2014

Modeling Errors in Biometric Surveillance and De-duplication Systems

Brian Matthew DeCann
West Virginia University

Modeling Errors in Biometric Surveillance and De-duplication Systems

by

Brian Matthew DeCann

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
Electrical Engineering

Arun A. Ross, Ph.D., Chair
Lawrence A. Hornak, Ph.D.
Bojan Cukic, Ph.D.
Xin Li, Ph.D.
Ashish Nimbarte, Ph. D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
May 2014

Keywords: Biometric System, Gait Recognition, Performance Model, Biometric Surveillance, Match Scores

Modeling Errors in Biometric Surveillance and De-duplication Systems

by

Brian Matthew DeCann

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
Electrical Engineering

Lane Department of Computer Science and Electrical Engineering

APPROVAL OF THE EXAMINING COMMITTEE

Lawrence A. Hornak, Ph.D.

Bojan Cukic, Ph.D.

Xin Li, Ph.D.

Ashish Nimbarte, Ph. D.

Arun A. Ross, Ph.D., Chair

Date

Modeling Errors in Biometric Surveillance and De-duplication Systems

Brian Matthew DeCann

Abstract

Modeling Errors in Biometric Surveillance and De-duplication Systems

by

Brian Matthew DeCann
Doctor of Philosophy in Electrical Engineering

West Virginia University

Arun A. Ross, Ph.D., Chair

In biometrics-based surveillance and de-duplication applications, the system commonly determines if a given individual has been encountered before. In this dissertation, these applications are viewed as specific instances of a broader class of problems known as *Anonymous Identification*. Here, the system does not necessarily determine the identity of a person; rather, it merely establishes if the given input biometric data was encountered previously. This dissertation demonstrates that traditional biometric evaluation measures cannot adequately estimate the error rate of an anonymous identification system in general and a de-duplication system in particular. In this regard, the first contribution is the design of an error prediction model for an anonymous identification system. The model shows that the order in which individuals are encountered impacts the error rate of the system. The second contribution - in the context of an identification system in general - is an explanatory model that explains the relationship between the Receiver Operating Characteristic (ROC) curve and the Cumulative Match Characteristic (CMC) curve of a closed-set biometric system. The phenomenon of *biometrics menagerie* is used to explain the possibility of deducing multiple CMC curves from the same ROC curve. Consequently, it is shown that a "good" *verification* system can be a "poor" *identification* system and vice-versa.

Besides the aforementioned contributions, the dissertation also explores the use of gait as a biometric modality in surveillance systems operating in the thermal or shortwave infrared (SWIR) spectrum. In this regard, a new gait representation scheme known as *Gait Curves* is developed and evaluated on thermal and SWIR data. Finally, a clustering scheme is used to demonstrate that gait patterns can be clustered into multiple categories; further, specific physical traits related to gender and body area are observed to impact cluster generation.

In sum, the dissertation provides some new insights into modeling anonymous identification systems and gait patterns for biometrics-based surveillance systems.

To my family

Acknowledgments

First and foremost, I would like to thank my advisor, Dr. Arun Ross, for giving me the opportunity to study within his research group. I first met Dr. Ross in a CITeR meeting in November of 2008 in Tucson, Arizona. One could say I was working on the "other side", evaluating research proposals on their scientific merit, but not actually doing any implementing of my own. It was around this time that I had realized that I wanted to be more than just an observer. It was difficult for me to read about various research topics, while not actively participating in any. I wanted to be in the forefront, working on challenging problems. But, I really did not have an "in" or a way to get there. At least not in my present position. Thanks to an introduction from my previous advisor, Dr. Stephanie Schuckers (of whom I am also indebted), I met with Dr. Ross who graciously took the time to discuss with me if there were any opportunities to pursue Ph.D study within his lab.

On paper, I never had impressive credentials for advanced graduate study. As an undergrad, I quit the honors program largely because I wanted to spend my summers working as an intern at a company rather than do on campus research with a faculty member. I took the GRE as a senior, but my scores were below average. I stayed on at Clarkson University to complete my M.S. degree, but I did not have any publications from my time there (at the time). Every graduate school I had applied to that Fall of 2008 essentially rejected me. Yet, for reasons unknown, Dr. Ross looked past my resume and gave me a chance where others did not. I am forever grateful for this. Sometimes the "what if" game is a funny thing. I have no idea how things might have shaped out otherwise if I had not been sent on that trip to Tucson.

One of the things I admire most about Dr. Ross is his drive for producing high quality research. This includes the problems solved within the lab and the preparation of research documents for publication (which includes presentation slides, posters, and handouts, in addition to the publication itself). Though in fairness, to the graduate student who just

wants to graduate, at times this same drive can be immensely frustrating. However, at the end of the day, I think this approach teaches students how to strive to be the best they can be. The numerous awards the lab has won and recognition across the world is a testament to that. Working with Dr. Ross, I learned how to take ownership of a problem and make it my own and to follow it where I saw fit; to follow my own creative path rather than follow a rigid assignment.

I would also like to thank Dr. Ross for putting up me. Namely, my stubbornness, my time away from the lab (often due to personal injuries), and my often unconventional schedule. Some academic advisors maintain a strict watch over the lab, including who is in and when. I do not always deal well with structure, and I will be the first to admit that I push boundaries. As such, I am grateful that Dr. Ross allowed me to work under the premise of: "it does not matter when you are in the lab (or even if you are in the same state as the lab!), as long as your work is done on time".

I would also like to thank my colleagues, including Raghavender (Raghav) Jillela, Asem Othman, Raghunandan (Raghu) Pasula, Yaohui (Eric) Ding, Cunjian Chen, Gizem Erdogan, Olaoluwa (Peter) Laseinde, Manisha Sam-Sunder, Aglika Gyaourova, whom I have spent countless hours with in the lab. In addition, Javier Galbally, Ajita Rattani, Emanuela Marasco, and Antitza Dantcheva, whom also had stints within the group. In particular, I would like to thank Raghav, for being a great friend during my time in the lab. For always being able to help out on a question, putting up with my rants, vents, and frustration. For the companionship whether it be 2:00 AM in the lab or in town for a beer.

I am grateful for my committee members, Dr. Lawrence Hornak, Dr. Bojan Cucic, and Dr. Xin Li, each of whom were key collaborators of the West Virginia Night Biometrics team. Also, Dr. Ashish Nimbarte for joining my committee as a specialist in gait, but from an alternative viewpoint. I thank each of my committee members for their collaboration and questions that helped refine this dissertation.

My dissertation would not have been possible without the contributions of several faculty members, students, and collaborators. In particular, the West Virginia Night Biometrics team, consisting of Dr. Lawrence Hornak, Dr. Bojan Cucic, Dr. Xin, Li, Dr. Jeremy Dawson, Dr. Thirimachos Bourlai, Nate Kalka, and Marco Piccirilli for all the questions and discussions during the Night Biometrics projects. Jeremy, Nate, and Marco also did tremendous work in collecting several gait datasets, of which my contributions were largely in the utilization of and not so much in the collection of.

Outside the lab, I would be remiss to not thank the great social group I built during my time in Morgantown (and brief time in Michigan). Sara Lake for being a great roommate and friend when I moved to the area, and for better or worse, introducing me to life in Morgantown. The gang at 304 Highland. All the great people I met through the West Virginia University Cycling team and the Morgantown cycling community. Jon Zerbe, Kyle Kukieza, Darren McNeil, J.R. Petsko, and many others have helped tremendously in keeping me balanced outside the lab and inspired a passion for a lifelong hobby. I can not count how many times an idea would pop into my head on a country road 30 miles outside of town. I will miss the Morgantown riding community more than anything. It is so dedicated and tight knit. A unique gem not to be found in any metropolitan area.

Finally, I would like to thank my family for their support over the years. My parents in particular for supporting me when I left my job to start again as a student at WVU, their continued support over the years. I also want to thank my grandparents, who were always very eager and interested to know how things were going. That genuine support helped me push through the harder times.

