptors can be different types of features. Authors in [138, 156] combine Gabor wavelets and LBP to achieve considerably better performance than either alone. The two features are complimentary in the sense that LBP captures small appearance details while Gabor wavelets encode facial shape over a broader range of scales. Fourier features, Gabor wavelets are combined in [73] to achieve better performance for face recognition. Global Fourier features describe the general characteristics of the holistic face and they are often used for coarse representation. Differently, local Gabor features reflect and encode more detailed variations within some local facial regions.

To investigate the effectiveness of the proposed feature fusion method for face recognition, this chapter explores 3 different feature fusion schemes: (1), fusion of pixel values in 3 color channels R, G, B ; (2), fusion of LBP features in 3 color channels R, G, B ; (3), fusion of pixel values and LBP features of a single color channel R . Many recent face recognition works conduct experiments on pixel values to evaluate the face recognition performance of their methods [25, 26, 63, 68]. LBP has been proven to be highly discriminative for face recognition [16, 18]. Thus these two features are used for the task of fusing features of different color channels R, G, B and the task of fusing different types of features in channel R . As R channel has been shown to perform better than other intensity images including Gray for face retrieval [1, 68], we take the R channel as an example channel for the fusion of different types of features.

4.2.3 Feature fusion with dimensionality reduction

Fusing multiple feature sets has many successful applications in face recognition. However, the fusion of multiple features inevitably causes the problem of high dimensionality. It is well known that high dimensionality degrades the classification performance (curse of dimensionality) [158, 159]. Thus, dimension reduction becomes an integrated part of feature fusion. PCA [160] is commonly used as a benchmark for the evaluation of the performance in FR algorithms [82] and it may significantly enhance the recognition accuracy [135, 136]. Plenty of color face recognition methods adopt the Enhanced Fisher Model (EFM) [25–27]. Therefore, PCA and EFM are used in this work as dimension reduction methods.

4.2.3.1 PCA and EFM

Suppose a face image is represented by a feature vector $\mathbf{x}$, its total covariance matrix Σ_t and within-class covariance matrix Σ_w are defined in equation (4.6) and equation (4.7), respectively. $\mathbf{x}_{ij}$ denotes j -th sample of class i , $i = 1, 2, \dots, p$, $j = 1, 2, \dots, q_i$. p indicates the number of classes and q_i indicates the number of samples for class i . $\bar{\mathbf{x}}_i$ indicates the mean of training samples in class i and $\bar{\mathbf{x}}$ indicates the mean of all training samples and T indicates transpose.

$$\Sigma_t = \sum_{i=1}^p \sum_{j=1}^{q_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}})(\mathbf{x}_{ij} - \bar{\mathbf{x}})^T. \quad (4.6)$$

$$\Sigma_w = \sum_{i=1}^p \sum_{j=1}^{q_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T. \quad (4.7)$$

PCA uses the Karhunen-Loeve Transform to produce the most expressive subspace for face representation and recognition. It factorizes Σ_t in equation (4.8) and obtain the eigenvector matrix Φ . We use eigenvectors matching to the d largest eigenvalues in Λ as the projection matrix P in equation (4.9) to compute the d -dimensional vector $\mathbf{y}$ in the PCA subspace.

$$\Sigma_t = \Phi \Lambda \Phi^T. \quad (4.8)$$

$$\mathbf{y} = P^T \mathbf{x}. \quad (4.9)$$

In order to use the Mahalanobis distance for similarity comparison between $\mathbf{y}$ rather than the Euclidean distance, we compute the within-class covariance matrix Σ_{wy} of $\mathbf{y}$ according to equation (4.7). Eigenvector matrix Φ_{wy} and eigenvalue matrix Λ_{wy} of Σ_{wy} are derived similarly as in equation (4.8). Then the whitening matrix Q is computed in equation (4.10).

$$Q = \Phi_{wy}(\Lambda_{wy})^{-\frac{1}{2}} \quad (4.10)$$

The final d -dimensional vector $\mathbf{z}$ for distance comparison is

$$\mathbf{z} = Q^T P^T \mathbf{x} = U^T \mathbf{x}. \quad (4.11)$$

Enhanced Fisher Model [134] is an example of discriminating subspace methods. It achieves high separability among the different pattern classes. The first step of EFM is the same as PCA in equation (4.9). After that, EFM computes the within-class covariance matrix Σ_{wy} , and the between-class covariance matrix Σ_{by} of $\mathbf{y}$ which is computed according to equation (4.12).

$$\Sigma_{by} = \sum_{i=1}^p q_i (\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})^T. \quad (4.12)$$

Eigenvector matrix Φ_{wb} and eigenvalue matrix Λ_{wb} are derived by solving the eigenvalue problem below

$$\Sigma_{wy}^{-1} \Sigma_{by} = \Phi_{wb} \Lambda_{wb} \Phi_{wb}^T. \quad (4.13)$$

Then a projection matrix H consisting of eigenvectors in Φ_{wb} corresponding to d' largest eigenvalues in Λ_{wb} is used to compute the final d' -dimensional vector $\mathbf{z}$

$$\mathbf{z} = H^T P^T \mathbf{x} = U^T \mathbf{x}. \quad (4.14)$$

Many other dimension reduction methods are modifications or extensions of the above two methods. Thus PCA and EFM are taken as two representative dimension reduction methods used in this work.

4.2.3.2 DR-Cat approach

Let $\mathbf{x}^1, \dots, \mathbf{x}^n, \dots, \mathbf{x}^N$ be N vectors of different features extracted from the same face. DR-Cat approach computes the covariance matrix Σ^n for each feature vector $\mathbf{x}^n$ separately, where Σ^n can be a total covariance matrix, within-class covariance matrix or between-class covariance matrix. From Σ^n , the projection matrix U^n is derived to project the high-dimensional feature vector $\mathbf{x}^n$ to a low-dimensional feature vector $\mathbf{z}^n$ as in equation (4.11) of PCA or equation (4.14) of EFM. Note that the covariance matrix Σ^n provides only within-feature information, which means that the dimension reduction is implemented independently on each feature. Then low-dimensional feature vectors $\mathbf{z}^1, \dots, \mathbf{z}^n, \dots, \mathbf{z}^N$ are concatenated into $\mathbf{z}$ for classification as in equation (10). Each low-dimensional feature vector is normalized to have zero mean and unit variance prior to their concatenation.

$$\begin{aligned}\mathbf{z} &= [\mathbf{z}^1; \dots; \mathbf{z}^n; \dots; \mathbf{z}^N] \\ &= [(U^1)^T \mathbf{x}^1; \dots; (U^n)^T \mathbf{x}^n; \dots; (U^N)^T \mathbf{x}^N].\end{aligned}\tag{4.15}$$

4.2.3.3 Cat-DR approach

Cat-DR approach concatenates different feature vectors $\mathbf{x}^n$ into an overall feature vector $\mathbf{x}$, $\mathbf{x} = [\mathbf{x}^1; \dots; \mathbf{x}^n; \dots; \mathbf{x}^N]$, to make use of their correlation information. Different feature vectors are normalized to have zero mean and unit variance before concatenation. Then the projection matrix U is derived from the covariance matrix Σ of the overall feature vector $\mathbf{x}$ as in equation (4.11) of PCA or equation (4.14) of EFM to project $\mathbf{x}$ to a low-dimensional feature vector $\mathbf{z}$ in equation (11) for classification.

$$\begin{aligned}\mathbf{z} &= U^T \mathbf{x} \\ &= U^T [\mathbf{x}^1; \dots; \mathbf{x}^n; \dots; \mathbf{x}^N].\end{aligned}\tag{4.16}$$

4.2.4 Covariance Matrix Regularization for feature fusion

4.2.4.1 Covariance matrices in DR-Cat and Cat-DR

The projection matrices, U^n in DR-Cat and U in Cat-DR, are derived from the covariance matrices of training data, Σ^n in DR-Cat and Σ in Cat-DR, respectively. Σ^n are in fact submatrices of Σ . The covariance matrix carries two different kinds of information: data variances of the variables and the covariances between each pair of variables. Σ^n consists of data variances and covariances within feature $\mathbf{x}^n$ while Σ possesses both within-feature covariances and cross-feature covariances. For a better understanding, we represent the covariance matrix Σ as:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \dots & \Sigma_{1n} & \dots & \Sigma_{1N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \Sigma_{n1} & \dots & \Sigma_{nn} & \dots & \Sigma_{nN} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \Sigma_{N1} & \dots & \Sigma_{Nn} & \dots & \Sigma_{NN} \end{pmatrix}. \quad (4.17)$$

As shown in equation (4.17), Σ can be represented as a block covariance matrix whose entries are partitioned into within-feature submatrices, denoted by $\Sigma_{nn}, n = 1, 2, \dots, N$, which are the same as Σ^n in DR-Cat, and cross-feature submatrices, denoted by $\Sigma_{nm}, n \neq m, n, m = 1, 2, \dots, N$, which are ignored in DR-Cat.

Within-feature submatrix Σ_{nn} is computed by feature vectors $\mathbf{x}^n$. It contains data variances and covariances within feature vector $\mathbf{x}^n$. Cross-feature submatrix Σ_{nm} contains data covariances between two different features $\mathbf{x}^n$ and $\mathbf{x}^m$. These cross-feature covariances have critical influence on the process of fusing different features. DR-Cat derives its projection matrix U^n from Σ_{nn} , ignoring the correlation between different features contained in Σ_{nm} . Cat-DR makes use of both Σ_{nn} and Σ_{nm} to derive U for feature fusion. In the ideal case of perfect training data which provides reliable and consistent information with the data population, Cat-DR achieves better performance than DR-Cat. However, in practice, the finite quantity of training images may lead to unreliable estimates of the cross-feature covariances, which causes overfitting and may make Cat-DR underperform DR-Cat.

4.2.4.2 Overfitting and Covariance Matrix Regularization

Overfitting is a modelling error which occurs when a function is too closely fit to a limited set of training data. In reality, the data being studied often has some degree of noise or error within it. Thus making a model conform closely to inaccurate data can affect the model with substantial errors and reduce its predictive power. The degree of overfitting depends on the level of noise in the training data.

In general, Cat-DR should deliver better performance than DR-Cat as it takes account of the correlation information between different features. However, the correlation information is estimated from the training data, which usually deviates from that of the data population, especially in the case of limited number of training samples. When the feature fusion model is trained to closely conform to the estimated correlation information from the finite training data, the resulting model will show overfitting and performance degradation on the data population or new data.

In order to reduce overfitting by regularization, authors in [48, 49] add a constant to diagonal elements of the covariance matrix. Another solution is to decompose the discriminant function into two parts and replace the small eigenvalues of the covariance matrix by a constant as in [50, 51]. Besides adding a constant to all eigenvalues or replacing the unreliable eigenvalues by a constant as discussed above, ERE in [52] replaces the unreliable eigenvalues with a model determined by the reliable eigenvalues. These three methods regularize the biased covariance matrix of training data by modifying its eigenvalues thus the regularized eigenspectrum can be closer to the population variances. However, modifying eigenvalues changes the trace of the covariance matrix and reduces the discriminating power of features themselves. In this chapter, we propose a covariance matrix regularization (CMR) method to solve the problem of unreliable estimates of cross-feature correlations in feature fusion. Instead of modifying eigenvalues, it assigns weights w_{nm} , $0 < w_{nm} < 1$, to cross-feature submatrices Σ_{nm} in the

covariance matrix Σ as shown below:

$$\Sigma^R = \begin{pmatrix} \Sigma_{11} & \dots & w_{1n} * \Sigma_{1n} & \dots & w_{1N} * \Sigma_{1N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{n1} * \Sigma_{n1} & \dots & \Sigma_{nn} & \dots & w_{nN} * \Sigma_{nN} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{N1} * \Sigma_{N1} & \dots & w_{Nn} * \Sigma_{Nn} & \dots & \Sigma_{NN} \end{pmatrix}. \quad (4.18)$$

The optimal value of w_{nm} depends on how much regularization is required for two different features $\mathbf{x}_n$ and $\mathbf{x}_m$, which can be estimated using some prior knowledge of feature properties and training data. For example, a relatively small weight is required in the case of large deviation between the estimated correlation and that of the data population. Experimental evaluation of the optimal value of weights in CMR can be found in section 4.2.5.2. By using CMR, the influence of correlation information estimated from the training data is suppressed. The feature fusion model learns from but does not adapt too much to the estimated correlation thus increases its generalization ability to unknown instances.

When fusing features, the technique of CMR defined in equation (4.18) is applied to the total covariance matrix and the within-class covariance matrix of training data. It is straightforward to compute Σ_t^R from Σ_t defined in equation (4.6) according to equation (4.18). The within-class covariance matrix Σ_{wy} is computed in the PCA subspace, where different original features are mixed in the low-dimensional feature vectors $\mathbf{y}$. Therefore, the within-class covariance matrix can not be directly regularized as in equation (4.18). In order to tackle this problem, the CMR is applied to the within-class covariance matrix of original feature vectors $\mathbf{x}$, Σ_w defined in equation (4.7). Thus Σ_w^R can be computed according to equation (4.18). Then, we apply P^R , which consists of d largest eigenvectors of Σ_t^R , to Σ_w^R as in equation (4.19) to compute the regularized within-class covariance matrix in the PCA subspace Σ_{wy}^R .

$$\Sigma_{wy}^R = (P^R)^T \Sigma_w^R P^R. \quad (4.19)$$

Details of the proposed CMR method are summarized the following algorithm.

- 1: Calculate the total covariance matrix Σ_t , and within-class covariance matrix Σ_w of $\mathbf{x}$ as in equation (4.6) and equation (4.7).
- 2: Apply CMR to Σ_t as in equation (4.18) and calculate P^R from Σ_t^R according to equation (4.8).
- 3: Apply CMR to Σ_w as in equation (4.18) and apply P^R to Σ_w^R to obtain Σ_{wy}^R as in equation (4.19).
- 4: Derive projection matrices using Σ_t^R and Σ_{wy}^R according to equation (4.11) for PCA or equation (4.14) for EFM.

4.2.5 Experiments

We assess the effectiveness of the proposed CMR technique for face recognition under 3 different feature fusion schemes: (1), fusion of pixel values from R, G, B channels; (2), fusion of LBP features from R, G, B channels; (3), fusion of pixel values and LBP features in the R channel. We conduct extensive experiments on three face datasets: MultiPIE [79], GT [80], AR [61] and FRGC [62].

The Multi-PIE database contains face images captured under variations of illumination, poses and expressions in four recording sessions. We use the largest variation subset, illumination subset, which consists of 105 subjects with 80 face images per subject across 4 sessions (20 images per subject in each session). Similar to [161], we randomly choose s samples from 20 samples per subject in session 1 as the training and gallery data. Remaining 6300 face images of 105 subjects in session 2 to session 4 serve as query data. The nearest neighbor classifier with mahalanobis distance is used for classification. The gallery image is obtained by averaging all training samples per person. Face regions are cropped from original images and resized to the resolution of 32×32 for extraction of pixel values, and LBP features. The patch size of LBP operator is set to be 4×4 . We show the exempling images on Fig. 5.6.

Similar to Multi-PIE database, we randomly choose s samples from 15 images for each subject as the training data on GT dataset. The rest $(15 - s)$ face images per subject serve

Figure 4.7: Exempling face images of the illumination variation subset from the Multi-PIE database.

as query data. The classifier is same as that used on Multi-PIE. The original face images are down-sampled to the resolution of 32×32 for extraction of pixel values, and LBP features. The patch size of LBP is set to be 8×8 .

Figure 4.8: Exempling face images from Georgia Tech face database.

