2013

Human metrology for person classification and recognition

Deng Cao
West Virginia University

Human Metrology for Person Classification and Recognition

Deng Cao

Dissertation submitted

to the Benjamin M. Statler College of Engineering and Mineral Resources

at West Virginia University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in

Computer and Information Science

Committee Members:

Donald Adjeroh, Ph.D, Chair

Arun Ross, Ph.D

Elaine Eschen, Ph.D

E. James Harner, Ph.D

Cun-Quan Zhang, Ph.D

Lane Department of

Computer Science and Electrical Engineering

Morgantown, West Virginia

Keywords: Human Metrology, Biometric, Classification, Recognition, Individuality, Capacity

Abstract

Human metrology generally refers to the geometric measurements extracted from humans, such as height, chest circumference or foot length. It provides an important soft biometric that can be used in challenging situations such as human identification at a distance, where hard biometric traits cannot easily be acquired. In this work, we first study the question of predictability and correlation in human metrology, using the tools of entropy. We show that various human metrological features are highly correlated with each other. Thus, partial or available measurements can be used to predict other missing measurements. We then investigate the use of human metrology for the prediction of other soft biometrics, viz. gender and weight. In particular, we consider geometric measurements from the head, and those from the remaining parts of the human body, and propose a copula-based model for their use in predicting gender and weight. For gender prediction, the proposed copula-based model results in a 0.7% misclassification rate using both body and head information, 1.0% using only body information, and 12.2% using only head information on the CAESAR 1D database [1] consisting of 2,369 subjects. For weight prediction, the proposed model gives 0.01 mean absolute error (in the range 0 to 1) using both body and head information, 0.01 using only body information, and 0.07 using only measurements from the head. This leads to the assertion that human body metrology contains enough information for reliable prediction of gender and weight. Furthermore, we investigate the efficacy of the model in practical applications, where metrology data may be missing or severely contaminated by various sources of noise. The proposed copula-based technique is observed to reduce the impact of noise on prediction performance.

We then study the question of whether face metrology and its use for reliable gender prediction. A new method based solely on metrological information from facial landmarks is developed. In this work, metrological features are defined in terms of normalized angle and distance measures, and computed based on a set of land-

marks on facial images. The performance of the proposed metrology-based method is compared with that of a state-of-the-art appearance-based method for gender classification. Results are reported on two standard face databases, namely, MUCT [110] and XM2VTS [108] containing 276 and 295 visible spectrum images, respectively. The metrology-based approach resulted in an accuracy of 86.83% on the MUCT database and 82.83% on the XM2VTS database. This was slightly lower than that of the appearance-based method by about 3.8% for the MUCT database and about 5.7% for the XM2VTS database. However, results on the WVUM Multispectral database consisting of 100 near infrared images and 100 multispectral images showed that the metrology-based method outperformed the appearance-based method (87.00% vs. 82.00%).

Furthermore, we study the question of person recognition (classification and identification) via whole body metrology. Using CAESAR 1D database as baseline, we simulate intra-class variation with various noise models. The experimental results indicate that given enough number of features, our metrology-based recognition system can have promising performance that is comparable to several recent state-of-the-art recognition systems. We propose a non-parametric feature selection methodology, called adapted k-nearest neighbor estimator, which does not rely on intra-class distribution of the query set. This leads to improved results over other nearest neighbor estimators (as feature selection criteria) for moderate number of features.

Finally we quantify the discrimination capability of human metrology. Generally, a biometric-based recognition technique relies on an assumption that the given biometric is unique to an individual. However, the validity of this assumption is not yet generally confirmed for most soft biometrics, such as human metrology. A scientific basis for establishing the uniqueness of human metrology will not only quantify the performance of an automatic recognition system, but will also result in the possible admissibility of metrology-based identification technique in various areas such as the

courts of law. Currently, only a few efforts have been made on theoretical studies of the discrimination capability of given biometric traits, such as individuality of fingerprints [120] and capacity of biometric systems [135]. We indicate the strengths and weaknesses of each approach. Following the review of prior work, we propose two schemes for theoretical analysis of the discrimination capability of human metrology.

Acknowledgements

I would like to thank my advisor, Dr. Donald Adjeroh, for his guidance and advice, and especially his continuous encouragement. I would like to thank my co-advisor Dr. Arun Ross, for his guidance and advice, and kind suggestions. It has been a honor to work under their supervision. I would like to thank my other committee members: Dr. Elaine Eschen, Dr. James Harner, and Dr. Cun-Quan Zhang for their excellent courses and their help during my studies. I would also like to thank my lab group members, Cunjian Chen, Marco Piccirilli and Richard Beal, for their help and support. And finally, I thank my parents and my wife for their selfless support. Without all of them, this dissertation would not have come about.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Soft biometric traits are those characteristics that provide some information about an individual, but lack the distinctiveness and permanence to sufficiently differentiate every pair of individuals [81]. Soft biometrics include physical or behavioral human characteristics, such as gender, ethnicity, age, eye color, weight and gait. Unlike the classical biometrics, also called "hard biometrics" such as fingerprint and iris, soft biometrics are usually applied to complement the identity information provided by hard biometrics. Can soft biometrics be solely used to distinguish individuals? What is the discrimination capability of a given soft biometric system? That is, for a given database, how many classes can the system successfully distinguish? To the best of our knowledge, only a few efforts have been made to address these questions. Our objective is to investigate whether a specific type of soft-biometrics, namely, human metrology, can be used to classify individuals. In particular, we consider the use of human metrology for the prediction of certain soft biometrics, including gender and weight. In order to do so, a series of metrology based recognition techniques are developed and their performances are evaluated both in a simulated environment

and in practice. Further, we investigate the discrimination capability of general soft biometric systems and human metrology in particular, from both a probabilistic and an information capacity perspective.

1.2 The Problem

1.2.1 Whole Body Metrology

Human classification and identification is a challenging problem, with diverse applications. Biometrics has thus become an active research field with many unresolved questions. For identification under confounding situations, such as night-time environments or identification at a distance, most traditional hard biometrics such as fingerprints, face, and iris may not be readily available. There is also the problem of poor quality for the video or image. An alternative is to exploit potential secondary or soft biometric traits[80] that could be automatically extracted from such typically poor quality video or images. Metrological features (such as human body shape, anthropometric measurements, and geometrical features) and other soft biometrics (such as gait, age, gender, weight, skin texture, etc), could be considered as evidence of human identity when hard biometrics are not available.

Whole Body Information as A Soft Biometric

Whole body metrology can be extracted via security surveillance in building entrances, parking lot, restaurants, supermarkets, airports, etc. One application of whole body metrology could be in gender and/or weight prediction. Gender prediction is a fundamental task for both humans and machines. As many social activities depend on precise gender identification, the problem has attracted considerable attention, and has been investigated from both psychological [17, 46] and computational [21] perspectives. Although most existing work has focused on assessing gender using

information from the human face[104, 66, 54], researchers also have considered using whole body information for gender prediction. Li et al. [95] attempted to perform gender classification using human gait. Cao et al. [19] studied the problem of gender recognition using whole body images using a part-based representation and an ensemble learning algorithm. They reported a 75.0% accuracy for predicting gender from either the front view or back view.

Guo et al. [69] used biologically-inspired features in combination with manifold learning techniques and achieved around 80% accuracy in gender prediction from body. Collins et al. [28] investigated several popular image representation methods, such as Histogram of Gradients (HOG), spatial pyramid HOG (PHOG) and bag of words model, with the goal of finding effective human body appearance models for gender prediction. Shan et al. [139] fused gait and face cues for automated gender recognition based on canonical correlation analysis. Their work demonstrated that the two sets of measurements from gait and face are complementary, resulting in improved recognition accuracy at the feature level.

Though the above approaches show that automatic gender prediction from human body is feasible, the methods are largely based on *appearance* or *texture* information. Our work, on the other hand, utilizes a set of *geometric measurements* to predict gender in the absence of textural details. A similar line of work was undertaken by Adjeroh et al. in [4] where the problem of gender prediction from whole-body human metrology was considered. However, in [4], the associations between different body measurements were neither characterized nor utilized to improve prediction performance. This work bridges this gap by introducing a novel copula-based prediction model that exploits the association structure between different human metrological features. In addition to gender prediction, the proposed model is also used to deduce an individual's weight from metrology. Weight prediction from metrology has been previously studied by Adjeroh et al. [4] and Velardo and Dugelay [153].

Whole Body Information for Human Recognition

The history of human metrology includes and spans various concepts such as cloth design, ergonomics, epidemiology and medical anthropology. But it was not applied to law enforcement until 1882, when a French police officer and biometrics researcher Alphonse Bertillon created a an identification system based on physical measurements of the human body, head and other personality characteristics [127]. Bertillon had been thinking of a better way to identify offenders and maintain their criminal records. He thought that it would be better to classify and file offender data according to their body size and measurements instead of their names, which were different every time they were arrested. Bertillon consulted the work of Lambert Quetelet, a Belgian statistician and mathematician, who had calculated that the chances against two people being roughly the same height were four to one. Bertillon figured that if more body measurements were added to the equation, the likelihood that any two people having the same dimensions would be rare. The uniqueness of human measurements became the basis of his identification system known as anthropometry or Bertillonage. The human measurements included standing height, sitting height, circumferences of head, width of head (between the cheek bones), length of ears, arm span, left forearm, left foot length, and length of left middle and little fingers [56]. He used the system in 1884 to identify 241 offenders, and the system was quickly adopted widely by American and British police forces.

In modern science, whole body metrology is not often used for human recognition. Collins et al. established a baseline method for human identification based on body shape and gait [29]. Hsin-Chun Tsai et al. [150] proposed a method that combines height and face information for long distance human identification. Unlike the above approaches, we study the question of whether or not human metrology can be solely used for person recognition, which includes person classification and identification.

1.2.2 Face Metrology

As mentioned in section 1.2.1, face information can also be used for gender prediction. Introduced in the 1990's, SEXNET was among the first automated systems capable of performing gender identification using human faces [65]. Since then, a number of studies investigated the problem as part of face recognition (FR). Modern FR systems typically combine textural information from the face with facial geometry. Popular examples include active appearance models (AAM) [94, 107], active shape models (ASM) [30], local feature analysis [89], and elastic bunch graph matching[158]. In such systems, the information about facial geometry is often extracted from specific landmarks on the face. In this work, we investigate whether only topological information extracted from facial landmarks can be used to perform gender classification reliably and efficiently, whether operating in the visible spectrum, or in the near infrared band.

Facial Landmarks in Gender Classification

There is still an on-going debate on whether facial landmarks (or information derived from such landmarks) can be used for reliable determination of the gender of a given individual. Farkas et al. [52] used 14 anatomical facial measurements to establish the morphological structure in 25 ethnic groups in both genders. They studied data measurements from 1470 subjects, aged 18 to 30 years, including 750 males and 720 females. The experimental results indicated that there exist a number of statistically significant differences across ethnic groups. In fact, in a recent study on inter-ethnic variability of facial dimensions, based on a review of the literature, Fang et al. [49] reported that no significant difference could be observed between gender when using neo-classical facial proportions, which include the heights and widths of the upper, middle, and lower face. Although one could argue that the conclusions in the recent report [49] clearly depend on the specific neo-classical facial proportions used, and

how measurements from the different ethnic groups were analyzed, it seems that there is still no consensus on whether facial measurements can in fact reliably distinguish between gender.

Our main goal in this work is to lay the above questions to rest.

1.2.3 Distinctiveness of Soft Biometrics

A fundamental requirement of any biometric recognition system is a specific human trait, which should have several desirable properties such as universality, measurability and uniqueness (or distinctiveness)[79]. Universality means every individual in the considered population should possess the trait. Measurability means it should be possible to acquire the biometric trait and transform it to digitized features without causing undue inconvenience to the individual. Distinctiveness means the trait should be sufficiently different across individuals in the population. For a given biometric, its distinctiveness is very important in quantifying to what extent the biometric can be relied upon to distinguish between individuals. However, compared to other properties, the distinctiveness is difficult to verify due to the huge number of individuals in the world. Some biometric traits, such as fingerprints and iris, are generally considered as being unique to an individual based on empirical results. Recently, the notion of individuality has been used to describe the uniqueness of fingerprint [120]. The related work on the individuality of several biometric traits have been studied using different methods. These methods are described in section 5.1. However, the underlying scientific basis for the distinctiveness of soft biometrics features is not yet well developed.

The distinctiveness can be compromised by the poor quality of the features. In practice, the obtained biometric information can be easily contaminated by various types of noise. From an information theoretic perspective, the quality problem can be considered as a noisy channel problem. The capacity of a channel describes the tight-

est upper bound on the amount of distinguishable information that can be reliably transmitted over the communication channel [34]. In Schmid and Nicolò's work [135], a parallel Gaussian channel model is adopted for analyzing the recognition capacity of a biometric system. Here the input is the accurate feature with Gaussian distribution and the noise is i.i.d. and also Gaussian. The statistics (such as mean and variance) of the input and the noise are considered as known. The Gaussian assumption is rather strong and might not always hold in practice. Yet, this work could provide a basis for developing a capacity-based model for analyzing the distinctiveness of soft-biometric traits.

1.3 Contributions

1.3.1 Classification and Prediction using Whole Body Metrology

In general, the use of whole body metrology for deducing soft biometric traits has several applications. In video-based surveillance systems, it may be easier to quickly extract geometric measurements of the human body for classification, rather than primary biometric traits such as face or iris. In recognition-at-a-distance applications, primary biometric traits may not be readily available thereby necessitating the use of the dynamic geometry of the human body for identification. In applications based on Microsoft Xbox Kinect, deducing gender or weight information from human anthropometric measurements may be useful for enhancing perceived user experience.

The dissertation contributes to the study of prediction using whole-body metrology in three ways. Firstly we investigate the issue of predictability and correlation in human metrology, using information theoretic notions of uncertainty. Secondly, we develop a copula-based gender and weight prediction model that accounts for associations between geometric attributes of the human body. Thirdly, we evaluate the

efficacy of the model on the CAESAR 1D database [1], both in the absence and presence of (simulated) noise. For gender prediction without noise impact, the proposed model yield 0.7%, 1.0%, and 12.2% misclassification rate using whole body information, body-only information and head-only information, respectively. For weight prediction without noise impact, the proposed model gives 0.01, 0.01, and 0.07 mean absolute error (in the range 0 to 1) using whole body information, body-only information and head-only information, respectively. This leads to the assertion that human body metrology contains enough information for reliable prediction of gender and weight. Furthermore, the proposed model is observed to reduce the noise impact on prediction performance.

1.3.2 Classification using Face Metrology

We investigate whether topological information extracted from facial landmarks can be used to efficiently perform gender prediction, and the experimental results show that it does.

The main challenges related to facial metrology include (a) difficulty in precisely localizing the landmark coordinates; (b) sensitivity of landmark localization to pose, expression, and other variations; and (c) sparsity of information encoded in landmarks for human identification. In spite of these challenges, landmarks from 2D faces can provide important cues for problems related to human recognition. Following the previous study on predictability and correlation in whole body human metrology [4], in this work we hypothesize that the information extracted from facial metrology alone can be used for gender recognition or facial classification. We assume that facial landmarks on a given face image are already provided and our research goal is to perform gender classification based solely on the information provided by facial landmarks. If gender classification can be successfully performed using these landmark points, then investment can be made in automating the landmark detection process. The

performance of our proposed facial metrology-based gender classification algorithm is compared to a benchmark appearance-based technique, namely, the Local Binary Patterns (LBP) method [5, 104]. The main contribution of our work is a demonstration that the classification performance using solely facial metrology is comparable to that of an appearance-based method when using either visible or near infrared (NIR) face images. Thus, we illustrate that by using only weak features, i.e., facial metrological features derived from facial landmarks, our approach results in only about 3.8-5.7% lower classification rate (on two different face databases) compared to a benchmark appearance-based method. On the other hand, using facial-metrology outperformed the appearance-based method by about 5% on NIR images.

1.3.3 Recognition using Whole Body Metrology

In this work, we provide an answer to the question: can we perform person recognition via human metrology? Using CAESAR 1D database as baseline, we simulate intra-class variation using various noise models. We propose a non-parametric feature selection methodology, called adapted k-nearest neighbor estimator, which does not rely on the intra-class distribution of the query set. This leads to improved results over other nearest neighbor estimators (as feature selection criteria) for moderate number of features. We then apply the selected features for person recognition. The experimental results indicate that given enough number of features, our metrology-based recognition system can have promising performance that is comparable to several recent state-of-the-art recognition systems.

1.3.4 Discrimination Capability of Human Metrology

In Chapter 5, we investigate the discrimination capability of soft biometric systems, namely how many classes the system can successfully distinguish. Inspired by the concept of individuality [120] and capacity [135], we first develop two schemes that

can address the individuality of human metrology, or any other biometrics that can be encoded as a collection of scalar numbers. Furthermore, our schemes are more general and realistic: for the individuality, the distribution of the features are not restricted. Also, the noise caused by small intra-class variation (errors) for each feature is explicitly controlled. For capacity, a Poisson channel model is proposed to analyze the recognition capacity of human metrology. Our study suggests that the performance of such a metrology-based system depends more on the accuracy and precision level of the ground truth or training set.

1.3.5 Publications Related to The Dissertation

1. Donald Adjeroh, Deng Cao, Marco Piccirilli, and Arun Ross. Predictability and correlation in human metrology. In *IEEE International Workshop on Information Forensics and Security*, 2010

2. Deng Cao, Cunjian Chen, Marco Piccirilli, Donald Adjeroh, Thirimachos Bourlai, and Arun Ross. Can facial metrology predict gender? In *IEEE International Joint Conference on Biometrics*, 2011

3. T. Bourlai, N. Kalka, D. Cao, B. Decann, Z. Jafri, F. Nicolo, C. Whitelam, J. Zuo, D. Adjeroh, B. Cukic, J. Dawson, L. Hornak, A. Ross and N. A. Schmid, Ascertaining Human Identity in Night Environments. Book Chapter in *Distributed Video Sensor Networks, Springer*, 2011

4. Deng Cao, Cunjian Chen, Donald Adjeroh, and Arun Ross. Predicting Gender and Weight from Human Metrology using a Copula Model. In *IEEE Biometrics: Theory, Applications and Systems*, 2012

5. Deng Cao, Cunjian Chen, Marco Piccirilli, Donald Adjeroh, Thirimachos Bourlai, and Arun Ross. Gender Prediction via Facial Metrology. Submitted to *IEEE Information Forensics and Security*, 2013

6. Deng Cao and Donald Adjeroh, On the Individuality of Human Metrology, Submitted to *IEEE Transactions on Systems, Man, and Cybernetics*, 2013

1.4 Organization

The dissertation is organized as follows. In Chapter 2, we presents a copula-based model for predicting gender and weight from human metrology, including body-only and head-only measurements. Chapter 3 introduces a fully automated system for gender prediction using frontal face images. The performance of the proposed metrology-based method is compared with that of a state-of-the-art appearance-based method, namely, local binary pattern (LBP). The performance under cross-database and/or cross-spectra conditions is also tested in practice. Chapter 4 gives initial person recognition results using solely human metrology, based on a proposed novel feature selection method called adapted k nearest neighbor estimator. Chapter 5 systematically investigates the problem of discrimination capability by formulating explicit expressions of individuality and capacity of a given biometric system. Chapter 6 draws some conclusions and also describes possible directions for future work.

Chapter 2

Whole Body Metrology for Classification[1]

2.1 Background

2.1.1 Predictability and Correlation

In this chapter, we first study the problem of predictability and correlation in human metrology. Our work is closely related to earlier work on single view metrology [35, 68]; session biometrics using height measurements [102]; and to efforts on analysis of human body shape and head sizes [6, 64, 16]. Other related work include general efforts on whole-body modeling [72], soft biometrics [80], and analysis of human gait [117, 74]. Our work differs from these in that none of the methods paid any specific attention to whole-body human metrology, beyond height or head dimensions. Those based on the CAESAR dataset [64, 6, 16] have all focused on the 3D data points. Here, we use only the 1D measurements in the CAESAR dataset, with potential

[1]Part of the work reported in this chapter has been published in the following papers:
[1] Donald Adjeroh, Deng Cao, Marco Piccirilli, and Arun Ross. Predictability and correlation in human metrology. In *WIFS*, 2010.
[2] Deng Cao, Cunjian Chen, Donald Adjeroh, and Arun Ross. Predicting Gender and Weight from Human Metrology using a Copula Model. In *BTAS*, 2012.

advantages in computation and automated acquisition from surveillance video. To our knowledge, this is the first attempt at a detailed and formal study of predictability and correlation in whole-body human metrology.

2.1.2 A General Prediction System

We then investigate the use of human metrology for the prediction of certain soft biometrics, viz. gender and weight. In particular, we consider geometric measurements from the head, and those from the remaining parts of the human body, and analyze their potential in predicting gender and weight.

An end-to-end biometric prediction system has several discrete stages. In the first stage (**feature extraction**), a person is characterized using a collection of biometric traits known as features. These features need to be properly extracted from an individual. In the second stage (**feature representation**), the raw features are transformed into a new feature space, which is expected to be suitable for further analysis. This stage usually involves various types of normalizations. In the third stage (**feature selection**), certain techniques are used to reduce the dimension of the feature space in order to enhance the generalization capability, or to speed up the learning process. After this stage, the raw feature vector $V = (x_1, \ldots, x_n)$ becomes $V' = (x'_1, \ldots, x'_m)$, where $m \leq n$. In the final stage (**feature prediction**), V' is sent to a classifier (for classification) or a regressor (for estimation). The output is a discrete class for classification or a continuous value for regression.

In this chapter, we do *not* focus on feature extraction. We assume that a set of features (measurements), referred to as the feature vector, is already provided. Our goal is to analyze these features and develop a prediction model based on these features. The input to our prediction system is a set of metrological features pertaining to an individual. These features correspond to the head and the body. The output

of the prediction system is the gender (classification) or weight (regression) of the individual.

2.2 Statistics of Human Metrology

We first consider the general statistics of human metrology, using available data on human body measurements.

2.2.1 Database

We used the Civilian American and European Surface Anthropometry Resource (CAESAR) 1D database [1] with 2400 US & Canadaian civilians, ages 18-65. There was an equal proportion of male and female subjects, and of people in the age ranges 18-29, 30-44, and 45-65. We used 43 human body measurements or attributes (see Table A.1 in the appendix) along with gender and weight. Measurements are in millimeters, while weight is in kilograms. After removal of samples with missing data, we obtained 2369 samples (1119 males and 1250 females). We randomly select 2000 samples as training set, and the rest 369 samples are used for testing. We also selected 10 measurements (SET-10M) for closer observation and ease of presentation. The measurements in SET-10M are as follows **1:arm length (AL); 2:armscye circumference (AC); 3:chest circumference (CC); 4:head breadth (HB); 5:head length (HL); 6:neck base circumference (NC); 7:shoulder breadth (SB); 8:stature (S); 9: waist circumference (WC); 10:weight (W)**. For each body dimension, the individual measurement X_i is normalized to the [0 1] range:

$$X_i = \frac{X_i - min\{X_i\}}{max\{X_i\} - min\{X_i\}}$$

2.2.2 Statistics

Table 2.1 shows the summary statistics on the measurements in SET-10M, using the 2000 people in the training set. Though height (stature) is easier to acquire, and has the largest values, it may not necessarily be the best for discriminating between individuals. It has a relatively low coefficient of variation (third to last), and low standard deviation for the normalized values (third to last). The circumferences, which typically cannot be extracted from a 2D video sequence, have higher first order entropy than one dimensional measures, such as height and shoulder breadth. This makes a strong case for methods that can predict such body circumferences reliably.

Table 2.1: Summary statistics on human metrology. *Std2: std. deviation for [0 1] normalized measures. CV: coeff. of variation; CRV coeff. of relative variation (See Section 2.4).

Measure	1 (AL)	2 (AC)	3 (CC)	4 (HB)	5 (HL)	6 (NC)	7 (SB)	8 (S)	9 (WC)	10 (W)
Min	237	293	739	123	166	344	346	1248	557	39.9
Max	416	606	1574	204	228	598	658	2084	1702	181
Median	324	415	981	150	194	435	460	1707	836	73.9
Std	23.7	51.6	121.2	7.3	9.4	39.7	48.3	101.6	143.2	19.3
*Std2	0.07	0.12	0.12	0.05	0.05	0.09	0.10	0.06	0.17	0.25
CV	0.13	0.17	0.15	0.09	0.11	0.16	0.16	0.12	0.13	0.14
CRV	0.13	0.17	0.15	0.09	0.15	0.16	0.16	0.12	0.13	0.14
Entropy	6.54	7.56	8.60	4.85	5.23	7.15	7.46	8.46	8.77	8.07

Figure 2.1A shows the probability distribution for the measurements and the scatter plots for pairs of measurements in SET-10M. The diagonal plots correspond to the probability densities, while the off diagonals contain the scatter plots for the corresponding pairs. The plots provide some idea on the nature of the measurements, and the potential dependence and/or correlation between them. Figure 2.1B shows the distribution of the ratio of the given measurement to the head length. This captures the traditional notion of "number of heads" in say height. (Here head length is the distance on a straight line from the glabella to the rearmost point on the skull).

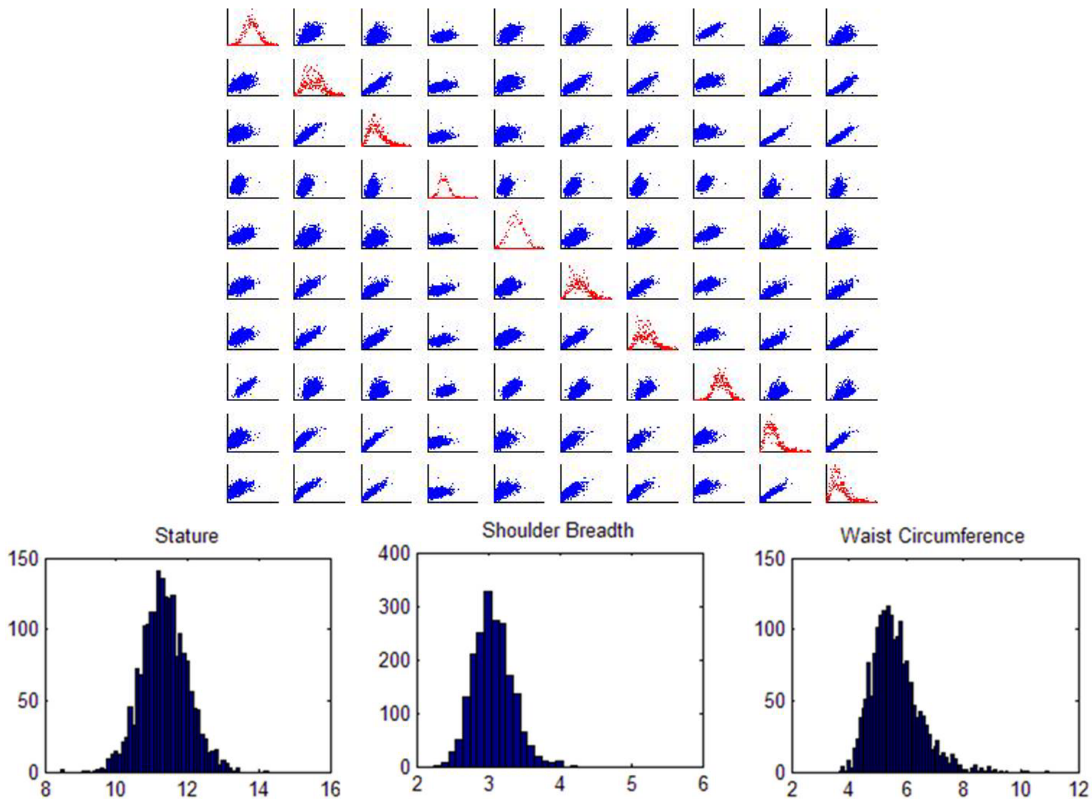


Figure 2.1: Statistics of human metrology. Top (A): Scatter plots and probability density for the 10 selected measurements; Bottom (B): Ratio plots: Distribution of the number of head lengths contained in other measurements, namely stature, shoulder breadth, and waist circumference.

2.3 Correlation in Human Metrology

Our goal is to build statistical models that can predict human metrology in the event of severe occlusion, missing body parts, poor segmentation, or other problems. In this work, we intend to follow the principle of Occam's Razor, which suggests the smallest model with minimum number of predictors. Unnecessary predictors will add noise to the estimation. Too many predictors may introduce redundancy and lead to multicollinearity, with added computational cost. Figure 2.1, however, shows that most of the measurements on the human body could be highly correlated. Thus, we need to study the potential correlation between the measurements, and see how this could be used in model building.

Significant correlation between human metrological features has been observed. For example, a tall person is *likely* to have long arms, long feet, and long fingers. However, previous gender prediction models [4, 153, 38, 69] do not explicitly consider the association between the features in the feature vector. Clearly, incorporating information on the association or correlation structure between human measurements is likely to lead to improved prediction performance. To better understand the interaction among different metrological features, we use a specific statistical tool. The term **association** should be considered as **statistical dependence**. We do not use the Pearson correlation coefficient since the linearity between features is not guaranteed. We also avoid the Chi-square goodness of fit test, since it requires prior knowledge of the distribution of samples (Here, a sample is a single feature vector). Instead, we use a non-parametric test which does not rely on any assumptions on the distribution of samples. The Kendall's tau rank correlation coefficient [87] is selected as our tool.

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a be a set of joint observations from two random variables X and Y, respectively, such that all the values of x_i and y_i are unique. A pair of observations (x_i, y_i) and (x_j, y_j) are said to be **concordant** if both $x_i < x_j$ and $y_i < y_j$ or if both $x_i > x_j$ and $y_i > y_j$. They are said to be **discordant**, if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant. Let n_c and n_d be the number of concordant and discordant pairs, respectively. Then, the Kendall's tau correlation coefficient is defined as follows:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}. \tag{2.1}$$

If X and Y are independent, the coefficient would be approximately zero. The coefficient is 1 for complete agreement between X and Y and -1 for complete disagreement. Figure 2.2 shows the absolute values of pairwise Kendall's tau coefficients for the 43 anthropometric measures in the CAESAR 1D database [1]. The features themselves are listed in Table A.1. The warmer color denotes a higher correlation and the cooler

color denotes a lower correlation. The figure shows that significant associations do exist between pairwise features. The average absolute coefficient value between features varies from 0.13 to 0.51.

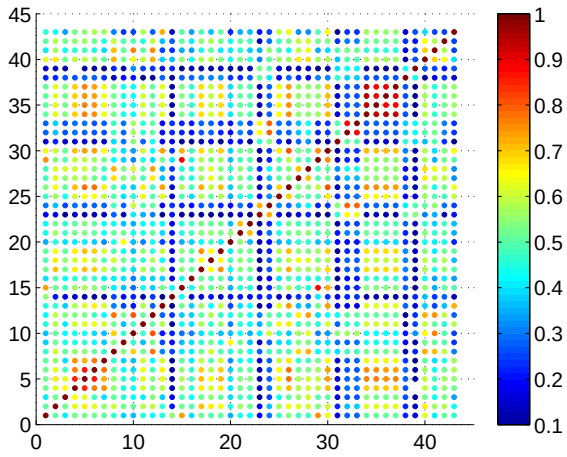


Figure 2.2: A display of the absolute values of the pairwise Kendall's tau coefficients between the 43 anthropometric measures found in the CAESAR database.

We wish to incorporate this association structure into our prediction model in order to boost its performance. Thus, a novel copula-based prediction system is proposed later in this chapter.

2.4 Predictability in Human Metrology

2.4.1 Uncertainty in Metrology

To investigate the potential predictability of human metrological data, we can use the uncertainty in body dimensions. The issue of uncertainty is related to two important questions with respect to human metrology as a soft biometric, namely: inter-class variability and intra-class variability. These are related to the predictability of such metrological features, and their identification capacity or discriminative ability. That is, given the measurements of the same body part from several people, how easily can

we predict the given measurement for a specific individual? Essentially, this question is related to the uniqueness of the individual measurements to a given person. Or, if the measurement is not unique, how many categories of people can it distinguish? Taken on an individual basis, single measurements may not be very discriminative. But when considered jointly, they may provide a reliable tool for grouping people into several defined categories, and hence a method for human identity profiling. Here, we focus on the issue of predictability of human metrology. This is more closely related to the inter-class variability than intra-class variability.

The variance of a random variable gives an idea of the uncertainty of the variable. We could assess the inter-class variability of the individual measurements using the coefficient of variability $CV(X)$ and the coefficient of relative variability $CRV(X)$, two simple measures related to the variance. For a given random variable X, these are defined as follows:

$$CV(X) = \frac{\sigma_X}{\mu_X};\ CRV(X) = \frac{\sigma_X}{max\{X_i\} - min\{X_i\}}$$

Perhaps, a better approach for studying the uncertainty of a random variable is by use of entropy. Let $X = x_1, x_2...x_n$ be a sequence, with symbols from an alphabet $\mathscr{A}$. The entropy of the sequence is defined as:

$$H(X) = -\sum_{i=1}^{|\mathscr{A}|} p(\sigma_i) \log p(\sigma_i),$$

where $p(\sigma_i)$ is the probability of the i-th symbol in the alphabet, $\mathscr{A}$. Given that our measurements are continuous variables, we can consider the differential entropy, rather than the discrete entropy. Let X be be a continuous random variable with probability density function $p(x)$, and support set $\Omega = \{x|p(x) > 0\}$. Assuming the integral exist, the differential entropy is then given by:

$$h(X) = -\int_{\Omega} p(x) \log_e p(x) dx.$$

For Gaussian variables this becomes,

$$h(X) = -\int_{\Omega} p(x) \log_e (\frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})) dx$$

$$= -\int_{\Omega} p(x) (\log_e (\frac{1}{\sigma\sqrt{2\pi}}) - \frac{(x-\mu)^2}{2\sigma^2}) dx$$

This can be evaluated to obtain: $h(X) = \log_2(\sqrt{2\pi e}\sigma)$ bits.

Thus, for Gaussian random variables, the differential entropy becomes a simple function of the variance. For most of the measurements on human body dimensions, the distribution can be approximated as Gaussian, and hence, the above can be applied to get an idea of the differential entropy. The joint entropy between two random variables shows the joint uncertainty between the two variables. Like the case of discrete entropy, we can define the joint differential entropy for two continuous random variables X and Y, with joint probability density function $p(x, y)$:

$$h(X, Y) = -\int_{\Omega} p(x, y) \log p(x, y) dx dy$$

The conditional differential entropy is given by:

$$h(X|Y) = -\int_{\Omega} p(x, y) \log p(x|y) dx dy = h(X, Y) - h(Y)$$

The mutual information between two random variables tells us how much information one contains about the other. A high mutual information implies some relative

redundancy between the variables. The mutual information is given by:

$$I(X;Y) = \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy$$

Equivalently, we can write:

$$I(X;Y) = h(X) - h(X|Y) = h(X) + h(Y) - h(X,Y).$$

The differential entropy provides an upper bound on the discrete entropy. The difference depends on the quantization step size used for the discretization step. In general, using a quantization step size q to convert a continuous variable to a discrete counterpart, we have the following relation between the discrete entropy and the differential entropy: $H(X) + \log q \to h(X)$ as $q \to 0$.

