2017

Face Image Modality Recognition and Photo-Sketch Matching

Bingjie Liu

Face Image Modality Recognition and Photo-Sketch Matching

Bingjie Liu

Thesis submitted
to the College of Engineering and Mineral Resources
at West Virginia University

in partial fulfillment of the requirements for the degree of

Master of Science in
Electrical Engineering

Guodong Guo, Ph.D., Chair
Xin Li, Ph.D.
Shuo Wang, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2017

Keywords: Face Images, Modality Recognition, Photo-sketch matching

ABSTRACT

Face Image Modality Recognition and Photo-sketch Matching

Bingjie Liu

Face is an important physical characteristic of human body, and is widely used in many crucial applications, such as video surveillance, criminal investigation, and security access system. Based on realistic demand, such as useful face images in dark environment and criminal profile, different modalities of face images appeared, e.g. three-dimensional (3D), near infrared (NIR), and thermal infrared (TIR) face images. Thus, researches with various face image modalities become a hot area. Most of them are set on knowing the modality of face images in advance, which contains a few limitations. In this thesis, we present approaches to face image modality recognition to extend the possibility of cross-modality researches as well as handle new modality-mixed face images. Furthermore, a large facial image database is assembled with five commonly used modalities such as 3D, NIR, TIR, sketch, and visible light spectrum (VIS). Based on the analysis of results, a feature descriptor based on convolutional neural network with linear kernel SVM did an optimal performance.

As we mentioned above, face images are widely used in crucial applications, and one of them is using the sketch of suspect's face, which based on the witness' description, to assist law enforcement. Since it is difficult to capture face photos of the suspect during a criminal activity, automatic retrieving photos based on the suspect's facial sketch is used for locating potential suspects. In this thesis, we perform photo-sketch matching by synthesizing the corresponding pseudo sketch from a given photo. There are three methods applied in this thesis, which are respectively based on style transfer, DualGAN, and cycle-consistent adversarial networks. Among the results of these methods, style transfer based method did a poor performance in photo-sketch matching, since it is an unsupervised one which is not purposeful in photo to sketch synthesis problem while the others need to train pointed models in synthesis stage.

Acknowledgements

First of all, I would like to thank my advisor Dr. Guodong Guo for his academic guidance, patience, and valuable advice during my master study. His passion and hard-work for research influence me a lot. With his empirical and inspirational guidance, I was able to complete this thesis.

I would like to thank Dr. Xin Li and Dr. Shuo Wang for being my committee as well as their precious time and help during the completion of my thesis. Dr. Li is always very kind to students and I have learned a lot from his classes as well as discussion after class. Dr. Wang has shared many ideas of computer vision used in health area, which gives me a new view of this area.

I would like to thank Dr. Thirimachos Bourlai, Dr. Natalia Schmid, Dr. Yanfang Ye and Dr. Katerina Goseva-Popstojanova for classes and projects' discussion, and their help during my master study.

I would like to thank Xudong Liu, Min Jiang, Na Zhang, Yufeng Yu, Wentian Zhuo, Yu Zhu, Jinge Wang and all members in computer vision lab for their help during my study.

I would like to thank the Lane Department of Computer Science and Electrical Engineering at West Virginia University for the excellent study environment.

Finally, I want to thank my parents and my family for their support and love.

Contents

List of Figures

List of Tables

Abbreviations

3D	Three **D**imensional Space
CASIA	**C**hinese **A**cademy of **S**ciences **I**nstitute of **A**utomation
CNN	**C**onvolutional **N**eural **N**etwork
CUFSF	**C**UHK **F**ace **S**ketch **F**ERET
CUHK	**C**hinese **U**niversity of **H**ong **K**ong
FERET	**F**ac**e** **Re**cognition **T**echnology
GANs	**G**enerative **A**dversarial **N**etwork**s**
HFB	**H**eterogeneous **F**ace **B**iometrics
HFR	**H**eterogeneous **F**ace **R**ecognition
HOG	**H**istogram of **O**riented **G**radients
LBP	**L**ocal **B**inary **P**atterns
LSIFs	**L**ight **S**ource **I**nvariant **F**eatures
MRF	**M**arkov **R**andom **F**ields
MrFSPS	**M**ultiple **R**epresentations-Based **F**ace **S**ketch-**P**hoto-**S**ynthesis
MWF	**M**arkov **W**eight **F**ields
NIR	**N**ear **I**nfra**r**ed

OTCBVS	**O**bject **T**racking, **C**lassification **B**eyond **V**isible **S**pectrum
PRIP-HDC	**P**attern **R**ecognition and **I**mage **P**rocessing **H**and-**D**rawn **C**omposite
RBF	**R**adial **B**asis **F**unction
RBMs	**R**estricted **B**oltzmann **M**achine**s**
SVM	**S**upport **V**ector **M**achine
TIR	**T**hermal **I**nfra**r**ed
VIS	**V**isible **L**ight **S**pectrum

Chapter 1

Introduction

1.1 Background and Motivation

In early days, people could easily identify individuals due to the sparsity of population and low mobility. Today, with a dramatic increase in world's population as well as the development of transportation leading to a high degree of mobility, identity management has become increasingly necessary. Many physical traits are widely used in this area, such as face, iris, fingerprints, and palm print. It is almost impossible for different people to have the same physical traits' information, so these body traits can provide unique information to accurately identify a person.

Unlike other physical characteristics, face images can be obtained easily and economically without the specific requirement of machine, environment, or distance [1]. Moreover, people often identify another person through their faces, and often share their own facial images with others in daily life. Therefore, it will reduce the feeling of being offended when using their face images for identification. Compared with other physical traits, human face contains other information such as gender, emotion, ethnicity, and age in addition to the identity [1]. For these reasons, face images are widely used criminal investigation, security access system, surveillance, human-computer interaction, and other key applications.

People have used smart phone or normal camera to take visible light spectrum (VIS) face images for many years. With the development of sensor technology, it is possible to acquire more different face image modalities, such as three-dimensional space (3D), near infrared

(NIR), and thermal infrared (TIR) face images. In fact, we can only get some given face image modalities in some specific cases. For example, VIS face images may not be useful in dark environment, whereas NIR and TIR human face images still provide useful information in this context. Another specific situation, which can only obtain specified face image modality, is locating potential suspects. Forensic sketches, described by witnesses or victims, may be the only way a suspect's face is imaged since capturing useful face photos during the criminal activities is nearly impossible. On the other hand, as we mentioned above, face images are widely used in today's crucial applications, so more than one modality may be required for those applications. For instance, unlocking a cell phone through the face requires to handle both natural light and dark surroundings, so different sensors and at least two modalities are used for this application.

As we know, more and more researchers have focused on areas with more than one modality of face image, such as recognizing subjects among multiple modalities. The face images in the database they used are pre-arranged based on their modalities, which facilitating the use of researchers. Sometimes, we want to verify whether the face image modality classification has human error, and a face image modality recognition method helps to check the accuracy of database arrangement again. On the other hand, most methods proposed in multiple modalities field focus only on certain modalities. To expend the possibility of using these methods, automatic recognition of the face image modalities is a basic matter.

Furthermore, the pre-arranged database cannot meet the sustainable growth of research demands [2]. Under certain conditions, most of face image modalities are acquired with the help of assistants which lack reality, and cost a lot of time as well as energy. As we mentioned earlier, volunteers can provide their raw face data on their own with the popularization of various face image modalities and the development of relative sensors.

For example, sensors in smartphones and digital cameras can acquire TIR, NIR, and 3D modalities in recent years. Based on these factors, a large number of modality-mixed images can be obtained directly from subjects. Therefore, the basic matter before subsequent research is to recognize these face image modalities. For these reasons, we present approaches for face image modality recognition in this thesis, which can provide the modality of input face image.

1.2 Related Work

Due to those realities mentioned in Section 1.1, an increasing number of researchers are focusing on areas related to various modalities of face images. In addition, there are many datasets produced for research demands.

Yi et al. [3] focused on Heterogeneous Face Recognition (HFR) between NIR and VIS face images. The datasets used in this paper are CASIA HFB database and CASIA NIR-VIS 2.0 database. They proposed a framework that combines Gabor features with a Restricted Boltzmann Machines (RBMs) to extract local representations, which is modality-free shared.

Liu et al. [4] used NIR and VIS heterogeneous database for face recognition. They extract invariance based on the Light Source Invariant Features (LSIFs). The basic idea of it is to collect a variety of band-pass filters. The database used in this paper is CASIA HFB database.

Xu et al. [5] used CASIA NIR-VIS 2.0 database for HFR. They transformed VIS face images into NIR modality as well as from NIR to VIS. A dictionary learning approach based on minimization of cross-spectral union is used to learn the mappings between NIR and VIS domains.

Bhowmik et al. [6] used IRIS Thermal/Visible Face database and Object Tracking, Classification Beyond Visible Spectrum (OTCBVS) database, and Terravic Facial IR database for their work. They used logarithmic polar transformations to handle face images under different conditions, such as different expressions, occlusions and various rotations. They used a fusion method based on weighted average of Daubechies wavelet transform (db2) coefficients to identify subjects.

Goswami et al. [7] focused on face recognition, and used 3D face images to improve the results. EURECOM Kinect Face Database and their novel database are used in this paper. Random decision forest is used as a classifier, which is based on entropy, saliency and Histogram of Oriented Gradients (HOG) of depth information feature extraction.

Galoogahi et al. [8] used CUFS database to demonstrate the performance of their face descriptor. This descriptor based on the gradient orientations can reduce the differences between sketch and VIS facial features with facial components' emphasizing coarse texture.

Klare et al. [9] focused on HFR with NIR, TIR, sketch and VIS face images. They directly calculate the similarity between modalities, and proposed a linear discriminant subspace called prototype random subspaces with the probe and gallery for recognition.

Based on these papers, most of researches have limitations because they are set on knowing the modality of face images in advance, which restricts the usage of their achievements. Due to this, we present approaches for face image modality recognition in this thesis to extend the possibilities of cross-modality research as well as deal with new modality-mixed face images.

1.3 Organization

In this section, we will briefly introduce the structure of this thesis, which is organized as

follows:

In Chapter 2, we will cover various face modalities, which are VIS, NIR, TIR, 3D and sketch. In addition, we will introduce several corresponding face image datasets with 1 to 3 face image modalities in each database. Then, a new database is assembled from the face images in those datasets. There are five face image modalities in total. After assemblage, the face images in the new database are uniformly renamed according to the logogram of face image's modality, the serial number of one subject, and the sequence number of his/her images. On the other hand, all the face images are cropped and resized to the same size, depending on the position of their eyes.

In the third chapter, we design experiments on face image modality recognition. Five feature descriptors are used in our experiments, which are Oriented Gradients (HOG), Gist, Gabor, Local Binary Patterns (LBP), and Convolutional Neural Network (CNN). The HOG descriptor can extract the shape features and the partial appearance of the face image by calculating gradient orientations' histograms. The Gist descriptor can extract global structure of a face image with low dimensional feature. Gabor descriptor captures a face image's properties with a family of Gabor kernels. LBP feature can be extracted after calculating certain small windows' histograms based on an encoded face image. High level features of a face image can be acquired by CNN descriptors. Additionally, Support Vector Machine (SVM) is used as a learning algorithm with both linear and Radial Basis Function (RBF) kernels. Then, the recognition results based on these feature descriptors and learning algorithm are compared to find the optimal method.

Chapter 4 focuses on photo-sketch matching. Due to the huge gap between VIS and sketch modalities, they cannot be directly matched. Before matching the face images of these two modalities, sketch synthesis is used to reduce the differences. Then, simple LBP descriptor and cos distance are used for matching. Three synthesizing methods are applied in this

thesis, which are style transfer, DualGAN, and cycle-consistent adversarial networks. All these methods provide end-to-end translation.

In the end, a conclusion of this thesis and future work will be presented in Chapter 5.

Chapter 2

Face Image Modalities

2.1 Introduction

As technology evolves, we can record our face images in a variety of methods, and the images can be presented in multiple modalities, such as three-dimensional space (3D), near infrared (NIR), visible light spectrum (VIS), thermal infrared (TIR), and sketch modalities. In addition, some of these modalities are acquired through specific sensors while sketch modality is generated manually. In this section, the basic idea of these five face image modalities will be mentioned below.

The portion of the human eyes' available electromagnetic spectrum is called visible light spectrum, and the wavelengths of it is about 400 nanometers to 700 nanometers [10]. VIS human face images capture the reflections of the human face under visible light, and they are often affected by illumination and pose variation. However, VIS facial images are the easiest modality people can acquire, thus, it has been widely used in our daily lives for many years.

Still exploring in electromagnetic spectrum, another region with wavelengths from about 750 nanometers to 1 millimeter is called infrared [11]. Infrared is an optimal choice for night vision. The wavelengths of infrared can be divided into many portions with different names. Near infrared is one part of infrared, and its wavelengths is about 750 nanometers to 1.4 micrometers, which is close to visible light in electromagnetic spectrum. On the other hand, thermal infrared is a portion with longer wavelengths, and it is close to microwave portion of electromagnetic spectrum [11]. This light can be emitted by any object whose

temperature is higher than absolute zero (-273.15 degrees centigrade). NIR and TIR are embranchment of infrared, and the face images in these two modalities can still provide useful information in dark environment.

A 3D image of face can represent its three-dimensional shape. 3D face images can be acquired based on a 2D intensity image with special sensors. Moreover, 3D face data is commonly rendered as shaded models, range images, and wire-frame meshes [12]. Face images in this modality can describe the facial properties while ignoring various illuminations and viewpoints [13].

Face sketches are the earliest modality among those modalities mentioned above. It can graphically show characteristics of human face. The face images in this modality are typically generated by an artist based on what he or she sees or others' description. Since they are created subjectively, the face sketch of one subject may be various based on different artists, and contains some shape exaggeration.