On AR, we randomly choose s samples from 7 samples per subject in session 1 as the training and gallery data. Remaining 700 face images of 100 subjects in session 2 serve as query data. The classifier is same as that used on Multi-PIE. Face portions are manually cropped from original images and resized to the resolution of 32×32 for extraction of pixel values, and LBP features. The patch size of LBP operator is set to be 8×8 . We show some exempling images on Fig. 5.8.

Figure 4.9: Exempling face images of the AR database.

On FRGC, the face region is cropped from the original images and resized to a spatial resolution of 64×64 . Other experimental settings are the same as those in Chapter 3. The patch size of LBP operator is set to be 8×8 . We show some exempling images on Fig. 4.10.

A number of experiments are conducted. To begin with, by decreasing the value of weights from 1 to 0, we validate that CMR solves the overfitting problem and improves the face recognition performance. After that, we show that when the number of training images per subject

Figure 4.10: Exempling face images of the FRGC database.

decreases, stronger regularization should be applied to the estimated cross-feature covariances by using smaller weights in CMR. Finally, the face recognition performance of the proposed CMR technique is compared with that of the best single feature, DR-Cat and Cat-DR by fusing features from multiple color channels, and multiple types of features. For the convenience and clarity of all experiments, we adopt the same value w for w_{nm} , $w_{nm} = w$, in CMR.

4.2.5.1 Evaluation of CMR against the level of regularization

Here, we conduct experiments to investigate how different levels of covariance matrix regularization make influence on the face recognition performance. Specifically, we vary the value of weights w in CMR from 1 to 0, so that its regularizing effect on the cross-feature covariances changes from weaker to stronger.

This experiment is carried out on Multi-PIE, GT and AR datasets. Face images in different color channels are arranged into column vectors as features of pixel values and LBP features are extracted from different channels separately. The radius and the number of sampling points in the LBP operator are set as 1 and 8 through this chapter. For features of pixel values and LBP, the numbers of training samples per subject s are 4 on Multi-PIE, 5 on GT and 4 on AR.

We report the face recognition performances of different weights in CMR by face recognition rate (FRR), which is the ratio of the number of correctly classified query images to the total number of query images. Note that, among all tested feature dimensions of PCA and EFM, the best found FRR is reported. We plot FRR against the value of weights in CMR for 3 different feature fusion schemes on Fig. 4.11 and Fig. 4.12. As we can observe, when the value of weights in CMR decreases from 1 to zero, the FRR increases to the maximum point and then decreases for all 3 feature fusion schemes on all three databases. The best performance

is achieved at weight of $0.5 \sim 0.9$. One clear and consistent conclusion summarized from Fig. 4.11 and Fig. 5.10 is that, applying CMR to feature fusion improves the face recognition performance consistently on all 3 feature fusion schemes and all 3 datasets.

Figure 4.11: Face recognition rates (%) of fusing features (pixel values or LBP) of 3 color channels (R,G,B) against the value of weights in CMR on Multi-PIE, GT and AR. Each column specifies one type of feature (pixel values or LBP) and each row specifies one dataset (Multi-PIE, GT and AR).

4.2.5.2 The optimal value of weights in CMR for different training data

In this section, we investigate how the optimal value of weights in CMR will change with the decreasing of the number of training samples per subject. The experiment is conducted on GT

Figure 4.12: Face recognition rates (%) of fusing different types of features (pixel values and LBP of channel R) against the value of weights in CMR on Multi-PIE, GT and AR.

and AR datasets, where CMR is used for fusing LBP features of different color channels (R , G , B) and PCA is used for dimension reduction. The FRR of CMR trained by s samples per subject are plotted against the regularization parameter w on Fig. 4.13 for GT and on Fig. 4.14 for AR. On GT, $s=10, 5, 3$ are tested. On AR, $s=6, 5, 4$ are tested. As we can observe from Fig. 4.13 and Fig. 4.14, when the number of training samples per subject decreases, the optimal value of weights in CMR (indicated by dotted lines) that achieves the best face recognition performance also decreases. As fewer training samples per subject are provided to the feature fusion model, the estimated cross-feature covariances from training data are less reliable and hence need more regularization. Thus lower weights should be assigned to the cross-feature covariances in CMR with smaller size of training data provided.

Figure 4.13: Face recognition rates against the value of weights of CMR for different numbers (s) of samples per subject on GT.

Figure 4.14: Face recognition rates against the value of weights of CMR for different numbers (s) of samples per subject on AR.

Table 4.3: Face recognition performances of the best single feature, DR-Cat, Cat-DR and CMR using pixel values of multiple color channels on Multi-PIE, GT, AR and FRGC.

Pixel values	Multi-PIE		GT		AR		FRGC	
	PCA	EFM	PCA	EFM	PCA	EFM	PCA	EFM
B.S. channel	<u>81.73</u>	<u>81.95</u>	79.20	79.60	79.86	80.00	43.35	60.45
DR-Cat	80.95	80.73	81.00	81.40	<u>83.14</u>	<u>82.86</u>	38.37	59.08
Cat-DR	81.29	81.13	<u>83.40</u>	<u>84.00</u>	81.29	81.57	<u>58.10</u>	<u>71.57</u>
CMR	83.08	83.27	85.00	85.20	84.29	84.14	63.39	74.25

Table 4.4: Face recognition performances of the best single feature, DR-Cat, Cat-DR and CMR using LBP of multiple color channels on Multi-PIE, GT, AR and FRGC.

LBP	Multi-PIE		GT		AR		FRGC	
	PCA	EFM	PCA	EFM	PCA	EFM	PCA	EFM
B.S. channel	82.40	82.24	79.20	79.60	80.71	81.14	61.68	70.82
DR-Cat	<u>84.79</u>	<u>84.71</u>	86.60	86.80	<u>85.86</u>	<u>86.14</u>	72.57	76.79
Cat-DR	83.63	83.49	<u>86.80</u>	<u>86.80</u>	85.57	85.43	<u>75.67</u>	<u>76.96</u>
CMR	85.94	85.89	88.20	88.00	88.29	88.71	76.35	77.27

4.2.5.3 Performance comparison of CMR against the best single feature, DR-Cat and Cat-DR

To systematically compare the performance of CMR with that of the best single feature, DR-Cat and Cat-DR, we conduct experiments on Multi-PIE, GT, AR and FRGC datasets. In CMR, we vary the value of w from 0 to 0.9 with step size of 0.1 and report the best classification performance. Training and testing protocols of Multi-PIE, GT and AR are the same as those in section 5.2.3.2. Training and testing protocols of FRGC are the same as those in Chapter 3. We show the FRR or FVR of the best single (B.S.) feature, DR-Cat, Cat-DR and CMR on Table 4.3 to Table 4.5 for the fusion of pixel values of R, G, B channels, the fusion of LBP of R, G, B channels, and the fusion of pixel values and LBP of channel R , respectively. We use **bold** texts and underline texts to highlight the highest and the second highest face recognition/verification accuracy among all methods, respectively.

As shown in Table 4.3 to Table 4.5, the best single feature performs worse than all feature fusion methods (DR-Cat, Cat-DR and CMR) in 20 of the 24 experiments, which indicates

Table 4.5: Face recognition performances of the best single feature, DR-Cat, Cat-DR and CMR using pixel values and LBP of channel R on Multi-PIE, GT, AR and FRGC.

R	Multi-PIE		GT		AR		FRGC	
	PCA	EFM	PCA	EFM	PCA	EFM	PCA	EFM
B.S. feature	83.08	82.83	83.60	83.40	78.57	78.86	61.68	70.82
DR-Cat	<u>88.51</u>	<u>88.29</u>	<u>86.20</u>	<u>86.60</u>	<u>82.86</u>	<u>83.14</u>	65.51	72.15
Cat-DR	87.97	87.98	84.60	85.00	81.86	81.71	<u>71.01</u>	<u>76.32</u>
CMR	90.16	90.03	88.00	87.80	84.86	85.14	71.42	76.86

that the fusion of multiple features is effective in promoting the face recognition performance. Although Cat-DR should perform better than DR-Cat in the ideal situation, it outperforms DR-Cat only in 12 of the 24 experiments. This shows that the full use of correlation information estimated from the training data causes overfitting that reduces the predictive accuracy. We propose the CMR technique to solve problems in DR-Cat and Cat-DR. In CMR, the correlation information is regularized and then used to train the feature fusion model. Experiments show that our proposed CMR consistently outperforms the best single feature, DR-Cat and Cat-DR for fusing features of different color channels, and different types of features in all 24 experiments.

4.2.6 Summary

In this chapter, we propose a covariance matrix regularization (CMR) technique to utilize the correlation between different features and reduce overfitting during the fusion of multiple features. It works by assigning weights to the cross-feature submatrices of covariance matrices of training data to suppress the influence of correlation between different features, which is estimated from the training data, in feature fusion. Extensive experiments conducted on four popular face datasets show that our proposed CMR technique consistently outperforms the best single feature, DR-Cat and Cat-DR for fusing features of different color channels and different types of features.

4.3 Color face descriptor TCLBP

4.3.1 Introduction

Color possesses discriminative information for face recognition [133] and considerable research efforts have been devoted to the efficient utilization of facial color information to enhance the face recognition performance [1, 2, 26, 63, 65, 114, 162]. Recently, local binary patterns (LBPs) [16, 29–31] have gained reputation as powerful face descriptors as they have shown robustness to variations such as facial pose, illumination, and misalignment, etc.

In color face recognition, some research efforts have been dedicated to incorporate color information into the extraction of LBP-based features. In [32], the authors proposed Color LBP based on color and LBP texture analysis. Color LBP was proven to be better than grayscale LBP and color features. Later, Choi proposed color local binary pattern (CLBP) in [25]. The authors incorporated the opponent LBP features which encode the texture patterns among pairs of color channels into CLBP. After that, local color vector binary patterns (LCVBPs) was proposed in [2]. The LCVBP contains two patterns: color norm patterns and color angular patterns.

The above reviewed works validate that color information combined with LBP-based features can greatly improve the face recognition performance. However, it should be noted that existing color LBP features are restricted to extracting inter-channel features from each pair of color channels using the same spatial structure as that used for the intra-channel features. Both CLBP and LCVBP suffer from the curse of high dimensionality and the redundant information. What's more, pixel values from different channels in certain color spaces are not quantitatively comparable. It is still an open problem that how to effectively combine color and texture information using LBP for the purpose of face recognition.

In this chapter, we propose a novel LBP-based color face descriptor, Ternary-Color LBP (TCLBP), for face recognition. It consists of Intra-channel LBP and Inter-channel LBP, which

are combined using weighted concatenation to balance the contribution of spatial structure and spectral structure to the face recognition. The main contribution of this chapter is to propose the Inter-channel LBP feature of extremely low dimensionality, which is generated by encoding the spectral structure of R,G,B color channels at the same location. We conduct extensive experiments to evaluate the effectiveness of the proposed TCLBP color feature on 3 public face datasets including Georgia Tech [80], FRGC [62] and LFW [81]. Experimental results show that TCLBP yields consistently better face recognition performance than other state-of-the-art color LBP descriptors.

4.3.2 The proposed color descriptor TCLBP

Here, we present how to extract the proposed TCLBP feature from color face images.

4.3.2.1 The LBP operator

The traditional LBP operator [16] was extended in [163] to use neighborhoods of different sizes by using circular neighborhoods and bilinearly interpolating. Another extension in [163] is the definition of so-called uniform patterns. A local binary pattern is called uniform if the binary pattern contains at most two bitwise transitions from 0 to 1 or vice versa when the bit pattern is considered circular. The uniform LBP operator is defined as below:

$$LBP_{P,R}^{u2}(i_c) = \begin{cases} \sum_{p=0}^{P-1} 2^p s(i_p - i_c) & \text{if } T \leq 2 \\ P(P-1) + 2 & \text{otherwise} \end{cases}, \quad (4.20)$$

where $i_p (p = 0, 1, 2, \dots, P-1)$ and i_c indicate the intensity values of neighboring points and the center pixel, respectively. $s(i_p - i_c)$ equals 1, if $i_p - i_c \geq 0$; and 0, otherwise. $T = |s(i_{P-1} - i_c) - s(i_0 - i_c)| + \sum_{p=1}^{P-1} |s(i_p - i_c) - s(i_{p-1} - i_c)|$ and $u2, P, R$ mean uniform LBP, P sampling points on a circle of radius of R respectively. For the uniform LBP operator used in our method $LBP_{8,1}^{u2}$, there are 59 local binary patterns in total.

Figure 4.15: Feature extraction process of Intra-channel LBP.

Figure 4.16: Dimensionality of LBP-based color features.

4.3.2.2 Intra-channel LBP

Different from the traditional gray-level images, color faces have 3 color-component images such as RGB, RQCr, YIQ, etc. Let $C_i, i = 1, 2, 3$ represent the 3 color-component images as shown in Fig. 4.15. Authors in [32] proposed to apply the uniform LBP operator (4.20) to each C_i separately. Thus each pixel value in C_i is now replaced with the LBP code defined in (4.20). Let $L_{C_i}, i = 1, 2, 3$ represent the resulted LBP-code images. The local LBP histograms $h_{C_i}^m, m = 1, 2, \dots, M$ are further extracted from L_{C_i} . Here m represents the m -th local region indicated by a rectangular box on L_{C_i} in Fig. 4.15 and M is the total number of local regions in a face image. All local LBP histograms $h_{C_i}^m, i = 1, 2, 3, m = 1, 2, \dots, M$ are concatenated together to form the Color LBP feature, which is also referred to as Intra-channel LBP in this chapter. Before concatenation, all LBP histograms have been normalized to unit norm.

4.3.2.3 Inter-channel LBP

Simply extending the LBP feature from a single gray-level image to three color-component images could suffer from the loss of correlation information of pixels from different color channels. Besides the Intra-channel LBP feature, which is used in Color LBP, LCVBP, CLBP and our proposed TCLBP as shown in Fig. 4.16, the inter-channel feature plays an important role in the color face recognition process. To capture the texture patterns of spatial interactions

Figure 4.17: Feature extraction process of Inter-channel LBP.

between each pair of color channels, authors in [1] proposed opponent LBP features, where the neighboring points i_p and center point i_c in (4.20) are taken from two different color channels. The opponent LBP extracted from each pair of C_i is then combined with Intra-channel LBP extracted from C_i to obtain CLBP. While in [2], the authors extracted Intra-channel LBP features from 4 new component images including one norm image n and three ratio images $r_{i,k}$, ($i < k, i = 1, 2, k = 2, 3$) computed by (4.21) [2] to form the LCVBP feature.

$$\begin{cases} n = \sqrt{C_1^2 + C_2^2 + C_3^2} \\ r_{i,k} = C_k / C_i, i < k, i = 1, 2, k = 2, 3 \end{cases} \quad (4.21)$$

One obvious limitation of existing methods is that inter-channel features are extracted by comparing the pixel values from each pair of color channels rather than from all three color

channels. Also, since the local spatial structure of pixel values has been encoded by Intra-channel LBP of each color channel, the comparison of pixel values at different locations is unnecessary for inter-channel features. To effectively explore the inter-channel correlation of pixel values of 3 color channels at the same location, we reconfigure the traditional LBP operator. Suppose a 3-dimensional vector $\mathbf{v} = [v_{C_1}, v_{C_2}, v_{C_3}]$ represents the 3 pixel values of 3 color channels at a certain location. Then the reconfigured LBP code for $\mathbf{v}$ taking C_i as the center channel is defined as below:

$$LBP(v_{C_i}) = \sum_{j=0}^1 2^j s(v_{C_j} - v_{C_i}), \quad (4.22)$$

where $C_j, j \in \{0, 1\}$ represents two neighboring channels of C_i . It is important to note that the number of bins in the reconfigured LBP is only 4, almost 1/15 of the number of bins in the uniform LBP operator (59), which is used in Intra-channel LBP. This low dimension cost helps in avoiding the curse of dimensionality and makes TCLBP contain much less redundant information.