Overall, it has been a tremendous journey and one I do not regret undertaking. I have been fortunate enough to travel across the United States to Florida, Arizona, Washington D.C., and Rhode Island as part of my graduate study. I have also had the opportunity to travel internationally to Spain, China, and Sweden. I think it is unlikely I would have had these opportunities in a post-graduate industry position.

Thanks for the ride.

Contents

List of Figures

List of Tables

Notation

The following notation and symbols are used throughout this document.

s	:	Biometric sample
p	:	Biometric probe (query to a matching algorithm))
$\mathbf{x}$	:	Biometric feature vector
x	:	Biometric match score
γ	:	Decision threshold for classifying match scores
$S(\mathbf{x}_i, \mathbf{x}_j)$	:	Function that generates a match score between feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$
$\mathcal{G}$	:	Reference sample database
$f_G(x)$	:	Genuine match score distribution
$f_I(x)$	:	Impostor match score distribution
N	:	Number of unique (true) identities in set of test data
N_T	:	Total number of samples in a set of test data
N_G	:	Number of genuine biometric samples per identity
N_{ref}	:	Number of reference biometric samples
N_{probe}	:	Number of probe biometric samples

Note: Bold upper case letters denote matrices and bold lower case letters denote vectors.

Chapter 1

Introduction

1.1 Biometric Recognition

Biometrics is the science of recognizing individuals based on the physical or behavioral traits of an individual [9, 10]. Physical traits are those which pertain to the appearance of an individual. Examples include face, fingerprint, iris, hand geometry and voice. Behavioral traits are those which individuals learn or acquire over time. Examples include gait, handwriting, and speech.

To be considered a biometric trait, a candidate physical or behavioral characteristic must satisfy at least the following requirements [10]:

Universality: The characteristic is present in a majority of individuals.

Distinctiveness: Any two individuals should exhibit a sufficient variation of the characteristic.

Permanence: The characteristic should not change significantly over time.

Collectability: The characteristic can be reliably acquired using sensors in a relatively non-invasive manner.

Generally, characteristics that satisfy all of the aforementioned criteria are denoted as *primary* biometric traits. Primary biometric traits are those which have a strong ability to discriminate between individuals. Examples of primary biometric traits include face, fingerprint and iris. However, not every biometric trait can be classified as primary. Some

biometric traits may be sufficiently distinct for some individuals, while being similar across others. Further, the traits themselves may not be permanent. Such a trait, which may not be permanent and unique, but can be used along with primary biometric traits for human recognition, are known as *soft* biometric traits. Examples of soft biometric traits include scars, marks, tattoos (SMT), gender, ethnicity, height, and age [11]. In some literature, traits such as gait are also viewed as being soft biometric traits.

1.1.1 Components of a Biometric System

In a biometric system, physical or behavioral traits are used to perform automated recognition of individuals. A biometric system is often described as a pattern recognition system, comprising of a sensor module, feature extraction module, matching module, and database module [10].

> ***Sensor Module***: The sensor module is an acquisition device, which captures the biometric data of an individual. An example of a sensor module is a camera (visible, infrared, etc.) that captures an image of an individual's face.
>
> ***Feature Extraction Module***: The feature extraction module converts the raw biometric data from the sensor module into a set of salient, or discriminatory features. For example, in a fingerprint recognition system, the feature set may consist of the position and orientation of minutiae points.
>
> ***Matching Module***: The matching module compares a set of probe (or query) biometric features against a set of reference data stored in a local database (ISO/IEC 2382-37 [12]), resulting in the generation of *match scores*. Match scores are created via the function $S(\mathbf{x}_1, \mathbf{x}_2)$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are two biometric feature sets. Match scores are scalar valued, often normalized between $[0, 1]$, where a value of $S(\mathbf{x}_1, \mathbf{x}_2) \approx 1$ represents a high degree of similarity between $\mathbf{x}_1$ and $\mathbf{x}_2$.[1] Conversely, a value of $S(\mathbf{x}_1, \mathbf{x}_2) \approx 0$ represents a low degree of similarity. Match scores are used by the matching module to make a decision regarding the identity of the user.
>
> ***Database Module***: The database module, defined as the reference database (ISO/IEC 2382-37 [12]), stores the reference sample data of enrolled users. During enrollment, an identifier representing an individual's identity (e.g., surname, username) is used to label the sample(s). The template and identifier are then stored in the database. It is not uncommon during enrollment to acquire multiple samples per individual in order to account for variations observed in the biometric trait.

[1] In this dissertation, unless otherwise stated, match scores are assumed to follow this convention.

1.1.2 Modes of Operation

A classical biometric system has two distinct operational stages: the *enrollment* stage, where biometric data acquired from an individual is stored in the database along with a label or an identifier denoting the identity; and the *recognition* stage when the input biometric data of an individual is compared against the enrolled data in order to recognize an individual. The recognition stage can be further categorized into one of two modes: *verification* and *identification* [10]. A visual summarizing the modules and operating modes of a biometric system is provided in Figure 1.1.

> ***Enrollment Mode***: In enrollment, a user submits their biometric data along with an identifier (e.g., name, user-id, etc.) and the data is added to the database for matching. The enrollment mode may include a *de-duplication* module, which evaluates whether the biometric data already exists in the database. Traditionally, enrollment is an overt process.
>
> ***Verification Mode***: In verification, the probe biometric data is submitted with a claim of identity. The system validates the claim of identity by comparing the probe biometric data *strictly* with similarly labeled templates in the reference database. In essence, the system is *verifying* the users claim of identity. This sort of matching is commonly referred as 1:1 matching, as a probe is compared against a single (or relatively small) number of reference entries. Verification systems are adept at performing *positive recognition* tasks, wherein a user claims to be a certain identity and the system either sustains or refutes this claim.
>
> ***Identification Mode***: In identification, the system only receives the probe biometric data (i.e., a claim of identity is not submitted). Therefore, in order to determine the identity of the probe, the system compares the features extracted from the probe to *every* reference sample in the reference database. This type of matching operation is commonly referred as 1:N matching, where N is the number of reference samples in the reference database. In this dissertation, the notation N_{ref} will denote the total number of reference samples.[2] The identification problem can further be described as either *open-set* or *closed-set*. In closed-set identification the identity of the probe is known to be present in the reference database. However, in open-set identification, the identity corresponding to the probe may or may not be in the reference database. In practice, most *operational* identification systems are open-set [13, 14]. In addition to performing positive identification tasks, identification systems are adept at performing *negative recognition* applications, wherein the system establishes whether an individual is who they deny (implicitly or explicitly) to be [10].

[2]The notation N will denote the number of identities in a dataset.