2.2 Database Collection

In this thesis, the more modalities of face images, the better. However, most databases contain only 2 or 3 modalities. Therefore, we aggregated many datasets into a new database which has 5 modalities of face images. These datasets are presented in this section.

2.1.1 CASIA HFB Database

This database is collected for heterogeneous face biometrics (HFB). It is consisted of 3D (acquired by Minolta vivid 910 laser scanner), NIR (850 nanometers wavelength), and VIS face images. These face images are captured with a flat background, as shown in Figure 2.1. This database is collected by Chinese Academy of Sciences Institute of Automation

(CASIA). There are 100 subjects in this database including 57 males and 43 females. Totally, 992 face images (400 VIS face images, 400 NIR face images, and 192 3D face images) are acquired in this database, and each subject has all those three kinds of face modalities. In addition, each subject has 4 VIS, 4 NIR, and 1 (or 2) 3D facial images [14, 15].

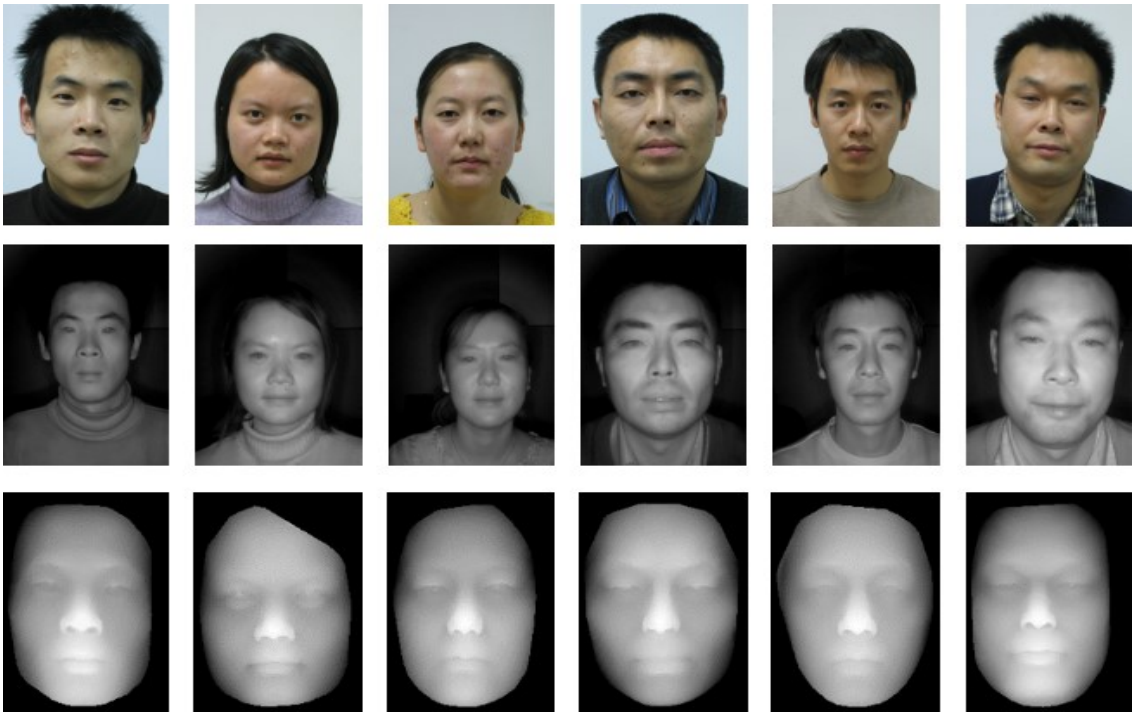

Figure 2. 1: Sample images of CASIA HFB Database. Each column contains three face image modalities of one subject. Images in the bottom row are 3D face images, images in the middle row are NIR face images, and the top row displays VIS face images.

2.1.2 CASIA NIR-VIS 2.0 Database

This database is built by the same group and the same devices as CASIA HFB database, but it only contains NIR (850 nanometers wavelength), and VIS facial images, as shown in Figure 2.2, where some of them are not captured in a flat environment. The subjects of this

database are increased to 725, and the age distribution of these subjects are bordered from young to old. Moreover, some subjects are captured more than one time since this database is collected with four sessions during several years. On average, each subject has 7 VIS and 17 NIR face images which means there are 17,580 face images (5,093 VIS face images, and 12,487 NIR face images) in total in this database [16].

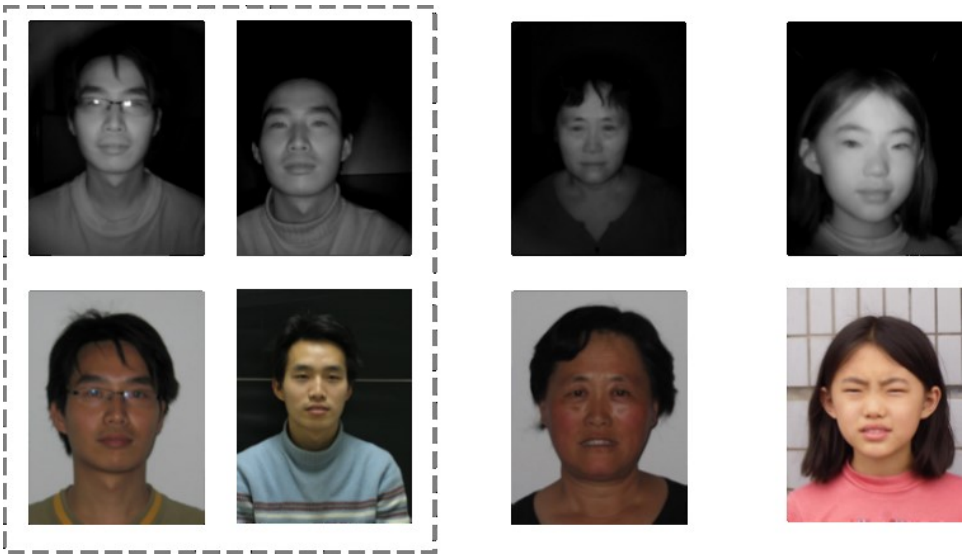

Figure 2. 2: Sample images of CASIA NIR-VIS 2.0 Database. Images in the top row are NIR face images, and images in bottom row are VIS face images. The images in the grey dotted box are the same subject's face images captured in different sessions, the other two columns are NIR-VIS-image pairs for two subjects per column.

2.1.3 Terravic Facial IR Database

This database only contains thermal images (with long-wavelength from 8-15 micrometers) of human face with 20 male subjects. Besides, there is a large number of TIR facial images for each subject. These TIR images are captured under different conditions, such as different rotations, wear/not wear glasses, and wear/not wear hat, as shown in Figure 2.3. In total, there are 22,767 TIR face images in this database [17].

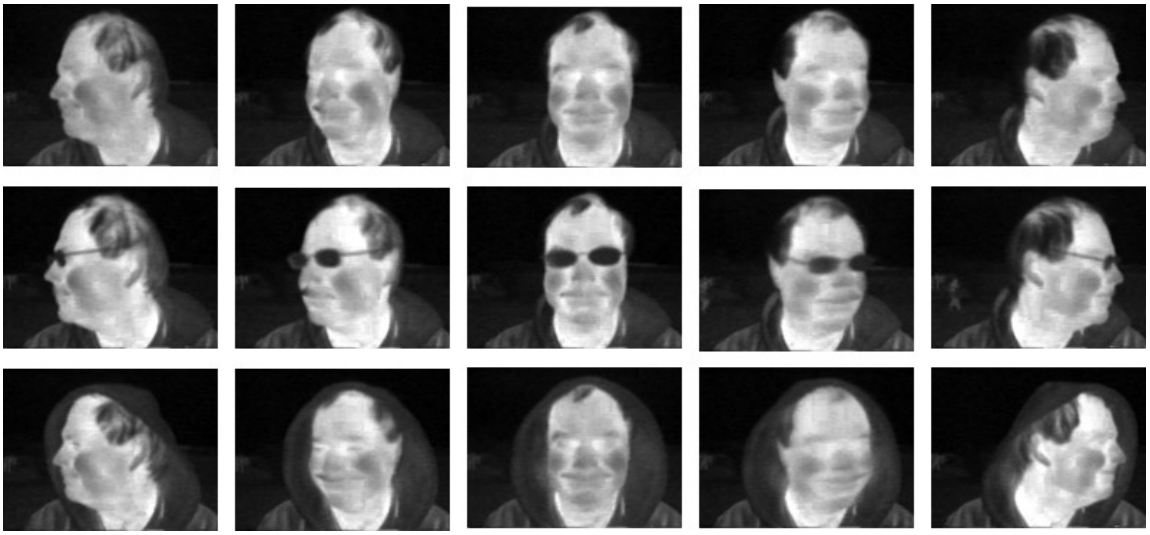

Figure 2. 3: Sample images of Terravic Facial IR Database. The top row images are captured only with different rotations. The middle row images are captured with various rotations and glasses wearing. The bottom row images are captured with rotation change and hat wearing.

2.1.4 IRIS Thermal/Visible Face Database

This database is another TIR face image (with long-wavelength from 8-15 micrometers) database we used in this thesis. There are 30 subjects in this database, and each subject has 88 to 125 VIS and TIR facial image pairs. They are captured with different expressions (e.g., surprised, happy, and angry), different rotations, and different illuminations (e.g., neutral, with left/right light, with both left and right lights, dark environment). In total, this database has 5,768 face images (2,893 VIS face images and 2,875 TIR face images) [18]. Some images are shown in Figure 2.4.

Figure 2. 4: Sample images of IRIS Thermal/Visible Face Database. The left three columns are TIR face images while the right three columns are VIS face images. From top to bottom row, there are different shooting conditions, which are surprised, happy, angry, left and right lights on, dark, left light on, and right light on.

2.1.5 EURECOM Kinect Face Database

This database is collected by Min et al. [19], composed by 3D (acquired by Kinect) and VIS face images. There are 52 subjects of which 14 subjects are female, and 38 subjects are male. The face images are captured with different facial expressions (e.g., expressionless, smile, open mouth), different illuminations (e.g., strong, neutral), different

poses (e.g., front, left profile, and right profile), with/without sunglasses, and different occlusions, as shown in Figure 2.5. Due to two collecting sessions, subjects are captured in the same postures twice in total. There are 18 VIS and 18 3D face images per subject, and 1,872 face images (936 VIS face images and 936 3D face images) in total.

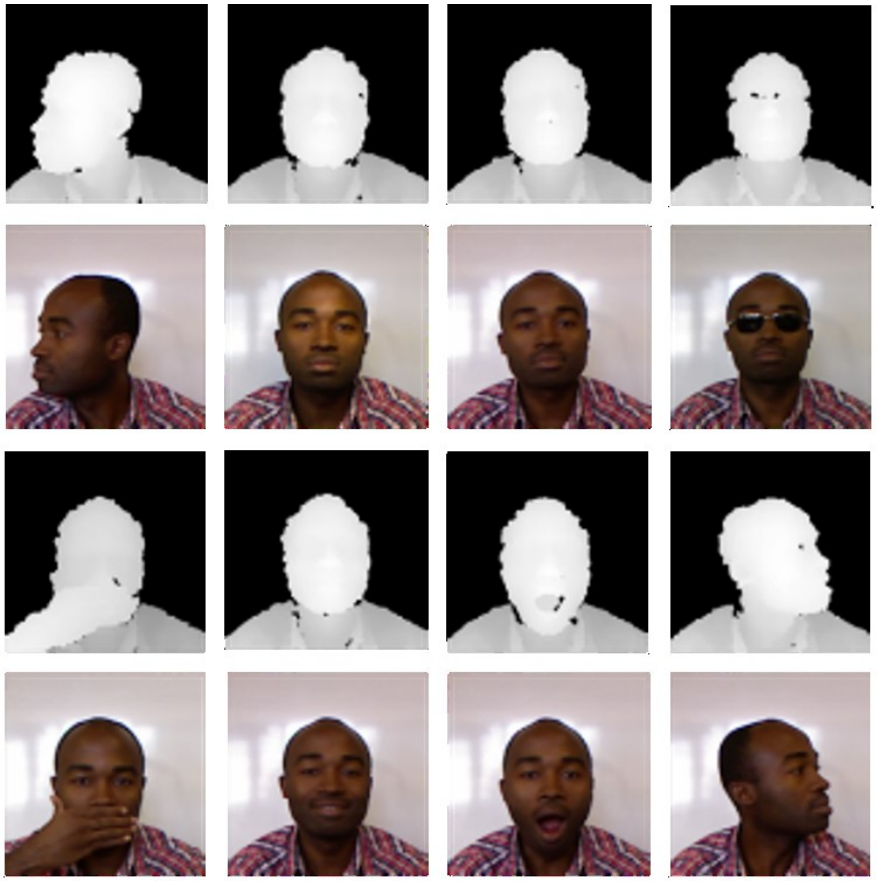

Figure 2. 5: Sample images of EURECOM Kinect Face Dataset. From top, the first and third rows' images are 3D face images, and the second and forth rows' images are VIS face images. Each VIS face images and their corresponding 3D face images have different shooting conditions, which are left profile, light on, neutral, glasses wearing, occlusion, smile, open mouth, and right profile.

2.1.6 CUFS Database

Chinese University of Hong Kong (CUHK) collected CUHK Face Sketch database [20], and this database is composed by CUHK student data set, AR database [21], and XM2VTS database [22]. There are 188 subjects collected from the first database, 123 subjects from the second one, and 295 subjects from the last one. Furthermore, these subjects are ranging from young to old, and from female to male. Based on the front VIS face image, all 606 subjects have a face sketch drawn by an artist. There are 794 face images (188 VIS face images and 606 face sketches) collected in this database, and some sample images shown in Figure 2.6.