This is particularly important for practical considerations, for instance, in our problem of human metrology. The probability distributions are likely to be quantized to discrete values before the entropy is computed. It is easy to extend the discussion on entropy of the measurements to higher order models, beyond the second order. Depending on the underlying distribution, some of the integrals involved may not exist. However, for some special cases, it may be possible to obtain close form solutions for some of the quantities. For instance, for multivariate normal variables $X_1, X_2..., X_n$, with mean vector μ, covariance matrix Σ, and probability density function

$$p(\mathbf{x}) = (\frac{1}{(\sqrt{2\pi})^n |\Sigma|^{1/2}}) \exp(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu))$$

where $|\Sigma|$ denotes the determinant of matrix Σ. The differential entropy can be evaluated to get $h(X_1, X_2, ..., X_n) = \frac{1}{2} \log(2\pi e)^n |\Sigma|$ bits. This is important. First, the Gaussian distribution provides an upper bound for differential entropy. Further, with several measurements of the same body dimension from many people, if we can

establish that their joint distribution can be approximated by a multivariate normal distribution, we can then compute their differential entropy by considering mainly their covariance matrix. The joint entropy then provides us with some idea on the ability of the measurements in distinguishing people (or at least in grouping people), since effectively, we will expect that about $2^{h(X_1,X_2,...,X_n)}$ people could be distinguished using the measurements directly. Table 2.1 showed the summary statistics on human metrology, including CV, CRV, and discrete entropy, as captured in the CAESAR dataset. Figure 2.3 shows the entropy plots using different number of bins. Waist circumference (measure 9 in Table 2.1) has the highest entropy with higher number of bins, followed closely by chest circumference (measure 3).

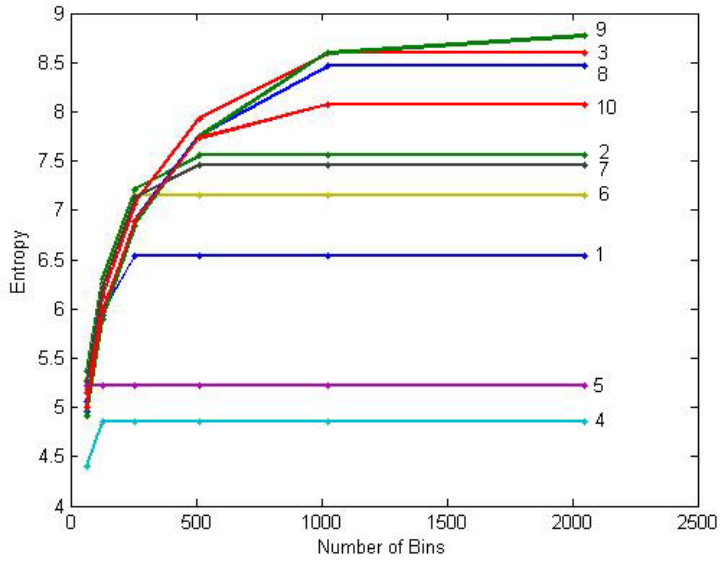
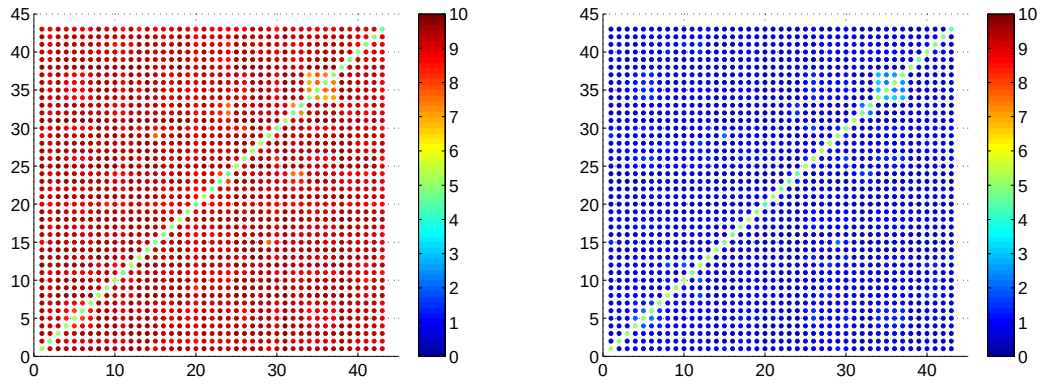


Figure 2.3: Entropy plots for the measurements in SET-10M.

Figure 2.4 shows the joint entropies (a) and the mutual information (b) between each pair of the 43 measurements, respectively. We observe that while a few measurements carry much information about others (high mutual information), most of the measurements contain relatively small portions of information about each other.



(a) (b)

Figure 2.4: The joint entropies (a) and the mutual information (b) between each pair of the 43 measurements using 64×64 bins.

2.4.2 Comparing Prediction Models

We first apply multiple linear regression to the CESAER data. The general form is as follows:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

where y is the response variable, X is the input variable (also called predictor variable) matrix, β is the parameter vector, ε is the error term, n is sample size and p is the number of parameters, excluding the constant term. Or more compactly:

$$y = \mathbf{X}\beta + \varepsilon$$

The least-square estimators of the elements in the parameter vector are obtained using the relation:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Since gender is a categorical variable, a logistic regression model is used to predict it separately.

To check the goodness of fit, R^2, the coefficient of determination is computed:

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

where y_i yield values in the range $0 \leq R^2 \leq 1$, and provide a measure of the proportion of variability in the response variable that is accounted for by the explanatory variables. We construct a family of prediction models, defined based on the order of the model, the number of variables involved, and the specific way in which the variables are combined. For a maximum order of 2, with a maximum of 2 predictor variables, this will lead to a family of 31 models, each member denoted with a binary code. For example, we have the following codes for three example members of the family:

2-predictor full model (Model # 31; code : **111111**) :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$

1-predictor full model (Model # 7; code : **110100**) :

$$y = \beta_0 + \beta_1 x_1 + \beta_{11} x_1^2$$

2-predictor, partial model (Model #27, code : **111001**) :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

We found that Model #31, the 2-repdictor full model produced the overall best results.

2.4.3 Human Metrology Predictability Network

Suppose we have various measurements available, and need to predict one or more other measurements, an important question is how we can choose the best subset of the measurements to be involved in the prediction. This is related to the issue of variable selection in pattern recognition [115]. Intuitively, we should choose the subset whose members have maximum correlation with the unknown measurement to be predicted, or one that minimizes the prediction error. Selecting the seed measurements using the first approach based on correlation can be performed using the correlation graph. To perform prediction for an unknown measurement, say Y, we involve only the measurements that share some edge with Y in the correlation graph, essentially the members in the subset: $X_{CG} = \{X | -T < \tau_{XY} < T\}$, where T is the threshold and τ_{XY} is the Kendall's tau correlation coefficient between X and Y.

The second approach would be to use the measurements that minimize the error when used to predict the unknown measurement. For each measurement X, we use the prediction models to perform an initial estimation, using the other measurements, and record the error that resulted. We then construct a bipartite graph whereby one set of nodes are for the predicted measurements, and the other for the predictors. An edge between a predictor node (say X) and a predicted node (say Y) indicates that the error from the prediction was less than a threshold.

We repeated the above using the 2-predictor model, and generated quite some interesting networks. Figure 2.5 shows an example for a threshold of $MAE \leq 0.04$. The key observation is that, for almost any given measurement, there is a set of pairs of other measurements that can predict it to within the error threshold. In most cases, there are several such pairs. The network therefore captures the overall predictability in human metrology. For most measurements, we need at most three hops to reach a node that can predict them to within the specified error threshold. Different thresholds will lead to different network configurations.

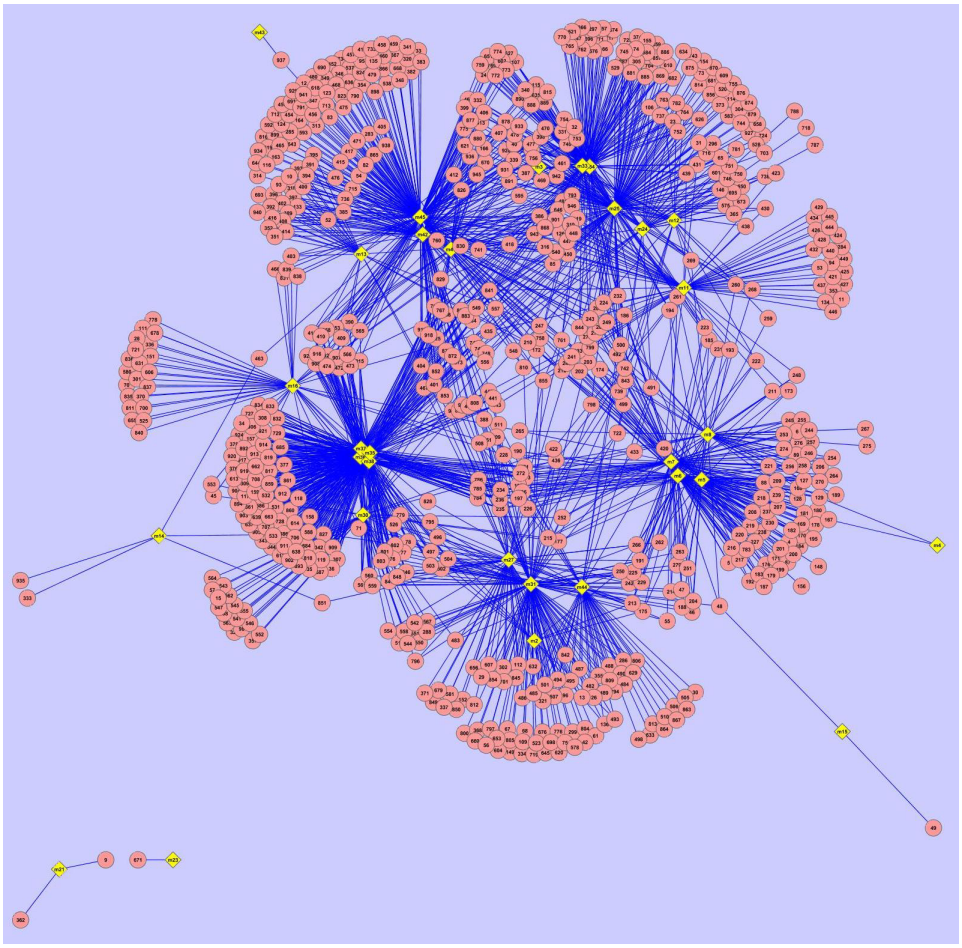


Figure 2.5: Predictability network in human metrology. Diamonds correspond to the predicted nodes; circular nodes denote pairs of measurements involved in the prediction. Number in a circular node denotes the index of the measurement pair involved. Results are for the 2-predictor model using a threshold of MAE$\leq$ 0.04.

2.5 Copula-based Prediction Model

In this section, we introduce a novel copula-based prediction model on gender and weight that is robust to heavy noise. The key characteristic of the model is the feature representation. We intend to construct a new feature type that have the following properties: (1) It contains association information between original features (in our case, the measurements); (2) It can be fused with the original raw features, and can be handled as well as the raw features, by further processors such as a SVM classifier; (3) The new feature representation should be optimized towards the corresponding

class; (4) It should be robust to possible errors in a realistic manner, so that the improved accuracy from association information will not be affected by minor errors in the measurements.

The process of constructing the new feature type is described in details in the following sections.

2.5.1 Bayesian Framework

Inspired by Kwitt et al [93], we construct the new feature type based on the class-conditional joint probability density. Let $1, \ldots, M$ be a set of classes, and $p(r)$ be the corresponding probability density function (*pdf*) for $r = 1, \ldots, M$. Let $z = (z_1, \ldots, z_B)$ be the B-dimensional feature vector in which the features may be dependent on each other. Assume that we are to classify a sample based on the evidence provided by z. An optimal decision rule is to choose the class that is most probable given observation z. Define a function $g(z) \rightarrow 1, \ldots, M$ that maps z to one of M classes. This decision rule can be formulated as a Bayes classifier [59]:

$$g(z) = \arg\max_r p(r|z). \tag{2.2}$$

However the posterior probability $p(r|z)$ is usually hard to obtain. So, we consider Bayes' theorem:

$$p(r|z) = \frac{p(z|r)p(r)}{p(z)}. \tag{2.3}$$

We may reasonably assume that each feature vector belongs to one and only one of the M classes with equal prior probability $p(r)$. Thus, by Bayes theorem, Eqn (2.2) can be rewritten as a maximum likelihood (ML):

$$g(z) = \arg\max_r p(z|r). \tag{2.4}$$

In practice, $p(z|r)$ can be estimated from a collection of training samples $z^1, \ldots, z^n$ from class r. Also, a classifier such as Support Vector Machine (SVM) can be used in lieu of Eqn (2.4) as is done in this work.

2.5.2 Copula Modeling

Our next objective is to compute the class-conditional likelihood $p(z|r)$. In this work, we choose the copula model to construct $p(z|r)$, which offers two advantages. Firstly, the copula representation does not require explicit mathematical relations between features, which are usually unknown in practice. Instead, it relies on the study of marginal distributions of components in z, which are substantially easier to obtain in practice. Secondly, the copula construction does not constrain the choice of marginal distributions, so the model can be adapted to different feature spaces.

Consider B uniform random variables, $u_1, \ldots, u_B$, where $u_i \in [0, 1]$ for $i = 1, \ldots, B$. Let $u = (u_1, \ldots, u_B)$. A copula is defined as follows:

$$C(u_1, \ldots, u_B) = Pr(U_1 \leq u_1, \ldots, U_B \leq u_B). \tag{2.5}$$

Thus, by Sklar's theorem [58], given a B-dimensional random vector $z = (z_1, \ldots, z_B)$, there exists a B-dimensional copula C such that:

$$C\left(F_1(z_1), \ldots, F_B(z_B)\right) = F(z_1, \ldots, z_B), \tag{2.6}$$

where $F_i(z_i), i = 1, \ldots, B$ are marginal cumulative distribution functions ($cdfs$): $F_i(z_i) = Pr(Z_i \leq z_i)$, and $F(z_1, \ldots, z_B)$ is the joint cdf. If $F_1, \ldots, F_B$ are given and they are continuous and non decreasing, we have [106]:

$$C(u) = F\left(F_1^{-1}(u_1), \ldots, F_B^{-1}(u_B)\right), \tag{2.7}$$

where $F_i^{-1}(u_i)$ denotes the inverse *cdf* of F_i. Eqn (2.7) is an important property of copulas that allows us to utilize the information about the marginals. Figure 2.6 and Figure 2.7 show the mapping from original data (Sitting Acromial Height and Ankle Circumference) to the unit square copula scale using a kernel estimator of the cumulative distribution function.

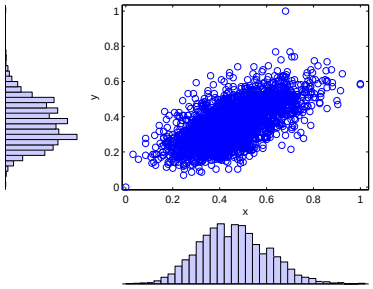


Figure 2.6: The scatter plot with marginal distributios for Sitting Acromial Height (x) and Ankle Circumference (y).

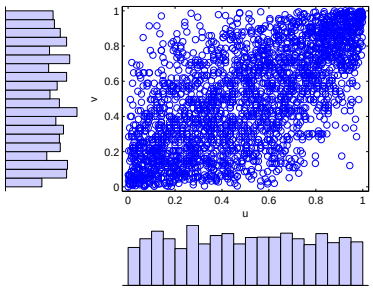


Figure 2.7: Transformed data in Figure 2.6 to the copula scale using the estimated *cdf*.

If the joint density f of F exists, it can be written as the product of the copula density c and the marginal densities [106, 93]:

$$
\begin{aligned}
f(z) &= \frac{\partial^B C(u)}{\partial z_1 \ldots \partial z_B} = \frac{\partial^B C(u)}{\partial u_1 \ldots \partial u_B} \prod_{i=1}^{B} \frac{\partial u_i}{\partial z_i} \\
&= c(u) \prod_{i=1}^{B} p_i(z_i). \qquad (2.8)
\end{aligned}
$$

Assuming that the class labels are provided in the training data (in our case, gender labels), the class-conditional copula density $c(u|r)$ can be easily calculated from Eqn (2.8) and has the form:

$$
c(u|r) = \frac{f(z|r)}{\prod_{i=1}^{B} p_i(z_i|r)}. \qquad (2.9)
$$

With enough training samples for each class, in Eqn (2.9), $f(z|r)$ and $p_i(z_i|r)$ for all i can be estimated from the corresponding training samples for each r. Thus, the association structure between B dependent features for a given class r is converted to a single variable $c(u|r)$. We denote $c(u|r)$ as our new feature and refer to this as the CFeature (C for copula). In our proposed algorithm, we do not consider a single association matrix. Rather, we deduce multiple association matrices corresponding to pairs of features. Consequently, we have multiple CFeatures corresponding to these pairs. To distinguish the set of original measurement features from the set of CFeatures, the former is referred to as MFeatures (M for metrology or measurement). In practice, computing $c(u|r)$ involves the estimation of the copula parameter matrix Λ^r, which is defined by the copula *type* that is chosen. There are several copula types possible: Gaussian copula, student t copula and various Archimedean copulas (e.g., Clayton, Frank, and Gumbel). For example, an Archimedean copula with Clayton generator is defined in Eqn (2.10), where ψ is the generator, ψ^{-1} is the generator inverse, and $\theta \in (0, \infty)$ is the only parameter. Since we consider pairwise associations, we can note that $C(u)$ has *two* components.

$$
\begin{aligned}
C(u) &= \psi_\theta \left[\psi^{-1}(u_1), \psi^{-1}(u_2)\right], \qquad (2.10)\\
\psi(t) &= (1+t)^{-1/\theta},\\
\psi^{-1}(t) &= t^{-\theta} - 1.
\end{aligned}
$$

2.5.3 Estimation of Parameters

The copula parameter matrix Λ^r in the copula model should be correctly estimated so that the probability of misclassification error is minimized. We use the local maximum likelihood estimation (MLE) method [93] to estimate this for each CFeature. Note there are two types of parameters: marginal parameters and copula parameters. To estimate these parameters, we can use a two-step procedure called the Inference Functions for Margins (IFMs) method [2]. Consider a pair of distinct features denoted by subscripts, i and j. For a given set of independent and identically distributed (iid) training samples, $z^1 = (z_i^1, z_j^1), \ldots, z^n = (z_i^n, z_j^n)$, $i, j \in \{1, 2, \ldots B\}$, the marginal parameters $\hat{\theta}_i$ and $\hat{\theta}_j$ are first estimated. In the second step, we transform z_i and z_j to their corresponding cumulative marginals F_i and F_j using the probability integral transform (PIT) [128] and estimate the copula parameter Λ using MLE [73]:

$$
\hat{\Lambda}^r = \arg\max_{\Lambda^r} \sum_{t=1}^{n} \log c\left(F(z_i^t; \hat{\theta}_i), F(z_j^t; \hat{\theta}_j) | \Lambda^r\right). \qquad (2.11)
$$

Although Sklar's theorem shows that a copula function always exists, Eqn (2.11) does not always have an explicit expression [93]. In practice, $\hat{\Lambda}$ can be obtained from the training data using Matlab's `copulafit` routine.

If a test vector $z \in r$, the copula density $c(u|r)$ should be high, otherwise it should be low. For gender prediction, since there are only two classes, we have one CFeature vector. We choose pairwise associations instead of higher dimensional associations

due to the consideration of the curse of dimensionality and error propagation. In practice, an "outlier" feature should not be associated with too many other features. To further reduce the redundancy and error propagation, a d-prime method [18] is used for feature selection:

$$d_k^{'} = \frac{\mu_k^m - \mu_k^f}{\sqrt{\left[(\sigma_k^m)^2 + (\sigma_k^f)^2\right]/2}}, \tag{2.12}$$

where $\left(\mu_k^m, \mu_k^f\right)$ and $\left(\sigma_k^m, \sigma_k^f\right)$ are the mean values and standard deviations of the distributions of the k-th CFeature given male (m) and female (f) classes, respectively. The d-prime value should be high when the two distributions are well separated. Only those CFeatures that are well separated between the two classes are selected. We set an empirical value $d^{'} = 0.2$ for all CFeatures in our experiments. The same procedure is used to select the CFeatures for weight prediction.

Our approach can be summarized in the following steps:

1. Input: Consider a set of training samples (feature vectors). Each sample is a B-dimensional feature vector (MFeatures).

2. Estimation: We first select K pairwise features, using d-prime method, from each training sample. Note that the maximum value of K is $\binom{B}{2}$. These pairs are used to construct K bivariate copulas. The copula parameter matrices are estimated for each class using the training set (see Eqn (2.11)). We represent this as $\hat{\Lambda}^r = \hat{\lambda}_1^r, \ldots, \hat{\lambda}_K^r, r = 1, \ldots, M$. (See below, the procedure for feature selection to reduce dimensionality).

3. Transformation: Given a test sample that has to be classified, we transform it into $M-1$ K-dimensional CFeature vectors $c(u|r)$, $r = 1, \ldots, M-1$ using the estimated parameter matrices (see Eqn (2.9)).

4. Classification: We concatenate these $(M-1)$ CFeatures with the original MFeatures and input the combined feature vector (MFeatures + CFeatures) to a SVM (that has been trained using training samples) for classification.

2.6 Experiments

The CAESAR 1D database [1] contains 1119 male and 1250 female subjects, and 43 measurements for each subject, after removing missing data. It also contains two major attributes which we considered as ground truths: gender and weight. 500 randomly selected males and 500 randomly selected females are included in the training set and the rest are included in the test set. For statistical validation, the experiment is repeated in a Leave-T-Out manner 50 times with replacement. Here T is the size of the test set. For classification purpose, we use lib-SVM classifier [22] with RBF kernel (Eqn 2.13):

$$Ker(v_1, v_2) = exp(-\frac{||v_1 - v_2||^2}{2\gamma^2}), \tag{2.13}$$

where v_1 and v_2 are the feature vectors, and γ is the width of the basis function. For gender prediction, the parameters we used are $C = 2000$ for soft margin [33], and $\gamma = 0.0001$ for the width of the basis function. For weight prediction, we used a nu-SVR regression scheme with the default parameter setting from lib-SVM ($C = 1$, $\gamma = 1/d$), where d is the number of features.

We separate the 43 measurements into two clusters: body cluster and head cluster. The reason for such a separation is to investigate if the measurements in the body cluster and head cluster can be independently used to predict soft biometrics, and to determine their variability in prediction performance. In practice, one of them may not be available. For instance, a webcam surveillance system may not be able

to capture the face (head) information, prompting the use of body information for prediction.

The features in the two clusters are non-overlapping (see Table A.1). That is, in each cluster, there is no measurement containing information or partial information from the other cluster. For the head cluster, the sitting height and sitting eye height are further combined into a single measure calculated as **(sitting height - sitting eye height)**. As a result, we have 35 measurements in the body cluster and 6 measurements in the head cluster.

We first observe how the prediction performance without CFeatures is affected by noise. Recalling that our features have been normalized to the range $[0, 1]$, we use a threshold $R = 3*\text{StdDev}$ to describe the Gaussian noise; this means 99.7% of the noise value will be in the range $(-R, R)$. We use 7 different R values: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6. In this representation, $R = 0$ means no noise is added and $R = 0.6$ means a large Gaussian noise, which could be up to 60% of the maximum feature value or more, is added.

Figure 2.8 shows the results for gender prediction using MFeatures. Here, performance is measured using the misclassification rate.

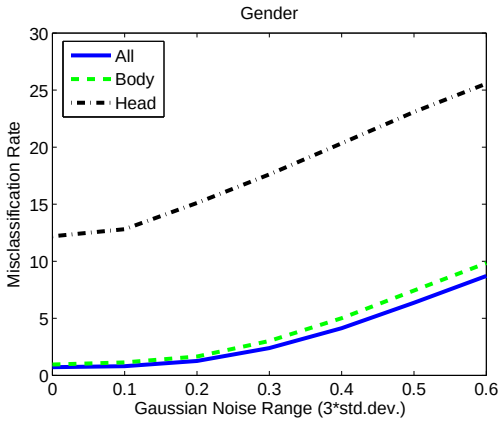


Figure 2.8: Misclassification rate (%) for gender prediction at various noise levels using MFeatures.

Figure 2.9 shows the results for weight prediction using MFeatures. The performance is measured using the mean absolute error (MAE).

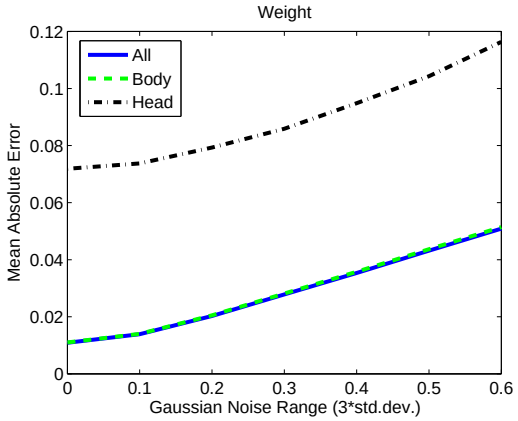


Figure 2.9: MAE for weight prediction at various noise levels using MFeatures.

Now we consider the pairwise CFeatures. In this work, an Archimedean copula with Clayton generator is used (Eqn (2.10)). Compared to the more complex student t copula, using Archimedean copula is computationally more efficient with almost the same classification performance. Compared to the Gaussian copula, the Archimedean copula yields more stable results, since the normality assumption is not always satisfied in practice. Based on the feature selection process described previously, the approximate number of CFeatures selected for gender and weight prediction is described in Table 2.2. The exact number varies depending upon the training set used.

The results for gender prediction using the copula model are shown in Figure 2.10. Corresponding results for weight prediction are shown in Figure 2.11. The results show that the CFeatures lead to a significant improvement in performance, especially when the noise is severe. We can also observe that not every possible pair of features are selected by the copula model (see Table 2.2)

To study the effect of the number of MFeatures in the copula model, we manually divide the 43 MFeatures into 3 categories by their measurability ranks (Table A.1). The rank 1 features are usually 1D measures and are larger compared to other fea-

Table 2.2: The number of MFeatures and CFeatures (numbers may vary depending on different iterations) selected by the proposed algorithm in our experiments.

		All	**Body**	**Head**
Gender	#MFeatures	43	35	6
	#CFeatures	169	165	3
Weight	#MFeatures	43	35	6
	#CFeatures	220	194	2

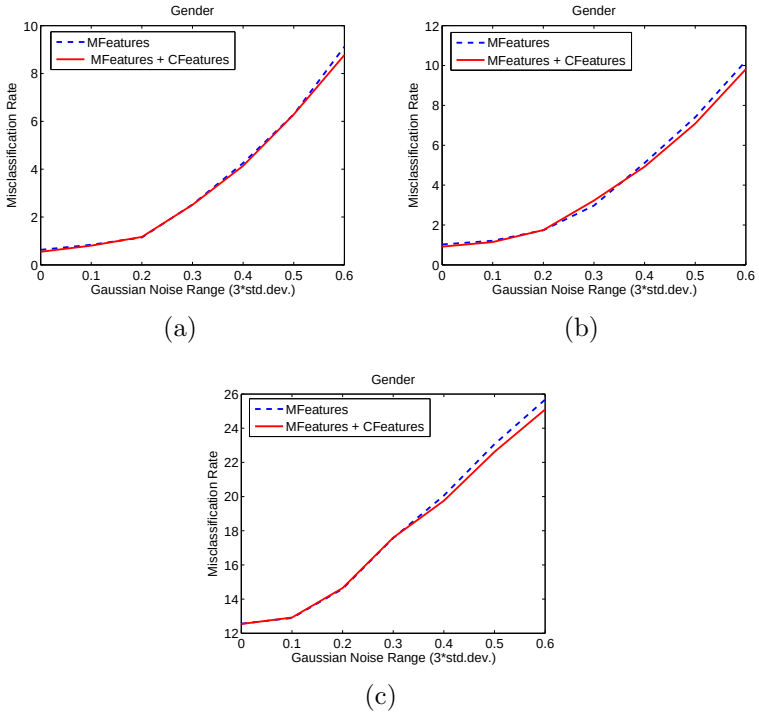


Figure 2.10: Gender prediction using: (a) all features; (b) body features only; (c) head features only.

tures. These include measurements such as stature and shoulder breadth. They are relatively easy to extract in practice. The rank 2 features are usually 2D measures such as chest circumference or head circumference. The rank 3 features are usually hard to extract automatically, such as hand circumference, triceps skinfold, or foot length. There are 25 features in rank 1, 10 in rank 2 and 8 in rank 3. We first randomly select 5 MFeatures from rank 1, and use the copula model to generate the corresponding CFeatures which are then input to the SVM. Next, we randomly select 10 features from the set of rank 1 features. We repeat this (i.e., increment number of

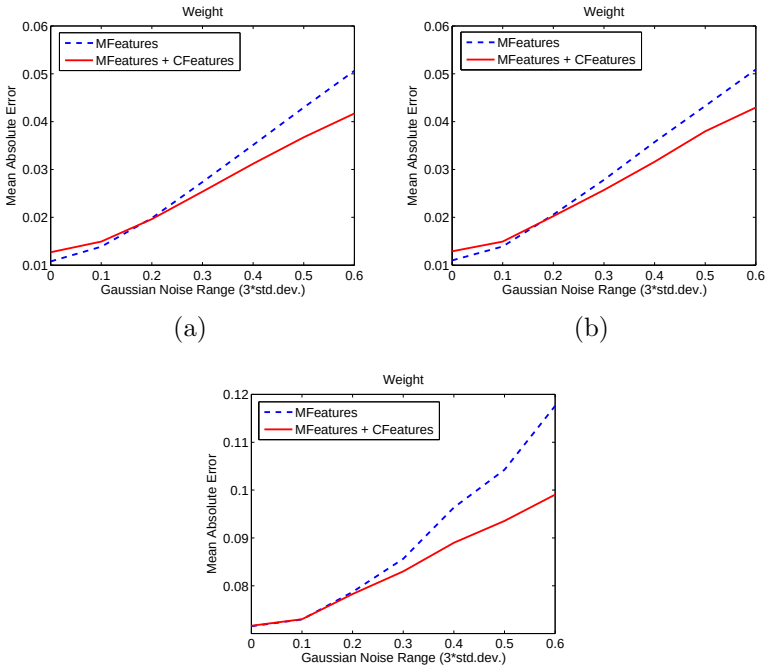
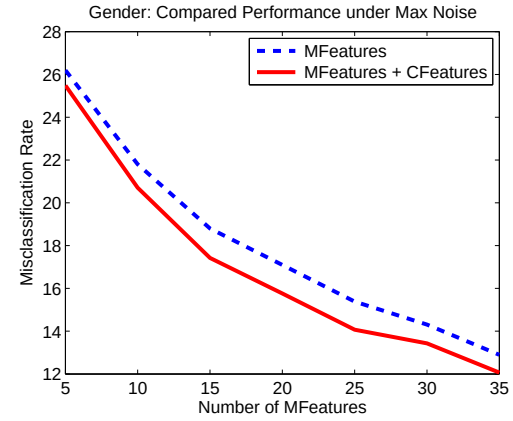
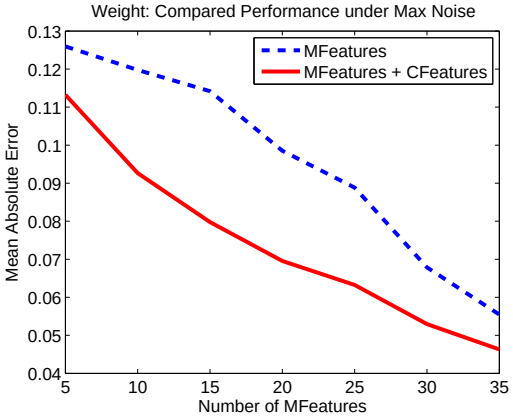


Figure 2.11: Weight prediction using: (a) all features; (b) body features only; (c) head features only.

features in steps of 5) until all 25 rank 1 features are selected. The next 5 features are then chosen from rank 2, until all 10 rank 2 features are selected. Each step is repeated 100 times with replacement for cross validation. In this experiment, we choose the first 500 males and 500 females in the database for training, and used the rest for testing. We do not use rank 3 features in this experiment, since they are usually difficult to automatically extract. Note that for every set of MFeatures selected, the corresponding set of CFeatures are computed using the proposed algorithm. Figure 2.12 shows the results under maximum noise ($R = 0.6$). We see that the performance of the copula model is affected by the number of features, especially for weight prediction. Although the copula model generally leads to a better performance, the maximum difference in performance relative to the original measurements (MFeatures) occurred when the number of MFeatures is moderate (between 10 to 25).



(a)



(b)

Figure 2.12: The performance using the copula model (MFeatures + CFeatures) under maximum noise: (a) gender; (b) weight. Notice the significant differences between the two plots in each figure around the regions with number of MFeatures between 10 to 25

Table 2.3: Comparison of recent studies on gender prediction

Data	#Subjects	Method	Perf.	Ref.
FERET	2,409	Raw Pixels+Adaboost	7.0%	[7]
FERET	994	SIFT+Bayesian	16.3%	[149]
LFW	7,443	LBP+Adaboost	5.6%	[138]
MUCT	276	Facial Metrology	13.2%	[18]
CAESAR	2,369	Metrology+Linear Reg.	9.0%	[4]
CAESAR	2,369	Metrology+SVM	0.6%	Proposed

2.7 Discussion

Due to the effectiveness of using copula to characterize the interaction between variables, our model combines the characteristics of copula and SVM to boost the gender and weight prediction performance on contaminated data. The results show that the body measurements (without head measurements) can provide comparable performance when using all the measures, including head measurements. The head cluster has lower performance, which is reasonable, since it only contains 6 features. When predicting with no noise or errors, we obtained a 0.63% misclassification rate for gender using all measures, a 0.93% error rate using only the body cluster, and a 12.6% error rate using only the head cluster. The above results are based on MFeatures only. Table 2.3 shows a comparison with other recent studies on gender prediction. Compared to previous work, precisely measured full body metrology (with no error or noise) provides a significant performance improvement in gender prediction.

For weight prediction, we have a 0.0108 mean absolute error (MAE) using all measures, a 0.0110 MAE using only body cluster and a 0.0719 MAE using head cluster only. To our knowledge, very little prior work has been done on automated weight prediction. One related work by Velardo and Dugelay [153] showed a 5%$\sim$10% error rate on weight estimation, using 7 measurements from the NHANES [57] database. Adjeroh et al's work [4] showed a 0.028 MAE using all measures. Table 2.4 sum-

Table 2.4: Comparison of recent studies on weight prediction

Data	#Subjects	Method	Perf.	Ref.
NHANES+Self-made dataset	28,000	Metrology+Linear Reg.	5%∼10%	[153]
CAESAR	2,369	Metrology+Linear Reg.	0.028 MAE	[4]
CAESAR	2,369	Metrology+SVM	0.0108 MAE	Ours

marizes these results on weight prediction. However, the impact of noise was not systematically considered in these previous studies.

The benefit of using the copula model is rather evident. Our study shows that the impact of Gaussian noise is moderate, even under severe contamination. Note that for the head cluster, the performance drops faster than the others: gender prediction drops by 13.11% from the no-noise case to the highest noise case. Compared to the head cluster, the body cluster shows a 9.04% drop in gender prediction, while using all-measures shows an 8.48% drop in gender prediction. This implies that the extra information provided by the association structure compensates for the effect of noise in the measurements.