By taking different color channels as the center channel, we can have 3 different combinations (R,G,B),(G,R,B) and (R,B,G) as shown in Fig. 4.17. Note that the color space $C_1C_2C_3$ for extracting Intra-channel LBP can be any effective color space such as RQCr [25], but the color space for Inter-channel LBP has to be RGB. R,G,B describe the quantitized strength of the light at different wavelengths, so pixel values of R,G,B can be directly compared. Even though some other color spaces perform better for face recognition than the RGB color space, their pixel values are not quantitatively comparable. Take RQCr as an example, R describes luminance information while Q,Cr describe chrominance information [26]. Thus R and Q, Cr have totally different physical meanings and direct comparisons between pixel values of R,Q,Cr may cause problems.

After going through the reconfigured LBP operator defined in (4.22), different combinations of R,G,B components are converted to LBP-code images $L_C, C \in \{R, G, B\}$. Similar to

L_{C_i} , L_C is partitioned into M local patches L_C^m , $m = 1, 2, \dots, M$ and a corresponding histogram h_C^m is extracted from each local patch L_C^m . All local LBP histograms h_C^m , $C \in \{R, G, B\}$, $m = 1, 2, \dots, M$ are normalized to unit norm and concatenated together to form the Inter-channel LBP feature.

By concatenating the Intra-channel LBP that describes local spatial structure of pixel values and the Inter-channel LBP that describes spectral structure of three color channels at the same location, we can obtain the final feature vector TCLBP. As the norm of both Intra-channel LBP and Inter-channel LBP is set to be 1 before concatenation, the average magnitude of each bin in Inter-channel LBP is $\sqrt{1/4} = 1/2$, which is much larger than that in Intra-channel LBP, $\sqrt{1/59} \approx 1/8$. To balance the contribution of spatial structure and spectral structure to the face recognition performance, we multiply Intra-channel LBP and Inter-channel LBP with weights 1 and 0.5, respectively before concatenation.

In face recognition, dimension reduction methods are applied to the extracted color features and the nearest-neighbour classifier using Mahalanobis distance is subsequently performed between probe and gallery low-dimensional features to determine the identity of probe images.

4.3.3 Experiments

The effectiveness of our proposed TCLBP feature for face recognition is verified by comparing it with 3 state-of-the-art color LBP features including Color LBP, CLBP and LCVBP. Extensive experiments are carried out on 3 publicly available datasets: Georgia Tech, FRGC and LFW.

4.3.3.1 Experimental settings

A translated, rotated, cropped and resized color face image is firstly transformed from the RGB space to the RQCr space. Face images are resized to the resolution of 64×64 and the patch size is set to be 8×8 . Also P, R in (4.20) are set to be 8 and 1, respectively. As for the dimension reduction methods, PCA is used on LFW and EFM [134] is used on Georgia Tech

Table 4.6: Results on Georgia Tech

Color feature	face recognition rate (%)	computation time (seconds)
Gray LBP	82.5714	4.7899
Color LBP [32]	90.8571	14.3557
CLBP [25]	93.4286	43.3938
LCVBP [2]	92.0000	19.0812
TCLBP	94.5714	24.2614

Table 4.7: Results on FRGC

Color feature	FVR (%) at FAR = 0.1%	computation time (seconds)
Gray LBP	53.3635	237.8086
Color LBP [32]	74.0783	712.7307
CLBP [25]	73.9662	2154.4122
LCVBP [2]	66.6004	947.3420
TCLBP	75.3234	1204.5282

and FRGC. PCA is commonly used as a benchmark for the evaluation of the performance in FR algorithms [82] and it may significantly enhance the recognition accuracy [135, 136]. Plenty of color face recognition methods employ EFM method for low-dimensional feature extraction [25, 26]. The face recognition rate/accuracy and the face verification rate (FVR) at the false accept rate (FAR) are used for measuring the identification or verification performance. As the face recognition performance relies on the dimension of low-dimensional features, we present the best found recognition rates/accuracies or verification rates.

4.3.3.2 Results on GT

On GT, the first eight samples of all persons are used as training data and the rest seven images of all persons serve as testing images. The face recognition rates using different color features are shown in Table 4.6.

4.3.3.3 Results on FRGC

The face verification rates using different color features on FRGC are shown in Table 4.7.

Table 4.8: Results on LFW

Color feature	face verification accuracy (%)	computation time (seconds)
Gray LBP	74.2666	76.2113
Color LBP [32]	77.4667	228.4111
CLBP [25]	76.3000	690.4315
LCVBP [2]	75.8500	303.5978
TCLBP	78.5667	386.0191

4.3.3.4 Results on LFW

The averaged face verification accuracies over 10 folds of View 2 on LFW are reported on Table 4.8. We follow the LFWs restricted protocol.

Experimental results show that the proposed TCLBP color feature outperforms the Color LBP, CLBP and LCVBP consistently over all 3 databases.

4.3.4 Summary

This chapter proposes a novel LBP-based color feature, TCLBP, which consists of intra-channel and inter-channel parts. As the image local spatial structure is encoded by the intra-channel features in 3 individual color channels, the proposed inter-channel feature encodes the spectral structure of the whole 3 channels at the same location. This leads to a more effective and efficient Inter-channel LBP feature than CLBP and LCVBP features. Moreover, the weighted concatenation of Intra-channel and Inter-channel LBP features balances the contribution of spatial structure and spectral structure to the face recognition. In addition, Intra-channel and Inter-channel LBP features are encoded in different color spaces because they compare the pixel values in different dimensions. Comparative experiments show that the proposed TCLBP yields visibly better face recognition performance than other state-of-the-art color LBP descriptors, Color LBP, CLBP and LCVBP, consistently on all 3 public face databases.

Chapter 5

Deep Learning Face Recognition

5.1 Enhance CNN performance using color pixel values

5.1.1 Introduction

Driven by its broad applications in human-computer interaction, homeland security, and entertainment [110, 111, 114, 162], face recognition becomes a more and more popular biometric topic in the community of computer vision and pattern recognition. Over the past decade, encouraging improvements have been achieved in face recognition performance by convolutional neural networks (CNNs) [3, 4, 8].

CNNs are high-capacity classifiers with very large numbers of parameters that must be learned from millions of training examples [33]. The main power of a CNN lies in its deep architecture [3, 4, 8], which allows for extracting a set of discriminating features at multiple levels of abstraction. However, it is impossible to collect millions of annotated images and implement the complex training process for each new face recognition task. In practice, very few people train an entire CNNs from scratch (or full training). Instead, it is common to pretrain a CNNs on a very large dataset (e.g. ImageNet, which contains 1.2 million images with 1000 categories), and then use the CNNs either as a fixed feature extractor or an initialization for the task of interest. Computer vision datasets have significant differences in image statistics [164]. When training and testing datasets have differences in viewpoints, image sizes, scene context,

illumination, expressions or other factors, the face recognition performance using pre-trained CNNs as a feature extractor will be affected inevitably. Thus the pre-trained CNN model is usually used as an initialization for fine-tuning on images from the application of interest [34–38]. However, our experimental results show that the use of pre-trained deep CNNs with fine-tuning cannot provide satisfactory face recognition performance, especially when there exist big differences between training and targeted face images.

To improve the face recognition performance of pre-trained CNNs with or without fine-tuning, we propose to combine image representations learned by CNNs with low-level features, color pixel values. Even though features extracted by CNNs from face images are discriminative, they contain high-level information only and are designated to solve the classification problem that is strictly restricted. Please note that features extracted by CNN in this chapter indicate high-layer CNN features which are usually used in classification tasks. As fine-tuning is still based on the pre-trained CNN model, its performance is constrained by the ability of the underlying pre-trained CNNs. The low-level features depict the characteristics of targeted face images from a different perspective, which is complementary to the high-level image representations learned by CNNs. Thus low-level features provide an effective way to enhance the recognition performance of CNNs. Whats more, the low-level features can also be used together with fine-tuning to further boost the face recognition performance. By conducting extensive experiments on LFW and FRGC datasets using the VGG-Face model [8], we show that the face recognition performance of the pre-trained VGG-Face model with or without fine-tuning can be improved significantly by low-level features, color pixel values. Furthermore, low-level features improve the performance of pre-trained CNNs much more than fine-tuning.

5.1.2 Convolutional Neural Networks

A deep CNN has multiple layers that progressively compute features from input images. It contains three basic components: convolutional, pooling and fully-connected layers.

The convolutional layer is the core building block of a CNN. It detects local features at all locations of the input image by learning a filter bank. Every filter is small along the width and height directions, but extends through the full depth of the input volume. We use $x_i^{(l)}$ to represent the i -th feature map at layer l . The convolutional layer computes as following:

$$x_j^{(l+1)} = s\left(\sum_i F_{ij}^{(l)} * x_i^{(l)} + b_j^{(l)}\right), \quad (5.1)$$

where $F_{ij}^{(l)}$ denotes the filter that connects feature map $x_i^{(l)}$ to output map $x_j^{(l+1)}$ at layer l , $b_j^{(l)}$ is the bias for the j -th output feature map, $s()$ is some element-wise non-linear function and $*$ denotes the discrete 2D convolution. The output feature maps along the depth dimension are stacked and produce the output volume.

The pooling layer is periodically inserted in-between successive convolutional layers in a CNN. Its function is to progressively reduce the spatial size of the representations to reduce the amount of parameters and computation in the network, and hence to control overfitting. Pooling is also used to introduce translation invariance of features.

In a fully-connected layer, output units have full connections to all input units. Their activations can hence be computed with a matrix multiplication followed by a bias offset, specifically:

$$x^{(l+1)} = s(W^{(l)}x^{(l)} + b^{(l)}), \quad (5.2)$$

where $x^{(l)}$ is the vectored input from layer l , $W^{(l)}$ and $b^{(l)}$ are the parameters of the fully-connected layer at layer l .

In order to provide training signals (class scores), the last layer of CNNs is normally associated with some loss, and the training for CNNs can be done by doing gradient descent on the parameters with respect to the loss.

5.1.3 Fine-tuning

Training a deep CNN from scratch (or full training) is not without complications [165]. First, CNNs require a large amount of labeled training data. Second, the training process requires extensive computational and memory resources. Third, training a deep CNN is often complicated due to overfitting and convergence issues. Therefore, it is common to pretrain a CNN on a very large dataset, and then use the pre-trained CNN model either as a fixed feature extractor or an initialization for the task of interest. As for the former method, we remove the last fully-connected layer outputting class scores and use the rest of the pre-trained model as a fixed feature extractor.

An alternative way is to fine-tune the weights of the pre-trained network by continuing the back-propagation process on targeted images. Fine-tuning is a procedure based on the concept of transfer learning [166]. It has been used successfully in many applications such as [35–37].

Several pre-trained CNN architectures have been proposed in the literature and some have been shown to produce better results than the most advanced state-of-the-art face recognition methods. In this chapter, we adopt the VGG-Face deep architecture [8] shown on Fig. 5.1. The VGG-Face model was learned from a large face dataset containing 982,803 web images of 2622 celebrities and public figures. It can be used as a feature extractor for classifying any arbitrary face image by running the image through the entire network. The 4096-dimensional vector of the first fully-connected layer fc6 is used as the feature.

To do fine-tuning, we remove the final softmax layer and initialize a new one with random values. The new softmax classifier is trained from scratch using the back-propagation algorithm with data from our new database. In addition, the number of output feature maps of the last fully-connected layer should be changed to the number of classes in the new dataset. We divide the training set of the new database into a smaller training set with 80% of the training data and a validation set with the remaining 20% of the training data. The training process will be terminated when the highest accuracy on the validation set is observed. Our implementation is

Figure 5.1: VGG-Face architecture, CONV indicates convolutional layers, POOL indicates pooling layers and FC indicates fully-connected layers.

based on the open source deep learning library Caffe [167], and we run the experiment using a NVIDIA K80 GPU, which is a dual GPU board that combines 24 GB of memory.

5.1.4 Enhance the CNN performance by color pixel values

The good performance achieved by pre-trained CNNs is strictly restricted to applications which have similar face images to the pre-training datasets. In practice, different applications have significant variations in viewpoints, image sizes, scene context, illumination, expressions or other factors. Although fine-tuning can offer some help in improving the performance of pre-trained CNNs, an effective fine-tuning needs to fine-tune all layers of the pre-trained models when the variations between the source and target applications are significant. However, when the pre-trained CNNs are fine-tuned too much, the limited number of fine-tuning data might cause the problem of overfitting. Given the “data-hungry” nature of CNNs and the difficulty of collecting large-scale image datasets, the applicability of CNNs to tasks with limited amount of training data appears as an important and open problem.

Here, we propose to use the lowest-level features, color pixel values, to enhance the face recognition performance of CNN models based on the framework shown on Fig. 5.2.

Color pixel values from the RGB color space and the LuC_1C_2 color space in chapter 3 are basic low-level features and they keep the most of the original appearance information of

Figure 5.2: Framework of the proposed method, DR indicates dimension reduction and RPs indicates raw pixels.

face images. The feature of pixel values has been used in many recent face recognition works [25, 68]. In order to provide complementary and raw information of the targeted face images to image representations learned by CNNs, we combine pixel values of face images with image representations learned by CNNs.

To get similarity scores from CNN features, we feed face images to the VGG-Face model as shown on Fig. 5.1. The 4096-dimensional vector obtained at the first fully-connected layer fc6 by eq. 2 is taken as the CNN feature. Suppose x_1 and x_2 are two learned feature vectors for the gallery and query face images, respectively, we use the following equation to calculate their similarity score:

$$s^{CNN} = \frac{x_1^t x_2}{|x_1| |x_2|}, \quad (5.3)$$

where t indicates transpose and $|\cdot|$ indicates L2 norm.

As for the similarity score of non-CNN features, color pixel values, we first apply PCA to pixel values of face images for dimension reduction. PCA is a standard technique for dimensionality reduction and has been applied to a broad class of computer vision problems. It has been explained in [135, 136] that PCA plays an important role in the classification. It alleviates the overfitting problem or improves the generalization capability by eliminating the subspace spanned by eigenvectors which match with the smallest eigenvalues while keeping the principal components. Though an unsupervised method that minimizes the data reconstruction error rather than maximizes the class discrimination, PCA improves the classification accuracy by removing the unreliable, problematic, or harmful dimensions. This ensures the minimum loss of the discriminative information in the extracted subspace. Let $z \in R^N$ be a column vector of pixel values, whose covariance matrix is Σ_z . The PCA factorizes the covariance matrix Σ_z into the following form:

$$\Sigma_z = \Phi \Lambda \Phi^t, \quad (5.4)$$

where Φ is an orthogonal eigenvector matrix and Λ is a diagonal eigenvalue matrix. PCA reduces the dimension of the vector z by:

$$z' = P^t z, \quad (5.5)$$

where P is the matrix of the m largest eigenvectors in Φ and z' is the resulting m -dimensional vector.

After dimension reduction, we use the Mahalanobis distance to compute similarity scores between low-dimensional feature vectors instead of the Euclidean distance. To start with, we calculate the within-class scatter matrix Σ^w of z' as following:

$$\Sigma^w = \sum_{i=1}^p \sum_{j=1}^{q_i} (z'_{ij} - \bar{z}'_i)(z'_{ij} - \bar{z}'_i)^t \quad (5.6)$$

where i indicates the number of classes, j indicates the number of images for class i and $\bar{z}'_i$ is the mean vector for samples of class i . Then we factorize the within-class scatter matrix Σ_w using eq. 5.4 to obtain the eigenvector matrix Φ^w and the eigenvalue matrix Λ^w . After that, we calculate the whitening matrix Q of Σ^w by

$$Q = \Phi^w (\Lambda^w)^{-\frac{1}{2}}. \quad (5.7)$$

The whitened low-dimensional feature vector is

$$z'' = Q^t z' = Q^t P^t z. \quad (5.8)$$

The Euclidean distance computed from z'' is the Mahalanobis distance of z' . Suppose we have two feature vectors z''_1 and z''_2 for gallery and query face images, respectively, their similarity score s^{nCNN} is also computed by eq. 5.3.