Figure 2. 6: Sample images of CUFS Database. From left side, images in the first and third columns are VIS images, and images in the second and forth columns are sketches which are drawn from corresponding VIS face images by an artist.

2.1.7 PRIP-HDC database

The Pattern Recognition and Image Processing Hand-Drawn Composite (PRIP-HDC)

database is a Sketch-VIS face image dataset. The face sketches of 265 subjects are hand-drawn, and these sketches are different from CHHK sketch database since they are based on eyewitness or victim's descriptions rather than drawing from a clear, front pose facial image. Moreover, this database is not able to release all images due to the intellectual property (IP) issue. Hence, we totally collected 47 pairs of Sketch-VIS facial images [23].

Figure 2. 7: Sample images of PRIP-HDC database. For each column, there is one subject's VIS-Sketch facial image pair. The sketches in the top row are drawn by eyewitness or victim's description. The VIS images in the bottom row are the real facial images of subjects.

2.1.8 CUHK Face Sketch FERET (CUFSF) Database

The face sketches in this database are hand-drawn with little shape exaggeration [24]. The researchers at CHUK found an artist to draw face sketches based on the front facial photos of Face Recognition Technology (FERET) database [25, 26], which has 1,194 subjects in total. Each subject in this database contains one VIS face image and its corresponding face sketch, thus, there are 2,388 face images in this database. Some samples of this database are shown in Figure 2.8.

Figure 2. 8: Sample images of CUFSF Database. Images in the top row are sketches drawn from the VIS facial images in the bottom row. There is one subject's VIS-Sketch facial image pair per column.

2.1.9 Summary

A new database is merged based on reorganizing the face images of those datasets mentioned above, and there are five face image modalities in total. Since eight datasets are collected, and some of them are created by the same group, there are some subjects that are repeated in different database with different naming conventions. For example, CASIA HFB Database and CASIA NIR-VIS 2.0 Database have repeated subjects with different nomenclature. In order to solve this problem, we manually conduct the face images of repeated subjects, to avoid the case that one subject's face images in different datasets are considered as different person's face images.

After dealing with the repeated subjects' face images, 51,455 face images are used in this thesis totally. The detailed information of this new database is shown below: 10,351 of them are VIS face images; TIR modality contains 25,642 face images; 12,487 are NIR face images; 1,847 are the number of face sketches; and the remaining 1,128 images are 3D face images.

For convenient management, all the face images in this database are renamed. The

logogram of facial image's modality, the serial number of one subject, and the sequence number of his/her images formed the naming convention, then arranging them into our new database. Furthermore, all the face images are cropped and resized to the same size (320*256) depending on the location of their eyes.

Chapter 3

Modality Recognition

As we mentioned before, face image modality recognition helps to recheck the arrangement of database, expend the usage of proposed methods focusing on specific face image modalities, and process modality-mixed data. Experiments on face image modality recognition is designed based on the reorganized database containing five face image modalities and over 50 thousand face images. The detailed information of the experiments and analysis of the results will be introduced in this chapter.

3.1 Methodology

To find an optimal recognition method, different feature descriptors are used in this thesis, which are Histogram of Oriented Gradients (HOG), GIST, Gabor, Local Binary Patterns (LBP), and Convolutional Neural Network (CNN). Shape features and local appearance of a face image can be extracted by HOG descriptor with calculating gradient orientations' histograms. Gist descriptor can extract global facial structure with a low dimension. Gabor descriptor captures facial properties with a series of Gabor kernels, and LBP feature can be generated after calculating certain small windows' histograms of an encoded face image. Gabor and LBP descriptors can extract texture features of a face image. CNN descriptor is widely used for feature extraction, and high-level features of a face image can be obtained based on it. For learning algorithm, Support Vector Machine (SVM) is used in this thesis. In this section, all these five features' extraction methods, and recognition methods will be discussed in detail.

3.1.1 Histogram of Oriented Gradients

Histogram of Oriented Gradients (HOG) [27, 28] is a descriptor which extracts local appearance and shape features from images. The basic idea of this feature is that dividing an image into cells, and calculating the histogram of gradient orientations within those cells. After normalizing cell histograms in each block, and the HOG feature vector is the combined vector of those histograms from all blocks [29]. Based on Dalal et al. [28], HOG descriptor extract features from an image, which is divided into 8*8-pixel cells, as follows:

1) A simple 1-D [-1, 0, 1] gradient filter with no smoothing is used to compute the gradient. Assume that the pixel at (x, y) needed to be computed has luminance value $\mathrm{I}(\mathrm{x}, \mathrm{y})$. Its horizontal gradient $G_x(\mathrm{x}, \mathrm{y})$ and vertical gradient $G_y(\mathrm{x}, \mathrm{y})$ are computed as

$$G_x(\mathrm{x}, \mathrm{y}) = \mathrm{I}(\mathrm{x}+1, \mathrm{y}) - \mathrm{I}(\mathrm{x}-1, \mathrm{y}) \tag{3.1}$$

$$G_y(\mathrm{x}, \mathrm{y}) = \mathrm{I}(\mathrm{x}, \mathrm{y}+1) - \mathrm{I}(\mathrm{x}, \mathrm{y}-1) \tag{3.2}$$

2) Compute the gradient's magnitude $\mathrm{m}(\mathrm{x}, \mathrm{y})$ and direction $\theta(\mathrm{x}, \mathrm{y})$.

$$\mathrm{m}(\mathrm{x}, \mathrm{y}) = \sqrt{G_x(\mathrm{x}, \mathrm{y})^2 + G_y(\mathrm{x}, \mathrm{y})^2} \tag{3.3}$$

$$\theta(\mathrm{x}, \mathrm{y}) = \tan^{-1} \frac{G_y(\mathrm{x}, \mathrm{y})}{G_x(\mathrm{x}, \mathrm{y})} \tag{3.4}$$

3) Compute an orientation histogram based on each pixel within the cell. Since there are 9 orientation bins evenly spaced over $0^{\mathrm{o}} - 180^{\mathrm{o}}$, accumulating the magnitude into an orientation bin which its direction belongs to.

4) Four connected cells combined into one block, then, normalize the histograms within it. The result of normalization can be computed as

$$v_i^n = \frac{v_i}{\sqrt{||v||_2^2 + \varepsilon^2}} \quad (3.5)$$

where v_i is unnormalized descriptor vector, i is a number from 1 to 36 (4 cells* 9 bins), $||v||_2^2 = v_1^2 + v_2^2 + \cdots + v_{36}^2$, and ε is a small constant avoiding divisor is 0 [30].

In this thesis, all face images are extracted HOG features based on those calculating steps mentioned above. Figure 3.1 shows an example of a VIS face image with its relative histogram of oriented gradients representation image.

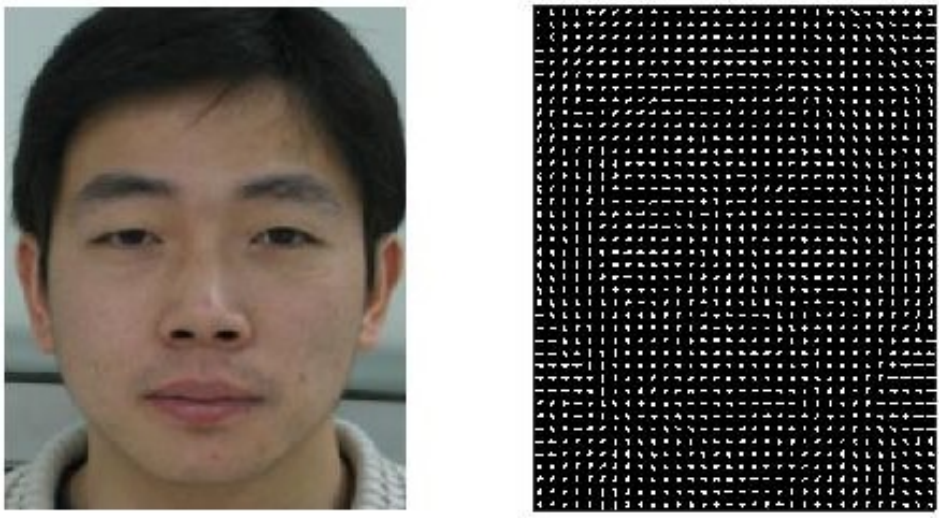

Figure 3. 1: An example of HOG descriptors based on a VIS face image with 8*8 pixels' cells, 2*2 cells' block, a simple 1-D [-1, 0, 1] gradient filter, and 9 orientation bins evenly spaced over 0 to 180 degrees. The left one is a VIS face image, and right one shows the orientation of the gradients.

3.1.2 Gabor

Gabor wavelet captures the properties of a facial image, such as spatial frequency, spatial localization, orientation, and quadrature phase relationship [31]. A convolution of an image with a family of Gabor kernels is called Gabor wavelet transformation. Then, a feature vector is derived after gathering all outputs of a facial image, which contains different

orientation, local, scale features [32]. The Gabor kernels can be computed as

$$\omega_{u,v}(\mathrm{p}) = \frac{\left||k_{u,v}|\right|^2}{\sigma^2} e^{\frac{-\left||k_{u,v}|\right|^2 ||p||^2}{2\sigma^2}} (e^{ik_{u,v}z} - e^{\frac{-\sigma^2}{2}}) \tag{3.6}$$

$$k_{u,v} = \frac{k_{max}}{f^v} e^{\frac{i\pi u}{8}} \tag{3.7}$$

where $u \in \{0, \dots, 4\}$ and $v \in \{0, \dots, 7\}$ are the scales and orientations of the Gabor kernels (as shown in Figure 3.2), $p = (x, y)$, k_{max} is the maximum frequency, f is the spacing factor between kernels in the frequency domain, and $\left||\cdot|\right|$ means norm operator [33].

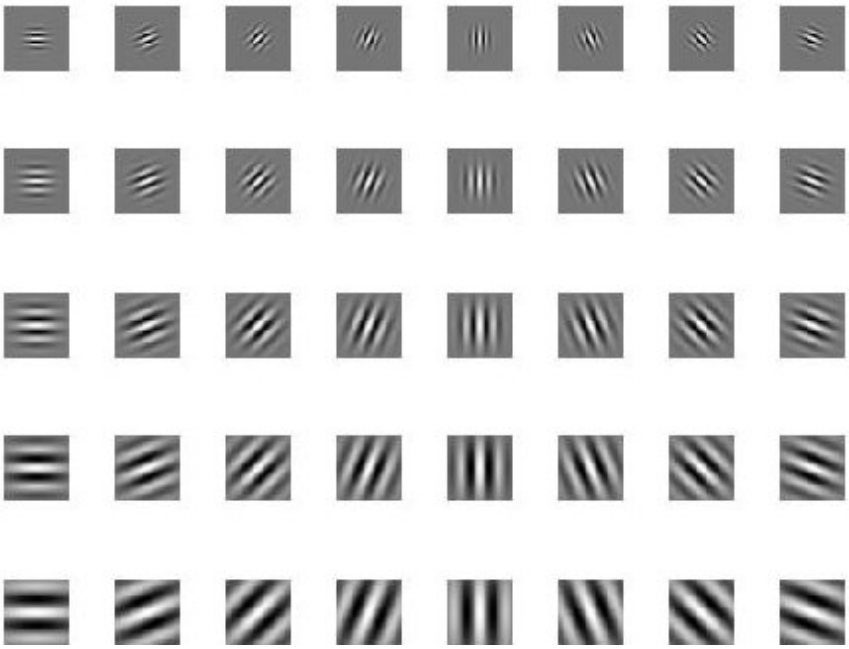

Figure 3. 2: The real parts of Gabor filters.

The results $O_{u,v}(p)$ of convoluting the image and Gabor kernels are downsampled by a factor λ, and normalized to 0 mean and unit variance before concatenated into a Gabor feature vector $g^{(\lambda)}$, which are defined as

$$O_{u,v}(p) = I(p) * \omega_{u,v}(\mathrm{p}) \tag{3.8}$$

$$g^{(\lambda)} = (O_{0,0}^{(\lambda)^t}\ O_{0,1}^{(\lambda)^t}\ \dots\ O_{4,7}^{(\lambda)^t})^t \tag{3.9}$$

where $I(p)$ is the gray level distribution of the image, $*$ is convolution operator, $O_{u,v}^{(\lambda)}$ is the down-sampled and normalized vector from $O_{u,v}(p)$, and t is the transpose operator.

In our experiments, Gabor feature vectors are acquired based on these extraction processes, and a magnitude Gabor representation of a VIS face image is shown in Figure 3.3.

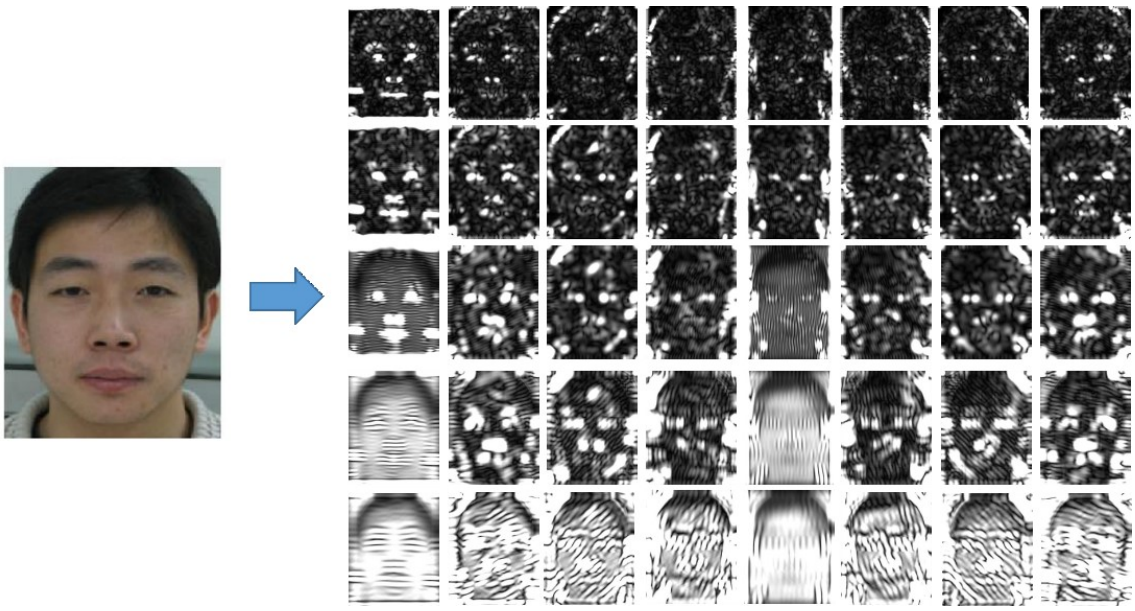

Figure 3. 3: Magnitude Gabor representation of a VIS face image, with 8 orientations and 5 scales of the Gabor kernels.