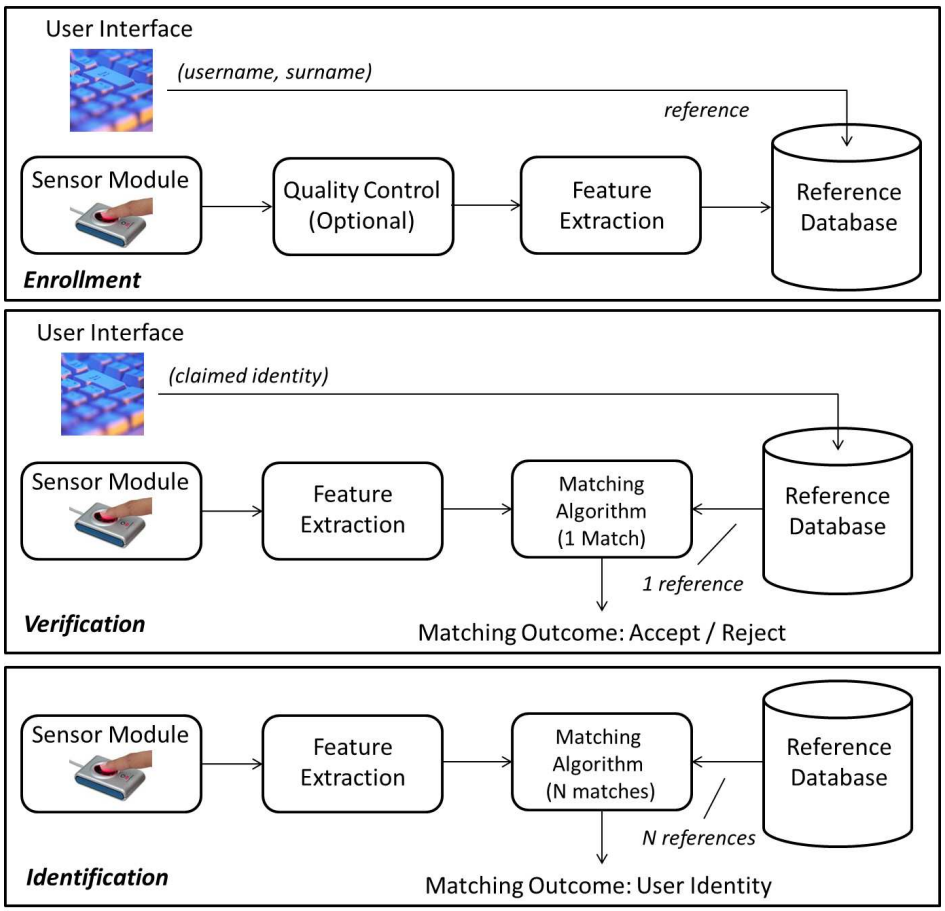

Figure 1.1: Simple flow diagram of a typical biometric system. During enrollment, the input biometric data is labeled with an identity and placed in the reference database. During matching, the input biometric data is compared against a set of reference samples in the reference database in order to determine if a matching identity exists.

Formally, the enrollment process can be described with an input feature vector $\mathbf{x}_E$, and an identifier, I. The feature vector and identifier are then added to the database.

Verification can be described with an input feature vector $\mathbf{x}_Q$, and a claimed identity, I. The system must determine if $(I, \mathbf{x}_Q)$ is either a *genuine* (true) or an *impostor* (false) identity claim. The two outcomes can be represented as ω_1 and ω_2, where ω_1 denotes the genuine category and ω_2 denotes the impostor category. Arriving at the decision of ω_1 or ω_2 is dependent on the similarity between $\mathbf{x}_Q$, the input feature vector, and $\mathbf{x}_I$, the feature vector in the database corresponding to user I. The degree of similarity is determined by generating the match score through the function $S(\mathbf{x}_Q, \mathbf{x}_I)$. The decision to classify the score as belonging to the genuine or impostor class is typically regulated by a predefined threshold,

denoted as γ. Thus a verification system classifies $(I, \mathbf{x}_Q)$ according to the rule in Equation (1.1).

$$(I, \mathbf{x}_Q) = \begin{cases} \omega_1, & \text{if } S(\mathbf{x}_Q, \mathbf{x}_I) \geq \gamma \\ \omega_2, & \text{otherwise} \end{cases} \tag{1.1}$$

In identification, the system attempts to deduce the identity of input vector $\mathbf{x}_Q$, based on the references present in the reference database. Let $I_1, I_2, \ldots, I_{N_{ref}}$ denote the identities that have been enrolled in the reference database, and $\mathbf{x}_{(I_k)}$ $(k = 1, 2, \ldots, N_{ref})$ denote the reference sample pertaining to identity I_k. The system then computes match scores for each reference and orders them from highest to lowest. The output is a set of L identities whose match scores exceed the decision threshold, γ. The identification process is described in Equation (1.2). If no match scores are generated with a value exceeding γ, the output is an empty set.

$$\mathbf{x}_Q = \begin{cases} I_k, & \textbf{if } max_k[S(\mathbf{x}_Q, \mathbf{x}_{I_k})] \geq \gamma, \quad k = 1, 2, \ldots, N_{ref} \\ NULL, & \text{otherwise} \end{cases} \tag{1.2}$$

If the system contains a de-duplication module, the input vector $\mathbf{x}_Q$ is compared against *every* reference sample in the database, not dissimilar from identification. If any of the generated match scores exceed the decision threshold, $\mathbf{x}_Q$ is flagged as a duplicate. The system administrator may then choose to deny or accept enrollment of the probe sample.

1.1.3 Biometric System Errors

A verification system can result in two types of errors from the matching module. A *false match* occurs when biometric samples from two different individuals generate a match score greater than threshold γ. Conversely, a *false non-match* occurs when biometric samples from the same individual results in the generation of a match score *below* threshold γ. As the value of γ changes, so does the probability of the system incurring a false match or false non-match. As the value of γ increases, the system is less tolerant of differences between two corresponding feature vectors, which can reduce the probability of observing a false match error, at the cost of increasing the probability of observing a false non-match. On the

other hand, as the value of γ decreases, the system is better able to handle noise and input variation. However, this comes at a cost of an increased likelihood that an impostor will be deemed genuine by the system. Thus, there is an inherent tradeoff between the propensity of the system to generate false-match and false non-match errors based on the selection of the decision threshold, γ.

An identification system can also result in a false match and false non-match, although these errors are defined differently. A *false positive identification* occurs when a probe, which *does not* have a matching entity in the reference database, incorrectly matches to some entity in the reference database [13, 14]. Conversely, a *false negative identification* occurs when a probe, which does have a matching entity in the reference database, is not observed in the output of L identities, or is observed at a position (e.g., rank) greater than R ($R = 1, 2, \ldots, L$).

1.1.4 Measuring Biometric System Performance

Since a biometric system is prone errors, it is necessary to develop techniques capable of providing a meaningful estimation regarding the performance of the system (e.g., matching algorithm). To facilitate this, a test database of biometric samples is required. Let N denote the number of identities in a test database, with N_G samples per identity. Denote the total number of samples as N_T (i.e., $N_T = N \cdot N_G$). By comparing each of the N_T samples against the remaining $N_T - 1$ samples, a total of $N_T(N_T - 1)$ match scores can be created. Assuming $S(\mathbf{x}_i, \mathbf{x}_j) = S(\mathbf{x}_j, \mathbf{x}_i)$ (i.e., a symmetric matcher), the number of distinct match scores is $\frac{1}{2}N_T(N_T - 1)$. Such a process can be defined as performing an "all-to-all" match test.

When computing match scores, two distinct types are computed: *Genuine* match scores and *impostor* match scores. Genuine match scores denote the scores generated when matching two samples belonging to the same identity. Impostor match scores denote the scores generated when matching samples belonging to different identities. For an "all-to-all" performance test with a symmetric matcher, the number of genuine and impostor match scores is given by Equations (1.3)-(1.4). Note that as the number of samples increases, the number of impostor scores that can be generated becomes polynomially larger than the number of

genuine scores.

$$\#GenuineScores = N\binom{N_G}{2} \tag{1.3}$$

$$\#ImpostorScores = N_G^2\binom{N}{2} \tag{1.4}$$

Using a histogram of the compiled genuine and impostor match scores, a pair of probability density functions can be estimated, denoting the probability of generating either a genuine match score or impostor match score with a specific value. These distributions are defined as the *genuine match score distribution*, $f_G(x)$, and *impostor match score distribution*, $f_I(x)$, respectively. Visually, these distributions offer meaningful information regarding the separability of genuine and impostor scores.

Measuring Verification Performance

In addition to serving as a visual aid of the separability of match scores, the distributions, $f_G(x)$ and $f_I(x)$ can be used to derive the *False Match Rate* (FMR) and *False Non-Match Rate* (FNMR), which are two measures for estimating verification performance. Mathematically, the FMR is defined as the integral of $f_I(x)$ for $x \in [\gamma, \infty)$. Similarly, FNMR is defined as the integral of $f_G(x)$ for $x \in (-\infty, \gamma]$. These expressions are provided in Equations (1.5) and (1.6). A looser interpretation of the FMR is the percentage of generated impostor scores that exceed γ. Similarly, the FNMR can be loosely characterized as the percentage of generated genuine scores that are less than γ. Since the FMR and FNMR are obtained based on estimates of $f_G(x)$ and $f_I(x)$, which in turn are estimated from the entirety of the match score data, these error rates can be defined as *aggregate-based* measures.

$$FMR(x) = \int_{\gamma}^{\infty} F_I(x)dx \tag{1.5}$$

$$FNMR(x) = \int_{-\infty}^{\gamma} F_G(x)dx \tag{1.6}$$

Note that the FMR and FNMR are a function of the decision threshold, γ. Using Equations (1.5) and (1.6), the Equal Error Rate (EER) can be derived, which denotes the

value of γ where the FMR and FNMR are equal. Graphically, the FMR and FNMR are often expressed by a Receiver Operating Characteristic (ROC) curve. The ROC curve plots 1-FNMR versus FMR for the range of γ.