Ignoring the minor fluctuations due to randomness, the results show that using the CFeatures has a positive impact, especially when the test set is severely contaminated. Among the three clusters, the impact of CFeatures is most obvious when using the head cluster (which has the least number of features). In our study, the Gaussian noise is "evenly" assigned to all test features. Under this condition, one can argue that the class-conditional probability density is unlikely to result in incorrect class prediction unless both features are significantly shifted in the same "direction" into an adjacent class due to noise. This property makes CFeatures less sensitive to noise than MFeatures. If the error bound for certain measures are known, we could pair features having low error bounds with features having high error bounds, so that the impact of error can be further mitigated. This issue warrants a more careful further study.

Chapter 3

Face Metrology for Classification[1]

In this chapter, we investigate the question of whether face metrology can be solely used for gender classification. A full automated prediction model is developed and the comparative performance is demonstrated.

3.1 Introduction

There are a number of studies that suggest significant variations in facial features among genders and/or ethnic groups. Farkas et al. [52] report that, compared to North American whites, Singaporean Chinese, Vietnamese and Thais have a wider mandible in both males and females. Turkish males and females have greater biocular width. Asian groups have wider nose width in both genders. Their study also showed some evidence of differences between gender for a given ethnic group. For example, the face width of Iranians and Turks was found to be similar to those of North American whites in males only. Greek females have greater nose height. Face width was similar for North American caucasians and Singaporean Chinese females; however in Singaporean Chinese males the face width was wider. Ibrahimagić-Šeper et al. [77]

[1]Part of the work reported in this chapter has been published in the following paper: Deng Cao, Cunjian Chen, Marco Piccirilli, Donald Adjeroh, Thirimachos Bourlai, and Arun Ross. Can facial metrology predict gender? In *IJCB*, 2011.

reported that male and females are significantly different in various anthropometric dimensions such as Frontotemporal-Frontotemporal width, Gonion-Gonion width, Zygion-Zygion width and Trichion-Gnathion length. That is, males have larger face dimensions than females. Note that, for the purpose of this study, the subjects were sampled from a population of Zenica in Bosnia and Herzegovina. Osunwoke et al. [119] reported 7 facial dimensions that are significantly different between gender for Bini adults in Nigeria. These dimensions include face length and width, nose length and width, bigonial breath, lip width and menton-subnasal length. Zhuang et al. [167] proposed a multivariate analysis method to investigate possible anthropometric differences among gender, ethnicity, and age groups using 18 facial measurements, including height, weight and neck circumference, collected using traditional anthropometric techniques. Their results showed statistically significant differences in facial anthropometric dimensions between males and females.

Ferrario et al. [55] considered a database of images with 51 females and 57 males. The subjects were all young white Caucasian dental students aged 20-27. Twelve (12) anatomical landmarks are directly traced on the face of each subject with a black eye-pencil, then all subjects are photographed with a standardized technique for frontal views of the face. After that, a set of 22 standardized points (manual landmarks) is traced on all images by the same operator. Even though gender classification was not the main objective of the study, the paper showed that a global facial shape difference exists between the genders. For example, the male faces tend to be generally more rectangular, while the female faces tend to be more square-like.

3.2 Related Work

Humans perceive gender not only based on the face, but also on the surrounding context such as hair, clothing and skin tone [96, 19], gait [95] and body [19, 4]. Below, we provide a review on gender classification using face images.

The problem of gender classification based on human faces has been extensively studied in the literature [114, 7]. There are two popular classifiers. The first was proposed by Moghaddam et al. [114] where a Support Vector Machine (SVM) is utilized for gender classification based on thumbnail face images. The second was presented by Baluja et al. [7] who applied the Adaboost algorithm for gender classification. Recently, due to the popularity of LBP in face recognition applications [5], Yang et al. [162] used LBP histogram features for gender feature representation and the Adaboost algorithm to learn the best local features for classification. Experiments were performed to predict age, gender and ethnicity from face images. A similar approach was earlier proposed in [146]. Other local descriptors have also been adopted for gender classification. Wang et al. [154] proposed a gender recognition method using Scale Invariant Feature Transform (SIFT) descriptors and shape contexts. Once again, Adaboost was used to select features from the SIFT and shape descriptors and form a strong classifier. A recent overview on the topic of gender classification from face images can be found in [104]. Among appearance-based descriptors that encode gender information such as LBP [146], SIFT [149] and HOG [19], the LBP has been observed to exhibit better discrimination capability while maintaining simplicity [104].

Geometry features have also been used as *a priori* knowledge to help improve classification performance [133, 161]. Gao and Ai [62] performed face-based gender classification on consumer images acquired from a multi-ethnic face database. To overcome the non-uniformity of pose, expression, and illumination changes, they proposed the use of Active Shape Model (ASM) to normalize facial texture. The work

concluded that the inclusion of ethnic labels can help improve gender classification accuracy in a multiethnic environment.

There are a few approaches focused explicitly and solely on facial metrology as a means for gender classification. Fellous [54] investigated the gender classification problem using a set of 109 frontal images from 52 females and 57 males selected from the ARPA/ARL and FERET databases. Each image has 256×256 pixels with 255 gray levels. 40 landmarks are manually extracted from each face image. Based on these landmarks, 24 horizontal and vertical distances were calculated. Metric information from the distances is used for gender classification and yielded a 90% accuracy on a test set consisting of 57 images from 26 females and 31 males. Based on these studies, one may be tempted to argue that gender information is embedded in the landmark coordinates. Given the above results, one can pose a related question: do pseudo-landmarks also contain reliable information about gender? In fact, pseudo-landmarks are often used for registration purposes [9] in traditional face recognition. The location of pesudo-landmarks may vary from user to user, or may vary in each acquisition by the same observer. Thus, it is possible that pseudo-landmarks may not be discriminative enough to clearly distinguish between gender. Burton et al. [17] used an image data set with 91 male and 88 female faces, and computed various geometric distances and ratios using key points in the images, including 3D distances derived by a combination of full-face and profile images and used a discriminant function for performing gender classification. The method however could not approach human performance in gender discrimination. In a similar work, Edelman et al. [46] attempted to use neural networks and facial subregions for gender classification. Their method achieved an accuracy of 66 - 78% on a database of 160 facial images (80 male, 80 female).

In our work, we take a more comprehensive look at the explicit use of facial geometry in solving the problem of gender classification. We use solely metrological

information based on landmarks, which may or may not have an anatomical underpinning. In our approach, a combination of local information from a few landmarks is used, rather than holistic information from all landmarks. To establish a base-line for comparison with appearance-based methods, we use LBP in combination with SVM to predict gender from face images.

We only consider face classification problems in this chapter. General face recognition techniques are reviewed in [166, 165]. We consider our work to be more closely related to earlier research by Shi et al. [141, 142] on face recognition using geometric features, where they used ratio features computed from a few anatomical landmarks.

3.3 Facial Landmark Categories

Following Shi et al.[142], we can divide facial landmarks into three broad categories: Type 1: anatomical landmarks; Type 2: manual image-based landmarks; and Type 3: automatic landmarks. Face based anatomical landmarks are biologically meaningful points defined as standard reference points on a human face or human head. They can be located by careful inspection and palpation, and can be traced on the skin using eye-pencils [55]. Thus, Type 1 landmarks involve face-based physical measurements or markers on a real face using specialized measuring devices, such as tapes and calipers. While they tend to be more abstract than other features of the skull (such as protuberances or lines), anatomical landmarks are considered very important in various scientific fields including cosmetic surgery, anthropology, and forensics [53, 51]. However, the main issue with the anatomical landmarks is that their manual extraction requires expertise and subject cooperation. Furthermore, the number of anatomical landmarks that can be extracted from a human face is rather limited. Thus, the exclusive use of anatomical landmarks in face recognition is not usually recommended.

Type 2 landmarks are landmarks manually extracted from 2D face images, or 3D face data. Measurements on these landmarks are not necessarily performed directly using tapes and calipers on the human body. They are marked according to certain physical properties of a human face, such as the contour of the face, eyebrows, mouth, etc. Compared to Type 1 landmarks, there are several factors that can lead to additional errors in detecting Type 2 landmarks. Typical examples include face distortion, pose, illumination, expression, and observer differences in manually localizing these landmarks. Even for the same face image, there could be variations if landmark annotation is repeated multiple times by the same observer, or recorded by different observers. Given these problems, Type 2 landmarks are generally less accurate and less consistent than Type 1 landmarks.

Type 3 landmarks are landmarks obtained by automatic techniques, using geometrical or mathematical properties of face images, such as extrema points and edges, or through the use of certain machine learning algorithms, such as ASM [30]. To the best of our knowledge, automatic landmarks are currently not as accurate as manual landmarks [111], and at times could have significant location errors. In computer vision research, manual landmarks are often considered as baseline ground truths for evaluating the precision of automatic landmarks.

Collectively, we refer to Type 2 and Type 3 landmarks as pseudo-landmarks, since their locations may not necessarily coincide with those of anatomical landmarks. Most landmarks used in computer vision are, however, pseudo-landmarks. Compared to anatomical landmarks, they are either already available in the database (for example, XM2VTS database [108] and MUCT database [110] contain 2D face images with manually annotated landmark points), or can be acquired using automatic detection or manual annotation.

3.4 Benefits of Facial Metrology

There are many advantages to the use of facial metrology. These include (i) *Memory Management*: compared to texture-based information from face images, landmarks require much less storage space, since we only need to store the coordinates of a limited number ($\sim 10^2$) of landmarks; (ii) *Information Privacy*: unlike the full face image, landmark information can be safely stored, transported, and distributed without potential violation of human privacy and confidentiality; (iii) *Prediction of Missing Information*: topological features (face coordinates) can be either global or local to specific facial regions. Thus, missing information can be approximately predicted, for example, using statistical approaches [4]; (iv) *Law Enforcement*: useful information from facial metrology could be used as forensic evidence in a court of law, where admissibility of quantifiable evidence is a major consideration.

3.5 Gender Classification via Facial Metrology

3.5.1 Facial Landmarks in Databases

Since our work on facial metrology is based on facial landmarks, two well-known databases with manually annotated landmarks were used, viz., MUCT [110] and XM2VTS [108]. The definition of the landmarks in MUCT is similar to the one used in XM2VTS, but face images in MUCT have 8 extra landmarks in the ocular region. For each subject in each database, only the first frontal face image and the corresponding landmark information were used for each subject. Figure 3.1 shows two sample faces with numbered landmarks, one from each database. The numbering system used in XM2VTS is the same as that of MUCT, except for the set of extra landmarks used in MUCT (i.e.,#69 - #76). More details on the databases can be found in the section on experiments.

(a) (b)

Figure 3.1: Sample faces with numbered manual landmarks from XM2VTS (a) and MUCT (b)

We first consider the spatial distribution of facial landmarks in the images in the databases. Such a distribution could shed some light on the potential of using landmarks for gender classification. Let n be the number of landmarks for each face. The k-th face F^k can be represented as a vector,

$$F^k = (x_1^k, y_1^k), (x_1^k, y_1^k), \ldots, (x_i^k, y_i^k), \ldots, (x_n^k, y_n^k) \tag{3.1}$$

where (x_i^k, y_i^k) is the Cartesian coordinate of the i-th landmark of F^k, $k = 1, 2, \ldots, M$, M is the number of faces.

Active appearance models (AAM) [107] can be used to automatically localize the facial landmarks. AAM imposes linear constrains on shape variation, and so an input face (or shape) can be represented as the linear combination of N base faces,

$$F = F_0 + \sum_{i=1}^{N} p_i F_i. \tag{3.2}$$

Here, F_0 is the mean face, F_i is the i^{th} base face, and p_i is the corresponding weight vector for this face. The texture is defined as the pixel intensities that are within the shape boundary. It can be defined as a vector of intensities $A(x)$:

$$A(x) = A_0(x) + \sum_{i=1}^{N} \lambda_i A_i(x), \tag{3.3}$$

where $A_0(x)$ is the mean texture and $A_i(x)$ is the i^{th} texture vector.

Unlike AAM, ASM seeks to only match the positions of the feature points, although some models may incorporate the texture information as well. Such a model is usually referred to as constrained AAM. The AAM fitting problem is usually defined by a cost function, which tries to minimize the following:

$$r(p) = \left[A_i(x) - A_0(x)\right]^T \left[A_i(x) - A_0(x)\right], \tag{3.4}$$

where, p are the parameters of the model, A_i is the i^{th} texture vector and A_0 is the mean texture. This classical optimization problem can be solved in an iterative manner. Matthews and Bakers [107] proposed a popular AAM fitting method within the framework of the Lucas-Kanade algorithm [99]. But it cannot generalize well on previously unseen subjects. Usually, the face detector is invoked first to provide a coarse location for the initialization of ASM, then the model would iteratively fit the face image until the convergence condition is satisfied.

In this work, we use the STASM library [111] for ASM fitting. Thus, we obtain automatic landmarks from MUCT and XM2VTS visible spectrum databases, and from WVUM Multispectral database [157]. Currently, the localization of feature points using ASM is not accurate for face images with large pose changes. Since our study is mainly constrained to near-frontal face images, the ASM is observed to localize features with good accuracy.

The illumination condition in the image may also affect the results of automatic landmark extraction. Several image pre-processing methods can be used to improve the image quality before automatic landmark extraction [140, 86]. However, these pre-processing techniques are not used for metrology-based features in this study, since our experimental analysis suggests that the quality of images is sufficient for automatic landmark extraction. However, before using the LBP method (for appearance-based gender recognition), histogram equalization was used to enhance the image contrast in the spatial domain.

3.5.2 Alignment and Normalization

The raw landmark coordinates are sensitive to translation, scaling, and 2D rotation caused by changes in the position and orientation of the camera, or of the subject (see Figure 3.2(a)). Thus, it is necessary to accurately detect the face [13] and pre-align it prior to landmark extraction. Alignment is done by applying an affine transformation to each face image, which moves the origin to the midpoint between the pupils, and rotates the image so that the line connecting the two pupils aligns with the horizontal axis. Also, the variation caused by scaling or distance to the camera is reduced by using the inter-eye distance as a reference measure (see Figure 3.2(b)).

After alignment, we determine if there are any significant difference in landmark distribution between genders. To this end, we first normalize each landmark $L_i^k = (x_i^k, y_i^k)$ as follows:

$$\hat{x}_i^k = \mu(x_i) + \alpha \left(\frac{x_i^k - \mu(x_i)}{\sigma(x_i)} \right) \tag{3.5}$$

$$\hat{y}_i^k = \mu(y_i) + \alpha \left(\frac{y_i^k - \mu(y_i)}{\sigma(y_i)} \right) \tag{3.6}$$

where α is a constant, $(\mu(x_i), \sigma(x_i))$ and $(\mu(y_i), \sigma(y_i))$ are the mean and standard-deviation of the i-th landmark for the x and y coordinates, respectively.

Figure 3.2 shows the resulting landmark distribution. The analysis is based on faces images in the XM2VTS database. Red crosses and red circles indicate the average landmark position across individuals for male subjects and female subjects, respectively. The blue and green scatter points are normalized landmark coordinates for each individual, with blue indicating male and green indicating female. As Figure 3.2 shows, some landmarks are significantly different between male and female subjects, while others do not exhibit much difference. Those normalized landmark positions with more separation between the male and female subjects are likely to lead to a better gender classification performance. We observe that when $\alpha = 5$, the normalization does not significantly affect the landmark distribution (see Figure 3.2(c)). As we reduce the value of α (e.g. at $\alpha = 1$), the male and female distributions start to become more clearly separated (see Figure 3.2(d)). At $\alpha = 0$, the results will correspond to the spatial distribution of landmarks for the average male face and the average female face (see Figure 3.2(e)).

3.5.3 Metrological Features

There are different ways to utilize the landmark information. We cannot directly use the landmark coordinates, since they will be sensitive to translation, scaling, and 2D rotation of face images. One could consider all distance ratios defined by sets of four landmarks, or triangular features defined by any three non-collinear landmarks [142]. The issue here is computational complexity. The dimensionality of the feature space will be $\Theta(n^4)$ for complete distance ratios, and $\Theta(n^3)$ for landmark triplets, where n is the number of landmarks. An alternative is to consider simple Minkowski distances between two arbitrary landmarks $L_i^k = (x_i^k, y_i^k)$ and $L_j^k = (x_j^k, y_j^k)$, given by:

$$D_{ij}^k = \left((x_i^k - x_j^k)^p + (y_i^k - y_j^k)^p\right)^{\frac{1}{p}}, \tag{3.7}$$

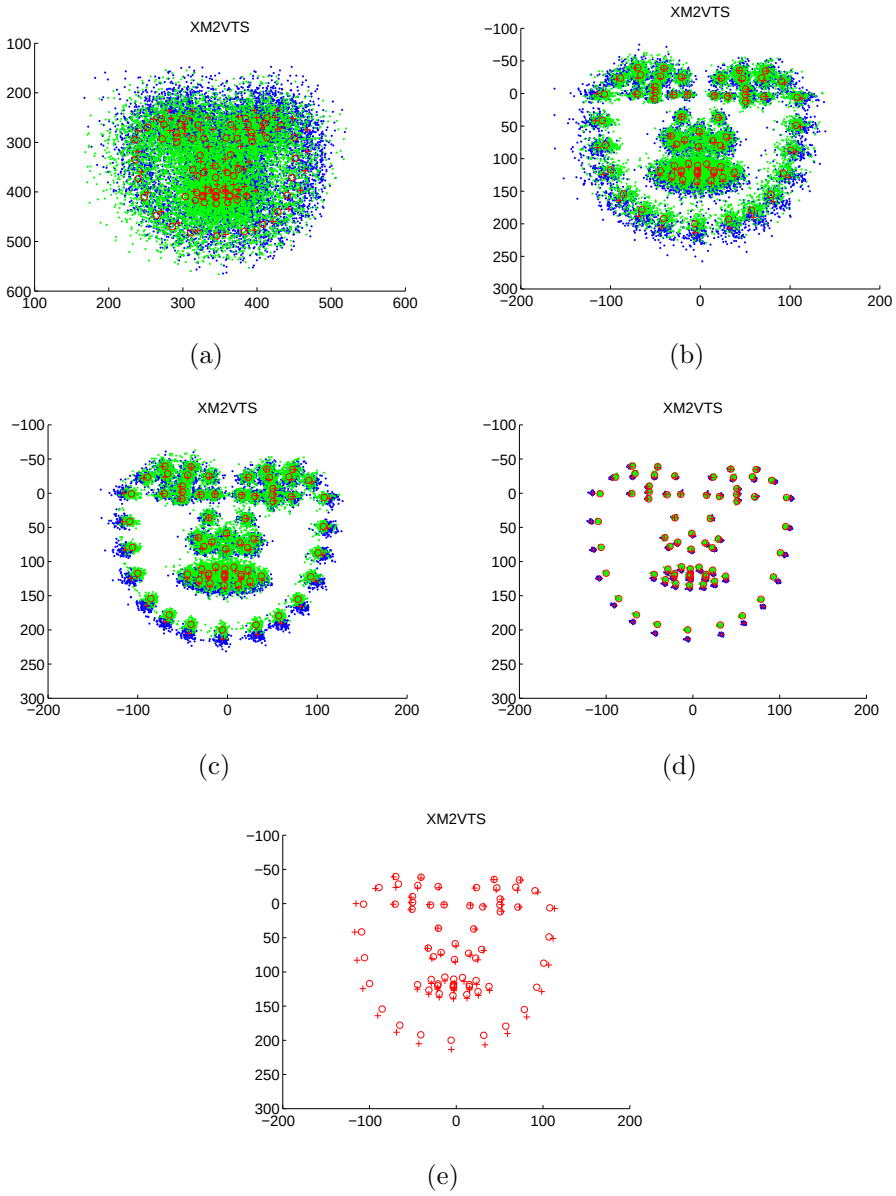


Figure 3.2: Spatial landmark distribution for the faces in XM2VTS database where x axis and y axis indicate the spatial coordinates. The red cross and red circle indicate average landmark positions across individual male and female subjects, respectively. The blue and green scatter points are normalized landmark coordinates for each individual: blue for male, green for female. (a): Without alignment; (b): After alignment; (c)-(e): After alignment and normalization, i.e. using α=5 (c), using α=1 (d), and using α=0 (e).

where p is the distance parameter. For a given p, the number of distances is thus $\Theta(n^2$). In this work we have considered the Euclidean distance ($p = 2$). The distances can be easily normalized to be scale-invariant by a reliable measure, such as inter-eye distance. The resulting ratios are also invariant to translation and 2D rotation. However, using only distance measures may not be reliable, since the orientation of the distances may be significant as well. For example, two individuals may have the same distance from the tip of the nose to the pupil, although one may have a longer face and the other may have more widely separated eyes. To improve the reliability of the features, we also use the horizontal angle subtended by each distance vector. The horizontal angle A_{ij}^k is computed from the pair-wise landmark coordinates:

$$A_{ij}^k = tan^{-1}\left(\frac{y_i^k - y_j^k}{x_i^k - x_j^k}\right) \tag{3.8}$$

3.5.4 Entropy Analysis

In order to assess the usefulness of individual landmark points, we appeal to the notion of measurement entropy. The entropy of a random variable $X = \{x_1, x_2, ...x_n\}$ with a probability mass function $p(x)$ is defined as

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i). \tag{3.9}$$

The joint entropy of a pair of random variables $X = \{x_1, x_2, ...x_n\}$ and $Y = \{y_1, y_2, ...y_m\}$ with a joint distribution $p(x, y)$ is defined by

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2 p(x_i, y_j). \tag{3.10}$$

The probability distributions are typically quantized to discrete values (bins) before the entropy is computed. In this work, we use 256×256 bins to compute the

Table 3.1: Summary statistics on the entropy (in bits) of landmark coordinates (C), entropy of Euclidean distances (D), entropy of horizontal angles (A), and joint entropy of distances and angels (DA).

Database	MUCT				XM2VTS			
Entropy	C	D	A	DA	C	D	A	DA
Mean	7.997	6.646	6.137	7.979	8.119	6.818	6.377	8.115
Max	8.065	7.123	7.065	8.072	8.149	7.160	7.154	8.170
Min	7.136	4.346	1.733	6.013	7.657	5.834	1.986	6.588
Std	0.143	0.300	0.950	0.125	0.081	0.119	0.860	0.065

joint entropy of pairwise landmark points, 256 bins to compute the entropy of Euclidean distances and the entropy of horizontal angles, and 256×256 bins to compute the joint entropy of distances and angles. The results (see Table 3.1) show that the entropy (in bits) associated with landmark coordinates is quite high, indicating that the landmark coordinates, or information derived from the coordinates could be the basis for discriminating between individuals, or groups of individuals. However, we may reasonably assume that the information also includes variations and possible perturbations caused by minor changes in pose, camera position, or expression, or by inconsistencies in localizing the landmark coordinates. These undesired effects are partially neutralized by transforming the landmark coordinates into Euclidean distances and horizontal angles. We notice that the average entropy of landmark coordinates is larger than that of distance, or angles, when treated individually. However, it is not larger than their sum, implying that the angles and the distances are perhaps capturing different types of discriminative information.

Figure 3.3(a)$\sim$3.3(f) show the joint entropies (H), the mutual information (I) and the absolute values of pairwise Kendall's tau coefficients (τ) for the x-coordinates and y-coordinates of the 76 manual landmarks in MUCT database, respectively. The results show quit high entropy levels and moderate mutual information. We also observe that the landmarks on face contour have relatively high correlation with the landmarks in mouth region. Also the landmarks in mouth region have relatively high correlation with each other. This might be caused by subjects' expression (neutral

or smile). Notice that all subjects have the same y-coordinates for the pupils due to the alignment.

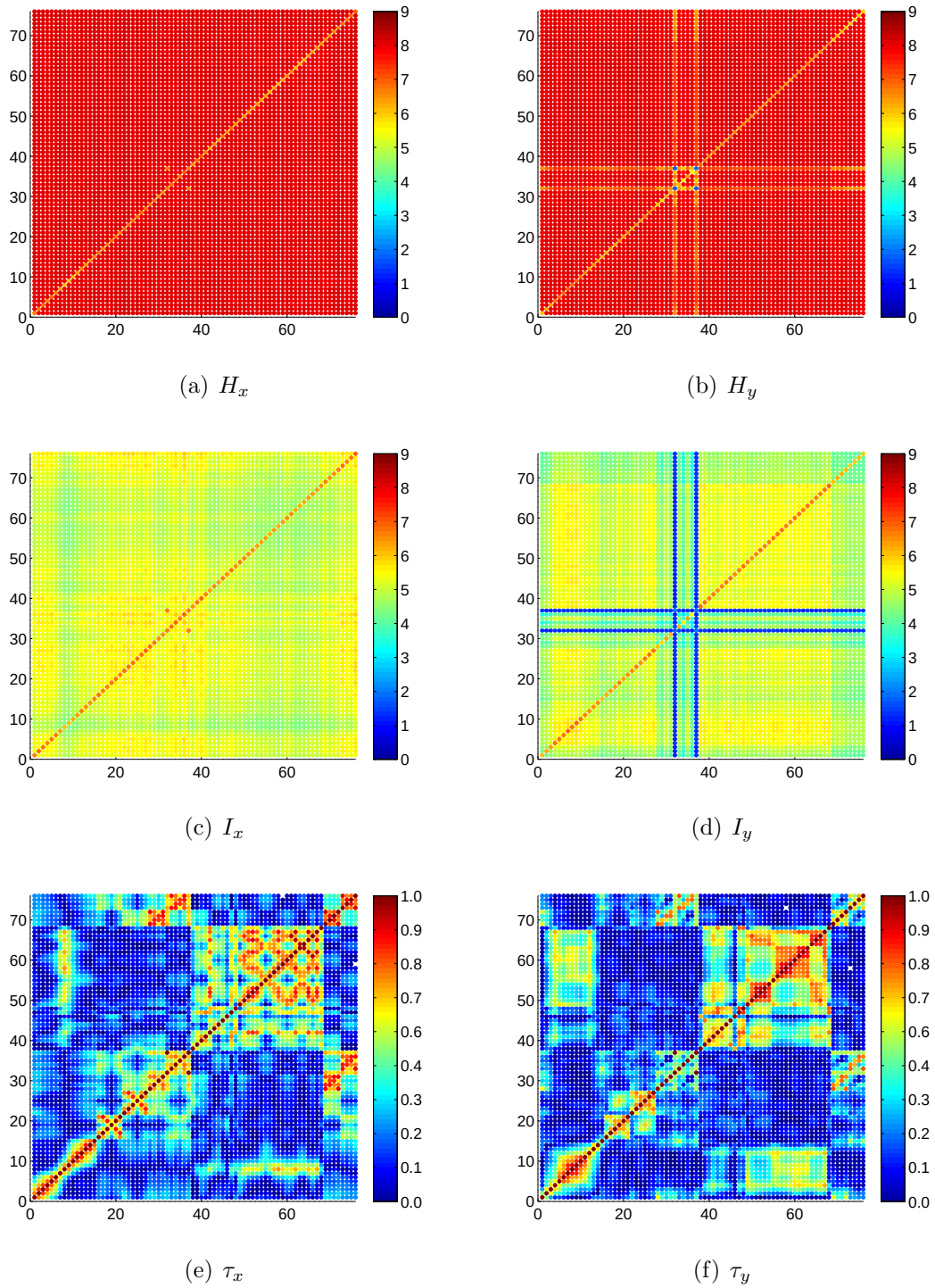
3.5.5 Feature Ranking and Selection

One problem in using the distances is the amount of computations involved. Using all pair-wise distances will lead to a very high dimensional feature space. For example, there are 5,700 features (distances and angles) for the MUCT database. Another problem is that the distance and angle measures may not always be robust. Some features may not be useful for gender discrimination and some others may be sensitive to errors caused by inconsistent landmark positions. Performance is expected to be compromised if such features are not removed or their impact is minimized.

To handle these issues, we apply a simple, yet efficient and robust (to outliers), d-prime-like scheme to rank the distances by their discrimination capabilities. For each pair-wise distance, across all the faces in our training set (see Section 3.7), we compute the d-prime as follows:

$$d'_{ij} = \frac{\mu(D_{ij}^M) - \mu(D_{ij}^F)}{\sqrt{([\sigma(D_{ij}^M)]^2 + [\sigma(D_{ij}^F)]^2)/2}} \tag{3.11}$$

where $\big(\mu(D_{ij}^M), \mu(D_{ij}^F)\big)$ and $\big(\sigma(D_{ij}^M), \sigma(D_{ij}^F)\big)$ are the mean values and standard deviations of the distance distributions between landmarks i and j, respectively, and M and F denote the gender. Similarly, we compute the d-prime-like value for each angular measurement. If the two distributions are well separated, the d-prime value should be relatively high. Otherwise, the measure results in a high inter-class error and should not be considered as a useful feature. The measures are then ranked in decreasing order based on their d-prime values, which corresponds to a decreasing order in their gender discrimination ability. Figure 3.4 shows two sample faces annotated with the Top-20 ranked Euclidean distances and horizontal angels (the angels



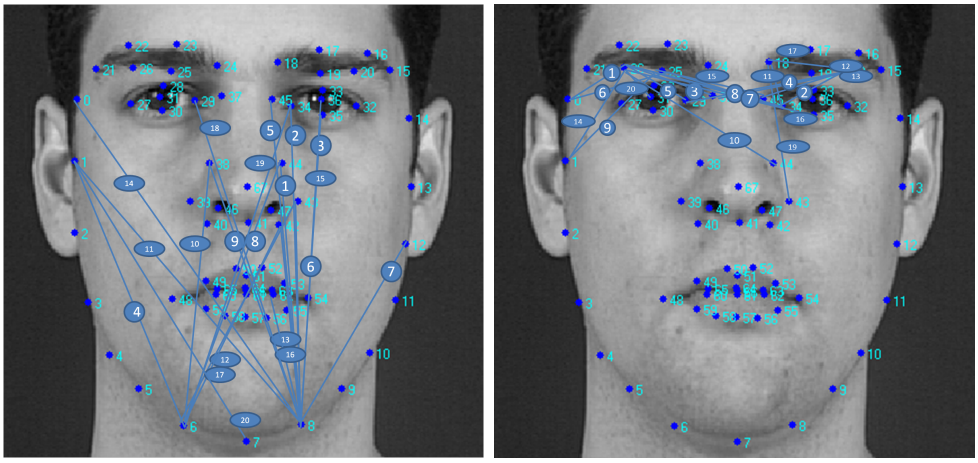
(a) H_x (b) H_y

(c) I_x (d) I_y

(e) τ_x (f) τ_y

Figure 3.3: The joint entropies (H), the mutual information (I) and the absolute values of pairwise Kendall's tau coefficients (τ) for the x-coordinates and y-coordinates of the 76 manual landmarks in MUCT database.

are represented by their corresponding distances). Our results show that only a few (generally less than one hundred) top-ranked features are needed for gender discrimination purposes. Note that the specific d-prime ranking for a given measurement could vary from database to database. However, the general trend is similar when both datasets are used (see also Table 3.2 in Section 3.7.3).



(a) (b)

Figure 3.4: Metrological features ranked by their discrimination abilities (Eqn (3.11). (a) Top 20 pairwise distances; (b) Top 20 horizontal angels. Sample faces are from the XM2VTS database. Numbers on the edges indicate the d-prime ranking.

3.5.6 SVM Classifier

We tested and compared results using three classifiers: support vector machine (SVM), k-nearest neighbor (KNN) and logistic models. We chose SVM for its superior performance and speed. For the SVM, we used the Gaussian radial basis function as the kernel:

$$K(u,v) = exp(-\frac{||u-v||^2}{2\gamma^2}) \tag{3.12}$$

where u and v are the feature vectors, and γ is the width of the basis function. We set γ=2. The SVM soft margin parameter C [33] is set to be 10.

The experimental results suggest that among the thousands of metrology-based features, only a few top-ranked features are significant for gender classification. However, the optimal number of features depends on the size and quality of the given database. More measures may not necessarily improve gender classification performance. Instead, too many measures may introduce more noise thereby compromising the performance (see Figure 3.7 in Section 3.7 on experimental results).

3.6 Gender classification via Appearance

To compare our results with state-of-the-art approaches, the use of appearance-based models for gender recognition was considered. In particular, we applied the LBP operator on the same datasets as we used for the study of metrological features. The basic LBP descriptor encodes micro-patterns of an image by thresholding 3×3 neighborhoods based on the value of the center pixel and then transforming the converted binary pattern sequence into a decimal value. It can be extended to accommodate neighborhoods of different sizes to capture textures at multiple scales [118].

To utilize LBP method for the extraction of gender features from facial images, the input image is first divided into non-overlapping blocks. Then, the spatial histogram features from each block is calculated and concatenated to form a global descriptor. Here, the LBP operator is denoted as $LBP^{u^2}_{P,R}$, where P refers to the number of equally spaced points placed on a circle with radius R and u^2 represents the uniform concept, which accounts for most of the patterns observed in the experiment. For instance, 11001111 is considered to be a uniform pattern since it contains no more than 2-bitwise transitions (1 to 0 and 0 to 1). When computing LBP histograms, every uniform pattern has a separate bin (58 bins in total) and all the other non-uniform patterns together have a single bin. In our experiments, the $LBP^{u^2}_{8,2}$ descriptor is used. The image is resized to 126×90, with each block consisting of 18×15 pixels.

The total number of blocks is, therefore, $7 \times 6 = 42$. For each block, we use LBP to extract 59 bin features, leading to a 2478-dimensional feature vector (see Figure 3.5). In order to handle lighting variations, histogram equalization is applied to reduce the illumination variation.

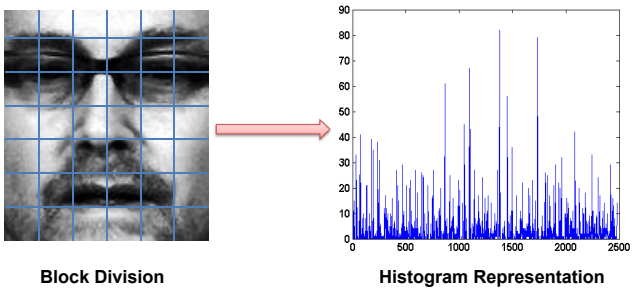


Figure 3.5: LBP gender feature representation. The face image is from MUCT database [110].

To design the gender classifier (i.e., predictor), we used the SVM. The SVM classifier is trained using a training set of labeled face images. The test sample is classified according to the sign of $y(s)$,

$$y(s) = w^T \phi(s) + b, \tag{3.13}$$

where $\phi(s)$ denotes the transformation of the original feature-space and b is the bias. w is the normal vector and determines the orientation of the hyper plane which is generated during SVM training. For classification, we use the histogram intersection kernel:

$$k(x, y) = \sum_{i=1}^{n} min(x_i, y_i), \tag{3.14}$$

where x_i and y_i are the i^{th} histogram bin for the feature vectors of x and y. The histogram intersection kernel was observed to be much more effective for classification than the linear or the RBF kernel when the LBP histogram features were used as input. Therefore, it is adopted in our appearance-based gender classification scheme.