3.1.3 GIST

Gist feature provides a global description of images, such as naturalness, openness, roughness, expansion, and ruggedness. In A. Oliva et al. [34], spatial envelope, a group of perceptual properties as mentioned before, represents the structure of an image in a low dimension. For a facial image, spatial envelope provides an overall facial description without detailed objects', such as mouth, nose, eyes, etc., information. Based on [35, 36],

a Gist descriptor used in this thesis can be computed as follows:

1) Using 32 Gabor filters, which contains 4 scales and 8 orientations, convolve with the image in Fourier domain, and the outputs were 32 feature maps as the same size as original image.

2) Divide those feature maps into 4*4 regions. For each region, calculate the average of feature values within it after switching back to time domain.

3) Combine all the 32 feature maps' average values into a Gist vector, and the size of the outputs is $(4 * 4) * 32 = 512$.

3.1.4 Local Binary Patterns

Local Binary Patterns (LBP) is a simple and powerful local structure descriptor by comparing each pixel with its neighbors, and perform well at monotonic illumination changing [37]. With developing of LBP, it is widely used in texture analysis, facial image analysis, environment modeling, and so on.

The original LBP operator compares each pixel in a decimal-valued gray level image with its 8 neighbors in 3*3 region, and 0 represents neighbors with smaller value while other neighbors are encoded with 1. Then, gather the comparison results of eight neighbors' around the central pixel into a binary number in clockwise order, which starts from the top-left neighbor, and the binary number can label the given pixel after being translated into a decimal number [38]. Figure 3.4 shows an example of this encoding method.

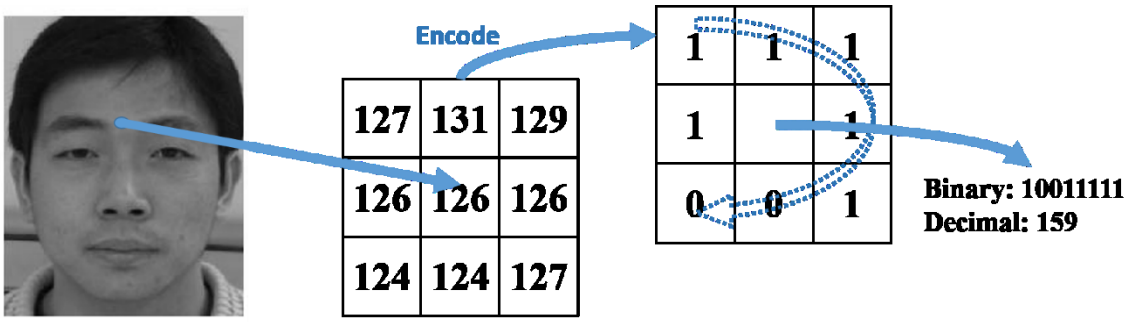

Figure 3. 4: An example of original LBP encoding. Select a pixel as the center, then compare the intensity values of its connected 8 neighbors with its value. A binary number can be acquired from top-left neighbors in clockwise order. The selected can be represented by the decimal value translated from the binary number.

At first, LBP operator cannot handle texture at vary scales since it has a fixed 3*3 region. An extension with different size neighborhoods is proposed by T. Ojala et al. [39] to deal with this problem. In addition, since some specific patterns have more information [39], they defined a binary LBP code with at most two bitwise transitions between 0 and 1 as a uniform LBP pattern. For example, 00010000 is a uniform pattern while 0001001 is not. The uniform pattern LBP can be computed as

$$LBP_{P,R}^{U^2}(x,y) = \sum_{P=0}^{P-1} s(g_P - g_c)\, 2^P \tag{3.10}$$

$$s(x) = \begin{cases} 1, & if\ x \geq 0 \\ 0, & if\ x < 0 \end{cases} \tag{3.11}$$

where (P,R) denote P sampling neighbor points on a circle of radius R, g_c is the gray-level value of point (x,y), g_P are the gray-level values of P neighbors.

After encoding all face image pixels, the face images are divided into several small size windows. Then, compute the uniform pattern LBP's histogram in each window. There are 256 patterns with 8 neighbors, and 58 of them are uniform patterns while other patterns are considered as one kind of pattern. Thus, there are 59 dimensions for one window's histogram sequence. The uniform pattern LBP feature vector is concatenated by all

windows' histogram sequences. For face image modality recognition, $LBP_{8,1}^{U^2}(x, y)$ and 64*64 non-overlapping windows are used. Figure 3.5 shows an example of this feature descriptor.

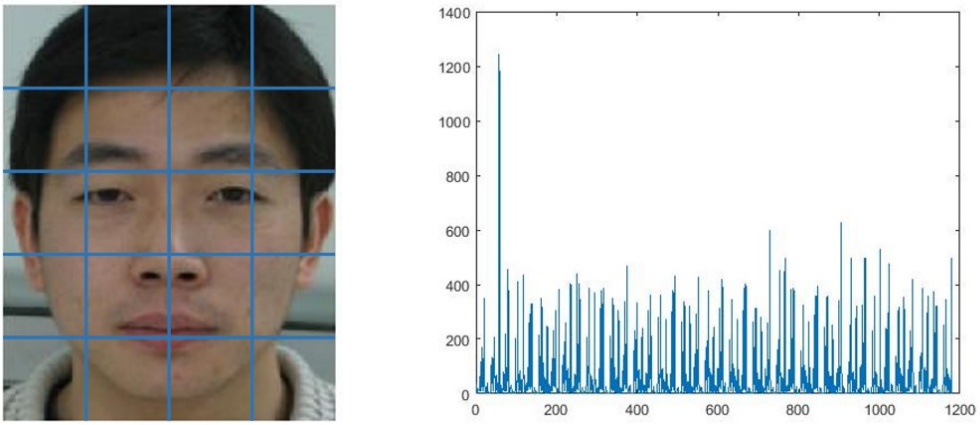

Figure 3. 5: An example of uniform pattern LBP descriptor of a VIS face image based on 64*64 non-overlapping windows, and 8 neighbors within one-pixel distance around central pixel. The left image shows the distribution of those small size windows, and the right plot is the histogram of the whole face image.

3.1.5 Convolutional Neural Network

Convolutional Neural Network (CNN) is inspired by [40], and is widely used in pattern recognition problems with continuous implement. With invariance models for certain inputs transformation, invariance properties are built into neural networks, which is the basic idea of CNN [41]. With continuous research on CNN, there are many pre-trained models, which using huge database for training. In this thesis, we used a pre-trained CNN descriptor, VGG-face [68], to extract high level features from face image. VGG- face CNN descriptor is trained with over two million face images and VGG-Very-Deep-16 CNN architecture. The network configuration contains 37 layers, and the convolution layer at

33^{rd} layer with 4096 dimensions has more information than the final one with 2662 dimensions [69]. Thus, 33^{rd} layer with 4096 dimensions is used for feature extraction.

3.1.6 Support Vector Machine

Support Vector Machine (SVM), a machine learning algorithm, is used for classification, distribution estimation, and regression [42, 43]. For basic two-category classification task, SVM constructs a hyperplane which can separate the data with two categories in a high dimensional space [44]. The proper hyperplane should have the best generalization capacity, and be in the middle of the maximum margin, in other words, the distances between the optimal hyperplane and the nearest points around it should be equal. According to [45, 46] the problem solution can be formulated as

$$\min_{W,b,\varepsilon} \quad \frac{1}{2}W^T W + C\sum_{i=1}^{l}\varepsilon_i\,,\ \mathrm{C} > 0 \tag{3.12}$$

$$\text{subject to} \quad y_i(W^T\phi(x_i) + b) \geq 1 - \varepsilon_i,\ \varepsilon_i \geq 0 \tag{3.13}$$

where (x_i, y_i) are l instance pairs, $x_i \in R^n$, $y_i \in \{1, -1\}$, $\phi(x_i)$ is the result of mapping x_i into high dimensional space, C is error term's penalty parameter, ε is the error term, (W, b) are the parameters of decision function.

In the dual space, function $W^T\phi(x_i) + \mathrm{b}$ is represented as $\sum_{i=1}^{l} a_i y_i K(x_i, x) + b$, where $K(x_i, x) = \phi(x_i)^T\phi(x)$ is called kernel function. There are four basic kernels as follows:

1) Linear: $K(x_i, x) = x_i^T x$

2) Polynomial: $K(x_i, x) = (\gamma x_i^T x + r)^d, \gamma > 0$

3) Radial basis function (RBF): $K(x_i, x) = \exp(-\gamma\|x_i - x\|^2), \gamma > 0$

4) Sigmoid: $K(x_i, x) = \tanh(\gamma x_i^T x + \mathrm{r})$

where γ, r, and d are kernel parameters.

In our experiments, a linear kernel and a non-linear kernel, which is the widely used RBF kernel, are used for face image modality recognition. The parameters used for these kernels will be presented in the following experiments section.

3.2 Experiments

The facial images used in our experiments are from the large database we assembled in chapter 2. There are five face image modalities of face images in total, which are 3D, NIR, TIR, sketch and VIS face image modalities. In addition, over 50 thousand face images are used in the experiments, and 26,719 of them (5,489 VIS face images, 13,336 TIR face images, 6,407 NIR face images, 924 face sketches and 563 3D face images) are selected as training data while others are set for testing data. Moreover, there is no overlapping subject between training data and test data, in other words, all face images of one subject with different modalities are arranged into either training or testing set.

Then, the five feature descriptors, mentioned in section 3.1, are used for feature extraction. Since the original high-dimension feature vectors generated by those descriptors slow down the speed as well as spend large memory, we reduced the dimension of those features for efficiency [47]. Principal component analysis (PCA) [48] method is used for feature dimensionality reduction in our experiments. The lowest-dimension results with at least 99 percentage representation of the original features can be acquired after dimensionality reduction. The reduced dimensions of each feature vectors and the corresponding representation percentages are shown in Table 3.1.

Table 3. 1: The results of features' dimensionality reduction by PCA.

	Gist	HOG	CNN	Gabor	LBP
Origin dimensions	512	43524	4096	51200	1180
Reduced dimensions	150	700	200	700	150
Percentage of origin	99.66%	99.78%	99.77%	99.75%	99.79%

After dimensionality reduction, we normalized those feature vectors, and used SVM as a learning algorithm for face image modality recognition. Furthermore, cross-validation method is used when finding the optimal parameters of SVM kernels, which can avoid overfitting. The kernels used in SVM are linear kernel and RBF kernel. Based on results, we set c (penalty parameter) to 1 for all features when using linear kernel SVM while set c and g (gamma) with different values for each feature when using RBF kernel SVM.

The RBF kernel's parameter c is set as 1 while g is 0.001 in HOG-RBF based recognition. To increase efficiency, various sizes (from 35 to 700 with 35-size intervals) of feature dimension are used to find the optimal one, which has better accuracy as well as lower dimension. Thus, an accuracy plot with multiple sizes of dimension is drawn in Figure 3.6. From this figure, we found that HOG feature has better performance with RBF kernel than with linear kernel in general, otherwise, the accuracy is not always increasing with feature dimension growth. Especially, in linear kernel SVM, obvious fluctuations from 100 to 455 dimensions can be observed.

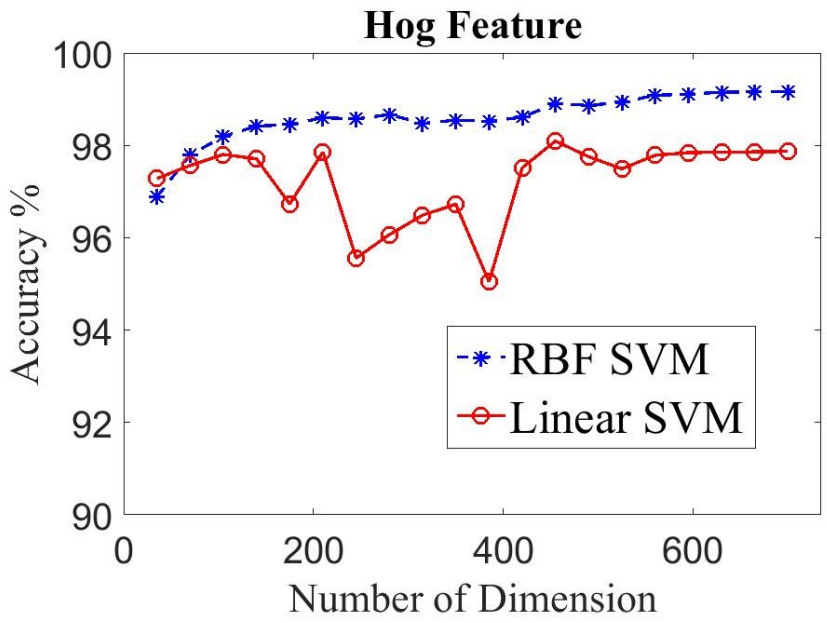

Figure 3. 6: The accuracy of HOG+SVM face image modality recognition in different number of feature dimension.