To fuse s^{CNN} and s^{nCNN} for face recognition, we adopt the following weighted sum rule:

$$s = w^{CNN} s^{CNN} + w^{nCNN} s^{nCNN}, \quad (5.9)$$

where w^{CNN} and w^{nCNN} indicate weights for CNN similarity score and non-CNN similarity score, respectively. The resulting similarity score s can be used for subsequent Nearest-neighbor classification. For different face recognition applications, the similarity between pre-training images and targeted images is different. Thus the performance of pre-trained CNNs on different databases also varies. We determine the value of w^{CNN} and w^{nCNN} by the similarity between VGG images and the targeted images.

5.1.5 Experiments

Extensive experiments are conducted on LFW [81] and FRGC [62] databases to show that the CNN performance, with or without fine-tuning, can be visibly enhanced by using simple pixel values and the performance improvement is larger than that achieved by the widely used fine-tuning.

We set the value of w^{CNN} and w^{nCNN} to be 0.9 and 0.1 on LFW database. The face recognition performance on FRGC is reported by using the face verification rate (FVR) with the false accept rate (FAR) equivalent to 0.1%. We set the value of w^{CNN} and w^{nCNN} to be 0.5 and 0.5 on FRGC database.

5.1.5.1 Results on LFW

We follow the “Unrestricted, Labeled Outside Data Results” protocol in [81] and compute the mean verification accuracy according to the 10-fold cross-validation scheme. Both images with and without backgrounds are used to test the face verification performance of CNN. Original images with backgrounds are resized to 140×140 . Images without backgrounds indicate that only the face region is cropped out and normalized to the size of 32×32 . As there is no label information to use for fine-tuning the VGG-face model on LFW, only the pre-trained VGG-face model is used. Results are listed on Table 5.1. As we can see on Table 5.1,

Table 5.1: Face verification accuracy (%) using CNN and pixel values on LFW

Background	Model	CNN	CNN+ RGB	CNN+ LuC_1C_2
Without	VGG-face	85.62	86.85	87.32
With	VGG-face	97.43	97.45	97.72

CNN+ RGB and CNN+ LuC_1C_2 increase the face verification accuracy of CNN by 1.2% and 1.7%, respectively, for images without backgrounds. Simple pixel values of RGB and LuC_1C_2 only marginally improve the face verification accuracy of CNN for images with backgrounds, because the VGG-face model was initially trained and tuned to achieve state-of-the-art results on LFW.

5.1.5.2 Results on FRGC

Both images with and without backgrounds are used for comparison. Original images with backgrounds are resized to 211×201 . Images without backgrounds indicate that only the face region is cropped out and normalized to the size of 64×64 . To adapt the VGG-face model to FRGC database, we fine-tune the VGG-face model using images from the FRGC training set. As there are 222 subjects in the training data, we change the output size of the last full connected layer from 2622 to 222. The base learning rate is set to 0.001 and it drops by a factor of 0.5. In order to magnify the number of training samples per user, we perform the following data augmentation, for each training sample of size $m \times m$, we extract all possible crops of size $(m - 2) \times (m - 2)$. Each crop is also flipped along its vertical axis yielding a total of $2 \times 3 \times 3 = 18$ crops. The crops are then re-sized back to 224×224 and used for training the CNNs. The batch size is set to be 50. We run the back-propagation algorithm for around 4,000 iterations. After fine-tuning, we obtain the fine-tuned VGG-face model, FVGG-face model. In the experiments, both the pre-trained VGG-face model and the fine-tuned FVGG-face model are tested. Results are shown on Table 5.2.

Directly applying the pre-trained VGG-face model to a new database produces unsatisfactory results as shown on Table 5.2. The widely used fine-tuning only increases the FVR of

Table 5.2: Face verification rate (%) at FAR = 0.1% using CNN and pixel values on FRGC

Background	Model	CNN	CNN+ RGB	CNN+ LuC_1C_2
Without	VGG-face	45.32	84.11	89.07
	FVGG-face	75.34	90.11	93.42
With	VGG-face	88.86	95.37	96.98
	FVGG-face	92.50	96.00	97.57

pre-trained CNN to 75.34% and 92.50%. The effect of fine-tuning is limited because fine-tuning is still based on the pre-trained model. Pixel values extracted from RGB and LuC_1C_2 color spaces are complementary to the CNN features. Therefore, CNN+ RGB improves the FVR of VGG-face model to 84.11% and 95.37% for images without and with backgrounds, respectively. CNN+ LuC_1C_2 improves the FVR of VGG-face model to 89.07% and 96.98% for images without and with backgrounds, respectively. These performance improvements are more significant than 75.34% and 92.50% achieved by fine-tuning. Furthermore, CNN+ RGB improves the accuracy of the fine-tuned VGG-face model, FVGG-face model, to 90.11% and 96.00% for images without and with backgrounds, respectively. CNN+ LuC_1C_2 improves the accuracy of the fine-tuned VGG-face model, FVGG-face model, to 93.42% and 97.57% for images without and with backgrounds, respectively. This shows that low-level features improve the face verification rate of the fine-tuned CNN models significantly.

5.1.5.3 Discussion

The face verification performance of the pre-trained VGG-Face model is unsatisfactory on LFW and FRGC datasets. Even after fine-tuning, the performance of CNNs is not good enough. By using color pixel values, the face verification performance of both pre-trained and fine-tuned CNN models can be improved significantly. Simple pixel values of RGB and LuC_1C_2 only marginally improve the face verification accuracy of CNNs on LFW database, because the VGG-face model was initially trained and tuned to achieve state-of-the-art results on LFW database [8]. What's more, the performance improvement on pre-trained models

achieved by color pixel values is much larger than that achieved by fine-tuning. The performance improvement made by color pixel values can be significant when the targeted images have large variations with pre-training images.

5.1.6 Summary

In this chapter, we show that pre-trained CNN models and widely used fine-tuned CNN models cannot provide satisfactory face recognition performance under the case that training and testing datasets have big differences. To address this problem, we propose to combine low-level features, color pixel values, with image representations learned by CNNs. Extensive experiments are conducted on LFW and FRGC databases using the pre-trained CNN model, VGG-Face. Results show that color pixel values can greatly improve the face verification performance of pre-trained CNNs and fine-tuned CNNs.

5.2 Feature fusion of VGG-Face and ResNetShort

5.2.1 Introduction

Recently, Convolutional Neural Network (CNN) has been widely used for feature learning in face recognition and very promising results have been obtained as in [8, 20]. The pre-trained VGG-Face model [8] was learned from a large face dataset containing 2.6M web images of 2,622 celebrities and public figures. It is widely used as a feature extractor for classifying face images as in [38–40]. Different from the architecture of VGG-Face, ResNet in [20] consists of residual modules which conduct additive merging of signals. The authors in [20] argue that residual connections are inherently important for training very deep architectures. It is natural to study the combination of VGG-Face with ResNet, which would allow two models to reap the benefits of each other. The fusion of features from different deep models has been used in many works to achieve the state-of-the-art FR recognition performance. Authors in [78] train 60 CNNs, and each of them extracts two 160-dimensional DeepID vectors from 60 face patches

with ten regions, three scales, and RGB or gray channels. The combination of 60 different deep models increases the face verification accuracy by 5.27% over the best single model. The deep learning structure proposed in [44] is composed of a set of elaborately designed CNN models, which extract complementary facial features from multi-modal facial data. Here, we train a ResNet-like CNN network by using face images from the recently released CASIA-WebFace dataset [41] and combine it with the pre-trained VGG-Face model by feature fusion. The method of Covariance Matrix Regularization (CMR) proposed in chapter 4 utilizes the correlation between different features and reduces overfitting during the fusion of multiple features. It works by assigning weights to the cross-feature submatrices of covariance matrices of training data to suppress the influence of correlation between different features, which is estimated from the training data, in feature fusion. This method is used here to fuse feature representations from VGG-Face and ResNetShort.

In the experimental part conducted on four public face databases including MultiPIE, GT, AR and LFW, we first show that our proposed ResNetShort model achieves state-of-the-art face verification performance on LFW. After that, we vary the value of weights in CMR to show how it solves the problem of overfitting and improves the face recognition performance of CNN features. Finally, we compare the performance of feature fusion of ResNetShort and VGG-Face by CMR against that of the best single feature.

5.2.2 VGG-Face and ResNetShort

Convolutional Neural Networks have significantly improved the state of the arts in face recognition [85]. VGG-Face is a deep neural network proposed by Simonyan et al. in [8]. This network is characterized by using 3×3 convolutional layers stacked on top of each other in increasing depth. The architecture of VGG-Face comprises 21 layers, which consist of 13 convolutional layers, 5 maxpooling layers and 3 fully connected layers. The first two fully connected layers are 4,096 dimensional and the dimension of the last fully connected layer

depends upon the loss functions used for optimisation. The pre-trained VGG-Face model was learned from a large face dataset (see Fig. 5.3 for sample images) containing 2.6M images of 2,622 celebrities and public figures. Faces are detected using the method described in [168] and a 2D similarity transformation is applied to map the face to a canonical position. VGG-Face is first trained as a multi-subject classification problem by minimizing the softmax loss and then fine-tuned by the recently proposed triplet loss [22]. The pre-trained VGG-face model has been widely used by researchers to extract CNN features from face images as in[38–40].

Figure 5.3: Sample face images from VGG Face database.

Unlike traditional sequential network architectures such as VGG, ResNet consists of “network-in-network” modules. First introduced by He et al. in [20], ResNet has become a seminal work, demonstrating that the degradation problem of deep networks can be solved through the use of residual modules. ResNet layers are formulated as learning residual functions with reference to the layer inputs. By referring to the CNN model used in [45] and residual modules, we train a model as shown in Fig. 5.4 and name it ResNetShort. The size of filters in convolution layers is 3×3 with stride 1, followed by PReLU [169] non-linear units. The max-pooling grid is 2×2 and the stride is 2. The number of feature maps in convolutional layers or the dimension of fully connected layers is indicated by the number on top of each layer. ‘ $\times h$ ’ represents a residual module that repeats for h times. Joint supervision of softmax loss and center loss [45] is adopted. The value of λ , which is used for balancing the softmax and center loss functions, is set as 0.005.

The recently released CASIA-WebFace [41] database is used to train the ResNetShort model. There are 494,414 images of 10,575 subjects in the CASIA-WebFace dataset. According

Figure 5.4: The ResNetShort architecture, where C, P, and F indicate convolutional, max pooling, and fully connected layers, respectively.

to [170], adding the individuals with only a few instances do not help to improve the recognition performance. Indeed, these individuals will harm the systems performance. Thus we take 434,793 images of 9,067 persons, which contain at least 14 images for each person, for training. All of the remaining images are discarded. In order to normalize the face images, we apply the affine transformation to them based on the positions of five facial points including two eye centers, the nose tip, and two mouth corners. All images are rescaled to the size of 112×96 . Exempling normalized images are displayed on Fig. 5.5. An off-the-shelf face alignment approach proposed in [143] is utilized for facial point detection. Also, we flip the training face images horizontally to increase the number of training images. We adopt Caffe [167] to implement the training of the ResNetShort model. The batch size of 256 is used. The learning rate of all layers starts from 0.1, and it drops by a factor of 0.9 after 16000 iterations of back-propagation, and it drops by 0.9 after 8000 iterations more to the learning rate of 0.001. The total number of iterations is 28000.

Figure 5.5: Normalized face images from CASIA-WebFace database.

Both the pre-trained VGG-Face model and our proposed and trained ResNetShort model achieve state-of-the-art face verification performance on challenging face datasets (refer to section 5.2.3.1). A comprehensive comparison between these two deep models is given on Table

Table 5.3: Comparison between the pre-trained VGG-Face model and our trained ResNetShort model, CONV and FC indicate convolutional and fully connected layers, respectively.

Model	VGG-Face	ResNetShort
Training data	VGG Face	CASIA-WebFace
Face alignment	vanilla DPM [168]	TCDCN [143]
Input size	$224 \times 224 \times 3$	$112 \times 96 \times 3$
Architecture	CONV + FC	Residual modules
Non-linear units	ReLU	PReLU
Feature size	4096	512
Supervision signals	softmax+triplet loss	softmax+center loss

5.3. From which we can observe that, these two models are trained from different face images by optimizing different loss functions through different deep architectures. This makes the learned discriminative information contained in VGG-Face features and ResNetShort features mutually complementary to each other. Therefore, we combine these two CNN models by feature fusion to effectively make use of their discriminative information. The feature fusion method of CMR proposed in chapter 4 is used here as it has been proved to outperform the best single feature, DR-Cat and Cat-DR for fusing features of different color channels and different types of features.

5.2.3 Experiments

We conduct extensive experiments on four public datasets: MultiPIE [79], GT [80], AR [61], and LFW [81] to assess the effectiveness of feature fusion of CNN features extracted by VGG-Face and ResNetShort models using CMR.

We use the largest variation subset of Multi-PIE dataset, illumination subset, which consists of 105 subjects with 80 face images per subject across 4 sessions (20 images per subject in each session). Similar to [161], we randomly choose s samples from 20 samples per subject in session 1 as the training and gallery data. Remaining 6300 face images of 105 subjects in session 2 to session 4 serve as query data. The nearest neighbor classifier with mahalanobis

distance is used for classification. The gallery image is obtained by averaging all training samples per person. Exemplar images are shown in Fig. 5.6.

Figure 5.6: Exemplar face images from the illumination variation subset of Multi-PIE.

Similar to Multi-PIE database, we randomly choose s images from 15 images for each person as training and gallery data on GT. The rest $(15 - s)$ face images per person serve as query data. The classifier is same as that used on Multi-PIE.

Figure 5.7: Exemplar face images from Georgia Tech face database.

On AR, we randomly choose s samples from 7 samples per subject in session 1 as the training and gallery data. Remaining 700 face images of 100 subjects in session 2 serve as query data. The classifier is same as that used on Multi-PIE. Exemplar images are shown in Fig. 5.8.

Figure 5.8: Exemplar face images from AR database.

On LFW dataset, we follow the “Unrestricted, Labeled Outside Data Results” protocol in [81] and compute the mean verification accuracy according to the 10-fold cross-validation scheme. Face images are normalized and aligned using the same method on CASIA-WebFace images. For face verification, we take the CNN features provided by the VGG-Face model and

Figure 5.9: Sample face images from LFW database.

the ResnetShort model as the feature representation of each probe image. We use the cosine distance for similarity calculation.

A number of experiments are conducted. To begin with, we evaluate the face verification performance of the pre-trained VGG-Face model and our proposed ResNetShort model on the challenging LFW dataset. Then, by decreasing the value of weights from 1 to 0, we validate that CMR solves the overfitting problem and improves the face recognition performance of CNN features. Finally, the face recognition performance of feature fusion using CMR is compared with that of the best single feature for fusing features extracted by VGG-Face and ResNetShort. For the convenience and clarity of all experiments, we adopt the same value w for w_{nm} , $w_{nm} = w$, in CMR.