The ROC itself has been extensively studied in the literature. Hanley and McNeil demonstrated that for a two-class problem (i.e., a classification problem with two outcomes), the area underneath the ROC curve (denoted by AUC) represents the probability that randomly selected data from both classes can be correctly classified [15]. Martin et al. defined the Detection Error Tradeoff (DET) curve, as a variant of the ROC curve [16]. The DET curve plots the false non-match rate versus the false match rate, directly visualizing the tradeoff between observing both types of errors. In addition, Green and Swets define the d' metric, which similar to the AUC, attempts to qualitatively measure the ROC using a single number [17].

Measuring Identification Performance

When evaluating identification performance, a set of probe samples (of size N_{probe}) is matched against a set of reference samples (of size N_{ref}), resulting in N_{probe} sets of match scores, with each set containing N_{ref} match scores. The match scores in each set are then ordered from highest to lowest.

In open-set identification, these sets are used to assess the *False Positive Identification Rate* (FPIR) and *False Negative Identification Rate* (FNIR), where the FPIR and FNIR are defined as the proportion of false positive identification and false negative identification errors observed from each of the N_{probe} ordered match score sets. Typically, the FPIR and FNIR are measured as a function of both γ and N_{ref}, the number of reference samples.

In closed-set identification, the ordered score sets from the N_{probe} probes are used to estimate the general probability that the correct matching identity pertaining to a probe is observed within the top K ($K \leq N$) ranks. In other words, the FNIR is computed with $\gamma = 0$. These probabilities are typically expressed visually through the Cumulative Match Characteristic (CMC) curve [8, 18, 19, 20, 21, 22]. Unlike the ROC curve, which is generated by looking at genuine and impostor scores all-at-once, the data in the CMC curve is obtained based on the explicit ordering of genuine and impostor scores in each ordered score set. As

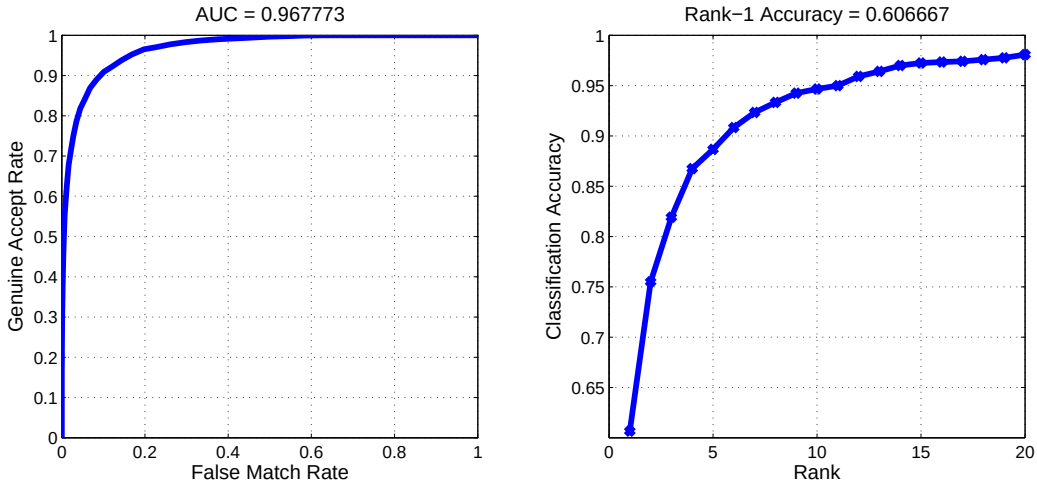

Figure 1.2: Example of the ROC curve (left) and CMC curve (right). Note in closed-set identification the CMC converges to a value of 1.0 as the number of ranks approaches the number of identities in the reference database.

such, the CMC curve can be defined as a *rank-based* metric. An example of both an ROC and CMC curve is presented in Figure 1.2.

In general, most biometric identification systems in real-world applications operate in the open-set mode [13, 14]. However, in the academic literature, most performance evaluations are conducted in the closed-set mode [8, 23, 24]. One possible explanation for this is that academic researchers do not have available the resources necessary to collect and maintain large-scale databases and thus by testing in closed-set, can utilize a maximum number of match scores for evaluation.

1.2 Biometric Surveillance Systems

1.2.1 Surveillance Systems

Surveillance has been a long used tool to protect private or commercial property. A simple surveillance system consists of a stationary camera, which records the events occurring within its field of view. Examples of such systems consists of Pan-Tilt-Zoom (PTZ) cameras and Closed Circuit Television (CCTV). Traditionally, these systems are used to deter criminal activity, as it is not unreasonable to suspect these cameras might be manually monitored

Figure 1.3: Sample images selected from the Performance Evaluation of Tracking and Surveillance (PETS) 2007 dataset [1]. Here, a scene is viewed by four surveillance cameras. Note the existence of correspondences between images.

or recorded. In the recent past, deployment of extensive networks [25, 26] of these cameras for the purposes of surveillance by businesses and local authorities has become increasingly popular. Today, it is increasingly common to view such networks within gas stations, commercial banks, casinos, shopping centers, schools, and in some cities, public street corners. Camera networks for major cities and municipalities is exceptionally large. For example, in Great Britain, more than four million cameras have been deployed, with at least 200,000 cameras in London [27]. Similarly, in the United States, cities such as Chicago (2,250 cameras) [28], New York City (2,500 cameras) [29] and New Orleans (1,000 cameras) operate extensive camera networks [30]. Figure 1.3 illustrates an example of surveillance imagery collected within an airport.

Given the ever increasing amount of surveillance information, it is near-impossible for it to be manually tended to at all hours of the day. This itself, presents a challenge. Media reports have indicated that despite the rise in deployed cameras, few are regularly watched

[31, 32, 33]. Moreover, studies have also demonstrated that overburdened human operators will profile, or selectively monitor individuals, based on factors such as race or age, which raises ethical questions [34, 35].

If there is an over abundance of observable data, how can it be reliably used to safeguard against criminal activity? The answer lies in automation of the surveillance process. Processing power of computer systems has been continually increasing, while the cost for it diminishes. This allows for implementation of reliable algorithms for automated surveillance systems [36, 37, 38, 39, 40]. An automated surveillance system consists of two components: an event detection scheme, which directs attention to potential threats and a tracking scheme, which can be used to rapidly fuse information from one person appearing in multiple cameras. An example of such a system is the problem of license plate recognition in traffic safety applications [41, 42].

Methods for event detection have also been characterized as activity recognition in the literature and have received extensive attention [43, 44, 45, 46, 47, 48, 49, 50, 51, 52]. Recent progress and maturation in this field is evidenced by the recent deployment of event detection algorithms in subway stations in San Francisco, California [53] and at the 2012 Republican National Convention in Tampa, Florida [54]. Regarding the tracking component, methods which consolidate information from individuals present in multiple cameras (at the same or different times) has also been an active area of study in the recent literature and is defined as the "re-identification" problem [55, 56, 57, 58, 59, 60, 61]. Information collected by the system for event detection or tracking could also be used to perform human identification. This can be accomplished by including a biometric recognition component into an automated surveillance system.

1.2.2 Formal Definition

A biometric surveillance system is therefore an automated surveillance system that uses video or images captured from a camera (or camera network) and uses the information acquired to perform biometric recognition. Such a system invokes the same architecture of a traditional biometric system (described in Section 1.1.1), wherein the sensor module

represents the camera systems used to collect surveillance data. The matching module of a biometric surveillance system operates *strictly* in identification mode. For this reason, a biometric surveillance system is sometimes referred to as a system performing "identification-at-a-distance".[3] By definition, a surveillance system aims to determine "who is present" in a scene, rather than individuals presenting a claim of identity.