Since the proportions of each face modality are not equal (TIR is about 49%, NIR is 25%, VIS is 20%, Sketch is 4%, and 3D is 2%), for further analysis, we calculated confusion matrices of each dimension size mentioned before. To find an optimal dimension for HOG+SVM classifier without modalities' proportion influence, we calculated each dimension's average accuracy over the five face image modalities. From the results, the optimal dimensions are 455 for linear kernel with 98.43% average accuracy and 630 for RBF kernel with 99.20% accuracy, and their confusion matrices are shown in Table 3.2 and Table 3.3.

From Table 3.2, it can be found that HOG+SVM with linear kernel relatively work better on recognizing 3D and sketch face image modalities. It cannot recognize VIS modality as well as others since there are both false positive and false negative conditions among other three modalities. From Table 3.3, we can have almost same observations with the linear kernel SVM, which performs better in 3D modality with less false positive conditions.

Table 3. 2: The confusion matrix of HOG feature with linear kernel SVM.

		Predicted				
		TIR	NIR	Sketch	3D	VIS
Actual	TIR	0.9906	0	0	0.0001	0.0093
	NIR	0.0025	0.9906	0	0.0023	0.0046
	Sketch	0.0012	0	0.9963	0	0.0025
	3D	0	0	0	1	0
	VIS	0.0363	0.0118	0.0047	0.0033	0.9439

Table 3. 3: The confusion matrix of HOG feature with RBF kernel SVM.

		Predicted				
		TIR	NIR	Sketch	3D	VIS
Actual	TIR	0.9973	0	0	0	0.0027
	NIR	0.0002	0.9985	0	0.0002	0.0011
	Sketch	0.0012	0	0.9951	0	0.0037
	3D	0	0	0	1	0
	VIS	0.0218	0.0077	0.0012	0.0002	0.9691

RBF parameters of Gabor feature are set as $c = 2$ and $g = 0.002$, and we can get an accuracy plot with different dimensions (from 35 to 700 with 35 dimensions' interval) as shown in Figure 3.7. From this figure, RBF SVM also performed better than linear one in general. For RBF kernel, the accuracy decreases when dimension increasing over 350. On the other hand, the accuracy plot of Gabor descriptor with linear kernel has an uptrend in general.

Based on the highest average accuracy (98.39% for linear kernel and 99.31% for RBF kernel) of five face image modalities, the confusion matrices are calculated with 350 dimensions for RBF kernel while 630 for linear, which are shown in Table 3.5 and Table 3.4. Based on these matrices, it can be explored that Gabor with both kernels performs well

for sketch modality recognition, and linear one performs a little better due to its higher accuracy. According to the RBF kernel matrix, Gabor feature with this non-linear kernel has a good performance of NIR and 3D modalities' recognition as well. Moreover, VIS still cannot be recognized as well as other modalities since there are both false positive and false negative conditions among the rest modalities.

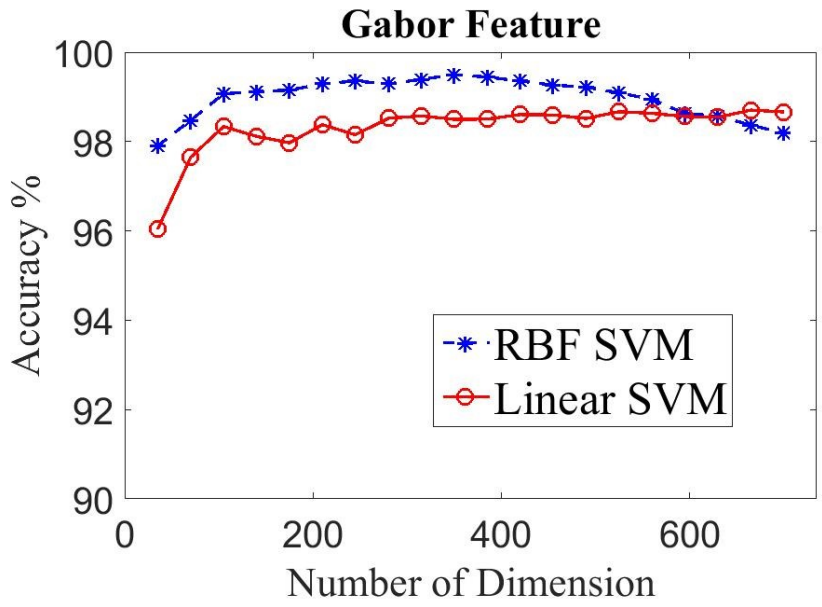

Figure 3. 7: The accuracy of Gabor+SVM face image modality recognition in different number of feature dimension.

Table 3. 4: The confusion matrix of Gabor feature with linear kernel SVM.

		Predicted				
		TIR	NIR	Sketch	3D	VIS
Actual	TIR	0.9868	0	0	0.0005	0.0127
	NIR	0.0077	0.9875	0	0.0005	0.0043
	Sketch	0	0	0.9988	0	0.0012
	3D	0.0212	0	0	0.9664	0.0124
	VIS	0.0147	0.0010	0.0029	0.0015	0.9799

Table 3. 5: The confusion matrix of Gabor feature with RBF kernel SVM.

		Predicted				
		TIR	NIR	Sketch	3D	VIS
Actual	TIR	0.9941	0	0	0	0.0059
	NIR	0.0023	0.9946	0	0	0.0031
	Sketch	0	0	0.9975	0	0.0025
	3D	0.0088	0	0	0.9805	0.0107
	VIS	0.0004	0.0006	0.0002	0.0002	0.9986

For Gist feature, we used $c = 1$ and $g = 0.005$ RBF kernel parameters. The dimensional accuracy result plot is shown in Figure 3.8. The range of dimensions used in this feature is from 10 to 150 with 10 dimensions' interval. Based on its dimensional accuracy plot, it is obvious that Gist descriptor with RBF kernel performs better than linear one in general. Moreover, the recognition accuracy generally rises along with dimension growth for both SVM kernels.

For this feature, we also calculated confusion matrices of both linear kernel (with 140 dimensions) and RBF kernel (with 130 dimensions) as shown in Table 3.6 and Table 3.7, and the average recognition accuracy of those five face image modalities respectively are 96.85% and 98.44%. From the confusion matrix tables, sketch is still a modality which can be easily recognized. In addition, Gist with linear kernel has a good recognition performance for TIR modality with little false negative and false positive conditions. For RBF kernel, there are false positive conditions among the rest four modalities while there is no false negative condition for TIR modality, based on these conditions, Gist with linear kernel is considered as a better TIR modality recognition method. Still explore in RBF confusion matrix, it can be found that NIR and 3D modalities' recognition results are better than VIS modality.

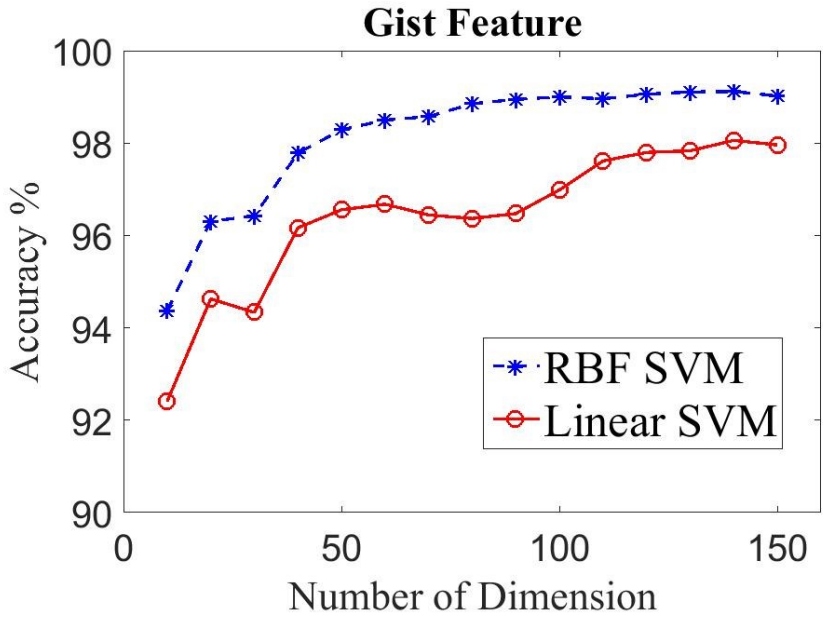

Figure 3. 8: The accuracy of Gist+SVM face image modality recognition in different number of feature dimension.

Table 3. 6: The confusion matrix of Gist feature with linear kernel SVM.

		Predicted				
		TIR	NIR	Sketch	3D	VIS
Actual	TIR	0.9996	0	0	0	0.0004
	NIR	0.0013	0.9635	0	0.0013	0.0339
	Sketch	0	0	0.9387	0	0.0613
	3D	0	0	0	0.9699	0.0301
	VIS	0.0049	0.0039	0.0170	0.0031	0.9711

Table 3. 7: The confusion matrix of Gist feature with RBF kernel SVM.

		Predicted				
		TIR	NIR	Sketch	3D	VIS
Actual	TIR	1	0	0	0	0
	NIR	0.0002	0.9880	0	0	0.0118
	Sketch	0.0012	0	0.9804	0	0.0184
	3D	0.0106	0	0	0.9735	0.0159
	VIS	0.0010	0.0019	0.0162	0.0006	0.9803

For the LBP feature, the RBF kernel's parameters are $c = 2$ and $g = 0.005$. Its accuracy plot vs. different dimensions (from 10 to 150 with 10 dimensions' interval) is shown in Figure 3.9. There is a sharp increase in range 10 to 30, and the possible reason leads to this result is that the representation percentage in this zone increased rapidly along with providing more information. Like former features' results, LBP feature with RBF kernel performs better than that with linear kernel in general. Besides, both kernels have an obvious uptrend with dimension growth.

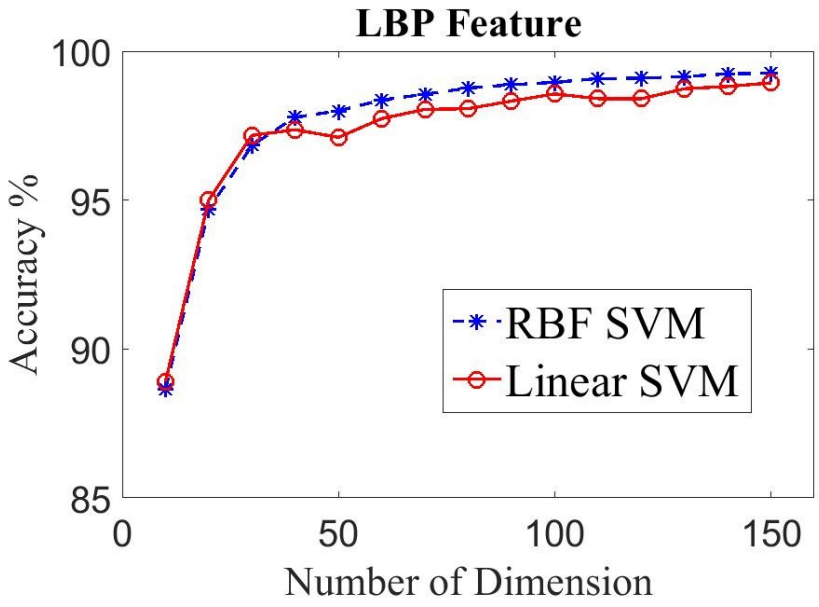

Figure 3. 9: The accuracy of LBP+SVM face image modality recognition in different number of feature dimension.

After calculating the dimensions' average recognition accuracy (98.93% is the highest accuracy for linear kernel while RBF kernel is 99.12%), size of one hundred and fifty is considered as the optimal dimension for both linear and RBF SVM. According to LBP descriptor with linear kernel confusion matrix, as shown in Table 3.8, this method performs well on 3D face image modality recognition. Table 3.9 shows the RBF kernel's confusion matrix, and it has a fine modality recognition result with TIR modality. On the other hand,

it does not perform as well as the linear one for 3D modality. For sketch modality, both linear and non-linear kernels have relatively good performance.

Table 3. 8: The confusion matrix of LBP feature with linear kernel SVM.

		Predicted				
		TIR	NIR	Sketch	3D	VIS
Actual	TIR	0.9974	0	0.0006	0	0.0020
	NIR	0.0005	0.9965	0	0.0002	0.0028
	Sketch	0.0025	0	0.9902	0	0.0073
	3D	0	0	0	1	0
	VIS	0.0258	0.0099	0.0021	0	0.9622

Table 3. 9: The confusion matrix of LBP feature with RBF kernel SVM.

		Predicted				
		TIR	NIR	Sketch	3D	VIS
Actual	TIR	0.9997	0	0	0.0003	0
	NIR	0.0002	0.9911	0	0.0020	0.0067
	Sketch	0	0	0.9927	0.0012	0.0061
	3D	0	0	0	0.9929	0.0071
	VIS	0.0099	0.0002	0.0004	0.0096	0.9799

RBF parameters $c = 1$ and $g = 0.002$ are chosen for CNN feature. The dimensional accuracy plot of this feature is shown in Figure 3.10, and the corresponding dimensions are from 10 to 200 with 10 dimensions' interval. Based on this accuracy plot, the results of both kernels are almost same when using 10 to 60 dimensions for face image modality recognition, and in this range the accuracy increased rapidly since the representation percentage of original features grew sharply. In general, CNN feature with RBF kernel performs better than that with linear kernel.