3.7 Experiments

3.7.1 Datasets and Setup

The MUCT hand-marked face database with facial landmarks [110] was created by researchers to generate data exhibiting diversity in pose, illumination, age, and ethnicity. We use 276 subjects from Category-A, consisting of 131 males and 145 females. The first sample of each subject (the near-frontal face) is selected and used in our experiments. Therefore, a total of 276 samples is used. The spatial resolution of raw images is 480×640 pixels. Since the landmark positions for the eyes are provided, for the appearance-based LBP method, the images are normalized and aligned based on eye coordinates [42]. Further, for the LBP method, histogram equalization preprocessing is used to reduce the effect of illumination. The final cropped image size is set to be 130×150. For the LBP method, the images were resized to 126×90. Sample images (before and after normalization) are shown in Figure 3.6.

Figure 3.6: Sample images from the MUCT database. Face images in the bottom row correspond to the cropped and geometrically normalized images, after face detection on the top row images.

The XM2VTS database [108] has 295 subjects. Each subject has one sample selected. There are 160 males and 135 females. Similar to the MUCT samples, the size of the cropped sample images is also 130×150.

To perform gender classification on the MUCT database, we randomly selected 50 males and 50 females for training. The remaining 176 samples were reserved for testing. This partitioning exercise was repeated 50 times without replacement. For XM2VTS, the same experimental design was applied, except the total number of test samples in this case was 195.

For the metrology-based approach, the d-prime-like feature ranking is applied separately on both the distance measures and the angle measures. Thus, we use both top-ranked distances and top-ranked angles for the analysis.

3.7.2 Performance of Facial Metrology

Figure 3.7 shows the performance of the metrology-based method for gender classification, using the proposed metrology-based features from facial landmarks. The performance using distances and angles separately varied somewhat with the database. In most cases, the performance on MUCT database is slightly better than that on XM2VTS. Further, the distance measures performed generally better than the angle measures. However, fusing the distance and angle measures at the feature level generally improved classification performance, especially when the feature space is small (less than 80 features). We observe that the metrology-based system can provide good results with only a few landmarks (around 10), suggesting that there is a possibility of using a lower-dimensional space, and hence lower computational cost. However, the experimental results did not indicate whether there exists an optimal number or combination of features. Since a large feature space will not necessarily lead to superior performance, we selected only the Top-10 ranked distances and the Top-10 ranked angles for subsequent experiments [2].

[2]Ranking was performed based only on the d-prime formulation. Performance could be further improved using standard feature selection methods, such as sequential forward (or backward) feature selection [83, 126]. We did not apply any standard feature selection schemes for determining the optimal set of features. Such an experiment maybe conducted in the future.

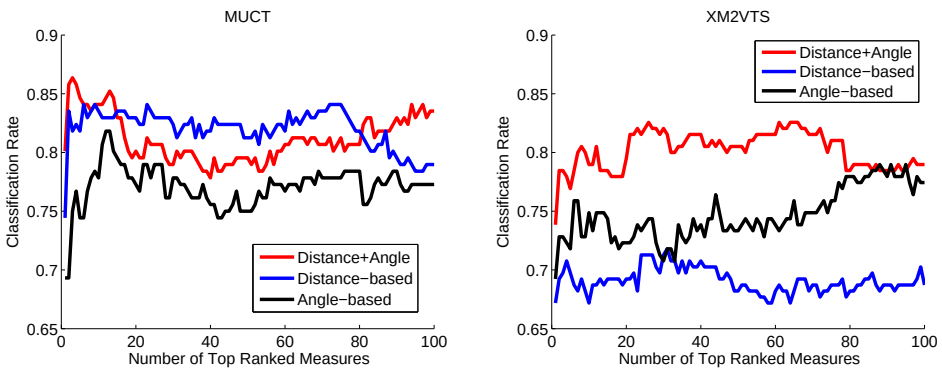


Figure 3.7: Performance comparison using the Top 1-100 angles, Top 1-100 distances and their fusion (2-200 features) in gender classification.

3.7.3 Landmark Discrimination Ability

The discussion so far has focused on the distance between pairs of landmarks, or the angles formed by such distance vectors. We also evaluated the discrimination capability of the *individual* landmarks. While we expect a landmark with a high discrimination ability to be involved in distance or angle measures with an equally high discrimination ability, this may not always be the case. An evaluation of the discrimination ability of individual landmarks is important in identifying landmarks that are major determinants of performance in a metrology-based method. Such landmarks can then become the focus of a more concerted effort at automated landmark detection. Consider the distances between landmark pairs as a matrix. To determine the discrimination ability of a single landmark, we simply compute the average d-prime value between that landmark and all the other landmarks. For the i-th landmark L_i, we have:

$$d_i^{'} = \frac{1}{N-1} \sum_{j \neq i} d_{ij}^{'}. \tag{3.15}$$

We call $d_i^{'}$ the marginal d-prime. We calculated the marginal d-prime values for distances and for angles separately.

Table 3.2 shows the Top-20 landmarks, ranked by their discrimination capability, as determined by their marginal d-prime values. Figure 3.8 shows an annotated view of the discrimination capability of the landmarks and their approximate regions on the face. Our results indicate that from a distance-based perspective, the landmarks on the face contour are crucial for gender classification. This is an important observation, especially given the original distribution of the landmarks on the face, as was shown in Figure 3.2. The landmarks in the eye region tend to have a low discrimination capability, perhaps due to the effect of the eyelids. The discrimination capability of landmarks in the nose region varied with different databases, probably due to the inconsistency in the annotation process. The landmarks in the mouth region also showed a low discrimination capability, because their positions are sensitive to the significant variability due to mouth expression. Overall, the top-ranking distances (with higher gender-discrimination ability) tended to be vertically-oriented measurements. Similar observations were reported in [55].

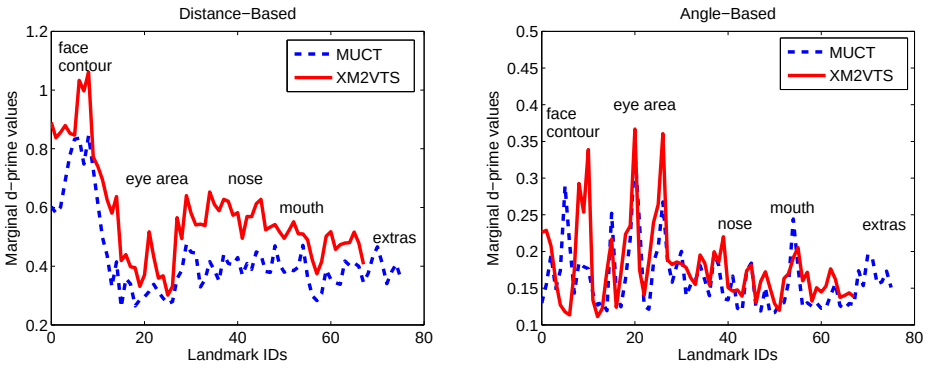
The angle-based marginal d-prime values are generally low, but they can still help in improving the gender recognition performance, as was shown in Figure 3.7. For the angular measurements, the landmarks on the face contour and in the eye region seem to be more significant than landmarks in other facial regions for the problem of facial metrology-based gender classification.

3.7.4 Comparative Performance

The experimental results show that facial metrology do have the potential to discriminate between genders. To place the results of the facial metrology-based approach in context, we compared it with the results obtained using an appearance-based method for gender identification.

Table 3.2: Top 20 landmarks ranked with respect to their discrimination ability using distance (D) and angle (A) measures.

	MUCT		XM2VTS			MUCT		XM2VTS	
Rank	D	A	D	A	Rank	D	A	D	A
1	9	21	9	21	11	2	38	11	19
2	7	6	7	27	12	12	71	12	40
3	6	27	8	11	13	30	3	35	16
4	5	16	1	9	14	55	72	30	3
5	8	55	4	26	15	49	28	15	56
6	10	22	3	10	16	71	9	38	38
7	4	20	5	25	17	45	25	46	35
8	11	7	6	20	18	31	46	13	55
9	1	26	2	2	19	32	30	39	28
10	3	31	10	1	20	38	35	45	30



(a) (b)

Figure 3.8: Discrimination ability of individual landmarks (based on marginal d-prime values), along with the approximate facial regions for the landmarks. (a) Distance-based; (b) Angle-based.

Table 3.3: Summary of the comparative performance results when using facial metrology (top 10 landmarks) and appearance-based models in gender classification. The summary statistics in this table are associated with the results in Figure 3.10.

	Metrology-Based		Appearance-Based	
Classification Rate	MUCT	XM2VTS	MUCT	XM2VTS
Mean	0.8683	0.8283	0.9063	0.8856
Max	0.9091	0.8718	0.9489	0.9282
Min	0.8295	0.7692	0.8750	0.8462
Std	0.0217	0.0251	0.0168	0.0191

Metrology-Based vs. Appearance-Based

As shown in Table 3.3, the current performance of the metrology-based approach is slightly lower than that of the appearance-based method. The major reason might be due to the limited nature of the information encoded in the landmarks, and the nontrivial human errors in the annotation process. Unlike in the classification of facial expression [101], we do not have prior knowledge about what local facial regions are most critical in determining gender. Yet, the performance of the metrology-based approach ($86.83 \pm 2.17\%, 82.83 \pm 2.51\%$) was only slightly inferior to that of the appearance-based method ($90.63 \pm 1.68\%, 88.56 \pm 1.91\%$) by about 3.8% for the MUCT database, and about 5.7% for the XM2VTS database. Also, compared to a 2478-dimensional feature space in LBP, the metrology-based method uses a 20-dimensional feature space. Thus the execution time at the test stage of our metrology-based method is lower than that of the LBP method: 0.02 ms vs. 1.8 ms per image for MUCT database, and 0.03 ms vs. 1.7 ms for XM2VTS database.

Manual Landmarks vs Automated Landmarks

Motivated by the results on the manually annotated landmark positions, we repeated our experiments on landmarks that were derived using automated methods, with no human intervention. Specifically, we generated facial landmarks in an automatic manner, using the ASM algorithm [111]. When using the automated method, we

have to first answer some some important questions, such as: How close are the automatic landmarks to manual landmarks? Do automated landmarks provide sufficient information for gender classification? The first question can be answered by computing the offset (in pixels) between the two types of landmarks. Here, we use the Euclidean distance between the positions of the automated landmark and the corresponding manual landmark as a measure of performance. As we can see from Figure 3.9, the offsets for faces in the MUCT database are relatively smaller than those from XM2VTS. This is not surprising though, since STASM used in the ASM algorithm was trained on MUCT data. The offsets for the XM2VTS dataset could go up to about 12 pixels. However, this result is still encouraging, especially considering that the size of the original image is 480×640 pixels.

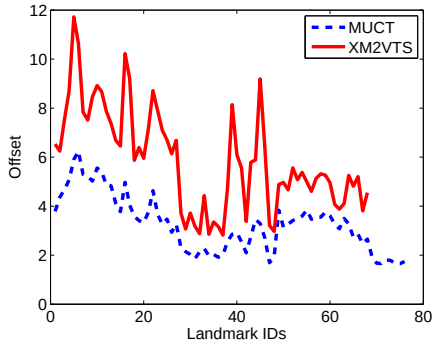


Figure 3.9: Average offsets in pixels between automated landmarks and manual landmarks.

The second question is answered by comparing the classification performance. The result shows that the performance drops by about 1% using automatic landmarks instead of manual landmarks on MUCT data, and by about 2% on XM2VTS data (Figure 3.10). We conclude that facial metrology can be used as a completely independent, yet fully automated, method for gender classification.

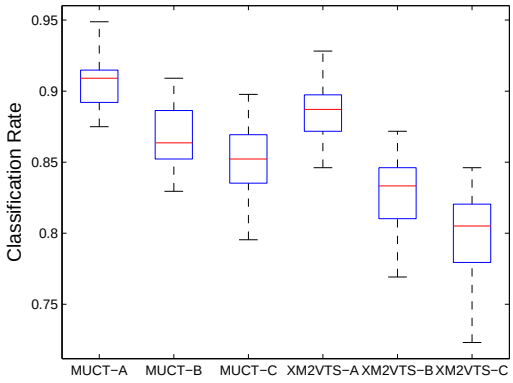


Figure 3.10: Box plots showing comparative performance of manual and automatic facial landmarks for metrology-based gender classification: A: Appearance based; B: Using manual landmarks; C: Using automatic landmarks.

Cross-Database, Cross-Spectra Considerations

To test the performance of our metrology-based method under cross-database and/or cross-spectra conditions, we use the WVUM Multispectral database [157] which contains 50 subjects, (30 males and 20 females). Each subject has 2 near-frontal multispectral facial images and 2 NIR images (see Figure 3.11) of size 1392×1040. A DuncanTech MS3100 camera was used to simultaneously capture four different wavelength bands (R,G,B and IR). We performed two experiments. The first experiment used all the images in the MUCT database (images in the visible spectrum) as the training set (145 males and 131 females), and then used multispectral images (100 images from 50 subjects) for testing. In our second experiment, we used all the images in the MUCT database as the training set, and the NIR images for testing. The two experiments thus show how the proposed metrology-based methods can work under cross-database considerations, where training is performed on one dataset, and testing is done on another dataset, This is more often the case in practice. The second experiment provides some idea on the performance of metrology-based methods on a different imaging modality, namely near-infra red (NIR) sources.

Table 3.4: Comparative performance on WVUM Multispectral Database

Multispectral	Features	Accuracy(All)	Accuracy(M)	Accuracy(F)
LBP+SVM	3717	83.00%	100%	57.50%
Metrology	20	83.00%	85.00%	80.00%
NIR	Features	Accuracy(All)	Accuracy(M)	Accuracy(F)
LBP+SVM	3717	82.00%	100%	55.00%
Metrology	20	87.00%	86.67%	87.50%

Table 3.4 shows the comparative performance on both tests. The facial-metrology method gives a well-balanced performance, while the performance of the LBP method appears to be highly biased towards male subjects. Perhaps, equally significantly, we can observe the superior performance of facial metrology over the LBP method, when testing on the NIR images. This is significant, which may be explained by the fact that, under NIR, the textural information that is exploited by the LBP are not as prominent as in visible light, while the facial metrology depends essentially on landmark points on the face, which are easier to acquire under NIR (see [130, 23]). We note that the results on cross-database and cross-spectra conditions are based on fully automated landmarks.

Figure 3.11: Sample images from the WVUM Multispectral Database. Top: Multispectral; Bottom: NIR

3.8 Discussion

The results show that facial metrology can indeed be used for gender classification. There are still several interesting open questions that need to be further studied. A key question would be how to improve the performance of the metrology-based method. How can the metrology-based method maintain its performance when confronted with increasing database sizes, and more variabilities in the face, say due to pose, expression, race, aging, etc? The above two questions might be effectively addressed by introducing a robust landmark detection technique, which can consistently localize the position of the landmarks under such variations. Another question has to do with the determination of the true capacity of facial metrology. To address this question we may need to consider a model with separable noise, for example, a 3D facial model that captures the structure of facial pose and expression at a more detailed level. The advantage of our proposed approach is that, due to its simplicity and independence, it could be combined with other more accurate (yet more computationally expensive) methods to improve the overall recognition performance. The performance difference in the challenging test on NIR using training data from other spectra and other databases makes this case more poignant. This work is a good starting point in addressing these questions, especially for gender classification, and perhaps for the more general problem of face recognition.

Chapter 4

Whole Body Metrology for Recognition

4.1 Introduction

In Chapter 2 and Chpater 3, We studied the problem of whether or not human metrology can be used for gender prediction. A more challenging follow-up question would be: can we do person recognition via human metrology? To answer this question, first we need to know whether or not human metrology is sufficiently different from person to person. Figure 4.1 shows the probability density histogram of pairwise Euclidean distances between 2,369 subjects (based on 43 manual measurements) in the CAESAR 1D database [1]. The figure suggests that most of the people are well separated in terms of such measures.

4.1.1 Remote Biometrics

Several different modalities such as fingerprints and iris have been used for person recognition. Although the current state-of-the-art fingerprints and iris recognition techniques have achieved very good accuracy [105, 20, 41], there is an explosion of

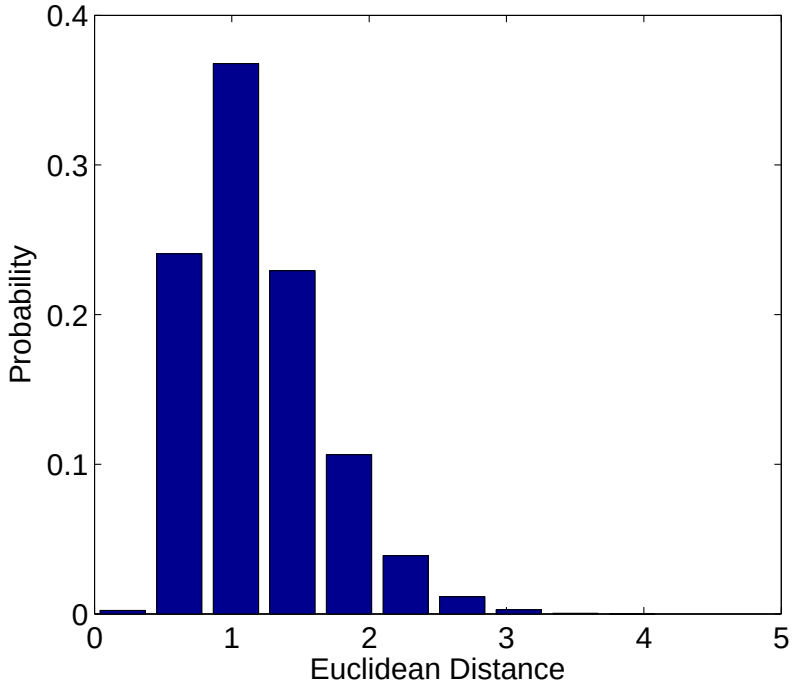


Figure 4.1: Probability density histogram of the pairwise Euclidean distance between 2,369 subjects in CAESAR 1D database. The measurements are normalized to range [0,1] using min-max normalization.

federal initiatives in the area of so-called remote biometric systems [45], which focus on person recognition in more challenging scenarios than traditional circumstances. In particular, remote biometric systems are expected to provide law enforcement and investigators the ability to ascertain the identity (1) of multiple people, (2) at a distance, (3) in public space, (4) without notice and consent, and (5) in a continuous and on-going manner [45]. Active topics in this area include gait recognition, face recognition at a distance, multi-biometric systems, etc. [148].

Unlike gait recognition [117, 74] or face recognition at a distance [156, 103], whole-body human metrology is a new modality that has not been well studied for person recognition. Our work is related to previous methods that have studied single view metrology [35, 68], session biometrics using height measurements [102], and human head or body shape analysis [6, 64, 16]. Our approach differs from the above studies in

two important aspects: (1) Instead of 3D points, we use only 1D measurements with potential advantages in terms of less computational requirements and affordability, and simpler automated extraction techniques. (2) A novel feature selection technique is developed to further reduce the number of required measurements while preserving the performance. This will lead to less computations at the matching stage. The new feature selection method can also be applied to other problems beyond human metrology.

4.1.2 A General Biometric Recognition System

Typically there are two different types of biometric recognition systems: *verification* systems and *identification* systems. A verification system is responsible for answering the question "Is he/she the person who he/she claims to be?", while an identification system answers the question "Who is this person?". Note that to answer the above questions, it is assumed that the person's identity has already been established and stored in a given database. Even though the verification problem is a one-to-one problem and the identification problem is a one-to-many problem, the solution to both problems relies on the same matching mechanism which measures the similarity between two individuals.

Similar to a prediction system, an end-to-end biometric recognition system consists of several stages. The first stage is **feature extraction**, in which a collection of biometric traits (features) is extracted from an individual. The second stage is **feature representation**. In order to characterize an individual, the raw features are transformed into a new feature space, which is expected to be more representative for further analysis. A common transformation is normalization, which will adjust the scale and location of each feature so that the contribution of each features to the final decision is comparable. By doing so, the biometric traits extracted from an individual can now be represented as a feature vector, e.g., $X = (x_1, \ldots, x_n)$ in

which every feature value $x_i, i = 1, 2, \ldots, n$ contains certain information about the individual. The third stage is **feature selection**. The goal of feature selection is to choose a minimal subset of the features, which will maintain or even improve the system performance when compared with using all features. Thus, in this stage, the feature vector $X = (x_1, \ldots, x_n)$ becomes $X' = (x'_1, \ldots, x'_m)$, where $m \leq n$. The next stage is the **matching stage**, in which the feature vector corresponding to a person of interest, say X' (sometimes called query), is compared against those in the given database (sometimes called templates) by using a matching score. For example, an Euclidean distance between the query and a template can be computed and used as a matching score. If the query and the template belong to to the *same* individual, the obtained matching score is called a *genuine* score. If the query and the template belong to two *different* individuals, the obtained matching score is called an *imposter* score. Figure 4.2 shows the distribution of genuine and imposter scores in CAESAR 1D database when all the 43 features are used ($m = 43$).

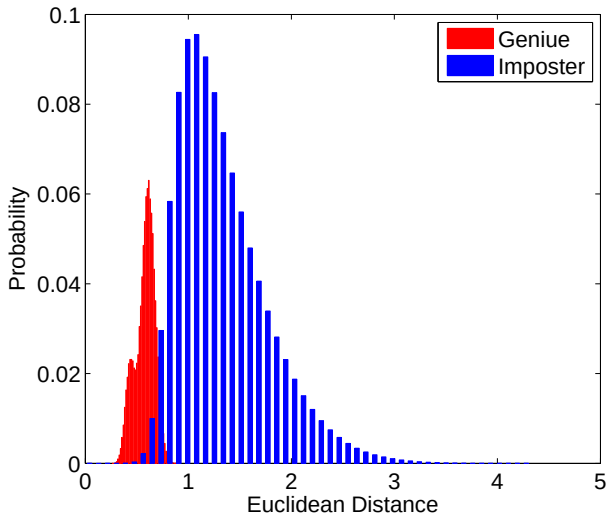


Figure 4.2: Distribution of genuine and imposter scores in CAESAR 1D database when all the 43 features are used ($m = 43$).

The final stage in recognition is the **decision stage**, where a person's identity is established based on the matching score. A threshold is used to decide whether to accept or reject a match. Two types of errors could occur: a query from the same individual as the template could be falsely rejected, or a query from a individual that is different from the template could be falsely accepted. In the context of biometric verification, the performance of the system can be determined by reporting its false accept rate (FAR) and false reject rate (FRR) at various thresholds. The plot that summarizes these error rates is known as the Receiver Operating Characteristic (ROC) curve. For an identification system with N enrolled identities, the output is a set of identities corresponding to the top t matches ($1 \leq t \leq N$). Similar to the verification system, The performance of the identification system is also established based on the match score, and the rank-t identification rate R_t for different values of t can be summarized using the Cumulative Match Characteristic (CMC) curve. In particular, the value of rank-1 identification rate R_1 is one of the most commonly used metrics to compare different biometric identification systems [131]. The detailed algorithms used at the different stages are described in Section 4.2.

4.2 Person Recognition via Metrology

4.2.1 Feature Extraction

As in Chapter 2, we do *not* focus on feature extraction in this chapter. We assume that subjects' features and their corresponding identities are already provided with sufficient accuracy. In particular, we use CAESAR 1D database [1] to generate our templates and queries. The templates are chosen from the 2369 subjects, in which each template consists of a collection of raw features (manual body measurements). In the original CAESAR database, each subject has one unique measurement set. The selected templates with one measurement set (as ground truth) per subject consist of

the training set (template set). In order to generate intra-class variation, independent random noise is simulated and added to each subject multiple times. The second dataset with artificial noise is then the test set (query set).

4.2.2 Feature Representation

In this study, a simple min-max normalization is applied to the new feature space:

$$x_i = \frac{x_i - min\{x_i\}}{max\{x_i\} - min\{x_i\}},$$

where x_i is the ith feature in the new feature vector $X = (x_1, \ldots, x_i, \ldots, x_n)$. We use min-max normalization for CAESAR database because (a) it is effective when the minimum and the maximum values of the data are known or can be estimated; and (b) it does not require the knowledge of the data distribution. Note that, min-max normalization is sensitive to outliers. Large outliers are not expected in our experiments due to the nature of the data. If large outliers are expected in the data, a more robust normalization technique, such as median normalization [78] or tanh normalization [82], can be considered as an better alternative to min-max normalization.

4.2.3 Feature Selection

Although only a limited number of measurements are available in our experiments, an appropriate feature selection stage is still necessary for the following reasons: (1) The complete feature set could be redundant due to the high correlation between feature components as we indicated in Section 2.3, and feature selection can not only reduce the computational complexity in the test process, but also reduce the workload during the feature extraction stage. (2)A high level of noise could be involved in practice. In that case, eliminating features with less information could lead to a more reliable

result. Also, a system that requires less features is potentially more robust to missing-data problem, and less computationally intensive at the matching stage.

In this work, a sequential forward selection technique is employed to perform feature selection [50]. That is, starting with no features in the model, we check the addition of each feature using a chosen model comparison criterion, add the feature (if any) that improves the model the most, and repeating this process until the number of desirable features is achieved, or no more feature improves the model. A key issue here is how to choose a reasonable and effective criterion. In the following sections, a number of feature selection criteria are evaluated and the reasons for choosing between different criteria are examined.

Equal Error Rate Criterion

For person verification, the Equal Error Rate (EER) criterion is commonly used in the biometric literature. However, it does not summarize the matching performance across all matching thresholds [131]. More importantly, it is prone to over-fitting the data [31]. In other words, it will often fit much better in training data than it does on test data that is from a different distribution. Similar problem occurs when rank-t identification rate is used as a selection criterion for person identification. For CAESAR database, the statistics of intra-class variation are not provided. To apply the EER criterion, the intra-class variation has to be artificially generated. However, in practice, as a remote biometric, it is expected that a person's metrological information will be extracted from a distance using surveillance cameras. Due to possibly complicated environmental factors and lack of prior knowledge, the intra-class variation caused by noise may remain unknown. As a consequence, the performance of the system could largely vary upon different test sets. Although EER and rank-t criteria may lead to very good accuracy, they are not particularly recommended for feature selection without intra-class information.

Entropy Criterion: Pairwise Joint Entropy

The over-fitting problem can be mitigated by constituting a more stiff selection criterion. We develop a feature selection criterion that does not rely on intra-class variation (noise) caused by complicated circumstances during the extraction stage. Inspired by the concept of information gain in decision trees [112], we view the problem as that of choosing a subset of given size m among n possible features ($m \leq n$), in which the inter-class information provided by these m features is maximized. We can establish such a criterion using the following algorithm:

1. **Initialization**: An empty pool $P = \{\}$ for containing selected features and a candidate list that includes all available features $C = \{x_1, \ldots, x_n\}$ are created. At the beginning, a feature with the highest entropy, or a pre-selected feature (e.g. stature) x_1 is removed from the candidate list and added into the pool. So we have $P = \{x_1\}$ and $C = \{x_2, \ldots, x_n\}$.

2. **Iteration**: for $i = |P| + 1$ to n, the next feature to be added into the pool, x_{add}, is chosen from the remaining candidate features in C:

$$x_{add} = \arg\max_{x \in C} H(f). \tag{4.1}$$

   Here, $H(f)$ is the joint entropy of all features in P plus the current candidate x_i from C:

$$H(f) = -\int_{\Omega} f(X) \ln f(X) dX, \tag{4.2}$$

   where $f(X)$ is the joint probability density function (pdf) of feature vector X, $X = P \cup \{x_i\}$; $dX = dx_1 \ldots dx_i$; and Ω is the region where the pdf $f(X)$ exists. The iteration runs $n - |P|$ times until all possible candidates in C are considered. x_{add} is then removed from C and added into P.

3. **Termination**: The above process repeats until there are m features ($1 \leq m \leq n$) in the pool or no more feature improves the performance, e.g., $P = \{x_1, x_4, x_{42}\}$ and $C = \{x_2, x_3, \ldots, x_{43}\} - P$.

One problem is that it is usually difficult to compute the joint entropy when $i > 2$ due to the complex or unknown form of $f(X)$. An intuitive solution is to compute average pairwise entropies among i features instead of the true joint entropy in the pool. The candidate that yields the highest average pairwise entropies is then selected. To apply this method, we can replace Eqn (4.2) in the above algorithm by:

$$H(f) = -\frac{1}{i} \sum_{j=n-(i-1)}^{n} \int \int f(x_i, x_j) \ln f(x_i, x_j) dx_i dx_j, \tag{4.3}$$

where $f(x_i, x_j)$ is estimated by splitting training data into a number of equal-area squares and counting number of features per square. The performance of a model based on such criterion is described in the experimental section.

The running time complexity of pairwise joint entropy (PJE) based forward selection can be estimated as follows: Assume the number of subjects in the training set is N, and each subject has n features. Starting from zero features in the pool, the forward selection tests $O(n)$ possibilities in each iteration, until there are $m(m \leq n)$ features in the pool ($O(m)$ iterations). For each candidate, the PJE between the candidate and every feature ($O(m)$) in the pool need to be considered. Assume a table of PJE for all features is given, the complexity of forward selection process is $O(nm^2)$. To compute the table of joint entropies, we estimate $f(x_i, x_j)$ by using a $b \times b$ (we assume $b^2 \sim N$) grid and counting number of features per square, which cost $O(N^2 + b^2) \sim O(N^2)$ time [32]. Since there are n^2 pairs of joint entropies, the cost of getting the table is $O(n^2 N^2)$. The overall time complexity of PJE based forward selection is $O(nm^2 + n^2 N^2)$. Compared to PJE criterion, the EER criterion requires c copies for each subject, thus the size of the training set becomes Nc. In each iter-

ation, we need to estimate the EER based on what is in the pool. The major cost involves computing Euclidean distances between every two training samples, which is $O(N^2c^2m)$ time. Thus the overall time complexity of EER based forward selection is $O(nm^2c^2N^2)$. This depends critically on the size of the database N, and will be much worse than PJE based selection when n is not very large.

As for the space complexity of pairwise joint entropy (PJE) based forward selection, assume that $O(b^2) \sim O(N)$ space is used for the grid when estimating the pairwise joint entropy; $O(N^2)$ space is used for the table that stores all pairwise entropies; $O(m)(m \leq n)$ space is used for the pool P and $O(n)$ space is used for the candidate list C, the overall space complexity of PJE based forward selection is then $O(N^2 + n)$. EER based forward selection will require more space, if we assume $O(N^2c^2)$ space is used to store genuine and imposter scores; $O(m)$ space is used when computing the Euclidean distance between two samples; $O(T)$ space is used to store FARs and FRRs under T different thresholds ($O(T) \sim O(N)$); and $O(m) + O(n) \sim O(n)$ space is used for the pool P and the candidate list C. The overall space complexity of EER based forward selection is then $O(N^2c^2 + n)$.

Entropy Criterion: k-Nearest Neighbor Estimators of Entropy

When we use pairwise joint entropy criterion, the mutual information among three or more features is ignored. In high dimensional feature space with possibly significant correlation between features, this method may not be able to capture the nature of the data accurately. Kozachenko and Leonenko [90] proposed a nearest neighbor estimator of entropy, given by:

$$\hat{H}_N = \frac{m}{N}\sum_{i=1}^{N} \ln \rho_i + \ln \frac{\pi^{m/2}}{\Gamma(m/2+1)} + \gamma + \ln(N-1), \tag{4.4}$$

where $\rho_i = min\|X_i - X_j\|, j \in 1, \ldots, N, j \neq i$ is the nearest neighbor of X_i and $\gamma = 0.5772$ is the Euler's constant. Singh et al. [143] extended the estimator using

k-nearest neighbors where k is a fixed integer (Eqn 4.5). Mnatsakanov et al. [113] studied k_n-nearest neighbor entropy estimators for a variable k_n that varies upon sample size N, as well as upon different data distributions. Since the above proposed estimators were shown to be asymptotically unbiased and consistent, it is reasonable to assume the entropy estimators to be more close to the true joint entropy and thus would be better alternatives than pairwise joint entropy criterion.

In this work, we first consider k-nearest neighbor estimator ($KNNE$) because it is a general form of nearest neighbor estimator, which does not require the knowledge of the data distribution. To apply $KNNE$, we replace Eqn (4.2) in our algorithm by

$$\hat{H}_k^N(f) = \frac{m}{N}\sum_{i=1}^{N} \ln R_{i,k,N} + \ln \frac{\pi^{m/2}}{\Gamma(m/2+1)} - \Psi(k) + \ln N, \tag{4.5}$$

where $R_{i,k,N}$ is the Euclidean distance from point X_i to its k^{th} nearest neighbor, $\Psi(k) = \Gamma'(k)/\Gamma(k)$ is the digamma function. The performance of a model based on such criterion is described in the experimental section.

The major cost of obtaining $\hat{H}_k^N$ is the computation of N terms of $R_{i,k,N}$'s, which is $O(mN^2)$. The overall time complexity of $KNNE$ based forward selection is $O(nm^2N^2)$. For space complexity, assume that $O(N)$ space is used to store distances between X_i and its neighbors when computing $\hat{H}_k^N$, and the remaining space requirement of $KNNE$ criterion is similar to that of PJE criterion. Thus the overall space complexity is $O(N^2 + n)$.

Entropy Criterion: Adapted k-Nearest Neighbor Estimators of Entropy

Although it is suggested that larger values of k are better for a more accurate estimator, in practice the choice of k is an issue. Singh et al. showed that the estimator achieved highest accuracy at $k = 4$ [143]. It remains unknown whether or not $k = 4$ is the best choice for every database. Considering that the choice of k could be af-

fected by fluctuations in the training data, we intend to design a more conservative estimator that would be more robust under small fluctuations. Note that in Eqn 4.5, the estimate of pdf $f(X_i)$ is given by:

$$\hat{f}(X_i) = \frac{k\Gamma(m/2+1)}{N\pi^{m/2}R_{i,k,N}^m}, i = 1, \ldots, N. \tag{4.6}$$

We can mitigate the fluctuation of the value of $f(X_i)$ by letting:

$$k = k_a = \arg\min_{k \in D} \|R_{i,k,N} - R_{avg}\|, \tag{4.7}$$

where $D = \{1, \ldots, K\}$ and $R_{avg} = \frac{1}{K}\sum_{k=1}^{K} R_{i,k,N}$. That is, to estimate $f(X_i)$ for a given feature vector X_i, the choice of k depends on the average distance between X_i and its first K nearest neighbors. This is significant, as previous approaches [90, 143] have chosen a single value of K. The kth nearest neighbor that is closest to the average of X_i's first K nearest neighbors will be selected for the calculation of the entropy estimator. We substitute k_a into Eqn 4.5 and obtain an adapted k-nearest neighbor estimator (K_aNNE):

$$\hat{H}_{k_a}^N(f) = \frac{m}{N}\sum_{i=1}^{N} \ln R_{i,k_a,N} - \frac{1}{N}\sum_{i=1}^{N} \Psi(k_a) + \ln \frac{\pi^{m/2}}{\Gamma(m/2+1)} + \ln N. \tag{4.8}$$

Note that the asymptotic unbiasedness and consistency of the k-nearest neighbor estimator does not change upon the k value. Since in the adapted k-nearest neighbor estimator, the k value for each feature vector is adaptively computed based on the average of first K nearest neighbors, the asymptotic unbiasedness and consistency of $\hat{H}_{k_a}^N$ still hold. In the experimental section, the performance of a model based on the adapted k-nearest neighbor estimator is compared with other models.