5.2.3.1 Performance evaluation of ResNetShort features

In this experiment, the face verification performance of ResNetShort is evaluated on the challenging LFW database. Following the “Unrestricted, Labeled Outside Data Results” protocol, we input aligned face images to ResNetShort models and the 512-dimensional feature vector after the first fully-connected (FC) layer is taken as the final deep representation. The unsupervised diagram is used here, where PCA and cosine distance are used to compute the similarity score between two features. We evaluate the covariance matrix of CNN features for PCA using the 9 training folds of LFW data in the 10-fold cross validation. The face verification performances of VGG-Face and ResNetShort are reported in Table 5.4, we also compare them with other state-of-the-art models DeepID [4] and Canonical View CNN [171], which have been peer-reviewed and published. FaceNet [22] is not included in Table 5.4 for comparison, as it is

Table 5.4: Face verification accuracy (%) of ResNetShort, VGG-Face, DeepID and Canonical CNN on LFW.

Model	ResNetShort	VGG-Face	DeepID [4]	Can. CNN [171]
Verif. metric	Cosine	Cosine	Joint Bayes	Joint Bayes
Mean accuracy (%)	98.72	97.93	97.45	96.45

trained from 260M images of 8M subjects and uses a complex triplets selection algorithm. It is not fair to compare it with other deep models trained using less than 0.5M images. We can observe from Table 5.4, the proposed ResNetShort model achieves comparable performance to other CNN models.

5.2.3.2 Evaluation of CMR against the level of regularization

Here, we conduct experiments to investigate how different levels of covariance matrix regularization make influence on the face recognition performance. Specifically, we vary the value of weights w in CMR from 1 to 0, so that its regularizing effect on the cross-feature covariances changes from weaker to stronger.

This experiment is carried out on Multi-PIE, GT and AR datasets. The number of randomly selected samples per subject, s , equals 2 on Multi-PIE, GT and AR to increase the difficulty of the face recognition task. We report the face recognition performances of different weights in CMR by face recognition rate (FRR), which is the ratio of the number of correctly classified query images to the total number of query images. Note that, among all tested feature dimensions of PCA and EFM, the best found FRR is reported. We plot FRR against the value of weights in CMR on Fig. 5.10. As we can observe, when the value of weights in CMR decreases from 1 to zero, the FRR increases to the maximum point and then decreases for CNN features on all three databases. The best performance is achieved at weight of $0.5 \sim 0.9$. One clear and consistent conclusion summarized from Fig. 5.10 is that, applying CMR to feature fusion of CNN features improves the face recognition performance consistently on all 3 datasets.

Figure 5.10: Face recognition rates (%) of fusing features extracted by different deep models (VGG-Face and ResNetShort) against the value of weights in CMR on Multi-PIE, GT and AR.

5.2.3.3 Performance comparison of feature fusion using CMR against the best single feature

To systematically compare the performance of feature fusion using CMR with that of the best single feature of VGG-Face and ResNetShort, we conduct experiments on Multi-PIE, GT, AR and LFW datasets. In CMR, we vary the value of w from 0 to 0.9 with step size of 0.1 and report the best classification performance. Training and testing protocols of Multi-PIE, GT and AR are the same as those in section 5.2.3.2. To increase the difficulty of the face verification task on LFW, only 1 training fold in the 10-fold cross validation is used to train PCA or EFM, remaining 9 training folds are used for testing. Other experimental settings remain the same as in section 5.2.3.1. We show the FRR of the best single (B.S.) feature and feature fusion using CMR on Table 5.5 for CNN features extracted by VGG-Face and ResNetShort. We use **bold texts** and underline texts to highlight the highest and the second highest face recognition/verification accuracy among all methods, respectively.

Table 5.5: Face recognition/verification performances of the best single feature and CMR using CNN features of VGG-Face and ResNetShort models on Multi-PIE, GT, AR and LFW.

CNN	Multi-PIE		GT		AR		LFW	
	PCA	EFM	PCA	EFM	PCA	EFM	PCA	EFM
B.S. model	94.89	95.08	97.85	97.38	91.43	91.57	97.14	97.14
CMR	98.02	97.94	98.92	98.77	95.57	95.71	97.80	97.83

As shown in Table 5.5, the feature fusion of VGG-Face and ResNetShort using CMR performs better than the best single feature in all experiments, which indicates that the feature fusion of VGG-Face and ResNetShort is effective in promoting the CNN-based FR performance.

5.2.4 Summary

In this chapter, we train a ResNet-like CNN model, ResNetShort, and combine its features with those of VGG-Face using CMR. The two models of ResNetShort and VGG-Face are trained from different face images by optimizing different loss functions through different architectures. This makes the learned discriminative information contained in ResNetShort features and VGG-Face features mutually complementary to each other. Extensive experiments conducted on four popular face databases show that the proposed feature fusion of VGG-Face and ResNetShort consistently outperforms the best single feature.

5.3 Deep Coupled ResNet Model

5.3.1 Introduction

Face recognition (FR) has been a very active research area due to increasing security demands, commercial applications and law enforcement applications[47, 114, 172]. Promising results have been achieved under challenging conditions such as occlusion [42], variations in pose and illumination [43]. While many FR approaches have been developed for recognizing HR face images[22, 44, 45], there are few studies focused on FR in surveillance systems, where HR

cameras are not available or there is a long distance between the camera and the subject. Under the condition of LR images, FR approaches developed for HR images usually decline [46, 47]. It is still a challenge to recognize faces when only LR probe images are available.

Here, we focus on the LR face recognition (LRFR) problem of matching LR probe face images with HR gallery images. Most of the approaches proposed for this task can be generally divided into two categories. One is to reconstruct the HR probe image from the LR one by super-resolution (SR) techniques and use it for classification. Although SR-based methods such as [92, 96–98] can generate visually appealing HR images, they are computationally expensive and not optimized for recognition purposes thus the results can be further improved [10, 47, 100, 101].

The other category is to simultaneously transform the LR probe and corresponding HR gallery images into a common feature subspace where the distance between them is minimized. Li et al. in [47] propose to learn two matrices which project the face images with different resolutions into a unified feature space where the difference between the LR image and its HR counterpart is minimized. Based on the idea of linear discriminant analysis, several discriminant subspace methods are proposed in [173–175]. Instead of using linear methods, Ren et al. in [95] project the LR and HR face images into an infinite common subspace by minimizing the dissimilarities captured by kernel Gram matrices. Multidimensional scaling (MDS) is employed in [100] to simultaneously transform the features from the poor quality probe images and the high-quality gallery images in the manner that their distances approximate those between gallery images. The same authors propose a reference-based approach for reducing the computational cost in [176]. Two discriminative MDS methods are proposed in [101] to make full use of identity information, including both inter-class and intra-class distances. Their new objective function is claimed to enlarge the inter-class distances to ensure discriminability.

In general, subspace-based methods achieve better recognition performance than SR-based methods. However, subspace-based methods usually extract pixel values or scale-invariant

Figure 5.11: Architecture of the proposed Deep Coupled ResNet (DCR) model. The trunk network learns discriminant features (indicated by v) shared by different resolutions of images, and the branch networks are trained as coupled mappings (indicated by x for HR features and z for LR features, respectively). C, P and F indicate convolutional layer, max-pooling layer and fully-connected layer, respectively. The number of output feature maps in convolutional layers and the number of outputs in fully connected layers are indicated by those on top of each layer. ‘ $\times h$ ’ represents a residual module that repeats for h times. k indicates the resolution of LR training images and β is a scaling parameter for center loss.

feature transform (SIFT) from images as feature representations. Their performance can be boosted by using feature representations which are robust to the resolution change. Motivated by the superior performance of convolutional neural networks (CNN), authors in [10] train a deep convolutional network, Resolution-Invariance CNN (RICNN), to learn resolution invariant features in a supervised way by mixing the real HR images with the upsampled LR ones. Although RICNN improves the performance of LRFR, it is sensitive to resolution change of probe images as indicated in [101]. In [177], the authors proposed a coupled mappings method for low resolution face recognition using deep CNN. But their approach uses two complex networks of 33 layers to extract features from fixed resolution of probe images.

In this chapter, we propose a CNN-based approach, the Deep Coupled ResNet (DCR) model, to solve above mentioned problems in LRFR. The novelty and contribution of the DCR model come from following 4 aspects. i) The DCR model consists of one big trunk CNN and two small branch networks. ii) We train the trunk CNN only once to learn discriminant

features shared by face images of different resolutions. It is constructed based on recently proposed residual modules [9]. iii) Two branch networks are trained to learn resolution-specific coupled-mappings so that HR gallery images and LR probe images are projected to a space where their distances are minimized. A Coupled-Mapping (CM) loss is proposed to optimize model parameters of branch networks; iv) In reality, there can be various resolutions of probe images to be matched with HR gallery images, the proposed DCR model solves this problem by training multiple pairs of small branch networks (2MB for each pair) while using the same big trunk network (105MB). Resolution indicator methods [178] can be employed to determine the resolution of probe images and which branch network to be used.

5.3.2 Deep Coupled ResNet model

5.3.2.1 Architecture of DCR

The key of low-resolution face recognition (LRFR) is to extract the feature that is robust to resolution change and to measure the similarity between the HR gallery image and the LR probe image. Recently, CNN has been widely used for feature learning in FR and excellent performances have been achieved as in [8, 9]. This motivates us to employ CNN in learning discriminative features shared by different image resolutions and in deriving coupled mappings to minimize the distance between the HR gallery image and the LR counterpart of a specific resolution. The architecture of the proposed Deep Coupled ResNet (DCR) model is shown on Fig. 5.11. It consists of a trunk CNN plus two branch networks, where the trunk CNN is designed to extract discriminant face features which are robust to the resolution degradation of face images. Two branch networks are trained to learn coupled mappings which minimize the distance between feature vectors extracted by the trunk branch from a HR image and its corresponding LR image of a specific resolution k .

The trunk network of the proposed DCR model is constructed based on the CNN model in [45] and the ResNet model proposed in [9]. Different from previous CNN architectures such as

VGG, ResNet in [9] consists of residual modules which conduct additive merging of signals. The authors in [9] argue that residual connections are inherently important for training very deep architectures. ResNet has become a seminal work, demonstrating that the degradation problem of deep networks can be solved through the use of residual modules. As shown on Fig. 5.11, the trunk network takes raw pixels of face images as input. C, P and F indicate convolutional layer, max-pooling layer and fully-connected layer, respectively. The kernel size of convolution layers is 3×3 with stride 1 and the kernel size of max-pooling layers is 2×2 with stride 2. Each convolutional layer is followed by a PReLU [169] non-linear unit. The number of output feature maps in convolutional layers and the number of outputs in fully connected layers are indicated by those on top of each layer. ' $\times h$ ' represents a residual module that repeats for h times.

The output feature vector $\mathbf{v}$ of the second last fully-connected layer in the trunk network forms input of the branch networks. The HR feature vector and its LR counterpart are fed into the branch networks in pair. Each branch network consists of two fully-connected layers and one PReLU unit in the middle. Thus the coupled mappings are formed as two non-linear mappings. Output feature vectors of the first fully-connected layers ($\mathbf{x}$ and $\mathbf{z}$) are taken as the feature representations for HR and LR images, respectively. The trunk network outputs feature representations which are robust to the resolution change of face images. Its model size is around 105MB. The branch network reduces the distance between the HR feature and its LR counterpart of a specific resolution. Its model size is around 2MB, 1.9% of the whole model.

5.3.2.2 Training strategy and Coupled-Mapping loss

The recently released CASIA-WebFace [41] database is used to train the proposed DCR model. The 434,793 images of 9,067 persons, which contain at least 14 images per person, compose the training set. In order to normalize face images, we apply the affine transformation to them based on the positions of five facial points including two eye centers, the nose tip, and two

mouth corners. All face images are rescaled to the size of 112×96 . We utilize a face alignment approach proposed in [143] to detect facial points. Some example face images are shown on Fig. 5.12. A two-step training strategy is adopted to effectively optimize parameters of the

Figure 5.12: Example face images from CASIA-WebFace.

DCR model. In the first step, the trunk network is trained using face images of 3 significantly different image resolutions, 112×96 , 40×40 and 6×6 , so that it can be applied to the feature extraction of images whose resolution varies from 112×96 to 6×6 . Similar to [10], different resolutions of face images are rescaled to the required input size of the network using bicubic interpolation before being fed to the network for training. In our case, the required size of input images is 112×96 . The joint supervision of softmax loss and center loss [45], L_t , is used to train the trunk network as below:

$$L_t = -\sum_{i=1}^{3m} \log \frac{e^{W_{y_i}^T \mathbf{v}_i + d_{y_i}}}{\sum_{j=1}^n e^{W_j^T \mathbf{v}_i + d_j}} + \beta \sum_{i=1}^{3m} \|\mathbf{v}_i - \mathbf{c}_{y_i}^v\|_2^2, \quad (5.10)$$

where m is the number of training samples of the same resolution and n indicates the number of subjects in the training data. $\mathbf{v}_i$ indicates the feature vector extracted by trunk network from the i th training image. W_j denotes the j th column of the weights W in the last fully-connected layers of trunk networks, and d is the bias item. y_i is the class label for the i th sample and $\mathbf{c}_{y_i}^v$ denotes the y_i th class center of deep features $\mathbf{v}$. β is a scaling factor.

In the second step, model parameters of the trunk CNN remain unchanged and two branch networks are trained using HR images (112×96) and LR images of similar resolution to that of the targeted LR probe images. The Coupled-Mapping (CM) loss function is proposed to supervise the training of branch networks.

In the training process of branch networks, we first expect to maximize the distances between each face image and its neighbors from different classes within HR or LR features. The softmax loss defined as in equation (5.11) facilitates interclass separation.

$$L_s = -\sum_{i=1}^m \log \frac{e^{U_{y_i}^T \mathbf{x}_i + b_{y_i}}}{\sum_{j=1}^n e^{U_j^T \mathbf{x}_i + b_j}} - \sum_{i=1}^m \log \frac{e^{V_{y_i}^T \mathbf{z}_i + a_{y_i}}}{\sum_{j=1}^n e^{V_j^T \mathbf{z}_i + a_j}}, \quad (5.11)$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ indicate the feature vectors extracted by branch networks from the i th HR and LR images, respectively. U_j, V_j denote the j th column of the weights U, V in the last fully-connected layers of branch networks, and a, b are the bias items, for HR feature and its LR counterpart, respectively.

In order to preserve intraclass compactness, we minimize the distances between each face image and its neighbors from the same class within HR or LR features. The center loss defined as in equation (5.12) minimizes the intra-class variations.

$$L_c = \sum_{i=1}^m \|\mathbf{x}_i - \mathbf{c}_{y_i}^x\|_2^2 + \sum_{i=1}^m \|\mathbf{z}_i - \mathbf{c}_{y_i}^z\|_2^2, \quad (5.12)$$

where $\mathbf{c}_{y_i}^x$ and $\mathbf{c}_{y_i}^z$ denote the y_i th class centers of HR features $\mathbf{x}$ and LR features $\mathbf{z}$, respectively.

For the task of LRFR, it is important to ensure the consistency between the LR feature and corresponding HR feature. Specifically, the final LR and corresponding HR feature vectors should be as close as possible. This can be achieved by minimizing the Euclidean loss defined in equation (5.13).

$$L_e = \sum_{i=1}^m \|\mathbf{x}_i - \mathbf{z}_i\|_2^2. \quad (5.13)$$

Considering the above three criteria, the Coupled-Mapping (CM) loss used for optimizing parameters of branch networks is constructed as below:

$$L_{CM} = L_s + \lambda L_c + \alpha L_e, \quad (5.14)$$

where λ and α are two scaling factors used for balancing three loss functions. In this way, not only the discriminability of HR and LR features are taken into account, but also the relationship between HR and corresponding LR features.