According to calculation result of average recognition accuracy, 120 dimensions (with

99.63% average recognition accuracy) is the optimal one for RBF kernel SVM while 200 with 99.58% accuracy is an optimal dimension size for linear kernel. According to their corresponding confusion matrices, as shown in Table 3.10 and Table 3.11, CNN descriptor with both linear and RBF kernels have good face image modality recognition performances for the sketch, 3D, TIR, and NIR face image modalities. On the other hand, this modality recognition method cannot handle VIS modality as well as other four modalities. According to the average accuracy, this feature with RBF kernel has a higher one than that with linear kernel. When analyzing the detailed results in confusion matrices, linear kernel has a better recognition performance for sketch, TIR and 3D modalities than RBF one. Thus, CNN descriptor with linear kernel is considered as the optimal method.

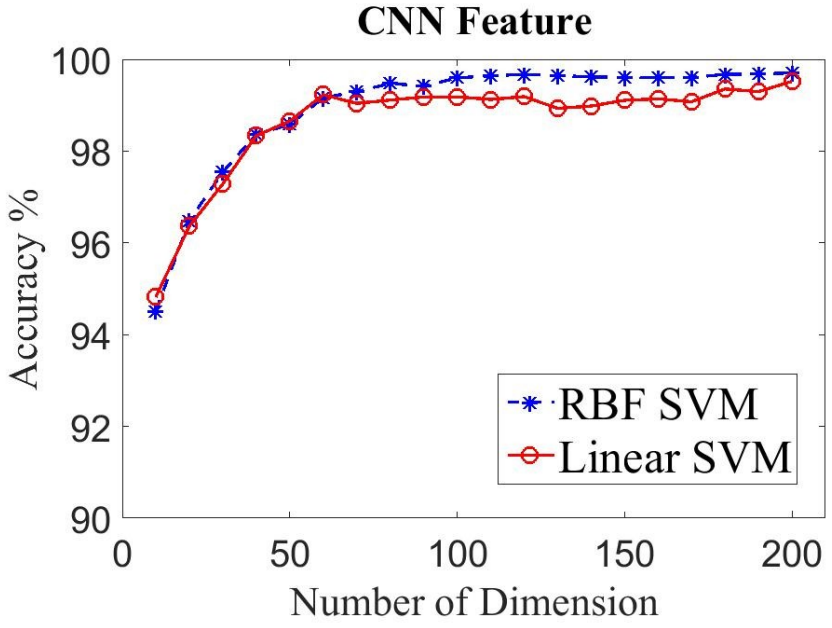

Figure 3. 10: The accuracy of CNN+SVM face image modality recognition in different number of feature dimension.

Table 3. 10: The confusion matrix of CNN feature with linear kernel SVM.

		Predicted				
		TIR	NIR	Sketch	3D	VIS
Actual	TIR	1	0	0	0	0
	NIR	0.0003	0.9951	0	0	0.0046
	Sketch	0	0	1	0	0
	3D	0	0	0	1	0
	VIS	0.0125	0.0017	0.0016	0.0002	0.9840

Table 3. 11: The confusion matrix of CNN feature with RBF kernel SVM.

		Predicted				
		TIR	NIR	Sketch	3D	VIS
Actual	TIR	0.9991	0	0	0	0.0009
	NIR	0.0007	0.9962	0	0	0.0031
	Sketch	0	0	0.9939	0	0.0061
	3D	0	0	0	0.9982	0.0018
	VIS	0.0056	0	0.0002	0	0.9942

Compared all the results we acquired, CNN descriptor with linear kernel is considered as the optimal method for face image modality recognition. Besides, sketch and 3D face image modalities have relatively huge differences with other modalities, since almost all methods used in our experiments have a good recognition performance for these two modalities, and the proportions of them are the lowest ones (sketch modality is 4% and 3d is 2%). On the other hand, all of the methods cannot perform the recognition of VIS face image modality as well as others due to some false positive and false negative conditions, which means there are some similarities between VIS modality and other modalities in local feature, texture feature and global feature. Table 3.12 shows the optimal dimensions' average accuracy of 3D, sketch, NIR, TIR and VIS modalities. Based on this table, it is explored that using RBF kernel can increase the recognition accuracy generally. However,

sometimes linear kernel performs better when analyzing the exact performance for each modality. In addition, Gist feature has the lowest accuracy with both linear kernel and RBF kernel. Due to this result, we considered that global feature is not an optimal one to solve this problem.

Table 3. 12: The optimal dimensions' average accuracy of each feature with both linear and RBF kernels.

	HOG	Gabor	Gist	LBP	CNN
Linear	98.43%	98.39%	96.85%	98.93%	99.58%
RBF	99.20%	99.31%	98.44%	99.12%	99.63%

3.3 Summary

In this chapter, five feature descriptors are used for face image modality recognition, which are HOG, Gabor, Gist, LBP, and CNN. HOG descriptor is used to extract shape features and local appearance of a face image. The basic idea of this feature descriptor is that calculating gradient orientations' histograms of local blocks of a face image. The global structure of a face image can be extracted by Gist descriptor with a low dimension will ignore detailed objects like mouth, nose, eyes. Gabor descriptor can capture facial characteristics with various scales and orientations' Gabor kernels while the basic idea of LBP feature is that calculating certain small windows' histograms of an encoded face image. Gabor and LBP descriptors can extract the texture feature of a face image. CNN feature is widely used to extract high-level facial features with invariance models for certain inputs transformation. In addition, SVM with RBF kernel and linear kernel are used as the learning algorithm in our experiments.

Then, over 50 thousand face images are used in this thesis, and they are divided into training and testing parts without overfitting. The size of training data and testing data are generally equal, while the proportions of the five face image modalities used in the

experiments, which are NIR, TIR, VIS, sketch and 3D modalities, are not the same. High dimensional features are acquired through those five feature descriptors. To reduce the large cost of memory as well as improve the speed, PCA is used for feature dimensionality reduction. After normalizing the low dimension facial features, cross-validation is used to avoid overfitting when finding the optimal parameters for SVM kernels.

Furthermore, we calculated confusion matrices for detailed modality recognition analysis. Based on the results, all of the feature descriptors have a higher recognition accuracy with RBF kernel than with linear kernel in general. Besides, almost all of the features with both linear and non-linear kernels have a good recognition performance on sketch and 3D face image modalities, even though, these two modalities have the lowest proportions among all modalities. Based on it, we believed that there are huge gaps between these two modalities and other modalities. Nevertheless, all the methods we used cannot handle VIS modality as well as other modalities, which means that the local, texture and global features of VIS face image modality have a little correlation to other four modalities'. From the experiments, all proposed approaches can get good results in face image modality recognition. According to detailed analysis, CNN feature with linear kernel SVM has the best performance for face image modality recognition, and is considered as the optimal method for this problem. Therefore, it is feasible to perform face image modality recognition with a large database.

Chapter 4

Photo-Sketch Matching

4.1 Background

In ancient times, people recorded their faces through hand-drawn ways, and face sketch is one representation of them. Sketch can graphically display structure of human face. Since it is a hand-drawn image modality, which based on what creator sees or the description given by others, face sketch contains shape exaggeration. Sketch modality is created subjectively, so the face sketch of one subject could be different based on various creators.

As we mentioned in chapter 1, face images are widely used in crucial applications. Nevertheless, we can only get some given modalities of facial images in special conditions. Sketch of suspect's face based on the witness' or victim's description is usually used for law-enforcement authority. Due to the reason that it is difficult to capture useful face photos of the suspect during a criminal activity, the only modality we can acquire is sketch. To locate the potential suspects of one crime, the forensic sketch described by eyewitness is used to automatically retrieve a large quantity of photos which may contain suspects'. Moreover, retrieved photos can assist modifying the forensic sketch interactively [49].

On the other hand, according to the analysis of face image modality recognition, there are huge differences (e.g. texture and shape) between photo and sketch modalities. Thus, they cannot be directly matched. One approach to solve this problem is to reduce the gap between these two modalities directly based on original images, and another one is to reduce the differences through photo-sketch synthesis. Additionally, photo-sketch synthesis method also can also be used in digital entertainment [50, 51, 52]. In this thesis,

we synthesized a sketch from a given photo, then matched the pseudo sketch with one drawn by an artist in database.

4.2 Related Work

Existing works of photo and sketch matching can be roughly categorized as mentioned before. One is directly using face images in two modalities, and another is synthesizing new images for recognition. The basic idea of both categorizations is reducing the differences between two modalities.

Zhang [24] and Galoogahi et al. [8] proposed descriptors which can reduce the gap between sketch and VIS modalities in feature extraction stage. Facial components with high contrast (e.g. mouth, nose, eyes, and eyebrows) have useful gradients which are more invariant than the gradients of low contrast parts (e.g. wrinkles and shadows) [8], thus, a feature descriptor based on emphasizing coarse texture is proposed to directly decrease the gap based on original face images.

In [24], a coupled information-theoretic encoding based descriptor, and mutual information maximization between photos and sketches is used to reduce the gap between sketch and VIS modalities without sketch synthesis. After extracting features through these gap-reduction descriptors, recognition results can just be generated from a normal method as photo recognition.

In [20, 49, 53-56], a pseudo sketch is synthesized before matching step, and this is another to reduce the huge difference between sketch and VIS modalities. Tang et al. [49] reduced the difference by transforming a VIS face image into sketch modality according to eigenvectors and eigenface weights.

In [20], a multiscale Markov Random Fields (MRF) model is used to jointly learn various

scales and locations' local overlapping patches for synthesizing a sketch based on them. Then, Zhou et al. [53] proposed a Markov Weight Fields (MWF) model which is superior to MRF model due to the reason that this model can handle new patches which are not included in training set.

Peng et al. [54] presented multiple representations-based face sketch-photo-synthesis (MrFSPS) method, which can adaptively learn various representations and candidate patches' weights based on Markov networks, to translate a VIS face image into sketch modality and vice versa.

Wang et al. [55] applied spatial neighboring constraint when selecting suitable neighbors, which have non-zero corresponding sparse representation coefficient, and calculating synthesis' combination coefficients. This is the basic idea of Bayesian face sketch synthesis method they proposed.

Unlike those synthesis methods that using image patches during modality translation. Zhang et al. [56] proposed end-to-end photo-sketch mapping, which is learned by convolutional network without pooling, local response normalization, etc. layers. Moreover, end-to-end means directly extract whole face features without using patches during modality transfer.

According to these correlative works, multiple synthesis methods can be divided into the one with image patches, and one directly synthesize with whole image, which often called end-to-end method.

4.3 Methodology

Most synthesis methods use a dictionary of exemplars to create pseudo sketches for matching stage. In this thesis, we transform photos into sketch modalities by three end-to-end methods like [56], which use whole input images and directly synthesis a sketch from it. These three methods are introduced in the following sections.

4.3.1 Style Transfer

Style transfer [57] separates an image into content and style parts by CNN. The basic idea of this method is combining the content of the input image and the style of the target modality image by jointly minimizing the losses of content and style reconstruction [58, 59]. The 19 layers VGG-Network [60], which uses average pooling instead of max one while not use fully connected layers, is applied for generating feature space, and each layer can be considered as a family of filters which are used for feature extraction. Furthermore, content and style reconstruction loss can be computed as

$$L_{content}(\mathrm{p},\mathrm{x},l) = 0.5\sum_{i,j}(F_{ij}^{l} - P_{ij}^{l})^2 \tag{4.1}$$

$$L_{style}(\mathrm{a},\mathrm{x}) = \sum_{l=0}^{L} w_l E_l \tag{4.2}$$

$$E_l = \frac{1}{4N_l^2 M_l^2}\sum_{ij}(G_{ij}^{l} - A_{ij}^{l})^2 \tag{4.3}$$

where p is the input image, x is generated image, a is the style image, l denotes layer, F_{ij}^{l} is the i^{th} filter's activation of x with position j in layer l, P_{ij}^{l} represents p's, A_{ij}^{l} is for a, G_{ij}^{l} is the inner product between feature map i and j in layer l, w_l is weighting factor in layer l, N_l denotes the number of filters in layer l, and M_l is the size of feature map

in layer l.

To jointly minimize the losses of content and style reconstruction, a general loss function can be calculated as

$$L_{general}(\mathrm{p},\mathrm{a},\mathrm{x}) = \alpha L_{content}(\mathrm{p},\mathrm{x}) + \beta L_{style}(\mathrm{a},\mathrm{x}) \tag{4.3}$$

where α and β are weighting factors.

4.3.2 DualGAN

DualGAN [61] is an image modality translator which is based on Generative Adversarial Networks (GANs) [62] and dual learning of natural language translation [63]. This method does not need preprocessed image pairs or huge amounts of training data. Figure 4.1 shows a flow chart of this end-to-end translation.