In contrast to a traditional biometric system, a biometric surveillance system (or an identification-at-a-distance system) is unique in that the acquisition of biometric data is passive. That is, an individual does not need to voluntarily interact with the image sensor. This property is advantageous as it enables such systems to operate covertly, which may aid in reducing the probability an individual can maliciously circumvent the system. However, passive acquisition of biometric traits reduces the pool of candidate biometric traits a system may use. For example, traits such as hand geometry, fingerprint, voice and vascular structure require precise interactions with a sensor to properly acquire the biometric data and cannot be used. In addition, since there are fewer constraints (in general) emplaced on the acquisition of a biometric trait, the data obtained in covert settings may be more noisy, which can adversely affect matching performance.

1.2.3 Challenges in Biometric Surveillance Systems

Research towards the development and deployment of a biometric surveillance system is not new. Perhaps the first open challenge in the United States to develop such a system was from the Defense Advanced Research Projects Agency (DARPA) in the early 2000's. The goal of the program was "to develop automated biometric identification technologies to detect, recognize and identify humans at great distances" [62]. A selection of the intended goals of the program included developing algorithms for locating and acquiring individuals at distances of 500 feet [62].

Successful realization of a biometric surveillance system requires addressing the challenges associated with each component of the system: Detection, tracking and recognition. As previously mentioned in Section 1.2.1, event detection via recognition of actions is well-

[3]In this dissertation, the terms biometric surveillance system and identification-at-a-distance convey the same meaning and are used interchangeably.

studied topic in the literature [43, 44, 45, 46, 47, 48, 49, 50, 51, 52], as is tracking of individuals [63, 64, 65, 37, 38, 39, 40, 66] and linking tracks between cameras (i.e., re-identification) [55, 56, 57, 58, 59, 60, 61]. However, the recognition component has many challenges that have not been sufficiently addressed. This includes subcomponents such as an appropriate biometric modality, the matching scheme, and analysis of errors.

Challenges in Biometric Modality

One of the major challenges of performing recognition at a distance is ensuring the biometric can be reliably collected, as a surveillance system operates in an unconstrained environment where individuals do not voluntarily interact with the camera system. That is, selection of an appropriate biometric modality must balance the trade-off between collectability and distinctiveness (Section 1.1). For example, biometric recognition via the iris has demonstrated extremely high discrimination power, with some large-scale system operators claiming no false match has been recorded [67]. However, conventional iris matching requires a *minimum* eye radius of 70 pixels [68]. In low-resolution images (such as those in a CCTV camera), this is simply not enough data to perform matching. While the camera resolution could be increased, this brings the enormous challenge of localizing an iris in a large and active background, which may be computationally infeasible.

Face recognition could be viewed as an alternative biometric suitable for surveillance applications, as compared to iris, the resolution required to match a face is much lower. As such, this allows for a greater distance for which recognition can be achieved, a property desirable for surveillance. In addition, the challenges of localization are less daunting, and several researchers have contributed positively to this problem. Examples of face detection algorithms include, but are not limited to: Viola and Jones [69], Rowley et al. [70] and Hsu et al. [71]. However, assuming localization can be accomplished, challenges such as matching different profiles (i.e., viewpoints of the face), facial expressions, and illumination remain difficult challenges, which can negatively impact matching accuracy.

Given the inherent detection and localization challenges in performing face and iris recognition, it may be worthwhile to consider "non-traditional" biometric modalities as well. Human gait, for example, is a biometric trait that may be advantageous for surveillance

applications. Gait recognition is typically performed by extracting features from a set of images, of which explicit subject interaction with the camera is not required, thus satisfying the collection criteria for a biometric surveillance system. In addition, extraction of features for recognition can be accomplished at resolutions much smaller than that of face and in images of lower contrast [72]. Further support for gait as a potential biometric for surveillance systems is evidenced in criminal investigations wherein forensic experts were able to successfully implicate potential suspects in robbery cases in the United Kingdom [73] and Denmark [74] using human gait patterns from CCTV cameras. Localization of a human within an image is also a challenge, but, as with face, this topic is actively studied and relatively sophisticated in the literature. Examples of human detection algorithms include the works from Dalal and Triggs [63], Mikolajczyk et al. [64] and Tuzel et al. [65]. Due to the potential advantages of using gait for human recognition, for the purposes of this dissertation, an emphasis is placed on the methods and applications of human gait recognition.

Challenges in Matching

An additional challenge for a biometric surveillance system has to do with the manner in which matching of encountered individuals can be achieved. In an operational environment, a biometric surveillance system observes some number of individuals and presumably, a subset of whom the system has not previously observed. That is, it is very likely the system does not have in its local database the corresponding biometric data of every individual it observes. This may be particularly true for "non-traditional" or "soft" modalities such as gait. Complicating the issue, performing a controlled enrollment such that biometric data is available for a majority of individuals the system is expected to encounter is unrealistic. Even if such a task could be reasonably accomplished, it would be both fiscally and time intensive. This presents a challenge in the act of performing recognition, as a biometric system cannon perform its primary function if it is not able to deduce any information regarding the individuals it encounters. For this reason, a biometric surveillance system should be able to *adaptively* and *dynamically* assemble and maintain a reference database over time. Such a property has been understudied in the literature.

In addition, a biometric surveillance system is likely to acquire large amounts of data over

time. Consequently, it may be necessary for maintenance purposes to remove or otherwise consolidate duplicate reference entities (i.e., any pair of reference entities that correspond to a single identity) [75]. This process is defined as de-duplication (previously defined in Section 1.1.2). In a de-duplication matching outcome, the result can also result in a dynamic change in the composition of the reference database (i.e., when a reference sample is flagged or "de-duplicated" from the reference database).

Challenges in Error Analysis

Note that in both of the previously defined matching challenges (matching non-enrolled identities and managing duplicate entries), the query to the system may or may not have a matching reference entry in the database. This loosely resembles the "open-set" identification problem, as defined in Section 1.1.2. However, while error analysis for open-set identification does account for reference databases of varied size, the error rates (FPIR and FNIR) assume reference entries are correctly labeled. Should reference entries be assigned labels (i.e., identifiers) *automatically* by the system, it is possible that labeling errors (via matching errors) can alter the matching dynamics such that FPIR and FNIR no longer resemble the actual error rates incurred.

1.3 Human Gait Recognition

1.3.1 Introduction to Gait Recognition

Gait recognition is defined as the pattern of locomotion in animals. Human gait therefore, is the manner in which people walk. Human gait is studied by researchers in computer vision, psychological, biomedical, and biomechanical fields. Traditionally, biomedical and biomechanical researchers study gait as a diagnostic indicator [76, 77, 78] and to study mechanical loads [79, 80], whereas researchers in computer vision and psychology study gait as it pertains to recognition (i.e., biometric) capability. Unlike traditional biometric modalities (e.g., face, fingerprint, iris) gait is sometimes classified as a *soft biometric*. Human gait recognition is perceived as an attractive solution for distance based recognition for a

number of reasons. First and most importantly, human gait is believed to be unique to the individual. Psychological studies by Cutting and Kozlowski demonstrated that humans are capable of perceiving gender and known individuals based on gait [81, 82]. Second, the gait biometric can be acquired passively, meaning physical interaction with the system is not required to collect the gait biometric from an individual. Passive acquisition beneficial as individuals unaware of the system may be less likely to conceal or otherwise spoof biometric information. Finally, discriminatory features of human gait can be perceived in low resolution video sequences. This suggests that expensive camera systems are not required for gait recognition [72]. In general, methods for extracting features for recognition of human gait can be categorized as *model-based* or *model-free* approaches.

1.3.2 Classes of Gait Recognition Approaches

Model-Based Approaches

Model-based approaches use information collected from known structure of individuals or through models of the human body. For example, biped models are common, but vary on the level of complexity and type of information extracted. Features extracted through a model-based approach include spectra of thigh inclination [83], thigh rotation [84], stride and elevation parameters [2, 85] and cadence [86]. The primary reason for classifying gait in this manner is that these models allow for robust feature extraction. An example of a biped model and features that can be extracted are presented in Figure 1.4.