Since computing $\hat{H}_{k_a}^N$ does not require major extra cost than computing $\hat{H}_k^N$ (simply an extra $O(K)$ cost per iteration), the overall time complexity of K_aNNE based

forward selection is $O(nm^2N^2)$, and the overall space complexity is $O(N^2+n)$. When n is not large, $KNNE$ and K_aNNE criteria are comparable with with PJE criterion, but less expensive than EER criterion in terms of time complexity. PJE, $KNNE$ and K_aNNE have similar space complexity.

4.2.4 Matching

In this work, the matching score is defined by the Euclidean distance between the query vector X and a template vector Y. We use Euclidean distance because it is a simple, nonparametric merit that does not rely on specific data distribution. More importantly, Euclidean distance is commonly used as a matching score for other biometrics in the literature. Since an optimal matching score definition for human metrology is yet to be discovered, using a common performance merit makes our results more comparable to different biometric systems.

4.2.5 Decision

In this work, the performance of a verification system is determined by using the ROC curve. The ROC curve indicates the FAR's and FRR's at various thresholds. The FAR/FRR error rates that are most close to the EER lines are to be reported. For an identification system with N templates, the performance is determined by the rank-t identification rate R_t in the CMC curve. The values of rank-1 identification rate and T value when $R_T = 100$ are to be reported.

4.3 Experiments

We use a random subset of 100 subjects from CAESAR 1D database [1] as our training set. Each subject has 43 manual measurements that are assumed to be sufficiently accurate. We also assume that in practice, the measurements would be extracted

using long-distance security surveillance systems. Thus the subjects' weight feature is currently not included in the feature space. The test set is generated by adding independent random noise z_i to each *normalized* feature x_i from selected subjects, so we have $x_i \leftarrow x_i + z_i$. We generate 9 copies for each subject, so the size of the test set is 900 and the total sample size in the experiment is 1000. Two types of noise models are simulated in order to generate intra-class variations: (1) Gaussian noise $z_i \sim Gaussian(0, (0.2/3)^2)$; (2)Uniform noise$z_i \sim Uniform(-0.1, 0.1)$. We consider a relatively high (around 20%) noise level, because currently it is difficult to extract accurate body measurements from individuals at a distance. When we start forward feature selection, the stature is used as the first selected feature. Person verification and identification are separately considered in this work. Unless otherwise indicated, the experimental results shown below are based on the average of 10 repetitions with random choice of noise to insure reliable outcomes.

4.3.1 Verification

We compare verification system performance based on different feature selection criteria against number of features: PJE=pairwise joint entropy, $KNNE$=k-nearest neighbor estimator, K_aNNE=adapted k-nearest neighbor estimator. For PJE, we use a 64×64 grid when estimating the pairwise joint entropies. For K_aNNE, we use $K = 5$ to compute k_a in Eqn 4.7. Comparison of system performance based on different feature selection criteria against 10, 20, 30 and 40 features is shown in Figure 4.3 (Gaussian noise) and Figure 4.4 (Uniform noise). We observe that using 40 features, K_aNNE performance is very close to that of using all the 43 features with no feature selection.

When the number of selected features m is larger than 10, our experiments suggest that the adapted k-nearest neighbor estimator generally leads to more promising outcomes than other feature selection criteria. A more detailed report on the per-

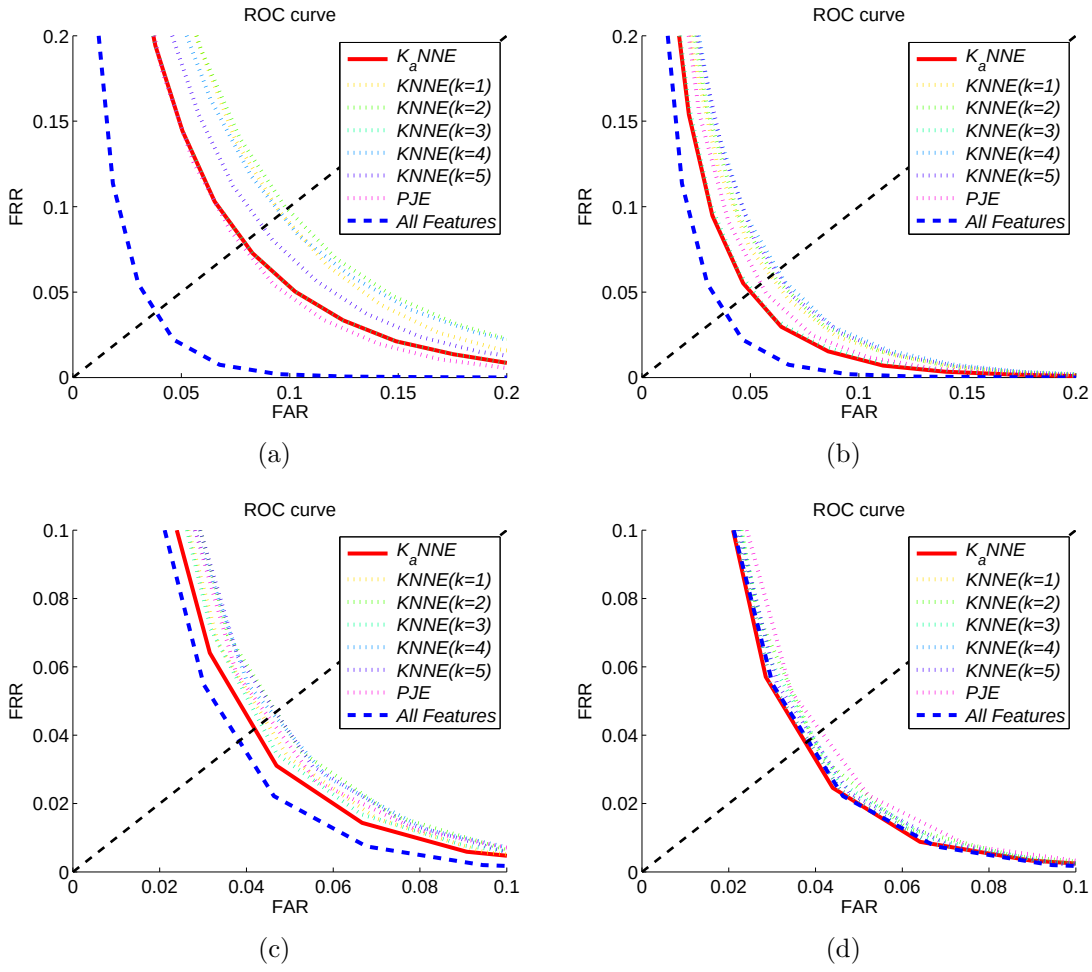


Figure 4.3: Comparison of verification system performance under Gaussian noise $Gaussian(0, (0.2/3)^2)$ based on different feature selection criteria against (a) 10, (b) 20, (c) 30 and (d) 40 features. The performance without feature selection (using all the 43 features) is shown in blue dash lines.

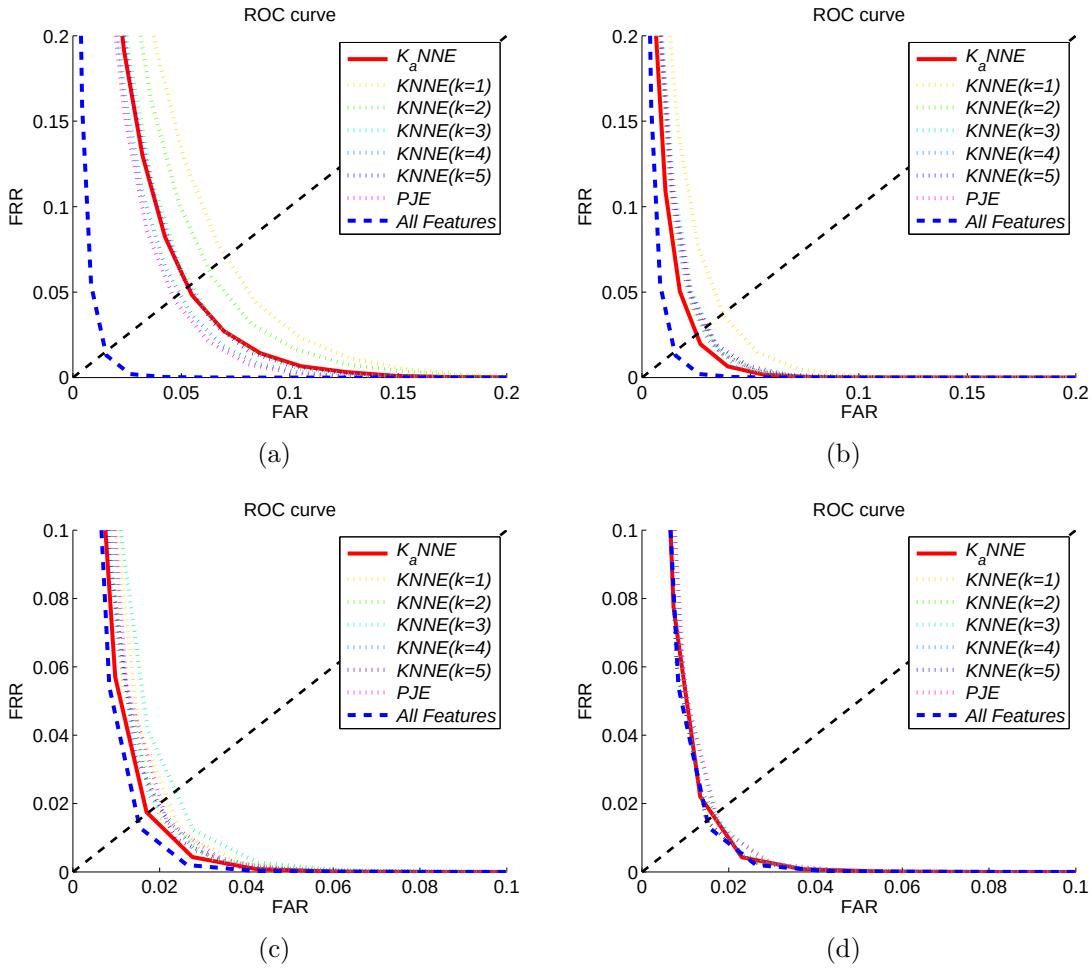


Figure 4.4: Comparison of verification system performance under Uniform noise $Uniform(-0.1, 0.1)$ based on different feature selection criteria against (a) 10, (b) 20, (c) 30 and (d) 40 features. As a reference, the performance without feature selection (using all the 43 features) is shown in blue dash lines.

Table 4.1: Comparison of verification system performance under (a) Gaussian noise and (b) Uniform noise based on K_aNNE against 10, 20, 30, 40, 43 features.

Noise	$Gaussian(0,(0.2/3)^2)$		$Uniform(-0.1,0.1)$	
#Features	Threshold	FAR/FRR	Threshold	FAR/FRR
10	20	0.09/0.09	17	0.05/0.05
20	17	0.06/0.06	13	0.02/0.04
30	16	0.05/0.03	12	0.02/0.02
40	14	0.04/0.04	11	0.02/0.02
43	14	0.04/0.03	11	0.01/0.02

formance of our verification system can be found in Table 4.1. The prioritization of the features based on K_aNNE in a forward selection scenario is shown in Table A.1. Note that the priorities of features may still vary upon different data sets.

In practice, some measurements may not be easily extracted. Thus it will be of interest to see how the system performance would be without the access to all measurements. Recall that in Chapter 2, we manually divide the 43 original measurements into 3 categories by their measurability ranks (Table A.1). There are 25 Category 1 features which are usually 1D measures and are larger compared to other features, such as stature and shoulder breadth. Compared to other measurements, it should be relatively easier to extract the category 1 features in practice. We test the system performance under the assumption that only Category 1 features are available. The outcomes under Gaussian noise $Gaussian(0,(0.2/3)^2)$ and Uniform noise $Uniform(-0.1,0.1)$ against 5, 10, 20, 25 features are shown in Figure 4.5. A more detailed report can be found in Table 4.2. The priorities of the features based on 25 Category 1 features are listed in Table A.1.

Compensating for Missing Measurements

Assume we have a system, in which there are more features available in the *training set* than in the *test set*. Can we utilize the information in the training set to compensate for the missing information in the test set? We know that a SVM prediction model

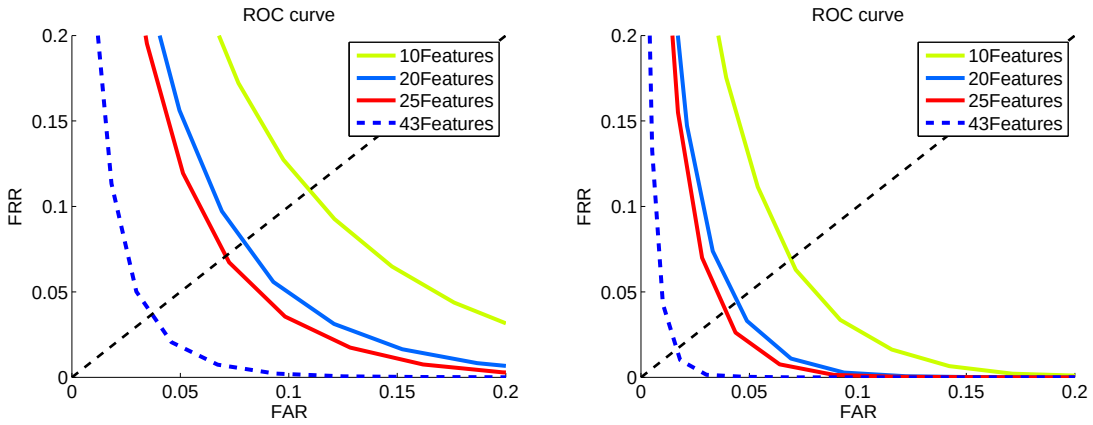


Figure 4.5: Comparison of verification system performance under (a) Gaussian noise $Gaussian(0, (0.2/3)^2)$ and (b) Uniform noise $Uniform(-0.1, 0.1)$ based on K_aNNE against 10, 20 and 25 Category 1 features. The performance when using all the 43 features is shown in blue dash lines.

Table 4.2: Comparison of verification system performance under (a) Gaussian noise and (b) Uniform noise based on K_aNNE against 5,10, 20, and 25 Category 1 features.

Noise	$Gaussian(0, (0.2/3)^2)$		$Uniform(-0.1, 0.1)$	
#Features	Threshold	FAR/FRR	Threshold	FAR/FRR
10	19	0.12/0.10	16	0.07/0.06
20	16	0.07/0.10	13	0.05/0.03
25	15	0.08/0.06	11	0.04/0.03
25+17 (predicted)	11	0.05/0.07	9	0.03/0.03

as described in Chapter 2 can be used to estimate the valued of the features that are not available in the test set. Assuming all the 43 features are available in the training set while only 25 Category 1 features are available in the test set, we compare the system performance between using 25 Category 1 features only and using 25 Category 1 features plus estimated 17 features. Each feature that does not exist in the test set is estimated by all 25 Category 1 features that do. Here, we use lib-SVM library with *nu*-SVR regression type and RBF kernel ($C = 1$, $\gamma = 1/d$, where d is the number of features. See Eqn 2.13) for prediction. The experimental results indicate that, if certain features are missing in the test set, using estimated features as

substitutions can improve the system performance. However, using estimated features does not yield better results than using actual features under a given noise condition in Figure 4.6.

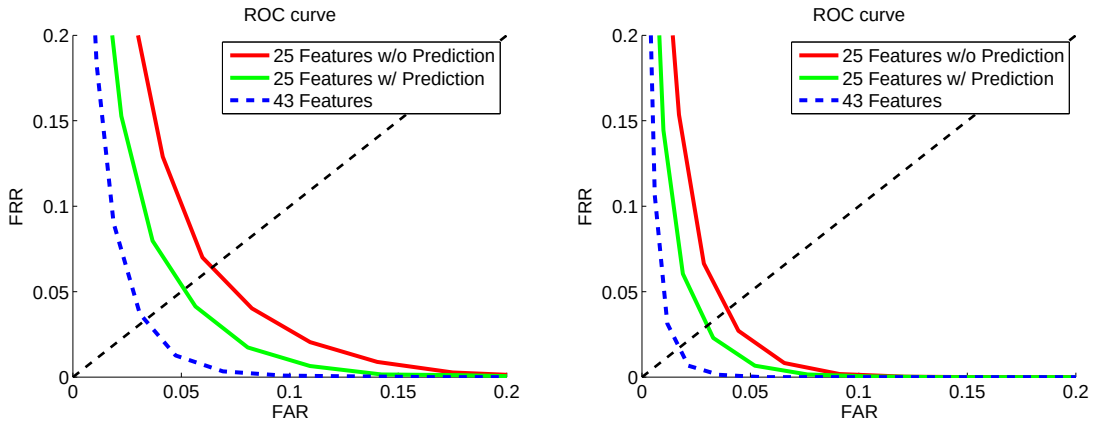


Figure 4.6: Comparison of verification system performance between with and without prediction under (a) Gaussian noise $Gaussian(0, (0.2/3)^2)$ and (b) Uniform noise $Uniform(-0.1, 0.1)$. The comparison is based on 25 Category 1 features. The performance when using all the 43 features is shown in blue dash lines.

Cross-Subject Verification Performance

In the above experiment, the subjects are the same between the training set and test set. We would like to see whether or not the K_aNNE criterion can generally lead to superior system performance on any subject. Another question is, if the prior knowledge of the inter-class variation is given, can we utilize it to achieve better accuracy? Both questions can be answered by choosing different subjects between training set and test set and adding EER criterion in comparing with other criteria. To apply EER criterion, the training set is generated using 100 subjects and 9 copies per subject for each subject with artificial independent random noise from certain distribution. The test set is generated using another 100 subjects and 900 simulated copies from same noise distribution. The experimental results are based on the aver-

age of 10 repetitions with random choice of test subjects. The comparison of system performance based on the different feature selection criteria against 10, 20, 30 and 40 features is shown in Figure 4.7 (Gaussian noise) and Figure 4.8 (Uniform noise). It is expected that EER criterion will lead to the best performance due to the ideal setup (training set and test set have identical intra-class distribution). However the experimental results do not always support such expectation. On the other hand, the K_aNNE criterion is still better than other entropy based criteria. For Gaussian noise, K_aNNE (with > 20 features) is better than using all original features. We can observe that under this more challenging test with more than 20 selected features, the proposed K_aNNE performs significantly better than using all the 43 features without feature selection.

4.3.2 Identification

In this work, an identification system is similar to a verification system. The verification system only verifies the identity of the query using one claimed template, while the identification system will compare the query with every template across the database. We use the same strategy to generate the matching score as in the above verification system. Under such condition, the CMC curve does not offer any additional information beyond the FAR and FRR information (ROC curve) [11]. Thus, the same framework for a verification system, including the proposed feature selection criterion K_aNNE, can be directly applied on the identification problem as well. Thus we test the performance of an identification system using the same selected feature sets based on K_aNNE (which have already been computed and tested in a verification system), as well as the same subjects and noise levels. The performance is measured using standard CMC curves when using 43 available features (see Figure 4.9). A more detailed report on our identification system can be found in Table 4.3. From Table 4.3 and Figure 4.9, we can see that using 30 or more selected features leads

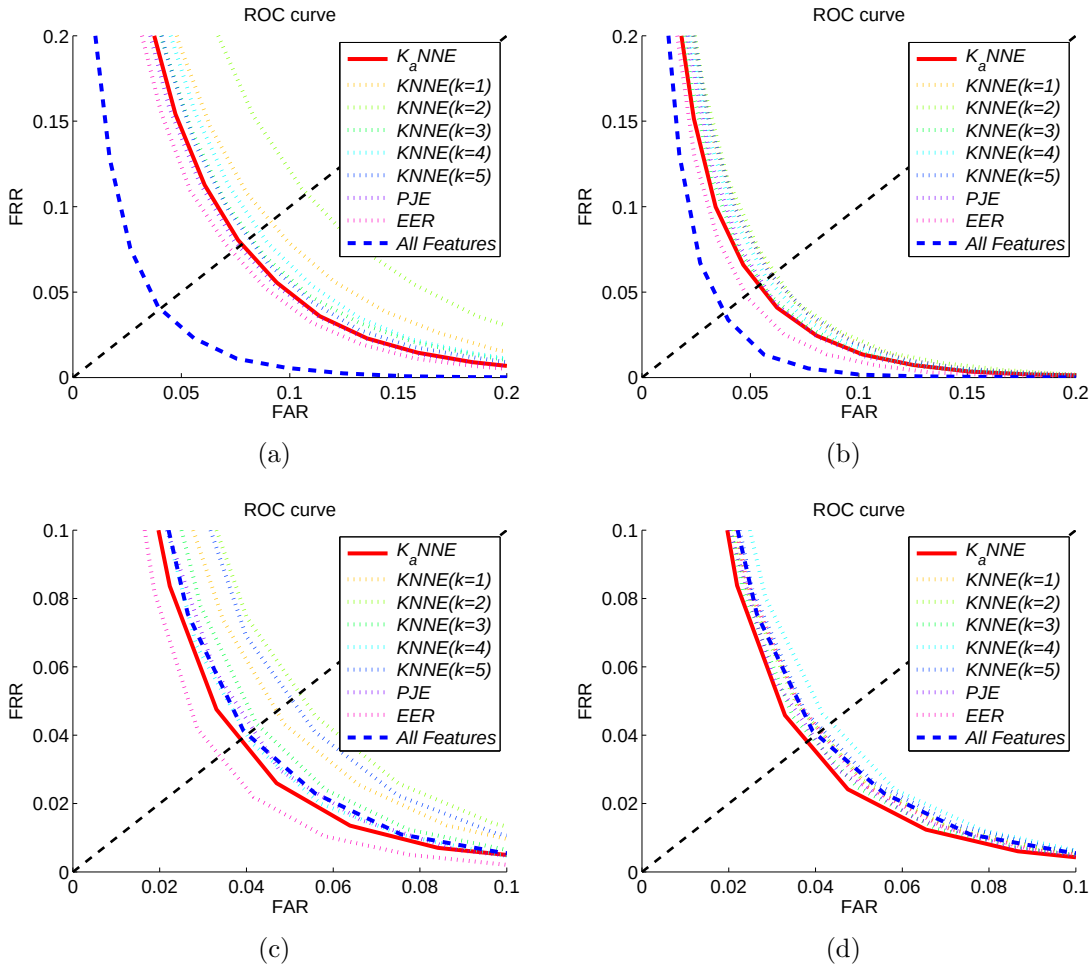


Figure 4.7: Comparison of cross-subject verification system performance under Gaussian noise $Gaussian(0, (0.2/3)^2)$ based on different feature selection criteria against (a) 10, (b) 20, (c) 30 and (d) 40 features. The performance without feature selection (using all the 43 features) is shown in blue dash lines.

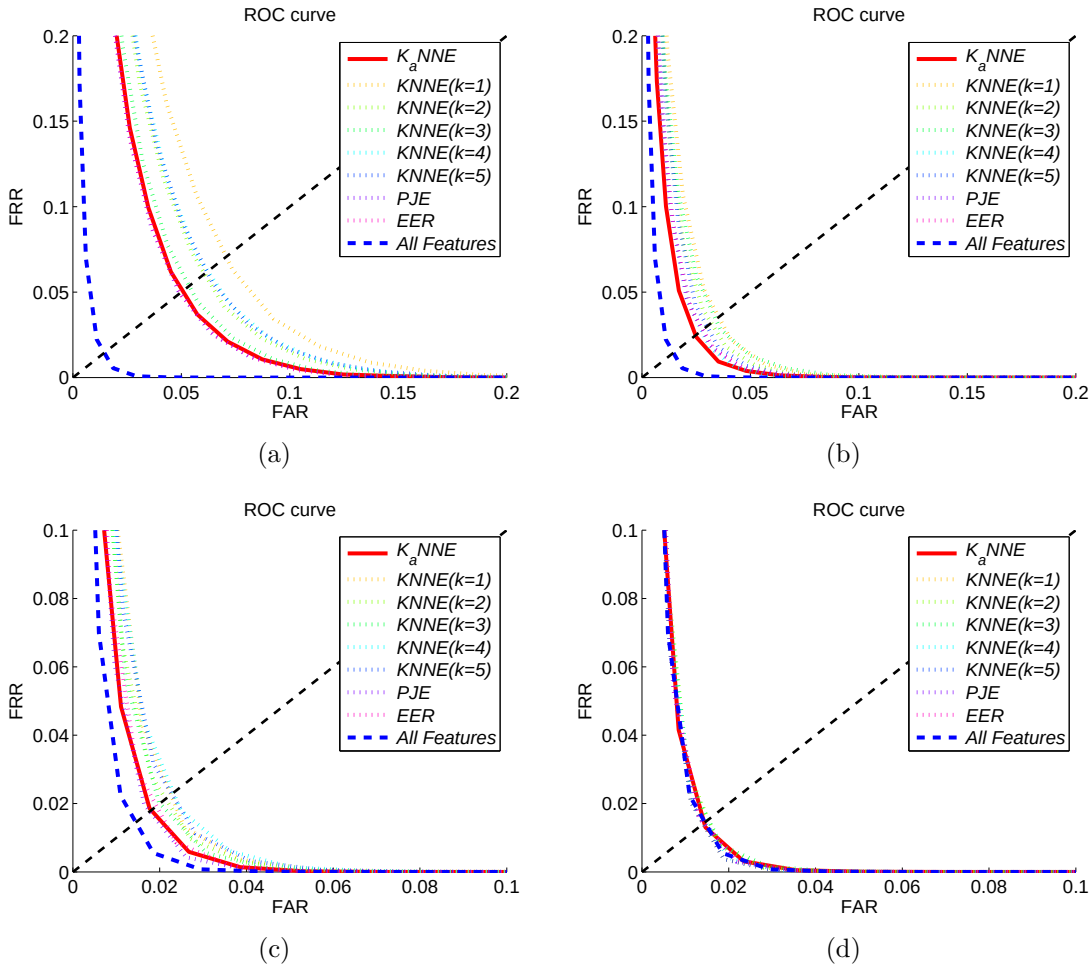


Figure 4.8: Comparison of cross-subject verification system performance under Uniform noise $Uniform(-0.1, 0.1)$ based on different feature selection criteria against (a) 10,(b) 20,(c) 30 and (d) 40 features. As a reference, the performance without feature selection (using all the 43 features) is shown in blue dash lines.

Table 4.3: Comparison of identification system performance under (a) Gaussian noise and (b) Uniform noise based on K_aNNE against 10, 20, 30, 40, 43 features. R_1 is the rank-1 identification rate, while T is the value when $R_T = 100$.

Noise	$Gaussian(0, (0.2/3)^2)$		$Uniform(-0.1, 0.1)$	
#Features	R_1	T (when $R_T = 100$)	R_1	T
10	82.98	26	89.98	8
20	96.67	10	99.07	5
30	99.11	3	99.78	3
40	99.62	3	99.91	2
43 (all)	99.82	2	100.00	1

to results that are very similar to using all the 43 features. This is significant given the expected improvement in computational cost at the matching stage, and hence system response time.

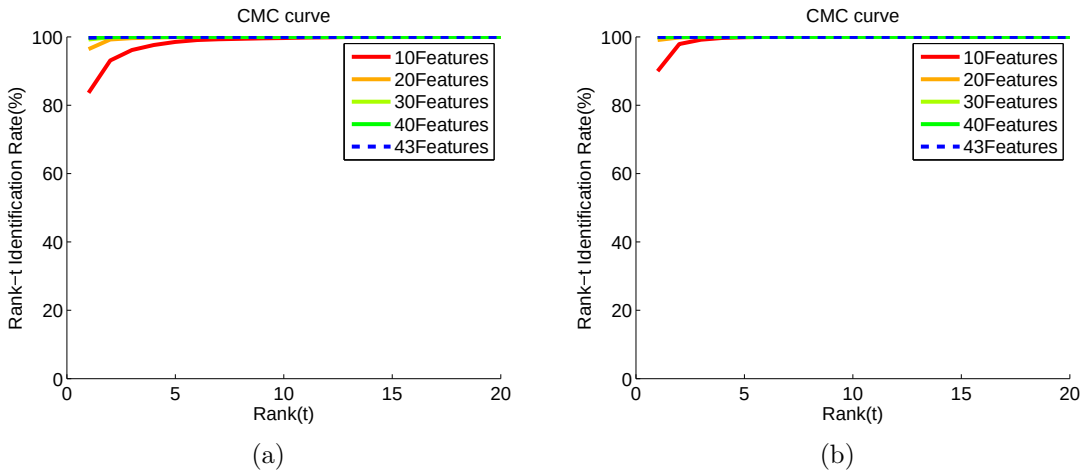
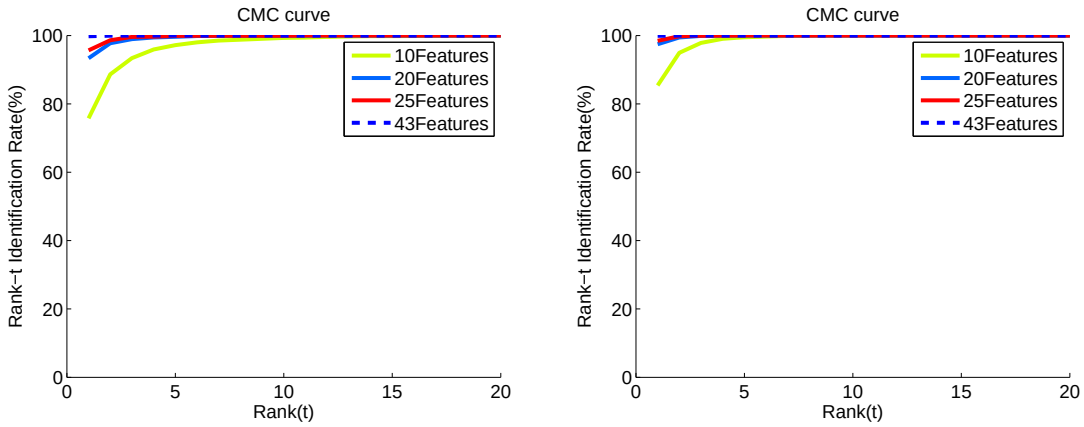


Figure 4.9: Comparison of identification system performance under (a) Gaussian noise $Gaussian(0, (0.2/3)^2)$ and (b) Uniform noise $Uniform(-0.1, 0.1)$ based on K_aNNE against 10, 20, 30 and 40 features. As a reference, the performance without feature selection (using all the 43 features) is shown in blue dash lines.

A similar analysis is applied on Category 1 features. The CMC curve (Figure 4.10) and a detailed report (Table 4.4) are shown as well.



(a) (b)

Figure 4.10: Comparison of identification system performance under (a) Gaussian noise $Gaussian(0, (0.2/3)^2)$ and (b) Uniform noise $Uniform(-0.1, 0.1)$ based on K_aNNE against 10, 20, and 25 features. Note that performance using all the 43 features is shown in blue dash lines.

Table 4.4: Comparison of identification system performance under (a) Gaussian noise and (b) Uniform noise based on K_aNNE against 10, 20,and 25 Category 1 features. R_1 is the rank-1 identification rate, while T is the value when R_T reaches 100.

Noise	$Gaussian(0, (0.2/3)^2)$		$Uniform(-0.1, 0.1)$	
#Features	R_1	T (when $R_T = 100$)	R_1	T
10	77.13	26	85.40	9
20	93.53	10	97.44	9
25	95.67	9	98.47	6

Impact of Noise Levels

We are also interested in the system performance under different noise levels. In Figure 4.11, the noise level (nl) is changing from 0 to 0.5 (x-axes) and the performance is measured by rank-1 identification rates (y-axes). Note that there are two different noise levels tested, Gaussian noise and Uniform noise, respectively. And each use a different way to define noise level.

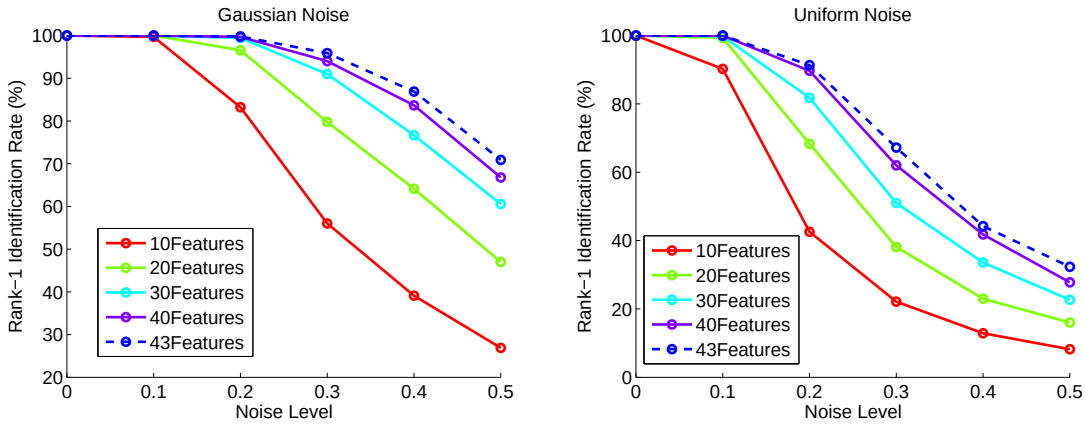


Figure 4.11: System performance under different noise levels (nl): (a) Gaussian noise $Gaussian(0, (nl/3)^2)$ and (b) Uniform noise $Uniform(-nl, nl)$ based on K_aNNE against 10, 20, 30 and 40 features. The performance without feature selection (using all the 43 features) is shown in blue dash lines.

4.3.3 Computational Time

We use a Dell XPS15Z laptop for our experiments. The laptop is equipped with Intel Core i5 2.3G CPU, 6GB memory, and Windows 7 operation system. All programs are written and executed using MATLAB R2011a. The training set has 100 samples and the test set has 900 samples. The training set is used for feature selection and the test set is used for verification and identification. All feature selection criteria are applied to a candidate list containing $n = 43$ features. The forward selection algorithm starts with one feature (Stature) and terminates when $m = 40$ features are selected. The computational time related to different feature selection criteria, verification, and identification are shown in Table 4.5. We observe that the EER criterion is the least efficient criterion, as expected. The results also suggest that the proposed K_aNNE criterion is just as efficient as the $KNNE$ criterion, and more efficient than the PJE criterion, if the table of joint entropies is not pre-computed.

Table 4.5: Computational time related to different techniques: feature selection criteria, verification, and identification. For PJE, 1.6 seconds is used to generate the table of joint entropies.

Technique	EER	PJE	$KNNE$	K_aNNE	Verification	Identification
Time (in seconds)	710.6	0.01+1.6	0.01	0.01	0.8	0.3

Table 4.6: Comparison between metrology-based recognition system against recent recognition systems using other biometrics.

Modality	Data (sample size)	Performance	Year	Ref.
Fingerprint	NIST (6,000)	$R_1 = 94\%$	2009	[116]
Multimodal (Fingerprint and Face)	NIST (517)	$R_1 = 100\%$	2009	[116]
Face	LFW (13,233)	$EER \simeq 15\%$	2009	[91]
Gait	CASIA-C (153)	$R_1 = 80.7\% - 99.0\%$	2009	[92]
Periocular	FRGC (2,272)	$R_1 = 87.32\%$	2011	[122]
Metrology (43 Features)	CAESAR (1,000)	$R_1 = 99.8\%/100\%$	2013	Ours
Metrology (30 Features)	CAESAR (1,000)	$R_1 = 99.1\%/99.8\%$	2013	Ours
Metrology (25 Category 1 Features)	CAESAR (1,000)	$R_1 = 95.7\%/98.5\%$	2013	Ours

4.4 Conclusion

This work provides initial person recognition results using solely human metrology. Using CAESAR 1D database as baseline, we simulate intra-class variation with considerable noise level. The system performance is tested from verification and identification prospective. The experimental results indicate that given enough number of features, our metrology-based recognition system can have promising performance that is comparable to several recent state-of-the-art recognition systems (see Table 4.6). We also propose a non-parametric feature selection criterion, K_aNNE, which does not rely on intra-class distribution of the query set. K_aNNE leads to more promising outcomes than other nearest neighbor estimators (as feature selection criteria) when number of features is larger than 10.