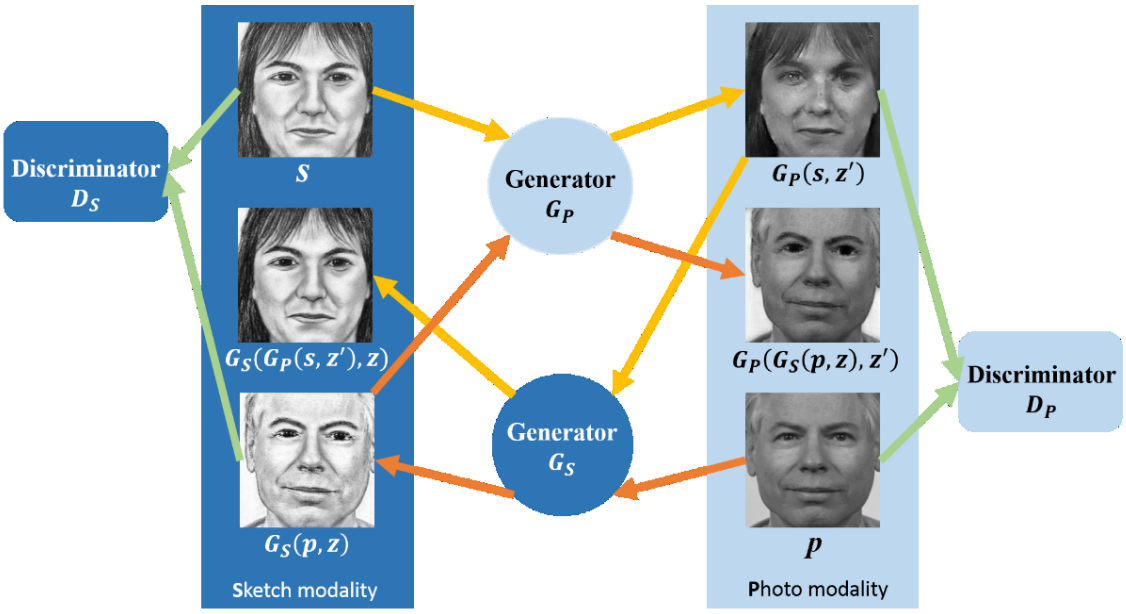

Figure 4. 1: The flow chart of DualGAN.

For photo to sketch synthesis, set photo modality as P while sketch modality as S. Since

this method contains the idea of dual translation, two generators (G_S and G_P) based on GANs are learned from both P to S and S to P. According to [64], the generators are fixed with the condition that using skip connections, which have the same number as down-sampling layers, between down-sampling layers and up-sampling layers. The reconstruction loss function for generators can be measured by L_1 or L_2 distance as follows

$$L_1^g(p,s) = \lambda_P\|p - G_P(G_S(p,z),\ z')\| + \lambda_S\|s - G_S(G_P(s,\ z'),z)\| \\ -D_S\big(G_P(s,\ z')\big) - D_P\big(G_S(p,z)\big) \tag{4.4}$$

$$L_2^g(p,s) = \lambda_P(p - G_P(G_S(p,z),\ z'))^2 + \lambda_S(s - G_S(G_P(s,\ z'),z))^2 \\ -D_S\big(G_P(s,\ z')\big) - D_P\big(G_S(p,z)\big) \tag{4.5}$$

where p denotes photo in domain P, s denotes sketches in domain S, G_S and G_P represent the generators for photo to sketch as well as sketch to photo, λ_P and λ_S are relative parameters, z and z' are random noises, and D_S and D_p are discriminators for different modality.

Moreover, two discriminators, based on Markovian PatchGAN architecture [65], are used independently for the distinction between real modality images and pseudo ones in the real modality's field. The loss functions based on [66] are calculated as

$$L_S^d(p,s) = D_S\big(G_S(p,z)\big) - D_S(s) \tag{4.6}$$

$$L_P^d(p,s) = D_P\big(G_P(s,\ z')\big) - D_P(p) \tag{4.7}$$

The aim of training step is to minimize the average of L_P^d and one of L_S^d while finding the minimum average of reconstruction losses.

4.3.3 Cycle-Consistent Adversarial Networks

Cycle-Consistent Adversarial Networks [67] aims to acquire the properties of one modality and point out how to translate an image to another modality. Like GANs, the idea of adversarial loss is used to make pseudo images indistinguishable from the real images in the real ones' modality. For cycle consistent adversarial networks, two generators and discriminators are applied as DualGAN. The adversarial losses can be computed as

$$L_{GAN}(G, D_Y, X, Y) = E_{y\sim data(y)}[(D_Y(y) - 1)^2] + E_{x\sim data(x)}\left[D_Y\left(G(x)\right)^2\right] \quad (4.8)$$

where G is a generator translating images from X to Y modality, D_Y is G's corresponding discriminator, $x\sim data(x)$ and $y\sim data(y)$ denote the distribution of images. The object of G is to minimize difference between $G(x)$ and Y while there is an adversary D_Y for distinguishing them [67].

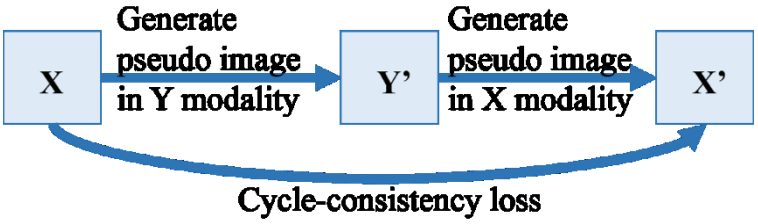

Figure 4. 2: The abridged general view of cycle-consistency loss. X denotes images in one modality, X' and Y' represent the pseudo images in corresponding modalities.

Cycle consistency losses, as shown in Figure 4.2, aim to make sure that images generated from X modality can be transferred back to original modality, which can help to synthesize desired pseudo images. The losses can be computed as

$$L_{cyc}(G, F) = E_{y\sim data(y)}[\|G(F(y)) - y\|] + E_{x\sim data(x)}[\|F(G(x)) - x\|] \quad (4.9)$$

where F is a generator translating images from Y to X modality.

In general, the aim of this method is to find the minimum cycle consistency losses which can lead to a maximum possibility that discriminators cannot distinguish the real images with pseudo ones. It can be computed as

$$\mathrm{G}^*, \mathrm{F}^* = \arg\min_{G,F} \max_{D_X, D_Y} L(G, F, D_X, D_Y) \tag{4.10}$$

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F) \tag{4.11}$$

where D_X is F's corresponding discriminator, and λ is a parameter that controls relative importance.

4.4 Experiments

The face images used in the photo-sketch matching are collected in chapter 2 with front pose while without illumination variant. Furthermore, the hand-drawn sketches do not have big shape exaggeration from the photo, and all of them are drawn from the relative front pose face photos. Totally, there are 1194 pairs of photo and face sketches used in our experiments.

When using the style transfer method, a face sketch is randomly selected. After extracting the style information from it, synthesize the pseudo sketches of the rest face photos based on the style and the content of each input image. The ratio of content and style weights used in our experiments are set as 1:30, besides, the layer weight is set as 1. Some synthesis results are shown in Figure 4.3.

Figure 4. 3: Some synthesis results based on style transfer. Each column contains three images for one subject. The top row shows VIS modality, the middle one displays the corresponding sketch drawn by an artist, and the bottom row contains the pseudo sketches generated by style transfer.

From Figure 4.3, we found that the pseudo sketches generated by style transfer method are not desirable. Some parts of the pseudo sketches are much brighter than other areas, and some contents are lost during synthesis. The most possible reason is that style and content just can be roughly separated by style transfer, and when combining the style and content, it will not generate precise sketches we want.

For DualGAN, we randomly selected 995 pairs of face photos and sketches. These pairs are used for training a generator which can translate a face photo into sketch modality. Thus, 199 pairs are used as testing data. In addition, λ_P and λ_S are set as 20. Both L_1 and L_2 distances are applied to measure reconstruction losses. Some results are shown in Figure 4.4. The pseudo face sketches based on this method seem to be closer to those drawn by an artist than style transfer in human vision, and some of face sketches used L_2 distances

have blurriness.

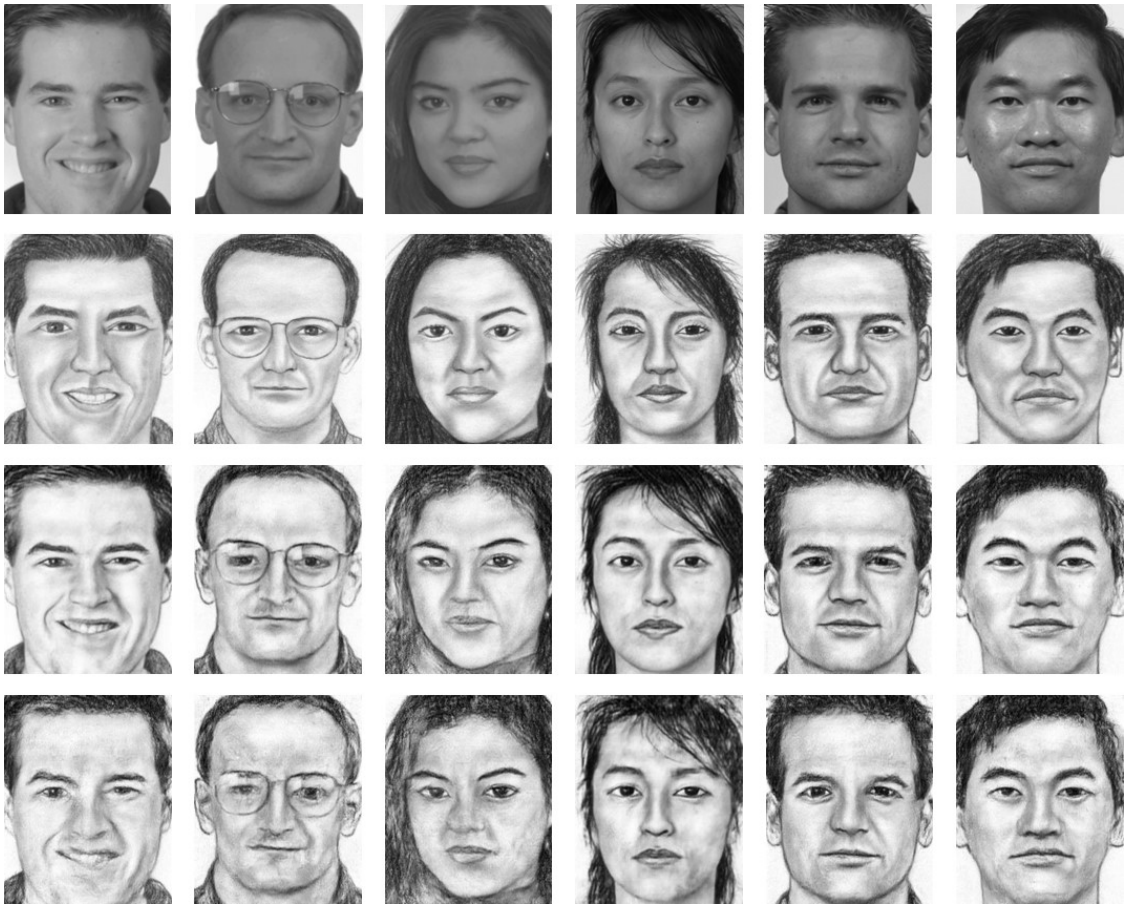

Figure 4. 4: Some pseudo sketches based on DualGAN. Each column contains four images of one subject. From top row to bottom row: face photos, sketches drawn by an artist, pseudo sketches based on L_1 distance, and pseudo sketches based on L_2 distance.

For cycle-consistent adversarial networks, we set $\lambda = 10$ for training, and used the same training set as in DualGAN. So, there are 199 testing images for this method, and some synthesis results are shown in Figure 4.5. The pseudo sketches of this method are considered as the best one among all three synthesis methods in human vision. Furthermore, these pseudo sketches contain more detailed information than former ones, for example, they keep more shadow information.

Figure 4. 5: Some pseudo sketches based on cycle-consistent adversarial networks. Each column contains four images of one subject. The top row contains face photos, the middle one displays sketches drawn by an artist, and the bottom one shows pseudo sketches synthesized by cycle-consistent adversarial networks.

After acquiring pseudo sketches, $LBP_{8,3}^{U^2}(x,y)$ and cos distance are used for matching stage. The rank-1 results are shown in Table 4.1. According to rank-1 accuracy results, all these three methods have a low value identification accuracy. Since the accuracy of matching sketch directly with VIS modality is considered as a baseline, these three synthesis methods have reduced certain differences between sketch and VIS modalities. Synthesis with style transfer has the worst performance among them since it has the worst pseudo face sketches. The reason caused this result is that pseudo sketches generated by style transfer have shape distortion, illumination variation, and content losses. Also, this method does not have pointed training stage, so it is difficult to separate content and style exactly, which lead to undesirable pseudo.

The other two methods have higher photo-sketch matching accuracy, nevertheless, the

values are not desired. The most possible reason leading to this result is that these methods cannot learn artist's shape exaggeration well, and only rank one accuracy is considered in the experiments. On the other hand, pseudo face sketches used L_1 distance just have a little higher accuracy than L_2 distance for photo-sketch matching even though pseudo face sketches with L_2 distance contain blurriness in their pseudo face sketches.

Table 4. 1: Rank-1 matching rates for style transfer, DualGAN, and cycle-consistent adversarial networks methods.

	Style transfer	DualGAN (L_1)	DualGAN (L_2)	Cycle-Consistent Adversarial Networks	Original
Accuracy	21.02%	49.75%	48.24%	60.80%	0.5%

4.5 Summary

In this chapter, we focus on photo-sketch matching. There are huge differences between sketch modality and VIS modality, so it is almost impossible to match these two modalities' face images directly. Besides, photo-sketch synthesis methods can reduce the gap between them. For synthesis methods, they can be divided into two aspects. One is synthesizing a pseudo sketch with certain patches, while another one is directly transforming the input image to another modality. In our experiments, synthesizing pseudo sketches from corresponding face photos is the basic idea of gap reduction. Furthermore, three end-to-end modality transfer methods, which are style transfer, DualGAN, and cycle-consistent adversarial networks, are used for synthesis stage. According to human vision, we find that pseudo sketches generated by style transfer are not desirable while other methods' pseudo sketches are closer to hand-drawn ones. Based on those pseudo sketch generated by the three synthesis methods, we calculate rank-1 accuracy of matching. As expected, matching based on style transfer did worst performance among these three methods. For other

methods, they have pointed training before synthesis. The photo-sketch matching accuracy results based on them are higher than the one of style transfer, but the values are not as high as we expected. Rank one identification accuracy and subjective shape exaggeration during creation stage are considered as the reason caused low values.