The primary benefit of a model-based approach is that if a model can accurately estimate the structural estimation of a human body, distortions or occlusions induced from the presence of objects are less likely to result in a decrease in recognition performance. However, model-based approaches are often complex and increased processing requirements may limit the application of these approaches in real-time environments.

Model-Free Approaches

Model-free approaches generally aim to extract features based on analyzing a moving shape. In general, the primary advantage of a model-free methodology is simplicity, as

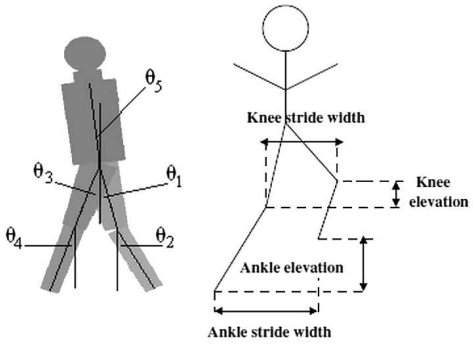

Figure 1.4: Example of a biped model used to extract featured in a model-based gait recognition algorithm [2].

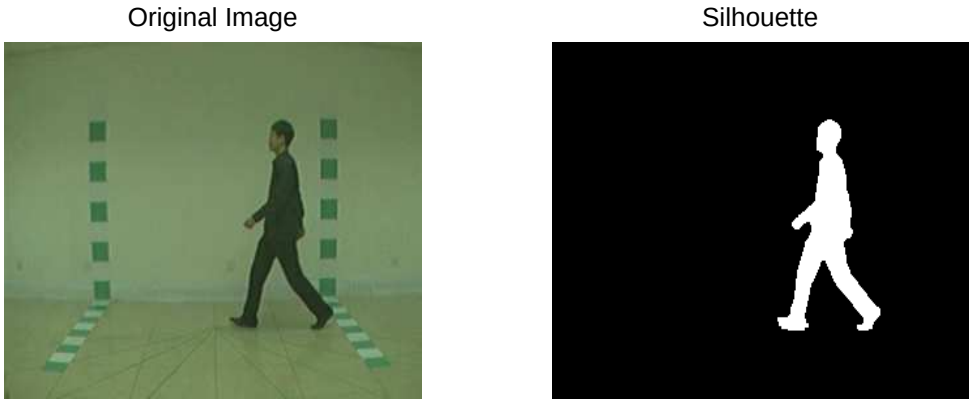

Figure 1.5: Example of a silhouette image (right) captured from raw video data (left). The silhouette image represents the shape or "contour" of a detected person.

features are entirely derived from silhouette shape dynamics. Typically, this is facilitated by the detection and conversion of the shape of a human individual into a binary "silhouette". An example of an extracted silhouette is illustrated in Figure 1.5.

The work by Niyogi and Adelson is recognized as one of the first approaches to automated gait recognition [87]. In their work, active contour methods were applied to the contour of a detected human and the change in the contour parameters was used as the feature for recognition [88]. Other early approaches were based on eigenspace projections using principle component analysis [89] or linear discriminant analysis [90].

Appearance based methods, which treat an image of the human body as the principal feature for recognition are perhaps the most popular approaches. Han and Bhanu first defined the Gait Energy Image (GEI) [91, 92], which is loosely defined as computation of the "average silhouette" and performing subspace optimization (e.g., PCA) and or discriminant analysis (e.g., LDA) to reduce the dimensionality of the image prior to matching. Due to its ease of computation, it is often cited as a benchmark for performance comparison and has spawned several variants. Several variants of the original GEI algorithm have been proposed, which denote pre-processing the "average silhouette" and or subspace optimization scheme. For example, the Enhanced Gait Energy Image (EGEI) by Yang et al., makes use of 2-D PCA to perform subspace optimization. Guan et al. also make use of 2-D PCA in conjunction with a random subspace projection as a method for subspace optimization [93]. Similarly, Tao et al. define a feature vector based on the Gabor response of the "average silhouette" and use a tensor discriminant for subspace optimization [94]. Zhang et. al define the Active Energy Image (AEI), which denotes an image representation based on the successive difference between frames and uses 2-D Locality Preserving Projections (2DLPP) to perform subspace optimization [95]. A similar image representation is the Gait History Image (GHI) by Liu and Zhang [96]. Chen et al. define a method that combines the positive difference of silhouette images to the GEI image to define the Frame Difference Energy Image (FDEI) [97]. Tan et al. also suggest cropping the "average silhouette" to just the "head" and "torso" regions [98]. This is defined as the Head Torso Image (HTI) [98].

Other approaches attempt to extract features from key poses or "stances" from the sequence of silhouette images. Sundaresan et al., define a method by which key poses can be directly used in a Hidden Markov Model framework for matching [99]. Kale et al. also define a Hidden Markov Model framework using the silhouette contour width as the primary feature for matching [100]. The evolution of the silhouette contour has also been defined as a Frieze pattern, which has been discussed as a feature descriptor by several authors [101, 102, 103, 104]. Alternatively, Wang et al., first show the evolution of the silhouette contour can be warped into the Procrustes shape-space, which is a method for computing the difference between 2-D shapes [105].

1.3.3 Segmentation and Silhouette Extraction

Traditionally, most gait recognition algorithms begin by converting the raw image data (which includes a detected human) into a set of binary "silhouette" images. In particular, this most commonly applies to recognition algorithms of the "model-free" variety (Section 1.3.2). The silhouette image is a binary image denoting the 2-D shape of a human as it occurs in a particular instance within the raw image data (Figure 1.5). In general, since the objective of gait recognition is to identify spatiotemporal features regarding human movement, the raw pixel intensities are not important. In fact, use of such information may be inappropriate as raw pixel data contains information such as clothing, which is very likely to vary with time.

In the gait recognition literature, the simplest and most common method for performing silhouette extraction is through background subtraction [106]. In simple background subtraction, an image I_0, denoting a "blank" scene (no individuals are present) is compared against a similarly aligned image, I_k, which has an individual present engaging in some type of action (in this case, walking). The absolute difference of $I_0 - I_k$ yields a difference image, I_{diff}, which has large intensity values in the pixel regions denoting the captured human. Using a threshold, all pixel values exceeding the threshold value can be assigned as foreground (i.e., belonging to the human) and the remaining pixels can be assigned as background (i.e., not belonging to the human). To eliminate the effect of spurious motion artifacts and noise, the silhouette image is often post-processed with a set of morphological filters. This process is summarized in Equations (1.7) and (1.8), where B denotes the silhouette image and λ denotes an intensity threshold.

$$I_{diff} = abs(I_k - I_0) \tag{1.7}$$

$$B = \begin{cases} 1, & I_{diff} > \lambda \\ 0, & I_{diff} \leq \lambda \end{cases} \tag{1.8}$$

To improve the silhouette extraction process, the image data is often pre-processed with a contrast enhancement or edge enhancement filter. An example of a contrast enhancement process is Contrast Limited Adaptive Histogram Equalization (CLAHE) [107]. CLAHE enhances the contrast of images by partitioning an image, I, into many sub-regions (or

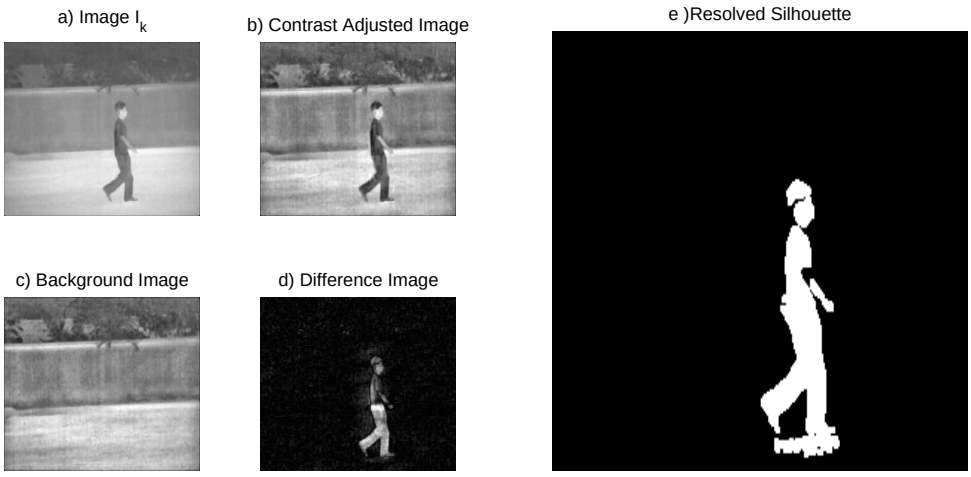

Figure 1.6: Segmentation process from a raw video image to a silhouette. a) Raw video image, b) Image following contrast enhancement, c) Background image, d) Difference image created by subtracting the contrast enhanced image from the background image, e) Resolved silhouette following thresholding and noise removal.

tiles) and applies histogram equalization to each region. Median filtering prior to contrast enhancement reduces the likelihood noise is amplified by the enhancement process. Figure 1.6 illustrates the process of resolving a complete silhouette image.