Chapter 5

Discrimination Capability of Human Metrology

The previous chapters have considered the use of human metrology in prediction, classification and recognition, including empirical performance evaluation. In this chapter, we quantify the theoretical discrimination capability of human metrology. A scientific basis for establishing the uniqueness of human metrology will not only quantify the performance of an automatic recognition system, but will also result in the admissibility of metrology identification technique in various areas such as the court of law. We develop several schemes to establish the limit of human metrology in recognition. We investigate two general approaches, one based on individuality model and the other based on channel capacity.

5.1 Related Work on Individuality Approach

5.1.1 Individuality of Fingerprints

The fingerprint individuality problem was first addressed by Galton in 1892 [60], which is defined as the probability of a specific fingerprint configuration. Galton

assumed that a full fingerprint can be covered by 24 independent square regions on average, each spanning 6 ridges. He further assumed 1/2 to be the probability to reconstruct any region by looking at the surrounding ridges; 1/16 to be the probability of occurrence of a specific fingerprint type; 1/256 to be the probability of occurrence of the correct number of ridges entering and exiting each of the 24 regions. Thus, the probability of a particular fingerprint configuration is:

$$P = \frac{1}{16} \times \frac{1}{256} \times \frac{1}{16} \times \left(\frac{1}{2}\right)^{24} = 1.45 \times 10^{-11}. \tag{5.1}$$

A number of subsequent models [71, 155, 36, 70] consider the probability of a particular fingerprint configuration based on the number of minutiae features n, and a fixed probability of their occurrence p. Assuming complete independence between the minutiae points, this gives:

$$P = p^n \tag{5.2}$$

Different p and n are used in different models. Although the above models are rather straight forward, a significant weakness is that they are based on ideal conditions, where the realistic problems such as partial matching and intra-class variations are not considered.

In Pankanti and Jain's work[120], the individuality is described in a more realistic manner: for a given input fingerprint containing n minutiae points, the individuality is the probability that an arbitrary fingerprint in a database containing m minutiae will have exactly q corresponding minutiae with the input (Eqn 5.3). It is easy to deduce that if there are q or more matches, the two fingerprints are considered sufficiently similar and thus should belong to the same person.

$$P(M, m, n, q) = \sum_{\rho=q}^{min(m,n)} \left(\frac{\binom{m}{\rho}\binom{M-m}{n-\rho}}{\binom{M}{n}} \times \binom{\rho}{q} l^q (1-l)^{\rho-q} \right). \tag{5.3}$$

In Eqn 5.3, $M = A/C$, where A is total area of overlap and C is the area of tolerance (Figure 5.1). l is the probability of two position-matched minutiae having a similar direction.

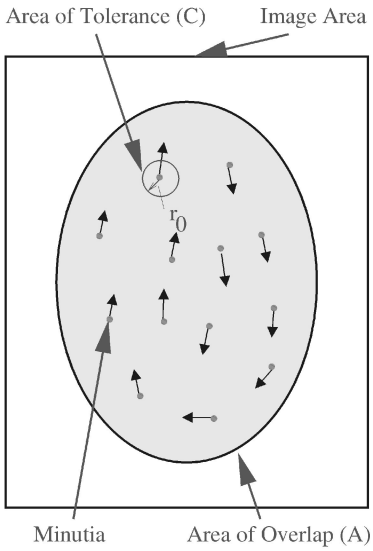


Figure 5.1: Parameters used in defining fingerprint individuality [120]. When an input fingerprint is matched with a template, an alignment is first established.

One weakness of Pankanti and Jain's work is that the assumption of uniform distribution of minutiae features may not always be satisfied in practice. This problem is later addressed by Dass et al. [39], using a family of finite mixture models which better represent clusters of features observed in fingerprint images compared to the uniform distribution. The estimates of fingerprint individuality are obtained using the probability of a random correspondence (PRC), which is defined as follows: Let Q denote the query fingerprint image and T denote the template fingerprint image. Let m be the total number of minutiae points in Q and n be the total number of minutiae points in T. Let p_m be the probability of a random minutiae feature from T matching one of the m minutiae features of Q. Then the PRC is the probability of obtaining exactly k matches between Q and T (Eqn 5.4):

$$PRC = \binom{n}{k} p_m^k (1 - p_m)^{n-k}. \tag{5.4}$$

A small PRC value indicates it is unlikely that the query and template fingerprint image belong to the same person. To calculate PRC, p_m has to be properly estimated based on the statistical distribution of the template database.

Another weakness of Pankanti's work is that it does not consider all possible discriminatory information that is embedded in fingerprints. Only ridge endings and ridge bifurcations are considered. Other level of fingerprint features, such as pattern type (Level 1) and pores (Level 3), are not included. In a later study, Chen and Jain [26] develop a more complex model to incorporate all three levels of fingerprint features. The correlation between features and the feature distribution are also considered. However, in above related work, the image quality is not explicitly taken into account for individuality.

5.1.2 Individuality of Iris

Iris is considered extremely individual [40, 41]. However, the individuality of iris is currently not well defined or quantified [129]. Unlike fingerprint, the iris information is usually represented as 2D binary code, called Iris Code. Two such codes can then be compared using certain distance measures (Hamming distance, Euclidean distance, etc). To address the individuality of iris, Yoon et al. [163] proposed a dichotomy solution, which transforms the distances into two categories: intra-class distances and inter-class distances. That is, given two Iris Codes, they either belong to the same person (thus their distance is intra-class) or not (thus their distance is inter-class). Regardless of the types of features, the feature distance vectors are numeric values that can be sent to a proper classifier for recognition. 11 models based on different features, distance types and classifiers are applied and compared, which

provide a strong background for future study. Unfortunately, the key question "what is the individuality of iris" is not explicitly answered.

Daugman [41] suggests that the iris recognition system could yield a zero false-match rate, on a large database that contains 632,500 iris images of 316,250 persons spanning 152 countries. However, this rate is predicated on high quality iris images, which are obtained under strict supervision. In practice, the image quality can be affected by various factors, which becomes a major concern that is related to the discrimination capability of an iris recognition system. In Kalka et al.'s work [85], the effect of various quality factors was analyzed, including de-focus blur, off-angle, occlusion/specular reflection, lighting, and iris resolution. A fully automated iris image quality evaluation block is developed to estimate the factors. This work shows that after removing the poor-quality images selected by their quality metric, a considerable improvement in recognition performance is achieved. They further provided an upper bound on the computational complexity required to evaluate the quality of a single image.

Kalka's work shows that the performance of an iris recognition system can be compromised by the image quality. Thus, to build a realistic model for the individuality of iris, the error impact should be taken into account. This interesting problem is still open for future study.

5.1.3 Individuality of Face

Unnikrishnan [152] used the notion of unusual features to study individuality in face recognition. Here, an unusual feature is defined as a feature whose metrics lie below the 5th or above the 95th percentiles for that feature. Those features could be nose length, inter pupillary distance, inter-alar width, upper lip length, shape of forehead, prominence of the chin, etc. Note that they are shape features, not appearance features. He further indicated that a face with 100 independent features will have 10

unusual features on average. It is easy to compute the probability of a particular face configuration with 10 unusual features:

$$P = 0.05^{10} = 9.8 \times 10^{-14}. \tag{5.5}$$

That is, the combination of these 10 unusual features can distinguish 10^{13} different faces. Perrett et al. [124] even identify 224 shape features on the frontal face. If all these features are acquirable by an automatic identification system, then this system can distinguish 10^{29} faces.

Although Unnikrishnan's work represents very promising results, it is still preliminary. The critical fact is that, when referred to face recognition, the facial features are usually not extracted from actual faces, but from 2D face images. Several issues need to be addressed before we can develop a realistic face individuality model: (1) Although there are many effective facial feature extraction techniques, no standard organization is currently established to group the facial information into feature categories. (2) The quality of image can be significantly compromised by pose, illumination, expression, and aging. (3) The dependence between facial features may not be negligible.

Klare and Jain [88] proposed a taxonomy which groups facial features into 3 levels: Level 1 features are those global features of a face that can be extracted from low resolution face images ($<$ 30 inter-pupilary pixel distance (IPD)), such as gender, ethnicity and general age group. Level 2 features are features that are explicit to face recognition and require more detailed face observations. These features are local and usually only relevant in face recognition, including features extracted using elastic bunch graph matching (EBGM) [159], local binary patterns (LBP) [5], SIFT feature descriptors [98, 109], metrological features [18], and so on. Level 3 features contain

micro level features on the face such as scars and facial marks [121]. Klare's work may serve as a prior to the studies on the individuality of facial features.

In the past two decades, a number of preprocessing methods have been developed to improve image quality. Blanz and Vetter [10] proposed a 3D morphable model that allows users to adjust the initial alignment between the input 2D image and the 3D morphable facial model, then change the pose of the input image to frontal and set the illumination to ideal ambient condition. The model is trained by a set of face images to learn the distribution of 3D facial shape and texture in a parameterized feature space. Gao et al. [61] proposed a pose normalization approach based on fitting active appearance models (AAM). In this work, profile faces with different rotation angles in depth were warped into shape-free frontal view faces. Bronstein et al. [14] present a 3D face recognition approach that is invariant to expressions. Their algorithm is a representation of the facial surface that is invariant to isometric deformations. Chen and Lovell [24] proposed a face recognition method which is robust to illumination and expression. In this work, adaptive principal component analysis (APCA) is used to construct a subspace of image representation, then warps the subspace according to inter-class and intra-class sample covariance, respectively. Park et al. [123] proposed a generative 3D aging model to simulate the facial aging process. In this work, the input image is projected into the parametric 3D aging pattern space. A new face image at target age is then simulated. For low-resolution face images, Bourlai et al. [12] proposed a method that applies a number of tools such as image filtering, linear de-noising, and thresholding-based nonlinear de-noising methods to enhance the quality of low resolution images. All these preprocessing methods can considerably improve the recognition accuracy.

In data analysis, we often assume that the data is drawn independently and identically from a certain distribution. However this assumption is not always true in practice. Sometimes we can accept an approximate independence. Sometimes, the

dependence can not be ignored. In that case we usually have two options. The first option is to eliminate the effect of dependence either by applying a de-correlation method [135], or by considering an informative feature subset, which involves a feature selection problem that can be solved in various ways [145, 18]. The second option is to incorporate the dependence information into the applied model. For fingerprint analysis, Dass et al. [39] proposed a mixture model in which minutiae are first clustered and then independently modeled in each cluster. A similar approach is applied by Chen et al. [26] when developing a mixture model based on 5 major fingerprint classes to evaluate fingerprint individuality. R. Kwitt et al. [93] proposed a joint statistical model for texture image retrieval problem, in which a copula-based method is applied to capture the associations among coefficients. These methods may be adapted for studies that involve different type of features.

5.2 Related Work on Channel Capacity Approach

One may argue that in Unnikrishnan's work [152], a specific number of rare features may not be guaranteed for each individual. Alternatively, if we represent each feature using a binary symbol (such as 'long (1)' or 'short(0)'), and consider each feature as i.i.d. Bernoulli random variables over the population with $Pr(f_i = 1) = 0.5$ for $i = 1, 2, \ldots, n$, then the probability of a particular face configuration is $1/2^n$. That means 8.59×10^9 individuals (which is more than the world population), can be distinguished using 33 features.

In practice, however, most human faces are remarkably similar, which means the variations in the relative sizes and distances among these features could be subtle. However the embedded noise in the face information could be overwhelming due to the large variations in pose, illumination, expression, occlusion, camera parameters, and background. Similar problem can apply to other biometrics, such as measurements on

the human body. Thus, to study the general performance of a biometric system, we need to address a more challenging problem: the impact of the noise. This problem can be addressed by adopting the concept of capacity from information theory.

5.2.1 Communication Channel and Capacity

In information theory, a communication channel (or channel), refers to a physical or logical transmission medium that can be used to transfer an information signal from one or more transmitters to one or more receivers. The transfer process is subject to uncontrollable ambient noise and the imperfection of the signalling process itself. The communication will not be successful unless the transmitter and receiver agree on what was sent. In information theory, the channel has a very important characteristic, called *channel capacity*, which is defined as the tightest upper bound on the amount of information that can be reliably transmitted over a communication channel. A channel is said to be memoryless if the probability distribution of the output depends only on the current input and is conditionally independent of previous channel inputs or outputs. The channel capacity of a memoryless channel is defined as [34]

$$C = \max_{p(x)} I(X;Y), \tag{5.6}$$

where $I(X;Y)$ is the mutual information of the input X and output Y and the maximum is taken over all possible input distributions. The mutual information is given by:

$$I(X;Y) = \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy, \tag{5.7}$$

or equivalently,

$$I(X;Y) = h(X) - h(X|Y) = h(X) + h(Y) - h(X,Y). \tag{5.8}$$

5.2.2 Recognition Capacity

The noise problem in a biometric authentication system can be considered as a noisy channel problem. The noise comes either from the errors that are inevitably involved during the feature extraction process, or from intended behaviors [151] such as spoofing. Thus, after a noisy feature extraction process, the subject is represented by a series of features. These features are further used to distinguish subjects. The quality, complexity, and variability of the features can be attributed to a recognition channel introduced and characterized by Schmid and O'Sullivan in [136, 137]. Similar to a communication channel, a recognition channel is also characterized by its capacity, called *recognition capacity*. The recognition capacity of a biometric system is considered as the maximum number of classes that can be successfully recognized asymptotically with probability of recognition error close to zero when the number of informative samples gets large. To achieve the expression of recognition capacity, the feature extraction process is modeled using a parallel Gaussian channel. In Schmid and Nicolò's work [135], the input $X = (x_1, \ldots, x_n)$ is considered as a set of independent features, which is obtained by feature selection and a de-correlation operation, such as PCA (Principal component analysis) [3] or ICA (Independent Component Analysis) [76]. Assume there is additive i.i.d. Gaussian noise $z_i \sim Gaussian(0, N_i)$ generated by the environment for each x_i $(i = 1, \ldots, n)$. It is also assumed that z_i is independent from x_i and z_i is independent from z_j when $i \neq j$. Finally, let the output be $Y = (y_1, \ldots, y_n)$. The original parallel Gaussian channel capacity for X is given by [34]:

$$C = \max_{\sum E[x^2] \leq P} I(x_i; y_i) = \sum_{i=1}^{n} \frac{1}{2} \log_2(1 + \frac{P_i}{N_i}) \quad bits, \tag{5.9}$$

where $P_i = E[x_i^2], P = \sum P_i$ are the power constraints. The equality is achieved if $x_i \sim Gaussian(0, P_i)$ for each i. Schmid and Nicolò [135] used a variation of the

channel capacity in Eqn 5.9 to derive the recognition capacity density for a biometric system based on PCA-encoding (Eqn 5.10):

$$C = \sum_{i=1}^{n} \frac{1}{2n} \log_2(1 + \frac{\lambda_i}{N_i}) \quad bits, \tag{5.10}$$

where the input X is encoded as a series of principal components and λ_i is the ith eigenvalue in the principal component analysis. Other efforts on channel capacity applications include Barni et al.'s watermark channel analysis [8] and Wyner's photon channel analysis[160].

5.3 Distinctiveness of Soft Biometrics

Soft biometric traits are those characteristics that provide some information about the individual, but lack the distinctiveness and permanence to sufficiently differentiate any two individuals [81]. Soft biometric traits include gender, ethnicity, age, eye color, hair color, weight, etc.

Jain et al. showed that 3 soft biometrics (gender, ethnicity and height) can improve fingerprint recognition by around 6% [81]. Other soft biometrics such as freckle, mole, scar, pockmark, skin color and wrinkle can also improve the face-recognition performance of a state-of-the-art commercial matcher [121]. Scheirer et al. show that the collection of 10 soft biometrics and 10 context attributes can boost the face identification system over the baseline by over 30% [134]. Furthermore, the possibility for human recognition based solely on a bag of soft biometric traits has been studied and promising preliminary results are shown by Dantcheva et al.[38]. Kumar et al. [91] also showed the collection of 65 attributes that are extracted from face images can be used as a stand alone feature model. Compared to the current state-of-the-art for the Labeled Faces in the Wild (LFW) data base, this model reduces the error rates by 23.92% in face verification.

One strength of the attribute traits is that they contain additional discriminatory information other than primary traits such as fingerprints and iris. The attributes are usually binary values, which means the computational time and space based on the attributes will be small. However, the measurability of a large number of attributes will be low. The automatic extraction of the attributes still remains a challenge. A large training sample may be required, which will be expensive and time consuming to collect.

How can we establish the discrimination ability of a soft biometric system? There are number of terms related to discrimination ability, such as individuality [120], recognition capacity [135], reliability [37], etc. To the best of our knowledge, only a few efforts have been made in theoretical study of the discrimination ability of given biometric traits. And the discrimination capability of soft biometric systems is currently neither well defined nor systematically studied. Given the characteristic of soft biometric traits, instead of trying to address the discrimination capability of single soft biometric trait, it may be more reasonable to consider the discrimination capability of a collection of a number of soft biometrics. In other words, our goal is to investigate whether a given number of features (not necessarily soft biometric features) is sufficient to distinguish individuals.

How can we address the discrimination capability of a biometric system? Even though we currently do not have a standard for measuring the individuality of soft biometrics, we observe that PRC [39] can be considered as a generic formulation for soft biometric traits, if the p in Eqn 5.4 is given and the feature set satisfies all or part of the following assumptions: (1) The features are scalar variables; (2) A match between two features is always aligned. That is, x_i will only be compared with y_i for all i; (3) All matches are independent and equally likely; (4) All features are sufficiently accurate and, as a consequence, no uncertainty should be associated with a match based on the quality of features. Based on these assumptions, we develop two

schemes to analyze the individuality of soft-biometric traits, which will be investigated in Section 5.5.

Schmid et al.'s capacity driven approach [135] could be adapted to certain biometric systems, for example, body measurements. However, this approach does not give the tightest upper bound if the feature distribution is not Gaussian. Unfortunately, in practice, the distribution of some soft biometric traits, such as gender and ethnicity, are not continuous, and thus are not Gaussian. Also, the distribution of some measurements might have long tails. Another issue is that the Gaussian channel requires Gaussian noise, while in practice we might need to handle non-Gaussian or unknown noise. Thus, we consider a different formulation using Poisson channel, which will be investigated in Section 5.6.

Another relevant consideration proposed by Dantcheva et al. [37] is the notion of reliability of a multi-trait soft biometric system (SBS). In practice, it is possible that the subjects will share similar facial and body characteristics. This is called cross subject interference. The reliability of a SBS captures the probability of false identification of a randomly chosen person out of a random set of N subjects. If we denote the number of categories by ρ, the feature space by $v = (v_1, \ldots, v_N)$, the number of non-empty categories by $F(v)$ (Obviously, we have $1 \leq F(v) \leq N$), the reliability is modeled by the probability $P(F)$ that a randomly drawn N-tuple of people will have F active categories out of a total of $min\{\rho, N\}$ possible active categories (Eqn 5.11):

$$P(F) = \frac{F^{N-F}}{(\rho - F)!(N - F)! \sum_{i=1}^{N} \frac{i^{N-i}}{(N-i)!(\rho-i)!}}. \tag{5.11}$$

Given ρ and N, the reliability of authentication averaged over the subjects in v is a function only of the number of non-empty categories $F(v)$, and independent of the distribution of categories.

5.4 Correlation Problem

5.4.1 Ideal Case

In ideal condition, the given database contains noise-free independent features only. The recognition capacity or the number of classes that can be successfully distinguished will then be:

$$2^{h(x_1,\ldots,x_n)} = 2^{\sum_{i=1}^{n} h(x_i)}, \tag{5.12}$$

and

$$h(x_i) = \sum_{j=1}^{J} p(x_i^j) \times \log_2(p(x_i^j)), \tag{5.13}$$

where $h(x_i)$ is the entropy of feature x_i with possible values $x_i^1, \ldots, x_i^J$ and probability mass function $p(x)$. If we assume the 43 anthropometric measures in the CAESAR 1D database [1] are noise-free and jointly normally distributed, we can apply PCA on the database and obtain its independent principal component representation. We can then substitute the principal components into Eqn 5.12 to compute the recognition capacity. The experiment results show that the average entropy of one principal component is 4.87 and thus the recognition capacity is $2^{209.49}$.

In practice, the recognition capacity would not be so high, because usually the given database will contain certain level of noise and the features are not independent of each other. For example, the CAESAR database [1] could have low level noise due to the measuring process. Also, the measurements are not independent variables [4]. In this work, we intend to address the dependence problem by pre-processing techniques, and use the proposed model to address the noise problem.

5.4.2 De-correlation

In some cases, significant correlation is observed between features in the given database [4]. If we assume independent features in our framework for discrimination

capability, it is necessary to apply de-correlation on the correlated features. If the data distribution is Gaussian, we can apply PCA to de-correlate the data. After PCA, the principal components are uncorrelated and independent. When the data is non-Gaussian, the statistical independence could be obtained using ICA. The key process in ICA is nonlinear de-correlation. There are a number of ways to define a suitable nonlinear function and more details can be found in [76]. A fast fixed-point algorithm can be used to compute the independent components [75, 2]. The algorithm converges, for example, when a quadratic (or skew) nonlinearity or a tanh nonlinearity is applied. Figure 5.2 shows the comparison between the Kendall' tau correlation (see Section 2.3) map for the original features and and that for the de-correlated features. Note that when using ICA, the dimension may be reduced due to the singularity of the covariance matrix.

5.5 Methodology for Individuality

5.5.1 Scheme 1: Binomial Model

Following [120] and [39], we can formulate the **individuality** of human metrology as the probability of getting k matches among n feature pairs for two given feature vectors $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$. Here an individual is described by a feature vector. We define a match:

$$m_i = \begin{cases} 1 & \text{if } \delta(x_i, y_i) \leq \epsilon_i \\ 0 & \text{if } \delta(x_i, y_i) > \epsilon_i, \end{cases} \tag{5.14}$$

where $i = 1, 2, \ldots, n$, ϵ_i is the tolerance term for the i^{th} match, and $\delta(x_i, y_i)$ is the distance function. In this work we use the simple form $\delta(x_i, y_i) = |x_i - y_i|$.

In scheme 1, we claim that the measurements in CAESAR database [1] satisfies all the four assumptions in Section 5.3. Assumption (1) is satisfied because the

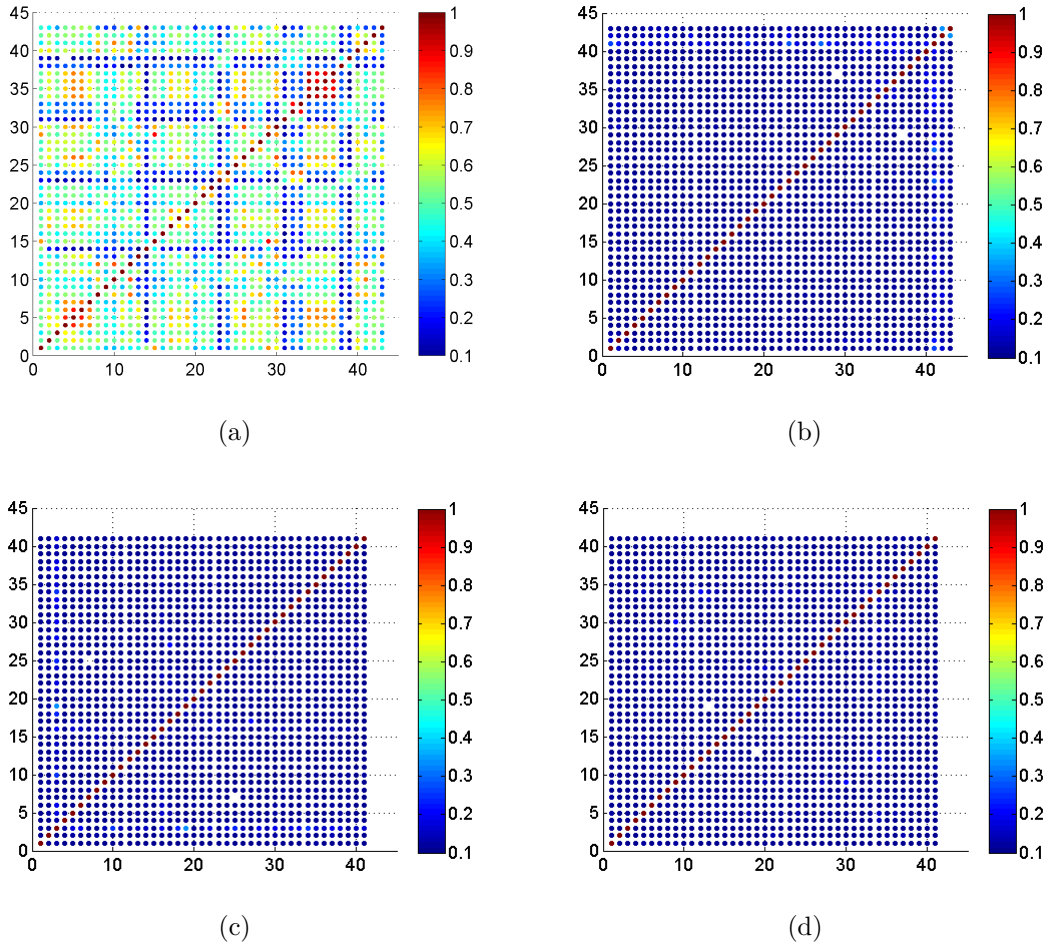


Figure 5.2: Comparing Kendall's tau correlation color maps between (a) original features, (b) principal components space, (c) ICA using quadratic nonlinearity and (d) ICA using tanh nonlinearity. Notice that the white spots in the figures are caused by near-zero correlation values.

original data is represented by scalar numbers (length and weight). Assumption (2) is satisfied by assuming all the feature values are correctly matched with the corresponding feature names. To maintain assumption (3), we could apply PCA or ICA on the original feature space to obtain the synthetic features that are statistically independent of each other. Note that the dimensionality of the feature space might be reduced due to the singularity of the covariance matrix. We assume that the manual measurements are sufficiently accurate and can be used as our baseline. Thus, assumption (4) is also satisfied.

Based on our definition and assumptions, the matches are Bernoulli random variables. Assumption (4) implies a small intra-class variation, thus we can reasonably let ϵ be fixed for all i. The probability of getting one successful match, we call it matching rate, is then

$$p = Pr(\delta(x_i, y_i) \leq \epsilon). \tag{5.15}$$

We can normalize every feature in the range $[\tau_{min}, \tau_{max}]$. If each feature follows a uniform distribution, by letting $\tau_{min} = 0, \tau_{max} = 1$, we have

$$p = \frac{\epsilon}{\tau_{max} - \tau_{min}} = \frac{\epsilon}{1 - 0} = \epsilon. \tag{5.16}$$

Similarly, if each feature follows a Gaussian distribution, the probability of getting one successful match between a pair of feature vectors will be:

$$p = \frac{1}{2}\left[1 + erf\left(\frac{\epsilon - \mu}{\sqrt{2\sigma^2}}\right)\right], \tag{5.17}$$

where μ and σ^2 are the mean and variance of the normal distribution. If we consider a standardized normal distribution with zero mean and unit variance, Eqn 5.17 becomes

$$p = \frac{1}{2}\left[1 + erf\left(\frac{\epsilon}{\sqrt{2}}\right)\right]. \tag{5.18}$$

Thus the probability of getting exactly k matches among n feature pairs is binomial, i.e.,

$$f(n,p,k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{5.19}$$

The probability of getting less than or equal to k matches is

$$F(n,p,k) = Pr(K \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}. \tag{5.20}$$

For example, if we have 40 measurements and the data is normalized in [0,1], with a 0.05 tolerance, the probability that we have exactly 40 matches is 9.1×10^{-53}. The probability that we have more than (but not equal) 30 matches is 8.2×10^{-33}. If we consider more than 30 matching features as yielding one reliable identification, then the number of classes (individuals) that can be distinguished is $\frac{1}{8.2 \times 10^{-33}} = 1.22 \times 10^{32}$.

Figure 5.3 shows $F(n,p,k)$ based on various values of n, k and p. The results show that if we allow a higher error tolerance (which leads to a higher matching rate), we need more matches to achieve a reliable identification.

5.5.2 Scheme 2: Poisson Binomial Model

In a more challenging case, for example, in using face metrology to distinguish between individuals, the measurements are usually extracted from face images. The quality of the extracted face measurements is usually not as good as the manual body measurements, since face images could have high level of noise or variation caused by pose, illumination, expression, 2D distortion or extraction technique used. In Cao et al's work [18], thousands of distances and angles based on 68 or 76 facial landmarks are extracted. The angles can be considered statistically independent of distances and thus can be normalized to $[\tau_{min}, \tau_{max}]$. However, the strong dependence among distances or among angles can not be easily eliminated by PCA or ICA de-correlation due to the potentially high level of noise.

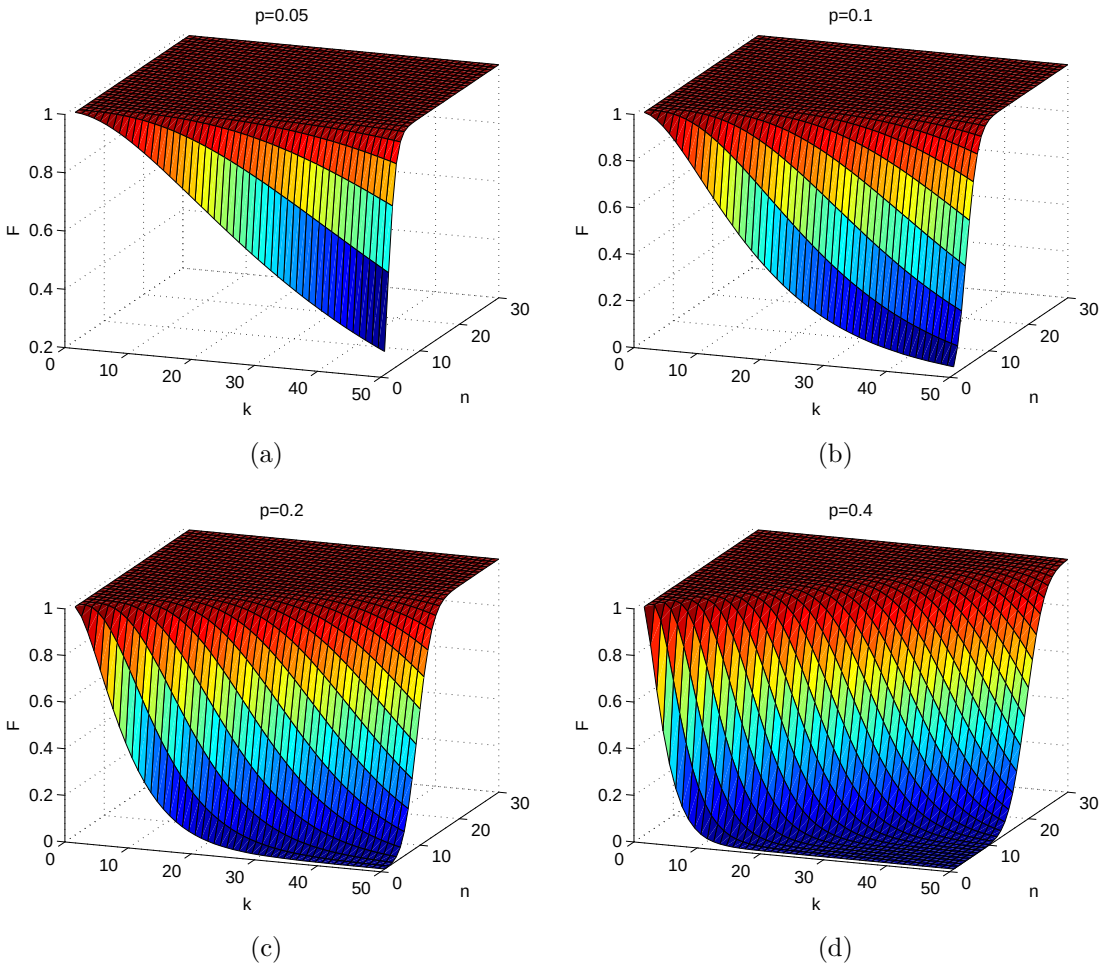


Figure 5.3: $F(n, p, k)$ based on: (a) $p = 0.05$, (b) $p = 0.10$, (c) $p = 0.20$ and (d) $p = 0.40$.

Thus, in Scheme 2, we claim that assumption (3) and (4) are not satisfied for some raw features, while assumption (1)-(2) remain valid. Although a strong feature selection method could remove most of the low-quality features [18], it also potentially reduces the overall discrimination information. Can we do better? One way is to relax the restriction on matching rate p, by assigning a specific p to each feature. Consequently, the individuality or the probability of getting exactly k matches among

n feature pairs becomes Poisson binomial:

$$f(n, p_1, \ldots, p_n, k) = \sum_{A \in F_k} \prod_{i \in A} p_i \prod_{j \in A^c} (1 - p_j), \tag{5.21}$$

where F_k is the set of all subsets of k integers that can be selected from $S = 1, 2, ..., n$, $A^c = S \backslash A$ is the complement of A.

The probability of getting less than or equal to k matches is

$$F(n, p_1, \ldots, p_n, k) = \sum_{l=0}^{k} \sum_{A \in F_l} \prod_{i \in A} p_i \prod_{j \in A^c} (1 - p_j). \tag{5.22}$$

We now need to compute the p_i's, the matching rates. Note that, since the matching rate is no longer fixed for all features, the tolerance terms ϵ_i's are also variables. Our strategy is to cluster the match of two or more dependent features and consider them as a single joint match, until all the joint matches that remain in consideration are mutually independent. A simple yet powerful k-mean clustering method could be used to achieve this purpose.

Let us consider the case of clustering two features (clustering more features can be computed recursively). Let x_i, x_j be two dependent features from a test individual, and y_i and y_j be the corresponding features from a individual in the gallery. Eqn 5.14 becomes

$$m_{ij} = \begin{cases} 1 & \text{if } \delta(x_i, y_i) \leq \epsilon_i \text{ and } \delta(x_j, y_j) \leq \epsilon_j \\ 0 & \text{Otherwise,} \end{cases} \tag{5.23}$$

Then the probability of getting a successful match is the joint probability

$$p = Pr(\delta(x_i, y_i) \leq \epsilon_i, \delta(x_j, y_j) \leq \epsilon_j). \tag{5.24}$$

We can use copula to represent the joint probability (see also Section 2.5). Let x_i and x_j be random variables with continuous marginal cumulative distribution

functions (*cdf's*) $F(x_i)$ and $G(x_j)$, respectively. By Sklar's theorem [144], their joint *cdf* $J(x_i, x_j)$ can be written as a copula C:

$$J(x, y) = C(F(x_i), G(x_j)). \tag{5.25}$$

Assume the distance function has the form $\delta(x_i, y_i) = |x_i - y_i|$, we can rewrite Eqn 5.24 as:

$$p = C(F(x_i + \epsilon_i), G(x_j + \epsilon_j)) - C(F(x_i - \epsilon_i), G(x_j - \epsilon_j)), \tag{5.26}$$

where the copula C and its parameters are either given, or can be estimated using experimental data. Here, we suggest the use of the Archimedean copula family, since (1) most Archimedean copulas admit an explicit formula for C and are efficient in complexity; and (2) Archimedean copulas support high dimensional structures. A copula C is called Archimedean if it has the representation

$$C(u, v) = \psi(\psi^{-1}(u) + \psi^{-1}(v)), \tag{5.27}$$

where ψ is called a generator. ψ is d-monotone on [0,∞).