Chapter 5

Conclusion and Future Work

In this thesis, we presented approaches to face image modality recognition to extend the possibility of cross-modality researches as well as handle new modality-mixed face images. Furthermore, a new database is assembled from eight datasets with five face image modalities and over 50 thousand face images. In addition, some face image modality recognition methods are presented in this thesis. Five feature descriptors are used for recognition, which are HOG, Gist, Gabor, LBP, CNN. SVM with both linear kernel and RBF kernel are used as learning algorithm in the experiments. Based on results' analysis, a CNN based feature descriptor with linear kernel SVM is an optimal approach for face image modality recognition. Furthermore, all approaches have good recognition performance for sketch modality due to the huge gap between sketch modality and other modalities.

Then, we focused on photo-sketch matching by synthesizing face photos into pseudo sketches. Style transfer, DualGAN, and cycle-consistent adversarial networks are the three end-to-end methods we used in this thesis. After acquiring all pseudo sketches, uniform pattern LBP is applied to extract feature from face sketches, and cos distance is used to measure the similarity between pseudo sketches and real ones. After comparing results of these three methods, we explored that the performance of style transfer is not as good as others, and the pseudo sketches generated by this method are not desirable, which leads to a low accuracy. On the other hand, cycle-consistent adversarial networks is the optimal synthesis method among these three methods due to its highest rank-one identification accuracy.

Currently, we focus on VIS to sketch modalities' matching by three synthesis methods. For the future works, we will compare more synthesis methods for photo-sketching matching, as well as, use these synthesis methods for various modalities matching. For example, transform a VIS face image into NIR modality or TIR modality, then, perform the matching stage based on pseudo NIR or TIR face image.

References

[1] A. K. Jain, A. A. Ross, and K. Nandakumar. “Introduction to Biometrics.” Springer, 2011.

[2] J. Choi, A. Sharma, D. W. Jacobs, and L. S. Davis. “Data insufficiency in sketch versus photo face recognition.” *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2012.

[3] D. Yi, Z. Lei, and S. Z. Li. “Shared representation learning for heterogenous face recognition.” *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE. 1 (2015): 1-7.

[4] S. Liu, D. Yi, Z. Lei, and S. Z. Li. “Heterogeneous face image matching using multi-scale features.” *2012 5th IAPR International Conference on Biometrics (ICB)*. IEEE. Pages: 79-84. 2012.

[5] F. Juefei-Xu, D. K. Pal, and M. Savvides. “NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction.” *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2015.

[6] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh. “On RGB-D face recognition using Kinect.” *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE. Pages: 1-6. 2013.

[7] T. Huynh, R. Min, and J. L. Dugelay. “An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data.” *Asian Conference on Computer Vision*. Springer, Berlin, Heidelberg. 2012.

[8] H. K. Galoogahi, and T. Sim. “Inter-modality face sketch recognition.” *2012 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2012.

[9] B. F. Klare, and A. K. Jain. “Heterogeneous face recognition using kernel prototype similarities.” *IEEE transactions on pattern analysis and machine intelligence*. 35.6 (2013): 1410-1422.

[10] C. Starr. “Biology: Concepts and Applications. Thomson Brooks/Cole.” ISBN 0-534-46226-X, referenced in http://en. wikipedia. org/wiki/Visible_spectrum, 2005.

[11] J. Byrnes. “Unexploded ordnance detection and mitigation.” Springer Science &

Business Media. 2008.

[12] K. W. Bowyer, K. Chang, and P. Flynn. “A survey of 3D and multi-modal 3D+ 2D face recognition.” University of Notre Dame. 2004.

[13] G. G. Gordon. “Face recognition based on depth and curvature features.” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1992.

[14] S. Z. Li, Z. Lei, and M. Ao. “The HFB Face Database for Heterogeneous Fac e Biometrics Research”. *In 6th IEEE Workshop on Object Tracking and Classi fication Beyond and in the Visible Spectrum (OTCBVS, in conjunction with C VPR 2009)*. IEEE. Pages: 1-8. 2009. http://www.cbsr.ia.ac.cn/english/HFB%20 Databases.asp

[15] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Z. Li. “Face Matching Between Near Infrared and Visible Light Images.” *In Proceedings of IAPR/IEEE International Conference on Biometrics (ICB-2007)*. Springer. Pages:523-530. 2007.

[16] S. Z. Li, D. Yi, Z. Lei, and S. Liao. “The CASIA NIR-VIS 2.0 Face Database.” *In 9th IEEE Workshop on Perception Beyond the Visible Spectrum (PBVS, in conjunction with CVPR 2013)*. Pages: 348-353. 2013. http://www.cbsr.ia.ac.cn/english/NIR-VIS-2.0-Database.html

[17] IEEE OTCBVS WS Series Bench; Roland Miezianko, Terravic Research Infrared Database. http://vcipl-okstate.org/pbvs/bench/Data/04/download.html

[18] IEEE OTCBVS WS Series Bench; DOE University Research Program in Robotics under grant DOE-DE-FG02-86NE37968; DOD/TACOM/NAC/ARC Program under grant R01-1344-18; FAA/NSSA grant R01-1344-48/49; Office of Naval Research under grant #N000143010022. http://vcipl-okstate.org/pbvs/bench/Data/ 02/download.html

[19] R. Min, N. Kose, and J. L. Dugelay. “KinectFaceDB: A Kinect Database for Face Recognition.” *IEEE Transactions on Systems, Man, and Cybernetics: Systems.* 44.11 (2014): 1534-1548. http://rgb-d.eurecom.fr/

[20] X. Wang, and X. Tang. “Face Photo-Sketch Synthesis and Recognition.” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 31.11 (2009): 1955-1967. http://mmlab.ie.cuhk.edu.hk/archive/facesketch.html

[21] A.M. Martinez and R. Benavente. “The AR Face Database.” *CVC Technical Report.* 1998.

[22] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. "XM2VTSDB: the Extended of M2VTS Database." *Proceedings of International Conference on Audio- and Video-Based Person Authentication*. Pages: 72-77. 1999.

[23] S. Klum, H. Han, B. Klare, and A. K. Jain. "The FaceSketchID System: Matc hing Facial Composites to Mugshots." *IEEE Transaction on Information Fore nsics and Security (TIFS)*. 9.12 (2014): 2248-2263. http://www.cse.msu.edu/rgr oups/biometrics/Publications/Databases/PRIP-HDC-Release.zip

[24] W. Zhang, X. Wang, and X. Tang. "Coupled Information-Theoretic Encoding for Face Photo-Sketch Recognition." *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. Pages: 513-520. 2011. http://mmlab.ie.cuhk.edu.hk/archive/cufsf/

[25] P.J. Phillips, H. Wechsler, J. Huang, and P. Rauss. "The FERET database and evaluation procedure for face recognition algorithms." *Image and Vision Computing J.* 16.5 (1998): 295-306.

[26] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi. "The FERET evaluation methodology for face recognition algorithms." *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 22.10 (2000): 1090-1104.

[27] D. G. Lowe. "Distinctive image features from scale-invariant keypoints." *International journal of computer vision.* 60.2 (2004): 91-110.

[28] N. Dalal, and B. Triggs. "Histograms of Oriented Gradients for Human Detection." *CVPR*. Pages 886-893. 2005.

[29] B. Li, and G. Huo. "Face recognition using locality sensitive histograms of oriented gradients." *Optik-International Journal for Light and Electron Optics.* 127.6 (2016): 3489–3494.

[30] P.-Y. Chen, C.-C. Huang, C.-Y. Lien, and Y.-H. Tsai. "An efficient hardware implementation of HOG feature extraction for human detection." *IEEE Transactions on Intelligent Transportation Systems*. 15.2 (2014): 656-662.

[31] C. Liu, and H. Wechsler. "Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition." *IEEE Transactions on Image processing.* 11.4 (2002): 467- 476.

[32] C. Liu, and H. Wechsler. "A Gabor feature classifier for face recognition." *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. 2 (2001): 270-275.

[33] M. Lades, J. C. Vorbruggen, J. buhmann, J. Lange, C. von der Malsburg, R.P. Wurtz, and W. Konen. "Distortion invariant object recognition in the dynamic link architecture." *IEEE Transactions on computers*. 42.3 (1993): 300-311.

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

[35] C. Siagian, and L. Itti. "Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 29.2 (2007): 300-312.

[36] http://www.researchgate.net/publication/262917557_LargeScale_Scene_Classification_Using_Gist_Feature.

[37] T. Ojala, M. Pietikäinen, and T. Mäenpää. "A generalized Local Binary Pattern operator for multiresolution gray scale and rotation invariant texture classification." *Second International Conference on Advances in Pattern Recognition*. Springer. Pages: 397-406. 2001.

[38] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen. "Local Binary Patterns and Its Application to Facial Image Analysis: A Survey." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. 41.6 (2011): 765-781.

[39] T. Ojala, M. Pietikäinen, and T. Mäenpää. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." *IEEE Trans. Pattern Anal. Mach. Intell.* 24.7 (2002): 971–987.

[40] D. H. Hubel, T. N. Wiesel. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex." *The Journal of Physiology*. 160.1 (1962): 106–154.

[41] BINA NUSANTARA. "Convolutional Neural Network."

[42] C.-C. Chang, and C.-J. Lin. "LIBSVM: a library for support vector machines." *ACM transactions on intelligent systems and technology (TIST)*. 2.3 (2011): 27. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[43] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. “A Practical Guide to Support Vector Classification.” *Tech. Rep*. 2003.

[44] M. M. Adankon, and M. Cheriet. “Support vector machine.” *Encyclopedia of biometrics*. Springer US. Pages: 1303-1308. 2009.

[45] B. E. Boser, I. Guyon, and V. Vapnik. “A training algorithm for optimal margin classifiers.” *In Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. ACM. Pages: 144–152. 1992.

[46] C. Cortes, and V. Vapnik. “Support-vector network.” *Machine Learning.* Springer. 20.3 (1995): 273–297.

[47] Z. Cao, Q. Yin, X. Tang, and J. Sun. “Face recognition with learning-based descriptor.” *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE. Pages: 2707-2714. 2010.

[48] I.T. Jolliffe. “Principal Component Analysis and Factor Analysis.” *Principal component analysis*. Springer New York. Pages:115-128. 1986.

[49] X. Tang, and X. Wang. “Face sketch recognition.” *IEEE Transactions on Circuits and Systems for Video Technology.* 14.1 (2004): 50-57.

[50] S. Iwashita, Y. Takeda, and T. Onisawa. “Expressive facial caricature drawing.” *Fuzzy Systems Conference Proceedings, 1999. FUZZ-IEEE'99. 1999 IEEE International*. IEEE. 3 (1999): 1597-1602.

[51] J. Fišer, O. Jamriska, D. Simons, E. Shechtman, J. Lu, P. Asente, M. Lukac, and D. Sykora. “Example-based synthesis of stylized facial animations.” *ACM Transactions on Graphics (TOG).* 36.4 (2017): 155.

[52] I. Berger, A. Shamir, and M. Mahler. “Style and abstraction in portrait sketching.” *ACM Transactions on Graphics (TOG).* 32.4 (2013): 55.

[53] H. Zhou, Z. Kuang, and K. Y. K. Wong. “Markov weight fields for face sketch synthesis.” *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. Pages: 1091-1097. 2012.

[54] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, and J. Li. “Multiple representations-based face sketch–photo synthesis.” *IEEE transactions on neural networks and learning systems*. 27.11 (2016): 2201-2215.

[55] N. Wang, X. Gao, L. Sun, and J. Li. “Bayesian face sketch synthesis.” *IEEE Transactions on Image Processing.* 26.3 (2017): 1264-1274.

[56] L. Zhang, L. Lin, X. Wu, S. Ding, and L. Zhang. “End-to-end photo-sketch generation via fully convolutional representation learning.” *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM. 2015.

[57] L. A. Gatys, A. S. Ecker, and M. Bethge. “A neural algorithm of artistic style.” *arXiv preprint arXiv:1508.06576* (2015).

[58] K. Simonyan, and A. Zisserman. “Very deep convolutional networks for large-scale image recognition.” *arXiv preprint arXiv:1409.1556* (2014).

[59] J. Johnson, A. Alahi, and L. Fei-Fei. “Perceptual losses for real-time style transfer and super-resolution.” *European Conference on Computer Vision*. Springer International Publishing. Pages: 694-711. 2016.

[60] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. “Caffe: Convolutional architecture for fast feature embedding.” *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. Pages: 675-678. 2014.

[61] Z. Yi, H. Zhang, and P. T. Gong. “DualGAN: Unsupervised Dual Learning for Image-to-Image Translation.” *arXiv preprint arXiv:1704.02510* (2017).

[62] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative adversarial nets.” *Advances in neural information processing systems*. 2014.

[63] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T. Liu, and W. Y. Ma. “Dual learning for machine translation.” *Advances in Neural Information Processing Systems*. 2016.

[64] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. “Image-to-image translation with conditional adversarial networks.” *arXiv preprint arXiv:1611.07004* (2016).

[65] C. Li, and M. Wand. “Precomputed real-time texture synthesis with markovian generative adversarial networks.” *European Conference on Computer Vision*. Springer International Publishing. Pages: 702-716. 2016.

[66] M. Arjovsky, S. Chintala, and L. Bottou. “Wasserstein gan.” *arXiv preprint arXiv:1701.07875* (2017).

[67] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. “Unpaired image-to-image translation using cycle-consistent adversarial networks.” *arXiv preprint arXiv:1703.10593* (2017).

[68] O. M. Parkhi, A. Vedaldi, and A. Zisserman. “Deep Face Recognition.” *BMVC*. 1.3 (2015): 6.

[69] F. Gurpinar, H. Kaya, H. Dibeklioglu, and A. Salah. “Kernel ELM and CNN based facial age estimation.” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Pages: 80-86. 2016.