In the event the silhouette extraction process proves to be erroneous or difficult, population based Hidden Markov Models have shown to be capable of silhouette enhancement [108]. For very long sequences, consisting of many complete gait cycles, usage of the foreground sum signal (sum of silhouette area) can be used to identify the best subset of frames to use for recognition [109].

1.3.4 Measuring Silhouette Quality

In the biometric literature, it is commonly believed that the *quality* of the raw biometric data is correlated with the separability of genuine and impostor match scores [110]. In other words, the variance in the genuine and impostor match score distributions is believed to be influenced by the quality of the data used for feature extraction. As such, researchers have developed methods for quantifying the quality of the raw biometric data. The operational impact of a quality assessment subroutine is that poor quality data is automatically rejected

from the system or flagged for intervention from a human operator. An example of an established quality metric is the NFIQ fingerprint quality measure developed by the National Institute of Standards and Technology (NIST) [111].

In the gait recognition literature, one such measure for measuring silhouette quality has been developed by Liu et al. [109]. The measure is based on an estimate of the noise in the "foreground-sum signal", which is defined as the sum of all foreground pixels in a silhouette image over a sequence of silhouette image data. A brief description of this measure is provided in the following paragraphs.

In a set of binary silhouette images, B_k $(k = 1, 2, \ldots)$, the foreground-sum signal is defined as the sum of foreground pixels in the k^{th} image. Mathematically, this is defined in Equation (1.9). In theory, $\phi(k)$ is periodic according to the rate at which an individual performs one half of a gait cycle.

$$\phi(k) = \sum_{\forall x} \sum_{\forall y} B_k(i, j) \tag{1.9}$$

Holes in the silhouette, or spurious foreground pixels caused from shadow artifacts can greatly impart noise to $\phi(k)$. Noise is measured in a three-step process. First, $\phi(k)$ is spatially normalized. Denote the spatially notmalized signal as $f(k)$. Spatial normalization is necessary to handle data collected at varying spatial resolutions. This also enables the quality metric to be compared across datasets. The normalization process is defined in Equation (1.10). Note the normalization parameters a_0 and a_2 in Equation (1.10) denote the DC component and amplitude of $\phi(k)$, respectively.

$$f(t) = \frac{\phi(k) - a_0(k)}{a_2(k)}, \text{where } a_0 = E[\phi(k)],\ a_2 = \frac{sup\{\phi(k)\} - inf\{\phi(k)\}}{2} \tag{1.10}$$

Next, the autocorrelation matrix of $f(k)$, $R_{f(k)}$, computed. An eigenvalue decomposition is then performed on $R_{f(k)}$. The resulting quality metric is denoted as ψ and is obtained by summation of the first two eigenvalues (λ_1, λ_2), followed by subtraction of the 5^{th} to d $(\lambda_5, \ldots, \lambda_d)$ eigenvalues, where d is the dimension of the autocorrelation matrix (Equation (1.11)). In general, the periodicity of a silhouette is captured by λ_1 and λ_2, while $\lambda_5, \ldots, \lambda_d$

denotes noise. Note that the 3^{rd} and 4^{th} eigenvalues are neglected. This is intentionally done as these components are likely to reflect both periodicity and noise[109]. Thus, a large value of ψ is indicative of higher quality silhouettes.

$$\psi = \sum_{i=1}^{2} \lambda_i - \sum_{i=5}^{M} \lambda_i \tag{1.11}$$

1.3.5 Perceived Challenges in Gait Recognition

Gait recognition has many perceived challenges, many of which are applicable in regards to a surveillance system. In response, the research community has been actively working to address these issues. Some of these issues are have received more attention than others.

Robustness to Silhouette Variations

One of the primary concerns is that most gait recognition approaches, in particular those of the model free variety, are less robust to silhouette variations that may arise in the silhouette. These variations can be a consequence of the person, as when an individual is observed carrying objects or with different clothing [112]), or a consequence of poor segmentation (i.e., the silhouette is degraded by erroneous holes or contains shadowing artifacts). In fact some research as demonstrated that the observed matching performance can be artificially *increased* if the matcher is encoding *silhouette errors* (e.g., shadow effects) [108].

Regarding the effect of clothing, the usage of part-based models has been proposed as a means to mitigate this challenge [113], but it remains an open challenge. Similarly, with regard to errors in silhouette extraction, with exception to the study by Liu and Sarkar [108], this problem is generally not studied in the gait recognition literature.

Though silhouette extraction via background subtraction is computationally simple and effective in a laboratory setting, its effectiveness in resolving silhouettes in less constrained environments (such as in an outdoor setting) is likely to be diminished. For example, in an outdoor setting, natural illumination changes via sunlight or swaying tree branches can impart the ability to quantify the background using a single image. Unfortunately, most researchers in the gait recognition literature neglect the problem of silhouette extraction

in difficult environments. Some of these challenges have been addressed by the general computer vision community. For example, Kim et al. [114] define a "Codebook Model" for background subtraction, which attempts to model the expected mean and variance in pixel intensities within a local neighborhood, thereby using an adaptive threshold for foreground classification. In addition, methods for modeling the background using gaussian mixture models [115] and fuzzy logic [116] have been proposed.

Though the aforementioned segmentation methods are more robust to illumination invariance, it is important to note that these algorithms are designed to operate in the *visible* spectrum. In a surveillance context, there is no assurance that images will have three channels of information. For example, CCTV images are often grayscale images. Aside from the work by Jacques et al. [117] denotes a background detection and shadow removal method explicitly for grayscale images, there is very little research focusing strictly on grayscale imagery. Further, there are no studies confirming whether methods applicable for grayscale imagery in the visible spectrum can also be extended to other image spectrums (e.g., thermal, short-wave infrared, etc.).

Robustness to Viewpoint

Viewpoint (i.e., the angle an individual is observed walking in the image plane) is also considered a challenge in gait recognition. Arguably, gait is best captured with respect to the Sagittal plane (side-profile) of the human body, since the dynamics of the legs, arms and body are most visible. Research has shown however, that matching can be performed when evaluating gait sequences from multiple viewpoints [118], but a different camera view between a probe sample and a reference entity will result in a loss of matching accuracy [24, 112]. Models for transforming the silhouette to a common domain have been proposed to rectify this issue. These models have shown promise when the difference in viewpoint is minimal [119] or large [87], but not for both cases inclusively.

Robustness to Time

One generally unanswered question regarding human gait recognition is the whether the gait biometric is stable over time (i.e., its permanence). To date, there is not sufficient

research to conclusively determine whether time impacts the stability of the gait biometric. If time is a factor, it is necessary to understand the rate at which the biometric trait degrades. Some large scale studies have attempted to investigate this issue. For example, the HumanID Gait Challenge dataset [120] does include data for temporal analysis and the baseline evaluation does demonstrate a loss of recognition performance over a period of six months. However, the temporal analysis was not an intended area of study and consequently the probe and reference data where time is a factor is also included with additional confounding factors. Conversely, a small study by Matovski et al. concluded that time *does not* have an impact on gait [121]. This study was performed using 25 individual identities in a controlled environment over a period of nine months, where an identification accuracy of 95% was reported between the two collection periods. Presently, no other studies or datasets in the literature seek to address this issue.