For example, the Clayton generator [27] is given by

$$\psi(t) = (1 + \theta t)^{-1/\theta}, \psi^{-1}(t) = t^{-\theta} - 1. \tag{5.28}$$

and the Gumbel generator [67] is given by

$$\psi(t) = exp(-t^{1/\theta}), \psi^{-1}(t) = (-\ln(t))^{\theta}. \tag{5.29}$$

After we obtain the matching rates, the individuality can be computed from Eqn 5.21.

5.5.3 Experimental Results

Due to the lack of the information on the noise level on any specific measurement, we only test the performance of an identification system based on the Binomial model in this work. As in Chapter 4, We use the same random subset of 100 subjects from CAESAR 1D database [1] as our training set. Each subject has 43 manual measurements. Each feature x_i in the training set is normalized to [0,1] using min-max normalization. Two types of noise models are simulated in order to generate intra-class variations: (1) Gaussian noise $z_i \sim Gaussian(0, (0.2/3)^2)$; (2)Uniform noise$z_i \sim Uniform(-0.1, 0.1)$. 9 copies are generated for each subject, so the size of the test set is 900 and the total sample size in the experiment is 1000.

Non-Euclidean Distance Measure

In Chapter 4, we used Euclidean distance measure as matching score. In this experiment, in order to be consistent with the formulation of the individuality, we define the matching score between a query vector $X = (x_1, \ldots, x_n)$ and a template vector $Y = (y_1, \ldots, y_n)$ as follows:

$$s = \sum_{i=1}^{n} m_i, \tag{5.30}$$

where m_i is defined by Eqn 5.14 with $\delta(x_i, y_i) = |x_i - y_i|$. s is also known as Thresholded absolute distance (TAD) [131].

Since we are considering a Binomial model, the tolerance terms in Eqn 5.14 is identical for each feature. Thus we have $\epsilon = \epsilon_i$ for all i. To specify ϵ, we compare the verification system performance using various ϵ values under both Gaussian $Gaussian(0, (0.2/3)^2)$ noise and Uniform noise $Uniform(-0.1, 0.1)$ and the results are shown in Figure 5.4. It is suggested that the system is not sensitive to the tolerance term when it changes from 0.001 to 0.02. Above 0.02 a larger tolerance term leads to worse performance. The system performance drastically drops when the

tolerance term is set to be zero. Also, using TAD instead of Euclidean distance as matching score will *not* significantly change the system performance.

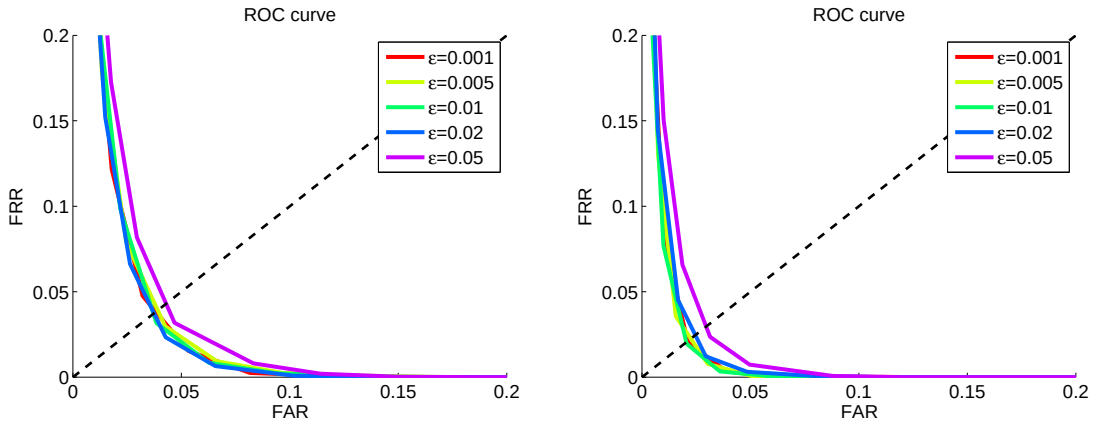


Figure 5.4: Comparison of verification system performance using min-max normalization and different tolerance term ϵ under (a) Gaussian noise $Gaussian(0, (0.2/3)^2)$ and (b) Uniform noise $Uniform(-0.1, 0.1)$. The comparison is based on all 43 features.

Z-score Normalization

Considering the fact that the distributions of most measurements in CAESAR 1D database are approximately Gaussian, we also apply z-score normalization:

$$x_i = \frac{x_i - \mu}{\sigma}, \tag{5.31}$$

where μ is the mean and σ is the standard deviation of the training set, respectively. We compare the verification system performance using different ϵ values under Gaussian $Gaussian(0, 0.4^2)$ noise and Uniform noise $Uniform(-0.6, 0.6)$. Since most of the original measurements should be in range [-3,3], the noise level retains approximately 20% in both cases. The experimental results that are shown in Figure 5.5 suggest that the system is not sensitive to the tolerance term when it changes from 0.01 to 0.2. Above 0.2 a larger tolerance term leads to worse performance. The system

performance drastically drops when the tolerance term is set to be zero. We observe that z-score normalization leads to better system performance than min-max normalization. However, we should be careful when using z-score normalization, because it is only optimal when the original data distribution is Gaussian. If the original data is not Gaussian distributed, z-score normalization can not retain the original data distribution [82]. Also, since z-score normalization does not guarantee a common numerical range [82], the simulated noise level tends to be slightly less than 20% in our experiment.

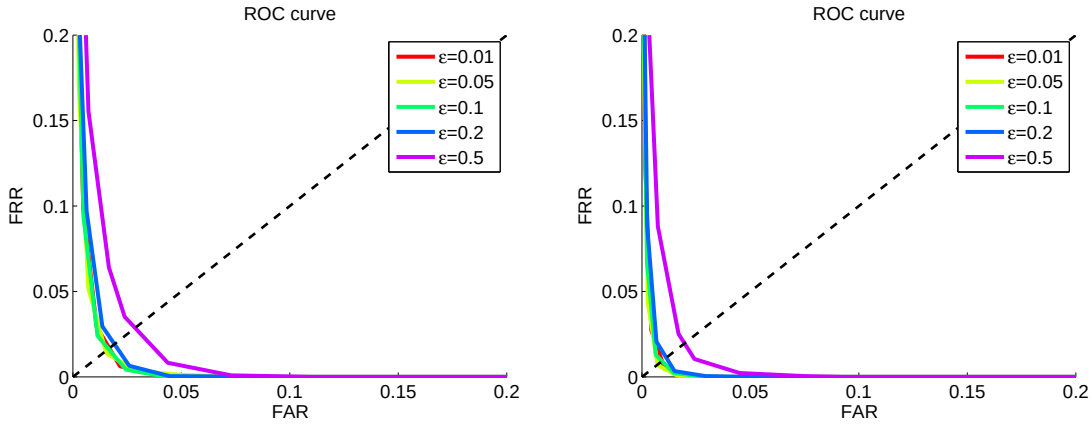


Figure 5.5: Comparison of verification system performance using z-score normalization and different tolerance term ϵ under (a) Gaussian noise $Gaussian(0, 0.4^2)$ and (b) Uniform noise $Uniform(-0.6, 0.6)$. The comparison is based on all 43 features.

Identification Performance

We then test the performance of an identification system using K_aNNE feature selection criterion that was introduced in Section 4.2.3. Figure 5.6 shows the system performance with respect to (a) min-max normalization under Gaussian noise $Gaussian(0, (0.2/3)^2)$ and $\epsilon = 0.02$; (b) min-max normalization under Uniform noise $Uniform(-0.1, 0.1)$ and $\epsilon = 0.01$; (c) z-score normalization under Gaussian noise $Gaussian(0, 0.4^2)$ and $\epsilon = 0.1$; and (d) z-score normalization under Uniform noise

$Uniform(-0.6, 0.6)$ and $\epsilon = 0.05$. A more detailed performance report can be found in Table 5.1. The experimental results indicate that there is no significant difference between Euclidean distance and TAD in terms of identification accuracy. And in our case, z-score normalization outperform min-max normalization and leads to better system performance.

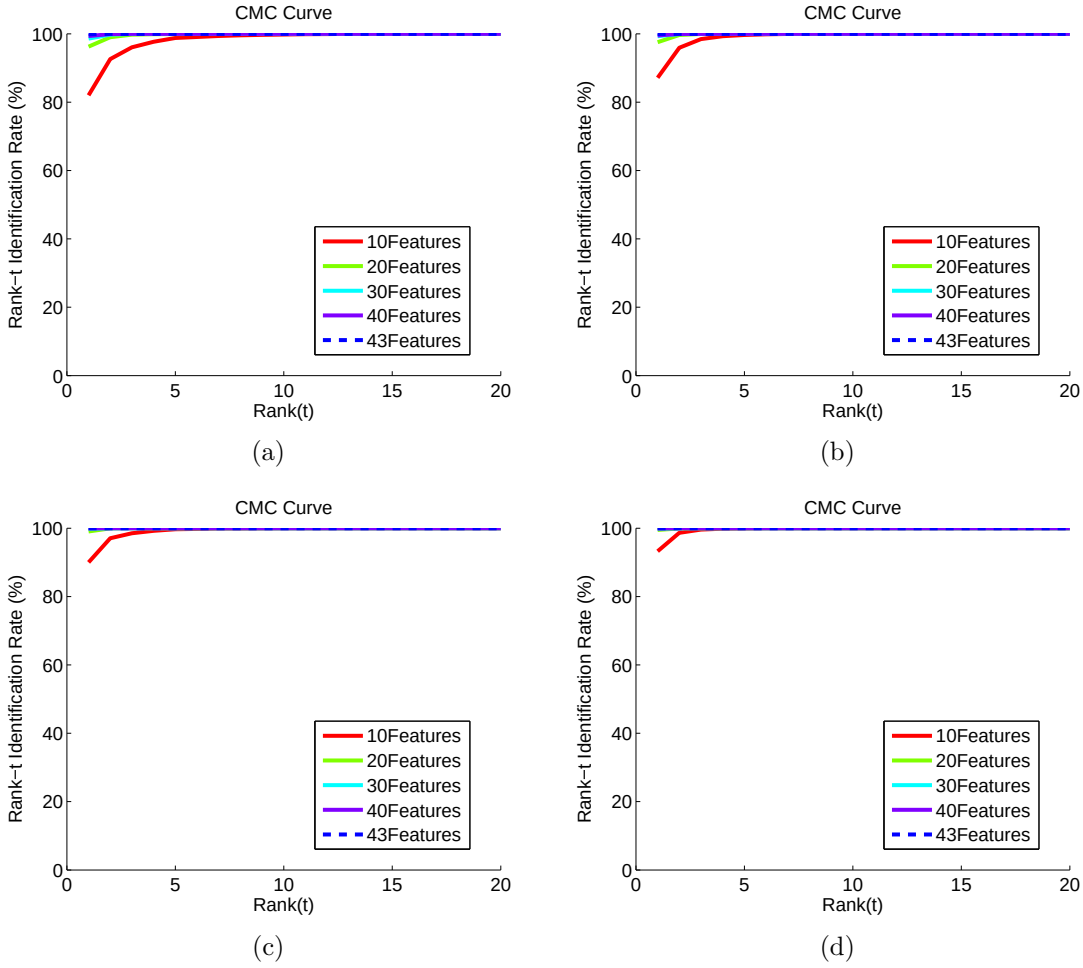


Figure 5.6: Comparison of identification system performance with respect to (a) min-max normalization under Gaussian noise $Gaussian(0, (0.2/3)^2)$ and $\epsilon = 0.02$; (b) min-max normalization under Uniform noise $Uniform(-0.1, 0.1)$ and $\epsilon = 0.01$; (c) z-score normalization under Gaussian noise $Gaussian(0, 0.4^2)$ and $\epsilon = 0.1$; and (d) z-score normalization under Uniform noise $Uniform(-0.6, 0.6)$ and $\epsilon = 0.05$. Experiments are based on K_aNNE feature selection criterion against 10, 20, 30 and 40 features. The performance without feature selection (using all 43 features) is shown in blue dash lines.

Table 5.1: Comparison of identification system performance with respect to different normalization and noise types indicated in Figure 5.6. Experiments are based on K_aNNE against 10, 20, 30, 40, and 43 features. R_1 is the rank-1 identification rate, while T is the value when $R_T = 100$.

	min-max & Gaussian		min-max & Uniform		z-score & Gaussian		z-score & Uniform	
#Features	R_1	T	R_1	T	R_1	T	R_1	T
10	82.00	13	87.16	8	90.00	10	93.37	7
20	96.24	8	97.56	6	99.04	7	99.50	3
30	98.60	5	99.31	5	99.82	4	99.87	2
40	99.33	4	99.53	3	99.96	3	99.93	2
43	99.84	4	99.96	2	99.98	2	100.00	1

5.5.4 Discussion

Scheme 1 has several advantages: (1) The noise caused by slight intra-class variation or small possible errors is explicitly controlled by the tolerance term ϵ; (2) The number of independent/reliable features is explicitly controlled by k; (3) It is flexible for different data distributions by switching the formulation of p.

Scheme 2 is a more general version of Scheme 1. When $p_1 = p_2 = \cdots = p$, it is reduced to Scheme 1.

Our current challenge is that, the copula model can only cluster a limited number of dependent features (two in general) and have to discard additional highly correlated features. In some cases, de-correlation or feature selection might be considered as helpful pre-processing.

We also need to consider the computational cost. The major cost involves computing the probability of getting at most n matches for one pair of individuals. In scheme 1, if the marginal distribution is known, the computational cost for Eqn 5.15 is $O(1)$. If the marginal distribution is unknown, p has to be estimated from the template database. Assume the database contains N templates, the cost for estimating p is in $O(N)$. The operation time for Eqn 5.19 is $O(n \times k)$ using dynamic programming, and

there are $O(k)$ terms in the summation in Eqn 5.20, thus the overall cost for getting at most n matches will be:

$$\begin{cases} O(n \times k^2) & \text{if } p \text{ is known} \\ O(N) + O(n \times k^2) & \text{if } p \text{ is unknown.} \end{cases} \tag{5.32}$$

In Scheme 2, if all the marginal distributions are known, the computational cost for getting all the p_i's is $O(n)$. If all the marginal distributions are unknown, the cost for estimating all p_i's is at least $O(n \times N)$. After a complete set of p_i's is obtained, we can then compute Eqn 5.21, which requires us to sum $\frac{n!}{(n-k)!k!}$ terms. However, Chen and Liu showed that this type of summation can be done in $O(n \times k)$ operations [25]. Considering that the number of terms in the first summation on the right hand side in Eqn 5.22 is $O(k)$, the overall cost for getting at most n matches is:

$$\begin{cases} O(n) + O(n \times k^2) \sim O(n \times k^2) & \text{if all } p_i\text{'s are known} \\ O(n \times N) + O(n \times k^2) \sim O(n(N + k^2)) & \text{if all } p_i\text{'s are unknown.} \end{cases} \tag{5.33}$$

Will the unusual features be more significant in human recognition than usual features? We can loosely address this question by utilizing the Chernoff bound, which is used to bound the success probability of majority agreement for n i.i.d.events. Let f_i be be independent Bernoulli random variables in (0, 1) for $i = 1, 2, \ldots, n$, each having probability $p > 1/2$ for outcome 1. A feature value is considered unusual when its metrics lie below the $\frac{(1-p)}{2}$th percentile or above the $\frac{(1+p)}{2}$th percentile for that feature. We further define the following:

$$f_i = \begin{cases} 1 & \text{if } f_i \text{ belongs to a specific individual (or class)} \\ 0 & \text{otherwise,} \end{cases} \tag{5.34}$$

The probability of simultaneous occurrence of more than $n/2$ of the events $f_i = 1$ has an exact value P which has the upper bound:

$$P \leq 1 - e^{-2n(p-1/2)^2}; \tag{5.35}$$

How many features do we need to extract to be confident to recognize a specific individual or class in a given database? It depends on p, which indicates how biased the features are. Figure 5.7 shows the Chernoff bound against feature number n with several different p values. For example, we can guess the answer with at most $P = 0.8647$ accuracy using 100 features, when these features are slightly biased ($p = 0.6$). We can achieve the same accuracy using 25 features when $p = 0.7$, or using only 6 features when $p = 0.9$. Note that we use the same p for each feature for a simple model. It is clear though, if the selected features are very unusual, the number of required features for accurate recognition will be significantly reduced.

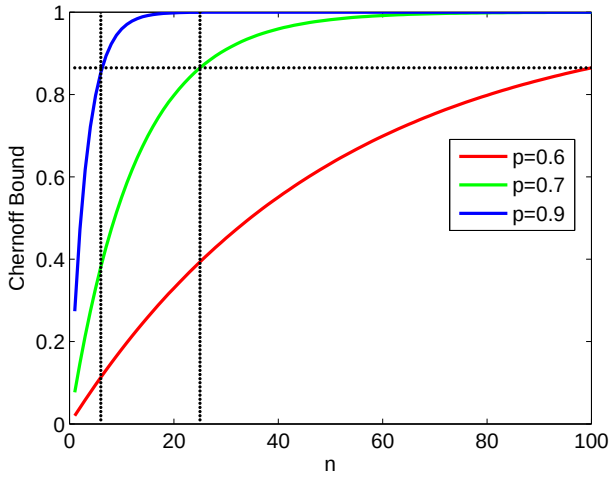


Figure 5.7: Chernoff bound against feature number n with different p.

5.6 Methodology for Capacity

5.6.1 Gaussian Channel Model

In this section, we study the discrimination ability of a soft biometric system from a capacity perspective. Inspired by Schmid's capacity approach [135], we can directly recall Eqn 5.9 to describe the discrimination capability of a soft biometric system:

$$C = \sum_{i=1}^{n} \frac{1}{2} \log_2(1 + \frac{P_i}{N_i}) \quad bits.$$

We do not use a capacity density expression here, because the noise level may vary upon different features. We can call Eqn 5.9 as recognition capacity of a soft biometric system. Depending on the value of C, a recognition system should be able to distinguish 2^C individuals per feature. However in practice, we may face two challenges when using such a model: (1) The basic assumptions of a Gaussian channel may not always be satisfied. That is, the noise distribution may not be Gaussian, and/or the data distribution (before noise) may not be Gaussian. In that case, a Gaussian channel model will be compromised. (2) We may be restricted by the accuracy and precision of the devices we use for measuring. High level of noise might be involved when we attempt to extract certain soft biometric features, especially when the subject is at a distance. As a consequence, even through the measurements could be continuous numbers, what is recorded could be discrete numbers. For example, a person's hight might actual be 188.567945... centimeters, but what is recorded could be 188 centimeters, or 19 decimeters, or 2 meters, or even simply a integer that indicates 'tall'. Note that, a discrete representation of features could be less accurate yet still captures the key characteristic of the measurements.

5.6.2 Poisson Channel Model

In order to address the above issues, we introduce a new channel model for studying the recognition capacity of human metrology. To establish such a model, we assume that the given database has the following properties: (1) The templates should be accurate enough to reflect the true values of the measurements of the subjects. This goal could be achieved by applying a supervised enrollment process. (2) The true values of the measurements should be non-negative real numbers. While after noise contamination, the measured outcomes are non-negative integers. This goal could be achieved with a thoughtful feature representation. Based on the above assumptions, we first establish the model for one single feature, and then extend it to multiple features. The channel we propose is a single-input single-output (SISO) Poisson channel. Recall that, a channel takes an input from a transmitter and produces a random output at the receiver according to a probability distribution conditioned on the input. A Poisson channel is a discrete memoryless channel (DMC) with input X and output Y, where X is a non-negative real variable and Y is a non-negative discrete variable. The conditional probability $Pr(Y = y|X = x)$ follows a poisson distribution, which predicts the degree of spread around a known average rate of occurrence [15]:

$$Pr(Y = y|X = x) = e^{-(x+\lambda)} \frac{(x + \lambda)^y}{y!}, \tag{5.36}$$

where λ is a non-negative constant called the dark current. The input x is a deterministic function (or rate function) in a non-homogeneous spatial Poisson process, where x takes the value on a one-dimension space $V = (0, A)$. Note that the poisson process does not have to be homogeneous. That is, we can use $x = x(t)$ to represent the value of x at position t. Based on our model assumption, the true values of the measurements from the subjects should be stored in the templates. The output are

the counts of the number of events inside each non-overlapping finite space-slot of V which are also independent to each other.

In our case, the input x is the true value of the measurement (or the expected value of a measurement with minimum error that we can possibly achieve). In other words, x is the ground truth that has been stored in a template. For instance, x could be the manual measurement obtained from a person by a medical expert, without the affect of clothing. The output y is a discrete value of the measurement that is extracted under unpredictable ambient noise. For example, y could be a measurement obtained from some footage of a surveillance camera, where the value of the measurement is affected by the distance, the view angle, the motion of the subject, the clothing and the resolution of the camera. In theory, the most possible value of y is the discrete value that is closest to x. However the actual extracted value of y is essentially random. We also need to define the output space $W = (0, T)$ where W is divided into a finite number of equal-length slots. The length and number of slots can be adjusted based on the accuracy and precision of the extraction process. If the measurements are extracted under poor conditions, we could reasonably define less number of slots, and vice versa. This strategy allows certain resilience to noisy data by dropping the small variation of the features that is most likely caused by noise.

Now we start to compute the capacity of the Poisson channel. There are two typical power constraints on channel capacity:

$$Pr(X > A) = 0, \tag{5.37}$$

and

$$E[X] \leq \alpha A, \tag{5.38}$$

where $0 < A \leq \infty$ is the peak-power constraint and $\alpha A > 0$ $(0 < \alpha < 1)$ is the average-power constraint. The capacity of the channel is denoted by $C(A, \lambda, \alpha)$. It

was first computed by Kabanov [84] and Davis [43] using martingale techniques. More recently, a more elementary and intuitively appealing method was developed by Wyner [160], who showed that $C(A, \lambda, \alpha)$ can be achieved by dividing the channel into small time-slots. In our case, the possible values of the measurement are divided into small space-slots (or simply slots) and the formulation remains unchanged. In particular, he first discretize the input X, which is the set of all possible values of a specific measurement, into small slots of size Δ. Second, the input is restricted to either A or 0 for each slot (This can be done by rounding up any non-zero input to A, or using a threshold). The receiver produces 0 if there is no count in the slot, or 1 if there is one or more counts (more than one counting is considered rare because of the small slot assumption). Then the channel reduces to a two-input two-output DMC and its transition probability can be approximated as:

$$Pr(1|x) = \begin{cases} \lambda\Delta e^{-\lambda\Delta} & \text{if } x = 0 \\ (A+\lambda)\Delta e^{-(A+\lambda\Delta)} & \text{if } x = A, \end{cases} \tag{5.39}$$

Let the capacity of the above channel be C_Δ. Our target channel capacity $C(A, \lambda, \alpha)$ is given by

$$C(A, \lambda, \alpha) = \lim_{\Delta \to 0} \frac{C_\Delta}{\Delta}. \tag{5.40}$$

When A, λ, α are given, $C(A, \lambda, \alpha)$ can be computed following Wyner [160]:

$$C(A, \lambda, \alpha) = A[q(1+s)\log_e(1+s) + s(1-q)\log_e(s) - (q+s)\log_e(q+s)] \quad nats, \tag{5.41}$$

where $s = \frac{\lambda}{A}$, $q = \min\{\alpha, q_0\}$, $q_0 = \frac{(1+s)^{1+s}}{s^s e} - s$, and $\log_e$ stands for natural logarithm.

Now we consider the extreme cases: (1) When $s = 0$ (no dark current), we have

$$C(A, \lambda, \alpha) = Aq\log_e \frac{1}{q} \quad nats, \tag{5.42}$$

where $q = \min\{\alpha, e^{-1}\}$.

(2) When $s \to \infty$, we have

$$C(A, \lambda, \alpha) = \frac{Aq(1-q)}{2s} + O(\frac{1}{s^2}) \quad nats, \tag{5.43}$$

where $q = \min\{\alpha, 1/2\}$.

For our study, we may reasonably let $s = 0$ since without input there will be no output. If we normalize the data in $(0, 1)$ using min-max normalization and multiply the data by A, then naturally the peak-power constrain becomes A and α becomes 0.5. Eqn 5.42 then becomes

$$C(A, \lambda = 0, \alpha = 0.5) = Ae^{-1} \log_e e = Ae^{-1} \quad nats. \tag{5.44}$$

Note that Eqn 5.44 gives the capacity for one feature only. To extend the formula to multiple features, we need to establish a multiple Poisson channel model. However, to the best of our knowledge, there is no close form for the capacity of dependent multiple Poisson channels. To simplify the underlying mathematics of our model, parallel Poisson channels are considered as a reasonable alternative. Note that to apply a parallel Poisson channel model, we need to assume that features are independent of each other, and the noise on different features are also independent. In practice, such an assumption may not be realistic. Thus, a de-correlation process could be used for generating synthetic independent features. Since the data is considered non-Gaussian, an ICA method that is mentioned in Section 5.4.2 could be applied. We do not further study the de-correlation topic in this work though. When using parallel Poisson channel model, the capacity is simply the sum capacity of n independent SISO Poisson channels [63] (given $\alpha = 0.5$ and no dark current):

$$C_p(\{A_1, \ldots, A_n\}, \lambda = 0, \alpha = 0.5) = \sum_{i=1}^{n} A_i e^{-1} \log_e e = e^{-1} \sum_{i=1}^{n} A_i \quad nats. \tag{5.45}$$

For example, if we have $m(m \leq n)$ features after de-correlation and $A = A_1 = \ldots = A_n$, a capacity of Ame^{-1} is achieved. The maximum number of people that could be distinguished will then be bound by $e^{Ame^{-1}}$.

5.6.3 Model Comparison

Compared to a Gaussian channel model [135], the Poisson channel model has the following characteristics: (1) Most importantly, Eqn 5.36 does not require a specific distribution of x or ambient noise on x. (2) Poisson channel is most often invoked for rare events, which implies that such a model is more suitable for data with high level of noise.

Both models will be affected by the data quality. For the Gaussian channel model, the recognition capacity is affected by the signal-to-noise ratio (SNR) defined in Eqn 5.46. A larger SNR will correspond to a higher recognition capacity. When there is no noise, we have SNR $\rightarrow \infty$ and Eqn 5.9 suggests a infinitely large recognition capacity. Figure 5.8 shows the relationship between SNR and recognition capacity (per feature) using Gaussian channel model.

$$SNR = \frac{1}{n} \sum_{i=1}^{n} \frac{P_i}{N_i}. \tag{5.46}$$

For Poisson Channel model, the recognition capacity is affected by the value of A, which essentially depends on the accuracy and precision of the output. A higher A value indicates a higher capacity and implies more possible output values. Figure 5.9 shows a plot of the Poisson channel capacity (per feature) against A values.

It is difficult to say which model is superior than the other, because they are based on different assumptions. However, we should note that when the number of features are equal in both models, the capacity of a Gaussian channel model depends on the noise level of the *test set*, while the capacity of a Poisson channel model depends

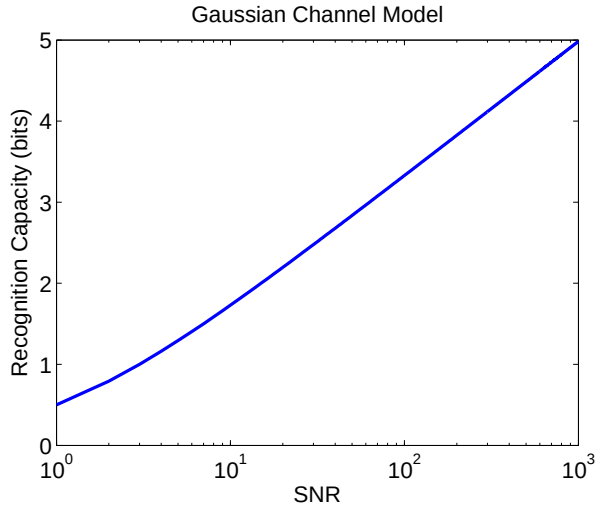


Figure 5.8: Recognition capacity (per feature) against SNR using Gaussian channel model.

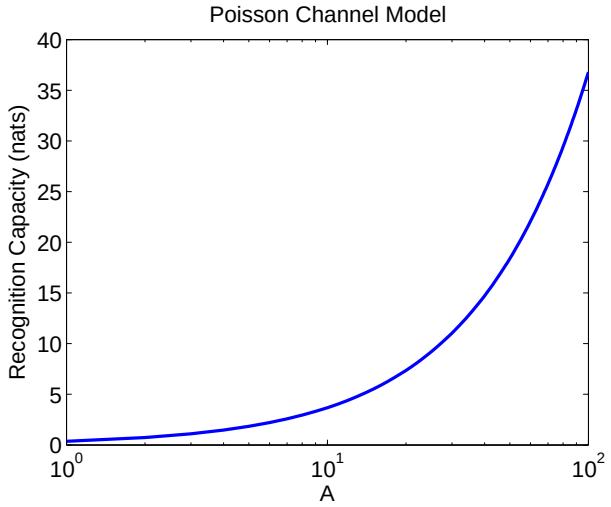


Figure 5.9: Poisson channel capacity (per feature) against A.

on the precision level of the *training set*. When peak-power constrain A increases, the capacity of a Poisson channel model also increases. Assume that in a Gaussian channel model, the power constrain $P = E[x^2] = 0.5^2$ and the noise level=0.2, then the capacity (per feature) is 2.92 *bits*. When $A = 10$, the capacity of a Poisson channel model (per feature) is 3.68 *nats*. When $A = 100$, the capacity (per feature) of a Poisson channel model is 36.8, which is larger than that of a Gaussian channel model with noise level $nl = 10^{-10}$. It seems that larger capacity would lead to better system performance, but in practice it may not be true. In Section 5.6.4, we will further study how the performance of an identification system under Poisson noise will be affected by the value of A.

5.6.4 Experimental Results

The performance of an identification system based on Poisson channel model is tested in this work. As in Chapter 4, We use the same random subset of 100 subjects from CAESAR 1D database [1] as our training set. Each subject has 43 manual measurements. Each feature x_i in the training set is first normalized to [0,1] using min-max normalization, and multiplied by a given peak-power constrain A_i. The average value of x_i is then $0.5 \times A$, which indicates $\alpha = 0.5$. In this study we assume $A = A_1 = \ldots = A_n$, thus the range of the entire training set becomes $(0, A)$. The test set is generated according to Poisson distribution conditioned with expected value x_i. That is, the output is a non-negative integer fluctuating around an average value x_i. 9 copies are generated for each subject, so the size of the test set is 900 and the total sample size in the experiment is 1000. The experimental setup in this work is similar to the setup in Section 4.3.2, we use min-max normalization, K_aNNE feature selection criterion and Euclidean distance as matching score. The major difference is that the data is now contaminated by Poisson noise instead of Gaussian or Uniform noise. Figure 5.10 shows the system performance with respect to (a) $A = 10$; (b)$A = 50$;

(c) $A = 100$ and (d) $A = 200$. A more detailed performance report can be found in Table 5.2.

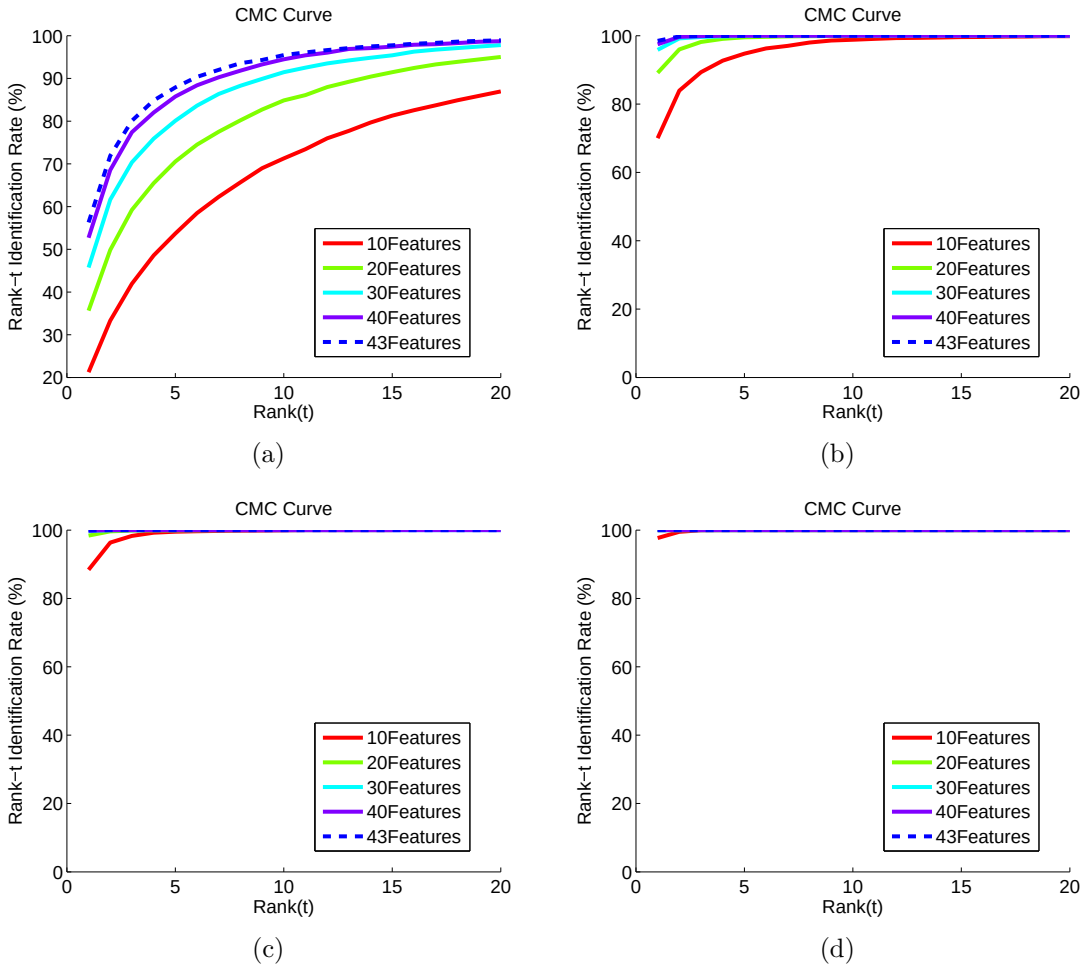


Figure 5.10: Comparison of identification system performance under Poisson noise with respect to (a) $A = 10$; (b)$A = 50$; (c) $A = 100$ and (d) $A = 200$. Experiments are based on K_aNNE feature selection criterion against 10, 20, 30 and 40 features. The performance without feature selection (using all 43 features) is shown in blue dash lines.