5.3.2.3 Matching face images

Once the training is finished, both HR gallery images and LR probe images can be fed into DCR to obtain their feature representations. Face verification or identification can be performed by computing their similarities. In real applications, probe images captured by surveillance cameras can be of various low resolutions. Thus multiple pairs of branch networks are trained in DCR to deal with different resolutions of LR probe images. As the big trunk network, which contains 98.1% of parameters of the whole DCR model, is fixed to extract resolution-robust features, only small branch networks, which contain 1.9% of all DCR parameters, need to be trained for matching HR images to probe LR images of a specific resolution. Hence, multiple branch networks greatly increase the effectiveness and efficiency of the DCR model to match HR images to different resolutions of probe face images. It is true that the final resolution of the face in the image might vary and not fit to the trained resolutions. As our branch network contains a small number of parameters and the model size is very small, we can train as many branch networks as needed to solve the problem of varying probe resolutions. In fact, one main purpose of our proposed small branch network is to handle many different low resolutions of probe face images.

5.3.3 Experiments

Extensive experiments are conducted on LFW [81] and SCface [99] databases to evaluate the proposed DCR model for matching LR probe images with HR gallery face images. Both LFW and SCface databases are widely-used benchmarks for FR in unconstrained environments. On LFW, the face verification performance of DCR is compared with VGGFace [8], LightCNN

[179], ResNet [45], and their fine-tuned versions, VGGFace-FT, LightCNN-FT, and ResNet-FT, respectively. Moreover, to evaluate the effectiveness of the proposed branch network for other networks, we replace the trunk network with LightCNN and VGGFace, and name the obtained models as Coupled-LightCNN and Coupled-VGGFace, respectively. Furthermore, the performance of the trunk network trained by face images of 3 different resolutions is also reported for comparison. Different sizes of LR probe images are used for testing. Please note that different sizes of face images are rescaled to the required input size of the network using bicubic interpolation before being fed to the network for feature extraction. On SCface, besides the above mentioned 9 CNN models, 4 state-of-the-art LRFR approaches, MDS [100, 176], DMDS, LDMDs [101], and RICNN [10], are also used for comparison with the DCR model. They are designed for LR face recognition and trained with LR images. The values of β in equation (5.10), λ and α in equation (5.14) are set to 0.008 in all experiments.

5.3.3.1 Experiments on LFW

On LFW, we follow the “Unrestricted, Labeled Outside Data Results” protocol in [81] and compute the mean verification accuracy according to the 10-fold cross-validation scheme. Face images are normalized and aligned using same methods as on CASIA-WebFace images. For two images in the face verification paradigm, we take the first one as HR (112×96) gallery image and down-sample the second one to a lower size of 8×8 , 12×12 , 16×16 or 20×20 , as LR probe image. Same sizes of CASIA-WebFace images are used for training of corresponding branch networks. The same face images are used for the fine-tuning of LightCNN, VGGFace, and ResNet models. PCA and cosine distance are used to compute the similarity scores between two CNN features. The total scatter matrix of PCA is computed using the 9 training folds of LFW data in the 10-fold cross validation. Face verification accuracies of VGGFace [8], VGGFace-FT, Coupled-VGGFace, LightCNN [179], LightCNN-FT, Coupled-LightCNN, ResNet [45], ResNet-FT, the trunk network and the proposed DCR model on LFW are shown on Table 5.6.

Table 5.6: Face verification accuracy of different approaches using different probe sizes on LFW

Probe size	8x8	12x12	16x16	20x20	Average
LightCNN [179]	67.7	78.3	86.9	92.7	81.4
LightCNN-FT	70.3	79.3	88.9	92.9	82.9
Coupled-LightCNN	80.0	85.1	90.2	93.5	87.2
VGGFace [8]	75.0	82.6	89.3	93.4	85.1
VGGFace-FT	82.3	88.6	92.7	94.8	89.6
Coupled-VGGFace	83.7	88.9	93.1	95.2	90.2
ResNet [45]	72.7	84.1	92.3	95.4	86.1
ResNet-FT	88.9	93.8	95.9	96.8	93.9
Trunk Network	92.2	93.6	95.5	96.8	94.5
DCR (Coupled-ResNet)	93.6	95.3	96.6	97.3	95.7

5.3.3.2 Experiments on SCface

Following the experimental settings in [101], frontal mugshot images are employed as gallery images and images taken by surveillance cameras at distance $d_i, i = 1, 2, 3$ are used as probe images. We take CASIA-WebFace images of size 112×96 as HR images and those of 112×96 , 30×30 and 20×20 as LR images for training of branch networks at distance of d_3 , d_2 and d_1 , respectively. Same as in [101], 50 out of 130 subjects in the SCface database are randomly chosen for fine-tuning of the branch networks and training of PCA. Rest of the subjects are for testing. Thus, there will be no identity overlapping between the training and testing sets. The same face images from CasiaWebface and SCface datasets are used for the fine-tuning of LightCNN, VGGFace, and ResNet models. The nearest-neighbor classifier is used to classify all probe images. We report the face recognition rates of MDS [100, 176], DMD-S, LDMDs [101], RICNN [10], VGGFace [8], VGGFace-FT, Coupled-VGGFace, LightCNN [179], LightCNN-FT, Coupled-LightCNN, ResNet [45], ResNet-FT, the trunk network and the proposed DCR model on Table 5.7.

We can observe from Table 5.6 and Table 5.7 that: (1), the branch network greatly increases the performance of LightCNN, VGGFace and the trunk (ResNet) networks; (2), the proposed DCR model achieves much higher face recognition accuracy than the state-of-the-art meth-

Table 5.7: Face recognition rates of different approaches at different distances on SCface

Distance	d1	d2	d3	Average
MDS [100, 176]	60.3	66.0	69.5	65.3
DMDS [101]	61.5	67.2	62.9	63.9
LDMDs [101]	62.7	70.7	65.5	66.3
RICNN [10]	23.0	66.0	74.0	54.3
LightCNN [179]	35.8	79.0	93.8	69.5
LightCNN-FT	49.0	83.8	93.5	75.4
Coupled LightCNN	50.5	85.0	94.0	76.5
VGGFace [8]	41.3	75.5	88.8	68.5
VGGFace-FT	46.3	78.5	91.5	72.1
Coupled-VGGFace	62.3	91.0	94.8	82.7
ResNet [45]	36.3	81.8	94.3	70.8
ResNet-FT	54.8	86.3	95.8	79.0
Trunk Network	52.0	89.5	96.3	79.3
DCR (Coupled-ResNet)	73.3	93.5	98.0	88.3

ods consistently for different resolutions of probe images on LFW and SCface datasets. The performance gain is significant for very LR probe images.

5.3.4 Summary

In this chapter, we propose a novel CNN-based approach named as the Deep Coupled ResNet model for the task of low resolution face recognition. It first extracts discriminative features shared by face images of different resolutions by a ResNet-like network, the trunk network. After that, coupled mappings are learned by branch networks to project features of HR images and corresponding LR images of a specific resolution into a common subspace where their distance is minimized. A Coupled-Mapping loss function is proposed to optimize model parameters of two branch networks. Experiments on LFW and SCface datasets show that the proposed DCR model achieves consistently and considerably better performance than the state-of-the-arts.

Chapter 6

Conclusions and Future Research

6.1 Color face recognition

Color information plays a discriminative and complementary role in FR process as proved in [1, 19, 26–28, 63, 65, 66, 69, 77, 112–114]. For color space conversion, we propose a color space LuC_1C_2 based on a framework of constructing effective color spaces for FR. It consists of one luminance component Lu and two chrominance components C_1, C_2 . The luminance component Lu is chosen among 4 different luminance candidates by analysing their R,G,B coefficients and the color sensor properties. In order to generate the two effective chrominance components C_1, C_2 , the directions of their transform vectors are derived by the discriminant analysis and the covariance analysis in the chrominance subspace of the RGB color space. The magnitudes of their transform vectors are derived according to the discriminant values of Lu, C_1, C_2 . Extensive experiments are conducted on 5 benchmark databases. The experimental results on Table 6.1 and Table 6.2 below show that both traditional descriptor-based features and CNN features extracted from the proposed color space perform better than those extracted from other color spaces such as $RQCr$ and RGB .

For color feature fusion, a Color Channel Fusion (CCF) method is proposed to make use of reliability and importance of features in different color channels. By integrating the dimension reduction rule of a single color channel across all three color channels, CCF selects more features from more reliable and discriminant color channels for DR-Cat methods. Extensive

Table 6.1: Performance comparison of LuC_1C_2 with state of the arts using complex features on FRGC

Color space	FVR (%) at FAR = 0.1%
Gray [74]	85.2
Gray [138]	88.1
Gray [139]	91.59
Gray [140]	92.89
RIQ [77]	80.03
RQCr [19]	92.43
ZRG-NII [2]	93.17
LuC_1C_2	93.38

Table 6.2: Averaged face recognition rate (AFRR) on GT database and averaged face verification accuracy (AFVA) on LFW database using ResNet- RGB and ResNet- LuC_1C_2 for images of resolution 4×4 , 6×6 , ..., 20×20 .

Database	ResNet- RGB	ResNet- LuC_1C_2
AFRR (%) on GT	76.5000	80.7885
AFVA (%) on LFW	84.7259	86.1630

experiments on three color face datasets are conducted to validate the effectiveness and robustness of the proposed CCF method. It outperforms the color channel concatenation method and the weighted decision-level fusion method. For Cat-DR methods, we propose a Covariance Matrix Regularization (CMR) technique to utilize the correlation between different features and reduce overfitting during the fusion of multiple features. It works by assigning weights to the cross-feature submatrices of covariance matrices of training data to suppress the influence of correlation between different features, which is estimated from the training data. Extensive experiments conducted on four popular face databases show that the proposed CMR technique consistently outperforms the best single feature, DR-Cat and Cat-DR for fusing features of different color channels and different types of features. Some of the experimental results are shown on Table 6.3. Besides fusing different color features, we also jointly use the information in three color channels during feature fusion by proposing a novel LBP-based color feature, TCLBP. It consists of intra-channel and inter-channel parts. The proposed inter-channel feature encodes the spectral structure of all 3 color channels at the same location. TCLBP is

Table 6.3: Face recognition performances of the best single feature, DR-Cat, Cat-DR and CMR using pixel values and LBP of channel R on Multi-PIE, GT, AR and FRGC.

R	Multi-PIE		GT		AR		FRGC	
	PCA	EFM	PCA	EFM	PCA	EFM	PCA	EFM
B.S. feature	83.08	82.83	83.60	83.40	78.57	78.86	61.68	70.82
DR-Cat	<u>88.51</u>	<u>88.29</u>	<u>86.20</u>	<u>86.60</u>	<u>82.86</u>	<u>83.14</u>	65.51	72.15
Cat-DR	87.97	87.98	84.60	85.00	81.86	81.71	<u>71.01</u>	<u>76.32</u>
CMR	90.16	90.03	88.00	87.80	84.86	85.14	71.42	76.86

more effective and efficient than existing LBP-based color features. Comparative experiments demonstrate that our proposed TCLBP yields visibly better face recognition accuracies than other state-of-the-art color LBP descriptors, Color LBP, CLBP and LCVBP.

6.2 Deep learning face recognition

Over the past decade, encouraging improvements have been achieved in FR by deep learning architectures [3, 4, 8]. However, we find in experiments that pre-trained CNN models and widely used fine-tuned CNN models cannot provide satisfactory face recognition performance under the case that training and testing datasets have big differences. To address this problem, we investigate the combination of the lowest-level features of color pixel values with image representations learned by CNNs. The low-level features depict basic characteristics of targeted face images from a different perspective, which is complementary to the high-level image representations learned by CNNs. Extensive experiments are conducted on LFW and FRGC databases using the pre-trained CNN model, VGG-Face. Results on Table 6.4 show that color pixel values can greatly improve the face verification performance of pre-trained CNNs and fine-tuned CNNs. In order to achieve better face recognition performance, we combine facial features learned by different deep models. A ResNet-like CNN model, ResNetShort, is firstly trained using images from the recently released CASIA-WebFace dataset [41] supervised by softmax loss and center loss. The two models of ResNetShort and VGG-Face are trained from different face images by optimizing different loss functions through different architectures. This makes

Table 6.4: Face verification rate (%) at FAR = 0.1% using CNN and pixel values on FRGC

Background	Model	CNN	CNN+ RGB	CNN+ LuC_1C_2
Without	VGG-face	45.32	84.11	89.07
	FVGG-face	75.34	90.11	93.42
With	VGG-face	88.86	95.37	96.98
	FVGG-face	92.50	96.00	97.57

Table 6.5: Face recognition/verification performances of the best single feature and CMR using CNN features of VGG-Face and ResNetShort models on Multi-PIE, GT, AR and LFW.

CNN	Multi-PIE		GT		AR		LFW	
	PCA	EFM	PCA	EFM	PCA	EFM	PCA	EFM
B.S. model	94.89	95.08	97.85	97.38	91.43	91.57	97.14	97.14
CMR	98.02	97.94	98.92	98.77	95.57	95.71	97.80	97.83

the learned discriminative information contained in ResNetShort features and VGG-Face features mutually complementary to each other. Their features are fused by CMR proposed in chapter 4. Extensive experiments are conducted on four popular face databases. The results on Table 6.5 show that the proposed feature fusion of VGG-Face and ResNetShort consistently outperforms the best single feature.

To improve the robustness and performance of CNN models for low-resolution FR tasks, we propose a novel approach named as the Deep Coupled ResNet model. It first extracts discriminative features shared by face images of different resolutions by a ResNet-like network, the trunk network. After that, coupled mappings are learned by branch networks to project features of HR images and corresponding LR images of a specific resolution into a common subspace where their distance is minimized. A Coupled-Mapping loss function is proposed to optimize model parameters of two branch networks. Different branch networks can be trained to deal with various resolutions of probe images while using the same trunk network. Experiments on LFW and SCface demonstrate that the proposed DCR model achieves consistently and significantly better performance than the state-of-the-arts. The results on SCface dataset are shown on Table 6.6.

Table 6.6: Face recognition rates of different approaches at different distances on SCface

Distance	d1	d2	d3	Average
MDS [100, 176]	60.3	66.0	69.5	65.3
DMDS [101]	61.5	67.2	62.9	63.9
LDMDs [101]	62.7	70.7	65.5	66.3
RICNN [10]	23.0	66.0	74.0	54.3
LightCNN [179]	35.8	79.0	93.8	69.5
LightCNN-FT	49.0	83.8	93.5	75.4
Coupled LightCNN	50.5	85.0	94.0	76.5
VGGFace [8]	41.3	75.5	88.8	68.5
VGGFace-FT	46.3	78.5	91.5	72.1
Coupled-VGGFace	62.3	91.0	94.8	82.7
ResNet [45]	36.3	81.8	94.3	70.8
ResNet-FT	54.8	86.3	95.8	79.0
Trunk Network	52.0	89.5	96.3	79.3
DCR (Coupled-ResNet)	73.3	93.5	98.0	88.3

6.3 Summary

In this section, we give a high-level discussion about the face recognition performance of all proposed methods in this thesis. In terms of accuracy, CNN-based methods including LuC_1C_2 -ResNet, CNN&RP (the combination of CNN features with raw pixels), CMR-CNN, and DCR perform much better than traditional methods including LuC_1C_2 -Tra, CCF, CMR-Tra, and TCLBP, where Tra indicates the traditional feature. LuC_1C_2 -ResNet extracts features from face images in the LuC_1C_2 color space, the extracted features can be fused with other CNN features or pixel values using CMR-CNN and CNN&RP, respectively. The computational speed of traditional methods depends on how complex the traditional feature is. Sometimes CNN-based methods can be faster than the traditional methods for feature extraction at the testing stage by using GPU cards. However, it takes quite a long time to train the CNN model from a large amount of face images. The methods of CMR-CNN and CNN&RP alleviate the problem of high time cost and achieve state-of-the-art recognition performance at the same time.

6.4 Future research

Building deeper and larger Convolutional Neural Networks has been a primary trend in the development of major visual tasks as in [9, 87]. In [9], the authors evaluate residual nets with a depth of up to 152 layers, which is 8 times deeper than VGG, a 16-layer network. GoogLeNet [87], used in the context of object detection and classification, is a 22-layer deep network. Deep and large networks require computation at billions of FLOPs. On the contrary, we should probably pay more attention to pursuing the best accuracy in very limited computational budgets at tens or hundreds of MFLOPs, focusing on common mobile platforms such as drones, robots, and phones.

We are going to conduct experiments on the LFW database to explore the smallest ResNet architecture that keeps the face verification accuracy as high as possible. The ResNetShort model used in chapter 4 will be taken as the baseline model. Different smaller versions of it will be designed by removing redundant residual modules, and reducing the number of feature maps. The performance of them will be tested using 112×96 images under the same experimental settings as in chapter 5 on the LFW database. Furthermore, we are also going to explore replacing the pooling layers with 1×1 convolutional layers to adjust the number of feature maps for residual modules.

Publications

- (i) **Ze Lu**, Xudong Jiang, and Alex Kot, “Color space construction by optimizing luminance and chrominance components for face recognition.” Major revision in *Pattern Recognition*, 2018.
- (ii) **Ze Lu**, Xudong Jiang, and Alex Kot, “Feature fusion with covariance matrix regularization in face recognition.” *Signal Processing*, 2018.
- (iii) **Ze Lu**, Xudong Jiang, and Alex Kot, “Deep coupled ResNet for low-resolution face recognition.” *IEEE Signal Processing Letters*, 2018.
- (iv) **Ze Lu**, Xudong Jiang, and Alex Kot, “A color channel fusion approach for face recognition.” *IEEE Signal Processing Letters*, 2015.
- (v) **Ze Lu**, Xudong Jiang, and Alex Kot, “A novel LBP-based color descriptor for face recognition.” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.
- (vi) **Ze Lu**, Xudong Jiang, and Alex Kot, “An effective color space for face recognition.” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.
- (vii) **Ze Lu**, Xudong Jiang, and Alex Kot, “Enhance deep learning performance in face recognition.” *IEEE International Conference on Image, Vision and Computing (ICIVC)*, 2017.

References

- [1] J. Y. Choi, Y. M. Ro, and K. N. Plataniotis, “Color local texture features for color face recognition,” *IEEE Trans. Image Processing*, vol. 21, no. 3, pp. 1366–1380, 2012.
- [2] S. H. Lee, J. Y. Choi, Y. M. Ro, and K. N. Plataniotis, “Local color vector binary patterns from multichannel face images for face recognition,” *IEEE Trans. Image Processing*, vol. 21, no. 4, pp. 2347–2353, 2012.
- [3] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708, 2014.
- [4] Y. Sun, X. G. Wang, and X. O. Tang, “Deep learning face representation from predicting 10,000 classes,” in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1891–1898, 2014.
- [5] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification-verification,” in *Advances in neural information processing systems*, p. 1988–1996, 2014.
- [6] Y. Sun, X. Wang, and X. Tang, “Deeply learned face representations are sparse, selective, and robust,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2892–2900, 2015.

REFERENCES

- [7] Y. Sun, D. Liang, X. Wang, and X. Tang, “Deepid3: Face recognition with very deep neural networks,” *arXiv preprint arXiv:1502.00873*, 2015.
- [8] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *British Machine Vision Conference*, vol. 1, p. 6, 2015.
- [9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *The IEEE Conf. Computer Vision and Pattern Recognition*, June 2016.
- [10] D. Zeng, H. Chen, and Q. Zhao, “Towards resolution invariant face recognition in uncontrolled scenarios,” in *International Conf. Biometrics*, pp. 1–8, IEEE, 2016.
- [11] Wikipedia, “Color vision — wikipedia, the free encyclopedia,” 2016. [Online; accessed 25-February-2016].
- [12] R. Kawakami, Z. Hongxun, R. T. Tan, and K. Ikeuchi, “Camera spectral sensitivity and white balance estimation from sky images,” *Intl. Journal of Computer Vision*, June 2013.
- [13] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, “Learning compact binary face descriptor for face recognition,” *IEEE trans. pattern analysis and machine intelligence*, vol. 37, no. 10, pp. 2041–2056, 2015.
- [14] M. Turk and A. Pentland, “Eigenfaces for recognition,” *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
- [15] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” *IEEE Trans. pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
- [16] T. Ahonen, A. Hadid, and M. Pietikainen, “Face description with local binary patterns: Application to face recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

- [17] C. Liu and H. Wechsler, “Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition,” *IEEE Trans. Image processing*, vol. 11, no. 4, pp. 467–476, 2002.
- [18] D. Chen, X. Cao, F. Wen, and J. Sun, “Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification,” in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3025–3032, 2013.
- [19] Z. Liu and C. Liu, “Fusion of color, local spatial and global frequency information for face recognition,” *Pattern Recognition*, vol. 43, no. 8, pp. 2882–2890, 2010.
- [20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE conf. computer vision and pattern recognition*, pp. 770–778, 2016.
- [21] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, “Bayesian face revisited: A joint formulation,” *Computer Vision–ECCV*, pp. 566–579, 2012.
- [22] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 815–823, 2015.
- [23] L. Torres, J. Reutter, and L. Lorente, “The importance of the color information in face recognition,” in *IEEE Intl. Conf. Image Processing*, vol. 3, pp. 627–631, 1999.
- [24] M. Rajapakse, J. Tan, and J. Rajapakse, “Color channel encoding with nmf for face recognition,” in *IEEE Intl. Conf. Image Processing*, vol. 3, pp. 2007–2010, 2004.
- [25] J. Y. Choi, Y. M. Ro, and K. N. Plataniotis, “Color face recognition for degraded face images,” *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 5, pp. 1217–1230, 2009.

REFERENCES

- [26] J. Yang, C. Liu, and L. Zhang, “Color space normalization: Enhancing the discriminating power of color spaces for face recognition,” *Pattern Recognition*, vol. 43, no. 4, pp. 1454–1466, 2010.
- [27] J. Yang and C. Liu, “A discriminant color space method for face representation and verification on a large-scale database,” in *IEEE Intl. Conf. Pattern Recognition*, pp. 1–4, 2008.
- [28] C. Liu, “Extracting discriminative color features for face recognition,” *Pattern Recognition Letters*, vol. 32, no. 14, pp. 1796–1804, 2011.
- [29] J. Ren, X. Jiang, and J. Yuan, “Noise-resistant local binary pattern with an embedded error-correction mechanism,” *IEEE Trans. Image Processing*, vol. 22, pp. 4049–4060, 2013.
- [30] A. Satpathy, X. Jiang, and H. Eng, “Lbp-based edge-texture features for object recognition,” *IEEE Trans. image Processing*, vol. 23, pp. 1953–1964, 2014.
- [31] J. Ren, X. Jiang, and J. Yuan, “Lbp-structure optimization with symmetry and uniformity regularizations for scene classification,” *IEEE Signal Processing Letters*, vol. 24, pp. 37–41, 2016.
- [32] J. Y. Choi, K. N. Plataniotis, and Y. M. Ro, “Using colour local binary pattern features for face recognition,” in *IEEE Intl. Conf. Image Processing*, pp. 4541–4544, 2010.
- [33] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and transferring mid-level image representations using convolutional neural networks,” in *IEEE conf. computer vision and pattern recognition*, pp. 1717–1724, 2014.
- [34] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *IEEE Intl. Conf. Computer Vision*, pp. 3730–3738, 2015.

- [35] S. S. Farfade, M. J. Saberian, and L. Li, “Multi-view face detection using deep convolutional neural networks,” in *ACM Intl. Conf. Multimedia Retrieval*, pp. 643–650, 2015.
- [36] F. Gurpinar, H. Kaya, H. Dibeklioglu, and A. Salah, “Kernel elm and cnn based facial age estimation,” in *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, pp. 80–86, 2016.
- [37] K. Zhang, L. Tan, Z. Li, and Y. Qiao, “Gender and smile classification using deep convolutional neural networks,” in *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, pp. 34–38, 2016.
- [38] X. Peng, Z. Xia, L. Li, and X. Feng, “Towards facial expression recognition in the wild: A new database and deep recognition system,” in *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, pp. 93–99, 2016.
- [39] F. Gurpinar, H. Kaya, H. Dibeklioglu, and A. Salah, “Kernel elm and cnn based facial age estimation,” in *The IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.
- [40] K. Zhang, L. Tan, Z. Li, and Y. Qiao, “Gender and smile classification using deep convolutional neural networks,” in *The IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.
- [41] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning face representation from scratch,” *arXiv preprint arXiv:1411.7923*, 2014.
- [42] J. Lai and X. Jiang, “Modular weighted global sparse representation for robust face recognition,” *IEEE Signal processing letters*, vol. 19, no. 9, pp. 571–574, 2012.
- [43] B. Wang, W. Li, W. Yang, and Q. Liao, “Illumination normalization based on weber’s law with application to face recognition,” *IEEE Signal Processing Letters*, vol. 18, no. 8, pp. 462–465, 2011.

REFERENCES

- [44] C. Ding and D. Tao, “Robust face recognition via multimodal deep face representation,” *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2049–2058, 2015.
- [45] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *European Conf. Computer Vision*, pp. 499–515, Springer, 2016.
- [46] B. Boom, G. Beumer, L. J. Spreeuwiers, and R. N. Veldhuis, “The effect of image resolution on the performance of a face recognition system,” in *International Conf. Control, Automation, Robotics and Vision*, pp. 1–6, IEEE, 2006.
- [47] B. Li, H. Chang, S. Shan, and X. Chen, “Low-resolution face recognition via coupled locality preserving mappings,” *IEEE Signal processing letters*, vol. 17, no. 1, pp. 20–23, 2010.
- [48] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, “Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition,” *Pattern Recognition Letters*, vol. 26, no. 2, pp. 181–191, 2005.
- [49] S. Ji and J. Ye, “Generalized linear discriminant analysis: a unified framework and efficient model selection,” *Trans. Neural Networks*, vol. 19, no. 10, pp. 1768–1782, 2008.
- [50] B. Moghaddam and A. Pentland, “Probabilistic visual learning for object representation,” *Trans. pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 696–710, 1997.
- [51] B. Moghaddam, “Principal manifolds and probabilistic subspaces for visual recognition,” *Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 780–788, 2002.

REFERENCES

- [52] X. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 383–394, 2008.
- [53] C. Jones and A. L. Abbott, "Color face recognition by hypercomplex gabor analysis," in *Intl. Conf. Automatic Face and Gesture Recognition*, pp. 6–pp, IEEE, 2006.
- [54] G. D. Finlayson, S. D. Hordley, and P. M. Hubel, "Color by correlation: A simple, unifying framework for color constancy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1209–1221, 2001.
- [55] J. Geusebroek, R. Van d. B., A. W. M. Smeulders, and H. Geerts, "Color invariance," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 12, pp. 1338–1350, 2001.
- [56] L. H. Wurm, G. E. Legge, L. M. Isenberg, and A. Luebker, "Color improves object recognition in normal and low vision.," *Journal of Experimental Psychology: Human perception and performance*, vol. 19, no. 4, p. 899, 1993.
- [57] A. Yip and P. Sinha, "Role of color in face recognition," 2001.
- [58] C. Wang, B. Yin, X. Bai, and Y. Sun, "Color face recognition based on 2dpca," in *Intl. Conf. Pattern Recognition*, pp. 1–4, IEEE, 2008.
- [59] P. Peer, "Cv1 face database," *Computer vision lab., faculty of computer and information science, University of Ljubljana, Slovenia. Available at <http://www.lrv.fri.uni-lj.si/facedb.html>*, 2005.
- [60] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (pie) database," in *Intl. Conf. Automatic Face and Gesture Recognition*, pp. 53–58, IEEE, 2002.

REFERENCES

- [61] A. M. Martinez, “The ar face database,” *CVC Technical Report*, vol. 24, 1998.
- [62] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, “Overview of the face recognition grand challenge,” in *IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 947–954, 2005.
- [63] C. Liu, “Learning the uncorrelated, independent, and discriminating color spaces for face recognition,” *IEEE Trans. Information Forensics and Security*, vol. 3, no. 2, pp. 213–222, 2008.
- [64] R. Lukac and K. N. Plataniotis, *Color image processing: methods and applications*. CRC press, 2006.
- [65] J. Y. Choi, Y. M. Ro, and K. N. Plataniotis, “Boosting color feature selection for color face recognition,” *IEEE Trans. Image Processing*, vol. 20, no. 5, pp. 1425–1434, 2011.
- [66] P. Shih and C. Liu, “Improving the face recognition grand challenge baseline performance using color configurations across color spaces,” in *IEEE Intl. Conf. Image Processing*, pp. 1001–1004, 2006.
- [67] J. Yang and C. Liu, “Color image discriminant models and algorithms for face recognition,” *IEEE Trans. Neural Networks*, vol. 19, no. 12, pp. 2088–2098, 2008.
- [68] P. Shih and C. Liu, “Comparative assessment of content-based face image retrieval in different color spaces,” *Intl. Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 07, pp. 873–893, 2005.
- [69] J. Yang and C. Liu, “A general discriminant model for color face recognition,” in *IEEE Intl. Conf. Computer Vision*, pp. 1–6, 2007.
- [70] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic press, 2013.

REFERENCES

- [71] P. Comon, “Independent component analysis, a new concept?,” *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.
- [72] J. Zou, Q. Ji, and G. Nagy, “A comparative study of local matching approach for face recognition,” *IEEE Trans. image processing*, vol. 16, no. 10, pp. 2617–2628, 2007.
- [73] Y. Su, S. Shan, X. Chen, and W. Gao, “Hierarchical ensemble of global and local classifiers for face recognition,” *IEEE Trans. Image Processing*, vol. 18, no. 8, pp. 1885–1896, 2009.
- [74] S. Xie, S. Shan, X. Chen, and J. Chen, “Fusing local patterns of gabor magnitude and phase for face recognition,” *IEEE Trans. Image Processing*, vol. 19, no. 5, pp. 1349–1361, 2010.
- [75] J. G. Daugman, “Two-dimensional spectral analysis of cortical receptive field profiles,” *Vision research*, vol. 20, no. 10, pp. 847–856, 1980.
- [76] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.
- [77] Z. Liu and C. Liu, “A hybrid color and frequency features method for face recognition,” *IEEE Trans. Image Processing*, vol. 17, no. 10, pp. 1975–1980, 2008.
- [78] Y. Sun, X. Wang, and X. Tang, “Deep learning face representation from predicting 10,000 classes,” in *IEEE Conf. Computer Vision and Pattern Recognition*, June 2014.
- [79] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, “Multi-pie,” *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.
- [80] A. Nefian and A. Nefian, “Georgia tech face database,” 2013.

REFERENCES

- [81] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” tech. rep., Technical Report 07-49, University of Massachusetts, Amherst, 2007.
- [82] P. J. Phillips, H. Moon, S. Rizvi, P. J. Rauss, *et al.*, “The feret evaluation methodology for face-recognition algorithms,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.
- [83] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [84] D. Williams and G. Hinton, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–538, 1986.
- [85] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *British Machine Vision Conference*, 2015.
- [86] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conf. computer vision*, pp. 818–833, Springer, 2014.
- [87] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *IEEE conf. computer vision and pattern recognition*, pp. 1–9, 2015.
- [88] L. Wolf, T. Hassner, and I. Maoz, “Face recognition in unconstrained videos with matched background similarity,” in *Computer Vision and Pattern Recognition*, pp. 529–534, IEEE, 2011.
- [89] J.-C. Chen, V. M. Patel, and R. Chellappa, “Unconstrained face verification using deep cnn features,” in *IEEE Winter Conf. Applications of Computer Vision*, pp. 1–9, 2016.