5.6.5 Discussion

The first observation from the experimental results is that we can not directly compare the capacity of a Poisson channel model with the capacity of a Gaussian model. Because even when A takes some large value, e.g., $A = 100$, the system performance

Table 5.2: Comparison of identification system performance under Poisson noise with respect to A=10, 50, 100 and 200. Experiments are based on K_aNNE against 10, 20, 30, 40, and 43 features. R_1 is the rank-1 identification rate, while T is the value when $R_T = 100$.

Noise	$A = 10$		$A = 50$		$A = 100$		$A = 200$	
#Features	R_1	T (when $R_T = 100$)	R_1	T	R_1	T	R_1	T
10	21.96	88	70.02	36	88.38	15	97.67	4
20	36.09	86	89.13	18	98.40	8	99.96	2
30	46.29	51	95.87	7	99.67	3	100.00	1
40	53.18	46	97.49	7	99.82	2	100.00	1
43	56.87	41	98.49	7	99.98	2	100.00	1

under Poisson noise is just slightly better than that under Gaussian noise when $nl = 0.2$. The second and more important observation, is that the system performance of a Poisson channel model will be affected by the value of A. In practice, the choice of A depends on the accuracy of the measuring method that is used to obtain the training data. For a given measurement X_i, we suggest that the corresponding value of A_i can be calculated as follows:

$$A_i = \frac{max\{X_i\} - min\{X_i\}}{acc_i}, \tag{5.47}$$

where acc_i is the accuracy of the measuring method that is used to obtain X_i in the training set. For instance, assume that we measure the stature using a tape and we are confident that the accuracy of our measuring method is up to 1 centimeter. After measured all subjects in the training set, we observe that the value of stature varies from 150 centimeters to 200 centimeters. Using Eqn 5.47, we then have $A_i = 50$. If we improve the accuracy of the measuring method from 1 centimeter to 0.5 centimeter by using more advanced measuring instrument, we can increase A from 50 to 100, and so on. Since ensuring *one* training set to be highly accurate is usually more accessible than ensuring *all* test sets to be highly accurate, our study suggests that generating

a template database with high accuracy could be a reasonable (and less expensive) alternative to upgrading all surveillance devices in a large area of interest.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Whole Body Metrology

In this work we investigate the use of human metrology. We first consider the whole body metrology and its subsets, head metrology and body metrology below neck, for the prediction of two soft biometrics: gender and weight. Experiments are performed using CAESAR 1D database [1] which contains 1119 male and 1250 female subjects after removing missing data. For each subject, the number of manual measurements considered for whole body, body under neck and head are 43, 35 and 6, respectively. For gender prediction, the proposed model results in a 0.7% misclassification rate using whole body information, 1.0% using only body (under neck) information, and 12.2% using only head information. For weight prediction, the proposed model gives 0.01 mean absolute error (in the range 0 to 1) using whole body information, 0.01 using only body (under neck) information, and 0.07 using only head information. This leads to the assertion that human body metrology contains enough information for reliable prediction of gender and weight. We further study the efficacy of the model in practical applications, where metrology data may be missing or severely contaminated

by various sources of noise. A novel copula-based technique is proposed to reduce the impact of noise on prediction performance. We observe that the copula-based model will boost the performance in general, especially when the noise is severe.

We also study the question of person recognition via whole body metrology. We test the performance of a verification system, as well as an identification system, using the same methodology. The experimental results based on CAESAR 1D database indicate that given enough number of features, our metrology-based recognition system can have promising performance that is comparable to several recent state-of-the-art recognition systems. A novel non-parametric feature selection criterion $KaNNE$ is developed in our work, which leads to more accurate outcomes when the number of features is larger than 10. Our experimental results also suggest that the missing information in the test set can be compensated by the information in the training set.

Face Metrology

We consider face metrology for the prediction of gender. A new metrology-based method is developed, which solely relies on metrological information from facial landmarks that can be manually or automatically extracted from face images. The performance of the proposed metrology-based method is compared with that of a state-of-the-art appearance-based method for gender classification. Results are reported on MUCT, XM2VTS and WVUM databases, respectively. The performance of the metrology-based approach was slightly lower than that of the appearance-based method by about 3.8% for the MUCT database and about 5.7% for the XM2VTS database. However, results on the WVUM database showed that the metrology-based method outperformed the appearance-based method (87% vs. 82%) on NIR images.

Discrimination Capability

Finally, we investigate the discrimination capability of the human metrology from individuality and capacity perspectives. Following the review of prior work, we propose several models to quantify the discrimination capability of the given biometric system. The dependence problem and the noise problem are considered in these models. In particular, a binomial model and a Poisson binomial model are developed to address the individuality, and a Poisson channel model is developed to address the recognition capacity. Note that the proposed models are suitable not only for human metrology, but also for any recognition system on noisy features that can be de-correlated or clustered.

6.2 Future Work

We briefly describe some potential future work, based on the material presented in this dissertation.

Noise Issue

Currently, our study of a human metrology based biometric system is based on CAESAR 1D database, in which the ambient noise is simulated based on certain simple distributions. However, in practice, the ambient noise could be more complex. To further study the accuracy of a human metrology based biometric system, or study the discrimination capability of such a system in a more realistic manner, a database that can provide both the ground truth and whole body images at a distance is necessary.

Missing Data

In practice, missing data is expected when features can only be extracted under poor conditions, such as long distance or low illumination. In such cases, certain data

may be either completely missing or too noisy to be useful. In our study, Support Vector Regression is applied to predict the missing data. However, the regression model predicts the most likely value of missing data but does not supply uncertainty about that value. This could cause an over-fitting problem. In future work, other imputation techniques that will add proper error terms to the estimation, such as Stochastic regression[47] or Multiple imputation[132], can be considered as better alternatives.

Matching Score

Although the Euclidean distance is simple and commonly used as a matching score, we may also consider other ways to generate matching scores. For instance, the Mahalanobis distance will take into account the correlations of the features and is scale-invariant. Thus it can be considered when the correlation between features can not be ignored and covariance matrix can be estimated. Another issue is that unusual features are not emphasized by the Euclidean distance. Even though one or a few unusual feature(s) should make a subject very different from other subjects, the contribution of the unusual feature(s) could be diluted by the cumulative effect of all other features. Thus, a more resilient distance measure can be investigated in future study.

Scaling Problem

we did not focus on feature extraction in this dissertation. However, in practice, raw measurements are usually sensitive to scaling (as well as other geometric transformations). To resolve the problem of constructing matches among subjects of unknown scale, a scale-invariant feature representation is necessary. Such a representation is not currently considered in our recognition system or individuality/capacity analysis. In future work, scale-invariant feature representation of 2D images or 3D images [48],

for instance, these based on scale-space theory [97] can be studied to reinforce our recognition system and discrimination capability models.

Ethnicity Classification

There has been increasing attention for ethnicity classification in recent years. For instance, Ding et al. [44] proposed a face-based ethnicity classification method using boosted local texture and shape descriptions from 3D face models; Lyle et al. [100] proposed a periocular-based gender and ethnicity classification technique using local appearance features extracted from the periocular region images; Zhang et al. [164] proposed a gait-based ethnicity classification method using multi-view fusion. However, there appears to be little prior work in the area of metrology-based ethnicity classification. We tested our metrology-based model introduced in Chapter 3 for ethnicity classification and the preliminary results show that our current method can perform at a correct classification rate of about 80%. A future question is how to increase the accuracy of the classification system, possibly by fusing metrology information and appearance information. Also, whole body metrology can be considered as another modality for ethnicity classification.

Association between Features

There are many different ways to represent the features other than using the measurements directly. For example, the copula-based features (CFeatures) introduced in Chapter 2 can be considered as an alternative feature representation. Also, in human history, there are many references about human body ratios in art, engineering or medicine, suggesting that certain body ratios can provide significant individual information. In Leonardo da Vinci's famous drawing **Vitruvian Man**, from below the chin to the top of the head is one-eighth of the height; the maximum width of the shoulders is a quarter of the height; the distance from the elbow to the armpit

is one-eighth of the height; the foot is one-seventh of the height of a man; and the palm is one-twenty-fourth of the height [125]. Although it was reported that certain human body ratios tend to be canonical[147], significant variation exists among different gender, age and ethnicity. And it can be easily observed that many possible body ratios vary on actual individuals. For example, we can consider the following body ratio features:

$$r_{AB} = \frac{A}{A+B}, \tag{6.1}$$

where A and B are two measurements from CAESAR database. Then the number of possible combinations in a database with n features is of size $O(n^2)$. We can concatenate the direct measurements and other feature representation, such as the CFeatures and/or the ratios together to yield a new feature space, which could providing richer information than merely direct measurements. Currently, our preliminary results show that such a simple concatenation does *not* increase the accuracy of the system. Thus, a future question is how to mitigate the error impact caused by combining noisy features and utilize the association information between features more efficiently so that the system performance can be improved.

Appendix A

Measurements in CAESAR Database

The original 43 measurements (excluding gender and weight) and their properties in the CAESAR 1D database are shown in the following Table A.1. The Measurability category column approximately indicates how difficult it is to automatically extract a measurement. That is, category 1 is for easy measurements; category 2 is for moderate measurements and category 3 is for difficult measurements. For the Cluster column, the measurements represented in red belong to the head cluster (H). The measurements represented in black belong to the body cluster (B). The others, represented in green, do not specifically belong to the body or the head, but are overall measurements (O). The K_aNNE rank gives the priorities of the features in forward selection in the verification system and identification system, based on our proposed adapted k-nearest neighbor estimator. K_aNNE rank I gives the priorities among all 43 measurements, while K_aNNE rank II gives the priorities among category 1 measurements (25 measurements).

Table A.1: The original 43 measurements (excluding gender and weight) and their properties in the CAESAR 1D database.

No.	Measurement	Measurability Category	Cluster	KaNNE Rank I	KaNNE Rank II
1	Acromial Height, Sitting (mm)	1	B	22	6
2	Ankle Circumference (mm)	3	B	32	
3	Spine-to-Shoulder (mm)	1	B	23	12
4	Spine-to-Elbow (mm)	1	B	35	23
5	Arm Length (Spine to Wrist) (mm)	1	B	41	18
6	Arm Length (Shoulder to Wrist) (mm)	1	B	33	22
7	Arm Length (Shoulder to Elbow) (mm)	1	B	11	11
8	Armscye Circumference (mm)	2	B	20	
9	Bizygomatic Breadth (mm)	1	H	4	2
10	Chest Circumference (mm)	2	B	14	
11	Buttock-Knee Length (mm)	1	B	13	5
12	Chest Girth at Scye mm)	1	B	39	17
13	Crotch Height (mm)	1	B	9	7
14	Elbow Height, Sitting (mm)	2	B	7	
15	Eye Height, Sitting (mm)	1	B	26	19
16	Face Length (mm)	1	H	3	4
17	Foot Length (mm)	3	B	12	
18	Hand Circumference (mm)	3	B	8	
19	Hand Length (mm)	3	B	24	
20	Head Breadth (mm)	1	H	21	13
21	Head Circumference (mm)	2	H	17	
22	Head Length (mm)	1	H	5	3
23	Hip Breadth, Sitting (mm)	1	B	27	10
24	Hip Circumference, Maximum (mm)	2	B	37	
25	Hip Circ Max Height (mm)	2	B	36	
26	Knee Height (mm)	1	B	40	25
27	Neck Base Circumference (mm)	2	B	15	
28	Shoulder Breadth (mm)	1	B	34	21
29	Sitting Height (mm)	1	O	30	14
30	Stature (mm)	1	O	1	1
31	Subscapular Skinfold (mm)	3	B	2	
32	Thigh Circumference (mm)	2	B	18	
33	Thigh Circumference Max Sitting (mm)	2	B	31	
34	Thumb Tip Reach (mm)	1	B	38	24
35	TTR 1 (mm)	1	B	25	9
36	TTR 2 (mm)	1	B	16	20
37	TTR 3 (mm)	1	B	28	15
38	Triceps Skinfold (mm)	3	B	6	
39	Total Crotch Length (mm)	3	B	10	
40	Vertical Trunk Circumference (mm)	3	B	43	
41	Waist Circumference, Pref (mm)	2	B	42	
42	Waist Front Length (mm)	1	B	29	8
43	Waist Height, Preferred (mm)	1	B	19	16

Bibliography

[1] Civilian American and European Surface Anthropometry Resource.

[2] Independent Component Analysis (ICA) and Blind Source Separation (BSS),http://research.ics.aalto.fi/ica/fastica/.

[3] H. Abdi and L. J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:433–459, 2010.

[4] D. Adjeroh, D. Cao, M. Piccirilli, and A. Ross. Predictability and correlation in human metrology. In *IEEE International Workshop on Information Forensics and Security*, 2010.

[5] T. Ahonen, A. Hadid, and M. Pietikäinen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–41, 2006.

[6] B. Allen, B. Curless, and Z. Popović. The space of all body shapes: reconstruction and parameterization from range scans. *Association for Computing Machinery's Special Interest Group on Computer Graphics and Interactive Technique*, 22(3):587–594, 2003.

[7] S. Baluja and H. A. Rowley. Boosting sex identification performance. *International Journal of Computer Vision*, 71(1):111–119, 2007.

[8] M. Barni, F. Bartolini, A. De Rosa, and A. Piva. Capacity of the watermark-channel: How many bits can be hidden within a digital image. In *Security and Watermarking of Multimedia Contents*, 1999.

[9] G. M. Beumer, Q. Tao, A. M. Bazen, and R. N. J. Veldhuis. A landmark paper in face recognition. In *Face and Gesture Recognition*, pages 73–78, 2006.

[10] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *Pattern Analysis and Machine Intelligence*, 25(9):1063–74, 2003.

[11] R.M. Bolle, J.H. Connell, S. Pankanti, N.K. Ratha, and A.W. Senior. The relation between the roc curve and the cmc. In *Automatic Identification Advanced Technologies*, pages 15–20, 2005.

[12] T. Bourlai, A. Ross, and A.K. Jain. Restoring degraded face images: A case study in matching faxed, printed and scanned photos. *Information Forensics and Security*, 6(2):371–384, 2011.

[13] T. Bourlai, C. Whitelam, and I. Kakadiaris. Pupil detection under lighting and pose variations in the visible and active infrared bands. In *IEEE International Workshop on Information Forensics and Security*, pages 156–161, Iguacu Falls, Brazil, Dec. 2012.

[14] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Expression-invariant 3D face recognition. In *Audio- and Video-Based Person Authentication*, pages 62–69.

[15] S. Bross, A. Lapidoth, and L. Wang. The Poisson channel with side information. In *Communication, Control, and Computing*, 2009.

[16] A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision, Lecture Notes in Computer Science 5303*, pages 15–29, 2008.

[17] A. M. Burton, V. Bruce, and N. Dench. What's the difference between men and women? Evidence from facial measurement. *Perception*, 22:153–176, 1993.

[18] D. Cao, C. Chen, M. Piccirilli, D. Adjeroh, T. Bourlai, and A. Ross. Can facial metrology predict gender? In *International Joint Conference on Biometrics*, 2011.

[19] L. Cao, M. Dikmen, Y. Fu, and T. S. Huang. Gender recognition from body. In *Association for Computing Machinery Multimedia*, pages 725–728, 2008.

[20] R. Cappelli and D. Maio. The state of the art in fingerprint classification. In *Automatic Fingerprint Recognition Systems*, chapter The State of the Art in Fingerprint Classification, pages 183–205. Spring New York, 2004.

[21] M. Castrillón-Santana and Q. C. Vuong. An analysis of automatic gender classification. In *Iberoamerican Congress on Pattern Recognition*, 2007.

[22] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *Association for Computing Machinery Trans. on Intelligent Systems and Technology*, 2:1–27, 2011.

[23] C. Chen and A. Ross. Evaluation of gender classification methods on thermal and near-infrared face images. In *International Joint Conference on Biometrics*, pages 1–8, Washington DC, USA, 2011.

[24] S. Chen and B. C. Lovell. Illumination and expression invariant face recognition with one sample image. In *International Conference on Pattern Recognition*, 2004.

[25] S. X. Chen and J. S. Liu. Statistical applications of the Poisson-Binomial and conditional Bernoulli distributions. *Statistica Sinica*, 7:875–92, 1997.

[26] Y. Chen and A. K. Jain. Beyond minutiae: A fingerprint individuality model with pattern, ridge and pore features. In *International Conference on Biometrics*, 2009.

[27] D. G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.

[28] M. Collins, J.G. Zhang, P. Miller, and H.B. Wang. Full body image feature representations for gender profiling. In *International Conference on Computer Vision*, pages 1235–1242, 2009.

[29] R. T. Collins, R. G., and J. Shi. Silhouette-based human identification from body shape and gait. In *Automatic Face and Gesture Recognition*, 2002.

[30] T. F. Cootes, C.J. Taylor, D.H. Cooper, and J Graham. Active shape models-their training and application. *CVIU*, 61:38–59, 1995.

[31] J. B. Copas. Regression, prediction and shrinkage. *J. Roy. Statist. Soc. Series B*, 45:311–354, 1983.

[32] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 2009.

[33] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[34] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 2006.

[35] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *Int'l J. Computer Vision*, 40(2):123–148, 2000.

[36] H. Cummins and C. Midlo. *Fingerprints, Palms and Soles*. Philadelphia: Blakiston, 1943.

[37] A. Dantcheva, J. Dugelay, and P. Elia. Soft biometrics systems: Reliability and asymptotic bounds. In *Biometrics: Theory, Applications and Systems*, 2010.

[38] A. Dantcheva, C. Velardo, A. D'Angelo, and J.-L. Dugelay. Bag of soft biometrics for person identification - new trends and challenges. *Multimedia Tools Appl.*, 51(2):739–777, 2011.

[39] S. C. Dass, Y. Zhu, and A.K.Jain. Statistical models for assessing the individuality of fingerprints. *Information Forensics and Security*, 2(3):391–401, 2007.

[40] J. Daugman. Recognizing people by their irispatterns. Technical report, University of Cambridge UK, 1998.

[41] J. Daugman. New methods in iris recognition. *IEEE Trans. Systems, Man, and Cybernetics, Part B*, 37(5):1167–1175, 2007.

[42] S. David, J. Ross, M. Teixeira, and A. Bruce. The CSU face identification evaluation system: Its purpose, features, and structure. In *International Conference on Vision Systems*, pages 304–313, Graz, Austria, Jun. 2003.

[43] M. H. A. Davis. Capacity and cutoff rate for Poisson-type channels. *IEEE Trans. Inform. Theory*, 26:710–715, 1980.

[44] H. Ding, D. Huang, Y. Wang, and L. Chen. Facial ethnicity classification based on boosted local texture and shape descriptions. In *Automatic Face and Gesture Recognition*, 2013.

[45] L. Donohue. Technological leap, statutory gap, and constitutional abyss: Remote biometric identification comes of age. *Minnesota Law Review, Forthcoming Georgetown Public Law Research Paper*, 12-123, 2012.

[46] B. Edelman, D. Valentin, and H. Abdi. Sex classification of face areas: how well can a linear neural network predict human performance. *Biological System*, 6(3):241–264, 1998.

[47] C.K. Enders. *Applied missing data analysis*. New York: Guilford Press, 2010.

[48] H. Fadaifard and G. Wolberg. Multiscale 3d feature extraction and matching. In *3D Imaging, Modeling, Processing, Visualization and Transmission*, 2011.

[49] F. Fang, P. J. Clapham, and K. C. Chung. A systematic review of inter-ethnic variability in facial dimensions. *Plast Reconstr Surg.*, 127(2):874881, 2011.

[50] J. J. Faraway. *Practical Regression and Anova using R*. University of Michigan, 2002.

[51] L. G. Farkas. *Anthropometry of the Head and Face*. New York: Raven Press, 1994.

[52] L. G. Farkas, M. J. Katic, and C. R. Forrest. International anthropometric study of facial morphology in various ethnic groups/races. *Journal of Craniofacial Surgery*, 16(4):615–46, 2005.

[53] L. G. Farkas and I. R. Munro. *Anthropometric Facial Proportions in Medicine*. Springfield, IL: Thomas Books, 1987.

[54] J.-M. Fellous. Gender discrimination and prediction on the basis of facial metric information. *Vision Res.*, 37(14):1961–73, 1997 Jul.

[55] V.F. Ferrario, C. Sforza, G. Pizzini, G. Vogel, and A. Miani. Sexual dimorphism in the human face assessed by Euclidean distance matrix analysis. *J. Anat.*, 183:593, 1993.

[56] J. Fisher. Alphonse bertillon: The father of criminal identification, 2008.

[57] Center for Disease Control and Prevention CDC. National health and nutrition examination survey, 1999-2005.

[58] E. W. Frees and E. A. Valdez. Understanding relationships using copulas. *North American Actuarial Journal*, 2(1):1–25, 1998.

[59] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. San Mateo, CA: Morgan Kaufmann, 1990.

[60] F. Galton. *Finger Prints*. London: McMillan, 1892.

[61] H. Gao, H. K. Ekenel, and R. Stiefelhage. Pose normalization for local appearance-based face recognition. In *International Conference on Biometrics*, 2009.

[62] W. Gao and H. Ai. Face gender classification on consumer images in a multiethnic environment. In *International Conference on Biometrics*, pages 169–178, Sardinia, Italy, Jun. 2009.

[63] S. A. M. Ghanem and M. Ara. The Poisson optical communication channels: Capacity and optimal power allocation. *IAENG International Journal of Computer Science*, 39, 2012.

[64] A. Godil and S. Ressler. Retrieval and clustering from a 3d human database based on body and head shape. *Computer Research Repository*, 2011.

[65] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *Neural Information Processing Systems*, pages 572–577, 1990.

[66] Arnulf B. A. Graf and Felix A. Wichmann. Gender classification of human faces. In *Biologically Motivated Computer Vision*, pages 491–500, 2002.

[67] E. J. Gumbel. Bivariate Exponential distributions. *J. Amer. Stat. Assoc.*, 55:698–707, 1960.

[68] F. Guo and R. Chellappa. Video metrology using a single camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1329–1335, July, 2010.

[69] G. Guo, G. Mu, and Y. Fu. Gender from body: A biologically-inspired approach with manifold learning. In *Asian Conference on Computer Vision*, 2009.

[70] S. R. Gupta. Statistical survey of ridge characteristics. *Int'l Criminal Police Rev.*, 218, 1968.

[71] E. R. Henry. *Classification and Uses of Fingerprints*. London: Routledge, 1900.

[72] A. Hilton, D. Beresford, T Gentils, R. Smith, W. Sun, and J. Illingworth. Whole-body modelling of people from multiview images to populate virtual worlds. *The Visual Computer*, 16(7):411–436, 2000.

[73] R. V. Hogg, J. W. McKean, and A. T. Craig. *Introduction to mathematical statistics*. Prentice Hall, 6 edition, 2004.

[74] M. A. Hossain, Y. Majihara, J. Wang, and Y. Yagi. Clothing invariant gait identification using part-based clothing categorization and adaptive weight control. *Pattern Recognition*, 43(6):2281–2291, 2010.

[75] A Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[76] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. New York: Wiley, 2001.

[77] L. Ibrahimagić-Šeper, A. Čelebić, N. Petričević, and E. Selimović. Anthropometric differences between males and females in face dimensions and dimensions of central maxillary incisors. *Medicinski glasnik*, 3(2), 2006.

[78] M. Indovina, U. Uludag, R. Snelick, A. Mink, and A. K. Jain. Multimodal biometric authentication methods: a COTS approach. In *Workshop on Multimodal User Authentication*, pages 99–106, 2003.

[79] A. K. Jain, R. Bolle, and S. Pankanti, editors. *Biometrics: Personal Identification in Networked Society*. Kluwer Academic Publishers, 1999.

[80] A. K. Jain, S. C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. *Proceedings of International Conference on Biometric Authentication*, pages 731–738, 2004.

[81] A. K. Jain, S. C. Dass, and K. Nandakumara. Can soft biometric traits assist user recognition? In *the International Society for Optics and Photonics*, 2004.

[82] A. K. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38:2270–2285, 2005.

[83] A. K. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:153–158, 1997.

[84] Y. M. Kabanov. The capacity of a channel of the Poisson type. *Theory of Probability and Its Appl.*, 23:143–147, 1978.

[85] N. D. Kalka, J. Zuo, N. A. Schmid, and Bojan Cukic. Estimating and fusing quality factors for iris biometric images. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 40(3):509–524, 2010.

[86] U. Kandaswamy, D. Adjeroh, S. A. C. Schuckers, and A. Hanbury. Robust color texture features under varying illumination conditions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 42(1):58–68, 2012.

[87] M. Kendall. A new measure of rank correlation. *Biometrika*, 30 (12):81:89, 1938.

[88] B. Klare and A. K. Jain. On a taxonomy of facial features. In *Biometrics: Theory, Applications and Systems*, 2010.

[89] S. G. Kong, J. Heo, B. R. Abidi, J. K. Paik, and M. A. Abidi. Recent advances in visual and infrared face recognition - a review. *Computer Vision and Image Understanding*, 97:103–135, 2005.

[90] L. Kozachenko and N. Leonenko. On statistical estimation of entropy of a random vector. *Probl. Inform. Transm.*, 23:95–101, 1987.

[91] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, 2009.

[92] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang. Automatic gait recognition using weighted binary pattern on video. In *Advanced Video and Signal Based Surveillance*, 2009.

[93] R. Kwitt, P. Meerwald, and A. Uhl. Efficient texture image retrieval using copulas in a Bayesian framework. *IEEE Trans. on Image Processing*, 20(7):2063–2077, 2011.

[94] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic face identification system using flexible appearance models. *Image and vision computing*, 13:393–401, 1995.

[95] X. Li, S. J. Maybank, S. Yan, D. Tao, and D. Xu. Gait components and their application to gender recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38:145–155, 2008.

[96] X.-C. Lian and B.-L. Lu. Gender classification by combining facial and hair information. In *International Conference on Neural Information Processing, Part II*, pages 647–654, 2008.

[97] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21(2):224–270, 1994.

[98] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, 1999.

[99] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, 1981.

[100] J.R. Lyle, P.E. Miller, S.J. Pundlik, and D.L. Woodard. Soft biometric classification using periocular region features. In *Biometrics: Theory Applications and Systems*, 2010.

[101] M. J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:1357–1362, 1999.

[102] C. Madden and M. Piccardi. Height measurement as a session-based biometric. *Proceeding, Image and Vision Computing Workshop*, 2005.

[103] H. Maeng, H.-C. Choi, U. Park, S.-W. Lee, and A. K. Jain. Nfrad: Near-infrared face recognition at a distance. In *International Joint Conference on Biometrics*, 2011.

[104] E. Makinen and R. Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):541–547, 2008.

[105] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition*. Springer, 2009.

[106] D. D. Mari and S. Kotz. *Correlation and Dependence*. Imperial College Press, 2001.

[107] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60:135–164, 2003.

[108] K. Messer, J. Matas, J. Kittler, and K. Jonsson. XM2VTSDB: The extended M2VTS database. In *Audio- and Video-Based Biometric Person Authentication*, pages 72–77, 1999.

[109] E. Meyers and L. Wolf. Using biologically inspired features for face processing. *Int. Journal of Computer Vision*, 76(1):93–104, 2008.

[110] S. Milborrow, J. Morkel, and F. Nicolls. The MUCT landmarked face database. *Pattern Recognition Association of South Africa*, 2010.

[111] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *European Conference on Computer Vision*, pages 504–513, 2008.

[112] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[113] R. Mnatsakanov, N. Misra, S. Li, and E. Harner. k_n-nearest neighbor estimators of entropy. *Math.Meth. Stat.*, 17:261–277, 2008.

[114] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *IEEE TPAMI*, 24:707–711, 2002.

[115] M. Nadler and E. P. Smith. *Pattern Recognition Engineering*. Wiley, New York, 1992.

[116] K. Nandakumar, A. K. Jain, and A. Ross. Fusion in multibiometric identification systems: What about the missing data? In *International Conference on Biometrics*, 2009.

[117] M. S. Nixon, R. Chelleppa, and T.-N. Tan. *Human Identification Based on Gait*. Springer, 2006.

[118] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:971–987, 2002.

[119] E. A. Osunwoke, F. S. Amah-Tariah, O. Obia, I. M. Ekere, and O. Ede. Sexual dimorphism in facial dimensions of the Binis of South-Southern Nigeria. *Asian Journal of Medical Sciences*, 3(2):71–73, 2011.

[120] S. Pankanti and A. K. Jain. On the individuality of fingerprints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 2002.

[121] U. Park and A. K. Jain. Face matching and retrieval using soft biometrics. *Information Forensics and Security*, 5(3):406–415, 2010.

[122] U. Park, R. R. Jillela, A. Ross, and A. K. Jain. Periocular biometrics in the visible spectrum. *Information Forensics and Security*, 2011.

[123] U. Park, Y. Tong, and A.K. Jain. Age invariant face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):947–954, 2010.

[124] D. I. Perrett, K. A. May, and S. Yoshikawa. Facial shape and judgements of female attractiveness. *Nature*, 368:239–242, 1994.

[125] V. Pollio. On symmetry: In temples and in the human body. In *Ten Books on Architecture, Book III*, chapter 1. Gutenberg.org, 2006. Translated by Morris Hicky Morgan.

[126] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern Recogn. Lett.*, 15(11):1119–1125, November 1994.

[127] H. T. F. Rhodes. *Alphonse Bertillon: Father of Scientific Detection*. London: George G. Harrap & Co., 1956.

[128] M. Rosenblatt. Remarks on multivariate transformation. *Ann. Math Stat.*, 23(3):470–472, 1952.

[129] A. Ross. Iris recognition: The path forward. *Computer*, 43(2):30 – 35, 2010.

[130] A. Ross and C. Chen. Can gender be predicted from near-infrared face images? In *International Conference on Image Analysis and Recognition*, pages 120–129, Burnaby, Canada, Jun. 2011.

[131] A. Ross and R. Govindarajan. Feature level fusion using hand and face biometrics. In *Proc. of SPIE Conference on Biometric Technology for Human Identification II*, volume 5779, pages 196–204, 2005.

[132] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley & Sons, 1987.

[133] Y. Saatci and C. Town. Cascaded classification of gender and facial expression using active appearance models. In *International Conference on Automatic Face and Gesture Recognition*, 2006.

[134] W. J. Scheirer, N. Kumar, K. Ricanek, P. N. Belhumeur, and T. E. Boult. Fusing with context: A Bayesian approach to combining descriptive attributes. In *International Joint Conference on Biometrics*, 2011.

[135] N. A. Schmid and Francesco Nicolò. On empirical recognition capacity of biometric systems under global PCA and ICA encoding. *Information Forensics and Security*, 3(3):512–528, 2008.

[136] N. A. Schmid and J. A. OSullivan. Performance prediction methodology for biometrics systems using a large deviations approach. *IEEE Trans. Signal Process., Supplement on Secure Media*, 52(10):3036–3045, 2004.

[137] N. A. Schmid and J. A. O'Sullivan. Performance prediction methodology for multi-biometric systems. In *Face Biometrics for Personal Identification, Multi-Sensory Multi-Modal Systems*, chapter Performance Prediction Methodology for Multi-Biometric Systems, pages 213–230. Berlin, Germany: Springer-Verlag, 2007.

[138] C. Shan. Gender classification on real-life faces. In *Advanced Concepts for Intelligent Vision systems*, page 323331, 2010.

[139] C. Shan, S. Gong, and P. W. McOwan. Fusing gait and face cues for human gender recognition. *Neurocomput.*, 71(10-12):1931–1938, 2008.

[140] S. Shan, W. Gao, B. Cao, and D. Zhao. Illumination normalization for robust face recognition against varying lighting conditions. In *Analysis and Modeling of Faces and Gestures*, 2003.

[141] J. Shi, A. Samal, and D. Marx. Face recognition using landmark-based bidimensional regression. In *International Conference on Data Mining*, pages 765–768, 2005.

[142] J. Shi, A. Samal, and D. Marx. How effective are landmarks and their geometry for face recognition? *Computer Vision and Image Understanding*, 102(2):117–133, May 2006.

[143] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk. Nearest neighbor estimates of entropy. *Am. J. Math. Manag. Sci.*, 23:301–321, 2003.

[144] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229231, 1959.

[145] Le Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13:1393–1434, 2012.

[146] N. Sun, W. Zheng, C. Sun, C. Zou, and L. Zhao. Gender classification based on boosting local binary pattern. In *Advnaces in Neural Networks*, pages 194–201, 2006.

[147] A. R. Tilley and H. Dreyfuss Associates. *The Measure of Man and Woman: Human Factors in Design*. Wiley, 2001.

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

[149] M. Toews and T. Arbel. Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1567–1581, 2009.

[150] H.-C. Tsai, W.-C. Wang, J.-C. Wang, and J.-F. Wang. Long distance person identification using height measurement and face recognition. In *Technology, Education and Networking Conference*, 2009.

[151] P. Tuyls and J. Goseling. Capacity and examples of template-protecting biometric authentication systems. *Biometric Authentication, Lecture Notes in Computer Science*, 3087:158–170, 2004.

[152] M. K. Unnikrishnan. How is the individuality of a face recognized? *Journal of Theoretical Biology*, 2009.

[153] C. Velardo and J. Dugelay. Weight estimation from visual body appearance. In *Biometrics: Theory, Applications and Systems*, 2010.

[154] J.G. Wang, J. Li, W.Y. Yau, and E. Sung. Boosting dense SIFT descriptors and shape contexts of face images for gender recognition. In *Computer Vision and Pattern Recognition Workshops*, pages 96–102, 2010.

[155] B. Wentworth and H.H. Wilder. *Personal Identification*. Boston: R.G. Badger, 1918.

[156] F. W. Wheeler, R.L. Weiss, and P. H. Tu. Face recognition at a distance system for surveillance applications. In *Biometrics: Theory, Applications and Systems*, 2010.

[157] C. Whitelam, Z. Jafri, and T. Bourlai. Multispectral eye detection: A preliminary study. In *International Conference on Pattern Recognition*, 2010.

[158] L. Wiskott, J. M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:775–779, 1997.

[159] L. Wiskott, J. M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching, in proc. ieee int. conference on image processing, vol. 1, p. 129, 1997. *Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[160] A. D. Wyner. Capacity and error exponent for the direct detection photon channel. *IEEE Trans. Inform. Theory*, 34(6):1449–1461, 1988.

[161] Z. Xu, L. Lu, and P. Shi. A hybrid approach for gender classification from face images. In *International Conference on Pattern Recognition*, 2008.

[162] Z. Yang and H. Ai. Demographic classification with local binary patterns. In *International Conference on Biometrics*, pages 464–473, 2007.

[163] S. Yoon, S.-S. Choi, S.-H. Cha, Y. Lee, and C. C. Tappert. On the individuality of the iris biometric. *Graphics, Vision and Image Processing*, 5, 2005.

[164] D. Zhang, Y. Wang, and B. Bhanu. Ethnicity classification based on gait using multi-view fusion. In *Computer Vision and Pattern Recognition Workshops*, 2010.

[165] X. Zhang and Y. Gao. Face recognition across pose: A review. *Pattern Recognition*, 42(11):2876 – 2896, 2009.

[166] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *Association for Computing Machinery Comput. Surv.*, 35(4):399–458, 2003.

[167] Z. Zhuang, D. Landsittel, S. Benson, R. Roberge, and R Shaffer. Facial anthropometric differences among gender, ethnicity, and age groups. *The Annals of Occupational Hygiene*, 54(4):391–402, 2010.