- [90] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [91] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. F., “Imagenet: A large-scale hierarchical image database,” in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
- [92] P. H. Hennings Y., S. Baker, and B. V. Kumar, “Simultaneous super-resolution and feature extraction for recognition of low-resolution faces,” in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1–8, 2008.
- [93] Y. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips, “A meta-analysis of face recognition covariates,” in *Biometrics: Theory, Applications, and Systems, 2009. BTAS’09. IEEE 3rd International Conference on*, pp. 1–8, IEEE, 2009.
- [94] Z. Wang, Z. Miao, Q. J. Wu, Y. Wan, and Z. Tang, “Low-resolution face recognition: a review,” *The Visual Computer*, vol. 30, no. 4, pp. 359–386, 2014.
- [95] C. Ren, D. Dai, and H. Yan, “Coupled kernel embedding for low-resolution face image recognition,” *IEEE Trans. Image Processing*, vol. 21, no. 8, pp. 3770–3783, 2012.
- [96] W. W. Zou and P. C. Yuen, “Very low resolution face recognition problem,” *IEEE Trans. Image Processing*, vol. 21, no. 1, pp. 327–340, 2012.
- [97] J. Wu, S. Ding, W. Xu, and H. Chao, “Deep joint face hallucination and recognition,” *arXiv preprint arXiv:1611.08091*, 2016.
- [98] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, “Studying very low resolution recognition using deep networks,” in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 4792–4800, 2016.

REFERENCES

- [99] M. Grgic, K. Delac, and S. Grgic, "Scface-surveillance cameras face database," *Multimedia tools and applications*, vol. 51, no. 3, pp. 863–879, 2011.
- [100] S. Biswas, G. Aggarwal, P. J. Flynn, and K. W. Bowyer, "Pose-robust recognition of low-resolution face images," *IEEE trans. pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 3037–3049, 2013.
- [101] F. Yang, W. Yang, R. Gao, and Q. Liao, "Discriminative multidimensional scaling for low-resolution face recognition," *IEEE Signal Processing Letters*, 2017.
- [102] G. D. Finlayson, S. S. Chatterjee, and B. V. Funt, "Color angular indexing," in *Computer Vision-ECCV'96*, pp. 16–27, Springer, 1996.
- [103] Y. D. Chun, N. C. Kim, and I. H. Jang, "Content-based image retrieval using multiresolution color and texture features," *IEEE Trans. Multimedia*, vol. 10, no. 6, pp. 1073–1084, 2008.
- [104] S. Liapis and G. Tziritas, "Color and texture image retrieval using chromaticity histograms and wavelet frames," *IEEE Trans. multimedia*, vol. 6, no. 5, pp. 676–686, 2004.
- [105] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Trans. Multimedia*, vol. 1, no. 3, pp. 264–277, 1999.
- [106] J. C. Nascimento and J. S. Marques, "Performance evaluation of object detection algorithms for video surveillance," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 761–774, 2006.
- [107] Y. Gao, M. Wang, Z.-J. Zha, Q. Tian, Q. Dai, and N. Zhang, "Less is more: efficient 3-d object retrieval with query view selection," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1007–1018, 2011.

REFERENCES

- [108] C.-K. Hsieh, S.-H. Lai, and Y.-C. Chen, "Expression-invariant face recognition with constrained optical flow warping," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 600–610, 2009.
- [109] Y. Wu, "On the security of an svd-based ownership watermarking," *IEEE trans. Multimedia*, vol. 7, no. 4, pp. 624–627.
- [110] N. Doulamis, A. Doulamis, G. Votsis, N. Tsapatsoulis, and S. Kollias, "An adaptive framework for multimedia information retrieval and emotionally rich human computer interaction," *IEEE Trans. Multimedia*, 2001.
- [111] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition," *Computer Vision and Image Understanding*, vol. 101, no. 1, pp. 1–15, 2006.
- [112] S.-J. Wang, J. Yang, N. Zhang, and C.-G. Zhou, "Tensor discriminant color space for face recognition," *IEEE Trans. Image Processing*, vol. 20, no. 9, pp. 2490–2501, 2011.
- [113] Z. Liu, J. Yang, and C. Liu, "Extracting multiple features in the cid color space for face recognition," *IEEE Trans. Image Processing*, vol. 19, no. 9, pp. 2502–2509, 2010.
- [114] Z. Lu, X. Jiang, and A. C. Kot, "A color channel fusion approach for face recognition," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1839–1843, 2015.
- [115] M. T. Sadeghi, S. Khoshrou, and J. Kittler, "Confidence based gating of colour features for face authentication," in *Multiple Classifier Systems*, pp. 121–130, Springer, 2007.
- [116] V. Georges, "System of television in colors," May 15 1945. US Patent 2,375,966.
- [117] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696–706, 2002.

REFERENCES

- [118] J. Kittler and M. T. Sadeghi, “Physics-based decorrelation of image data for decision level fusion in face verification,” in *Multiple Classifier Systems*, pp. 354–363, Springer, 2004.
- [119] W. H. Buchsbaum, *Color TV servicing*. Prentice Hall, 1975.
- [120] P. Colantoni *et al.*, “Color space transformations,” *see* <http://www.radugaryazan.ru/files/doc /colorspacettransform95.pdf>, 2004.
- [121] Y. Ohta, *Knowledge-based interpretation of outdoor natural color scenes*, vol. 4. Morgan Kaufmann, 1985.
- [122] J. Lu and K. N. Plataniotis, “On conversion from color to gray-scale images for face detection,” in *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, pp. 114–119, 2009.
- [123] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, “Two-dimensional pca: a new approach to appearance-based face representation and recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 131–137, 2004.
- [124] LibRaw, “Channel noise and raw converters,” 2008.
- [125] D. Wu, J. Tian, B. Li, Y. Wang, and Y. Tang, “Recovering sensor spectral sensitivity from raw data,” *Journal of Electronic Imaging*, vol. 22, no. 2, pp. 023032–023032, 2013.
- [126] X. Jing, Q. Liu, C. Lan, J. Man, S. Li, and D. Zhang, “Holistic orthogonal analysis of discriminant transforms for color face recognition,” in *Intl. Conf. Image Processing*, pp. 3841–3844, IEEE, 2010.
- [127] C. Zhao, D. Miao, Z. Lai, C. Gao, C. Liu, and J. Yang, “Two-dimensional color uncorrelated discriminant analysis for face recognition,” *Neurocomputing*, vol. 113, pp. 251–261, 2013.

REFERENCES

- [128] Y. Ohta, T. Kanade, and T. Sakai, “Color information for region segmentation,” *Computer graphics and image processing*, vol. 13, no. 3, pp. 222–241, 1980.
- [129] P. L. Duren, *Theory of Hp spaces*, vol. 38. IMA, 1970.
- [130] H. Stokman and T. Gevers, “Selection and fusion of color models for image feature detection,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, p. 371–381, 2007.
- [131] B. V. Funt and G. D. Finlayson, “Color constant color indexing,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 522–529, 1995.
- [132] G. Snedegor, W. G. Cochran, *et al.*, “Statistical methods.,” *Statistical methods.*, no. 6th ed, 1967.
- [133] X. Jiang and J. Lai, “Sparse and dense hybrid representation via dictionary decomposition for face recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 1067–1079, 2015.
- [134] C. Liu and H. Wechsler, “Robust coding schemes for indexing and retrieval from large face databases,” *IEEE Trans. Image Processing*, vol. 9, no. 1, pp. 132–137, 2000.
- [135] X. Jiang, “Linear subspace learning-based dimensionality reduction,” *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 16–26, 2011.
- [136] X. Jiang, “Asymmetric principal component and discriminant analyses for pattern classification,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, p. 931–937, 2009.
- [137] J. Y. Choi, W. De Neve, K. N. Plataniotis, and Y. M. Ro, “Collaborative face recognition for improved face annotation in personal photo collections shared on online social networks,” *IEEE Trans. Multimedia*, vol. 13, no. 1, pp. 14–28, 2011.

REFERENCES

- [138] X. Tan and B. Triggs, “Enhanced local texture feature sets for face recognition under difficult lighting conditions,” *IEEE trans. image processing*, vol. 19, no. 6, pp. 1635–1650, 2010.
- [139] C. H. Chan, M. A. Tahir, J. Kittler, and M. Pietikäinen, “Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors,” *IEEE trans. pattern analysis and machine intelligence*, vol. 35, no. 5, p-p. 1164–1177, 2013.
- [140] C. Ding, J. Choi, D. Tao, and L. S. Davis, “Multi-directional multi-level dual-cross patterns for robust face recognition,” *IEEE trans. pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 518–531, 2016.
- [141] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, p-p. 1097–1105, 2012.
- [142] Z. Lu, X. Jiang, and A. Kot, “Feature fusion with covariance matrix regularization in face recognition,” *Signal Processing*, vol. 144, pp. 296–305, 2018.
- [143] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning,” in *European Conference on Computer Vision*, pp. 94–108, Springer, 2014.
- [144] H. Demirel and G. Anbarjafari, “Pose invariant face recognition using probability distribution functions in different color channels,” *IEEE Signal Processing Letters*, vol. 15, pp. 537–540, 2008.
- [145] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, “Regularized discriminant analysis for the small sample size problem in face recognition,” *Pattern Recognition Letters*, vol. 24, no. 16, pp. 3079–3087, 2003.

REFERENCES

- [146] X. Jiang, B. Mandal, and A. Kot, “Enhanced maximum likelihood face recognition,” *Electronics Letters*, vol. 42, no. 19, p. 1, 2006.
- [147] X. Jiang, B. Mandal, and A. Kot, “Complete discriminant evaluation and feature extraction in kernel space for face recognition,” *Machine Vision and Applications*, vol. 20, no. 1, pp. 35–46, 2009.
- [148] O. Ikeda, “Segmentation of faces in video footage using hsv color for face detection and image retrieval,” in *Intl. Conf. Image Processing*, vol. 3, pp. III–913, 2003.
- [149] G. Paschos, “Perceptually uniform color spaces for color texture analysis: an empirical evaluation,” *IEEE trans. Image Processing*, vol. 10, no. 6, pp. 932–937, 2001.
- [150] Y. K. Park, S. L. Park, and J. K. Kim, “Retinex method based on adaptive smoothing for illumination invariant face recognition,” *Signal Processing*, vol. 88, no. 8, pp. 1929–1945, 2008.
- [151] Z. Zheng, F. Yang, W. Tan, J. Jia, and J. Yang, “Gabor feature-based face recognition using supervised locality preserving projection,” *Signal Processing*, vol. 87, no. 10, p-p. 2473–2483, 2007.
- [152] T. Mandal, Q. J. Wu, and Y. Yuan, “Curvelet based face recognition via dimension reduction,” *Signal Processing*, vol. 89, no. 12, pp. 2345–2353, 2009.
- [153] B. Xiao, X. Gao, D. Tao, and X. Li, “A new approach for face recognition by sketches in photos,” *Signal Processing*, vol. 89, no. 8, pp. 1576–1588, 2009.
- [154] D. V. Jadhav and R. S. Holambe, “Radon and discrete cosine transforms based feature extraction and dimensionality reduction approach for face recognition,” *Signal Processing*, vol. 88, no. 10, pp. 2604–2609, 2008.

REFERENCES

- [155] C. Liu and H. Wechsler, “A shape-and texture-based enhanced fisher classifier for face recognition,” *IEEE trans. image processing*, vol. 10, no. 4, pp. 598–608, 2001.
- [156] X. Tan and B. Triggs, “Fusing gabor and lbp feature sets for kernel-based face recognition,” in *International Workshop on Analysis and Modeling of Faces and Gestures*, pp. 235–249, Springer, 2007.
- [157] Z. Liu and C. Liu, “Fusion of the complementary discrete cosine features in the yiq color space for face recognition,” *Computer Vision and Image Understanding*, vol. 111, no. 3, pp. 249–262, 2008.
- [158] R. O. Duda, P. E. Hart, D. G. Stork, *et al.*, *Pattern classification*, vol. 2. Wiley New York, 1973.
- [159] A. K. Jain, R. P. W. Duin, and J. Mao, “Statistical pattern recognition: A review,” *IEEE Trans. pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, 2000.
- [160] D. Pissarenko, “Eigenface-based facial recognition,” *December 1st*, 2002.
- [161] P. Zhu, L. Zhang, Q. Hu, and S. C. Shiu, “Multi-scale patch based collaborative representation for face recognition with margin distribution optimization,” in *European Conference on Computer Vision*, pp. 822–835, Springer, 2012.
- [162] Z. Lu, X. Jiang, and A. Kot, “An effective color space for face recognition,” in *IEEE Intl. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2019–2023, 2016.
- [163] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, pp. 971–987, 2002.
- [164] A. Torralba and A. A. Efros, “Unbiased look at dataset bias,” in *Computer Vision and Pattern Recognition*, pp. 1521–1528, IEEE, 2011.

REFERENCES

- [165] D. Erhan, P. Manzagol, Y. Bengio, S. Bengio, and P. Vincent, “The difficulty of training deep architectures and the effect of unsupervised pre-training,” in *Artificial Intelligence and Statistics*, pp. 153–160, 2009.
- [166] Y. Bengio, “Deep learning of representations for unsupervised and transfer learning,” in *ICML Workshop on Unsupervised and Transfer Learning*, pp. 17–36, 2012.
- [167] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in *ACM international conf. Multimedia*, pp. 675–678, 2014.
- [168] M. Mathias, R. Benenson, M. Pedersoli, and L. Van G., “Face detection without bells and whistles,” in *European Conf. Computer Vision*, pp. 720–735, Springer, 2014.
- [169] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *IEEE international conf. computer vision*, pp. 1026–1034, 2015.
- [170] E. Zhou, Z. Cao, and Q. Yin, “Naive-deep face recognition: Touching the limit of lfw benchmark or not?,” *arXiv preprint arXiv:1501.04690*, 2015.
- [171] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, “High-fidelity pose and expression normalization for face recognition in the wild,” in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [172] W. Zhang, S. Shan, X. Chen, and W. Gao, “Local gabor binary patterns based on kullback–leibler divergence for partially occluded face recognition,” *IEEE Signal Processing Letters*, vol. 14, no. 11, pp. 875–878, 2007.
- [173] C. Zhou, Z. Zhang, D. Yi, Z. Lei, and S. Z. Li, “Low-resolution face recognition via a simultaneous discriminant analysis,” in *Intl. Joint Conf. Biometrics*, pp. 1–6, IEEE, 2011.

REFERENCES

- [174] S. Siena, V. N. Boddeti, and B. V. Kumar, “Coupled marginal fisher analysis for low-resolution face recognition,” in *European Conf. Computer Vision*, pp. 240–249, Springer, 2012.
- [175] J. Shi and C. Qi, “From local geometry to global structure: Learning latent subspace for low-resolution face image recognition,” *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 554–558, 2015.
- [176] S. P. Mudunuri and S. Biswas, “Low resolution face recognition across variations in pose and illumination,” *IEEE trans. pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 1034–1040, 2016.
- [177] E. Zangeneh, M. Rahmati, and Y. Mohsenzadeh, “Low resolution face recognition using a two-branch deep convolutional neural network architecture,” *arXiv preprint arXiv:1706.06247*, 2017.
- [178] Y. Wong, C. Sanderson, S. Mau, and B. C. Lovell, “Dynamic amelioration of resolution mismatches for local feature based identity inference,” in *Intl. Conf. Pattern Recognition*, pp. 1200–1203, IEEE, 2010.
- [179] X. Wu, R. He, Z. Sun, and T. Tan, “A light cnn for deep face representation with noisy labels,” *arXiv preprint arXiv:1511.02683*, 2015.