1.4 Motivation

The motivation of this dissertation is to formally introduce and discuss some of the aforementioned challenges facing a biometric surveillance system (or a biometric system performing identification-at-a-distance). In particular, human gait recognition in the short-wave infrared (SWIR) spectrum is highlighted as a potential solution for a suitable recognition modality and image spectrum. In addition to an evaluation regarding the ability to perform gait recognition in an operational environment, a subsequent analysis is performed regarding whether gait patterns (on a silhouette-level) can be clustered.

Beyond the selection of image and recognition modalities, it is also important to address the operational challenges of a covert biometric surveillance system. Namely, how might the system dynamically update its database, and given a method to facilitate this, is the probability of error affected? In addition, an operational surveillance system may collect an abundance of data. Consequently, it may be necessary to de-duplicate (e.g., consolidate) identity profiles. However, the errors surrounding the (general) de-duplication task have not been described in the literature. Thus, an analysis is performed to ascertain whether traditional measures can predict de-duplication error. Finally, an analysis is performed with

regard to *academic* evaluations of the identification problem. Typically, such evaluations are performed in closed-set and measured with the CMC curve. However, the performance values presented by the CMC curve may not reflect a true indication regarding the separability of match scores (which is depicted in the ROC curve). Thus an analysis is performed relating the various outcomes that can occur from both curves from the same match score data.

1.5 Contributions

The primary contributions of this dissertation are as follows:

1. A method for performing automated human gait recognition is presented. Inspired by the work by Wang et al. [105] and defined as the "Gait Curves" algorithm, the algorithm denotes a model-free method for matching a sequence of silhouette images by treating the "left" and "right" contours of the silhouette as 1-D signals and warping them into the Procrustes shape-space for matching. The matching performance of the Gait Curves algorithm is comparable to existing methods in the literature on benchmark datasets.

2. Matching performance of the Gait Curves algorithm (and benchmark algorithms) is also evaluated on a new, novel dataset. Defined as the WVU Outdoor SWIR Gait (WOSG) dataset, the dataset is distinct from traditional gait recognition datasets in that it is designed to be more closely related to the type of data that may be encountered in an operational setting. In addition, the WOSG dataset is the only large gait recognition dataset that has been collected in the Short-Wave Infrared (SWIR) spectrum, which may be an operationally advantageous spectrum for an operational gait recognition system.

3. By treating individual gait curves as 1-D signals, it is also possible to detect the presence of carried objects that can distort the shape of the silhouette and correct for them. As such, a method for accomplishing this is demonstrated in the case of backpack detection.

4. An analysis is performed demonstrating that human gait can be clustered using a variety of feature descriptors (i.e., methods) and clustering algorithms.

5. A matching framework is defined which may be advantageous for surveillance-based applications. Defined as "Anonymous Identification", the framework enables the automatic enrollment of probes into the reference database when a match to an observed probe cannot be found.

6. Since Anonymous Identification denotes a *variant* of the classical open-set identification problem, where the contents of the reference database can be altered following a matching decision, a comprehensive error analysis is performed. The result of the error analysis suggests that (a) the sequential order in which probes are observed can profoundly impact the observed error rate of the system and (b) traditional measures for reporting biometric performance (i.e., FMR, FNMR, FPIR and FNIR) do not accurately model anonymous identification error.

7. To account for the failure of traditional measures to describe anonymous identification error, an error model is developed to act as a better prediction of these errors. The proposed error model is also relevant to matching schemes where the contents of the reference database can change depending on a matching outcome, as in biometric de-duplication and re-identification.

8. A follow-up study is performed into the errors of biometric de-duplication, which is relevant for the maintenance of large biometric databases. The analysis demonstrates that in a simple problem space, de-duplication error cannot be accurately assessed using traditional measures for error analysis.

9. An in-depth analysis regarding the relationship of ROC and CMC curves is performed. The principle results demonstrate that a single ROC curve can be mapped to many CMC curves. Consequently, the CMC curve should always be accompanied by the ROC curve in order to better characterize performance.

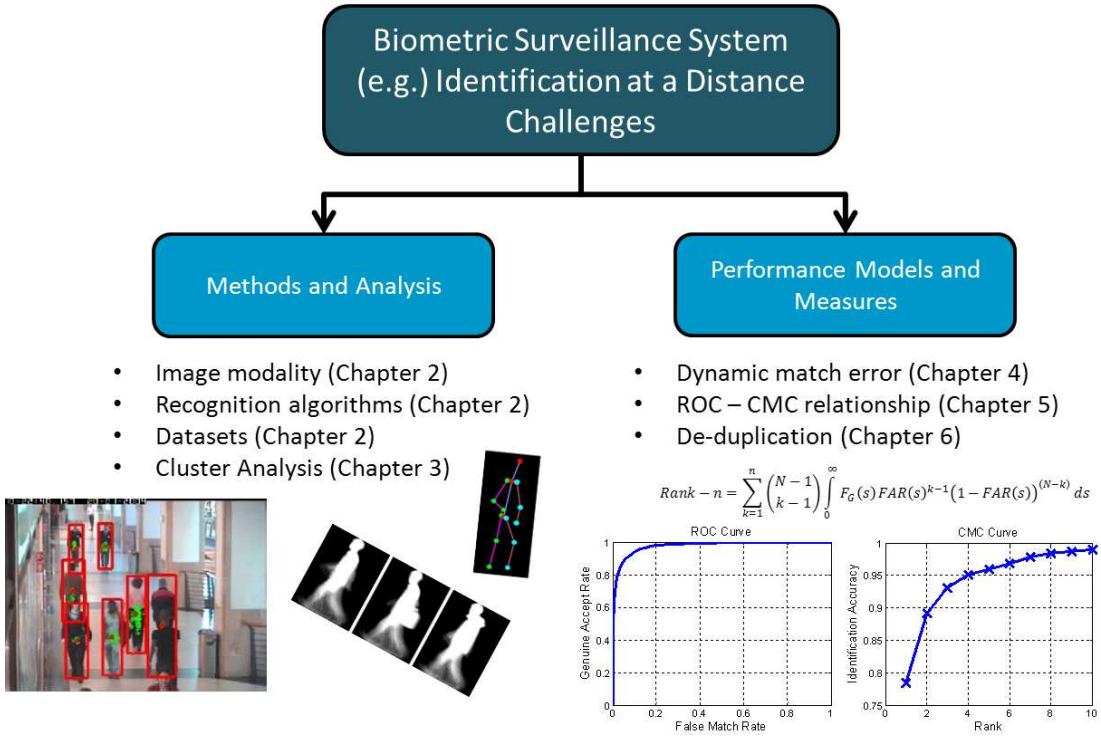

Figure 1.7: The organization of this dissertation can be viewed into two components: "Methods and Analysis" and "Performance Models and Measures".

10. The analysis regarding the ROC and CMC relationship is supported by a method for creating *faux identities* from an input set of genuine and impostor match scores. The method leverages the Doddington's Zoo concept to create different inter- and intra-class relationships between identities.

1.6 Thesis Organization

This dissertation can be characterized into two components: "Methods and Analysis" and "Performance Models and Measures". In the "Methods and Analysis" component, the principle aim is to discuss the challenges surrounding the hardware (e.g., image sensor) and matchers (e.g., recognition algorithms) in a biometric surveillance system performing automated gait recognition. Chapter 2 discusses much of this, providing an algorithm for performing automated gait recognition and performs an evaluation on a new and challenging dataset, which acts as a better representation of operational data. Chapter 3 provides an

investigative look into gait recognition as a clustering problem, where the objective is to determine if the gait biometric can be clustered, and if generated clusters can be explained in a physiological context.

In the "Performance Models and Measures" component, the principle aim is to investigate the functionality of traditional biometric measures (e.g., ROC and CMC curves, FPIR, FNIR, etc.) in describing error in biometric surveillance systems; particularly systems that engage in a dynamic matching process (e.g., the matching outcome impacts the composition of the reference database). This is primarily expanded upon in Chapter 4, in the context of an "Anonymous Identification" system, which generates "identity profiles" by dynamically enrolling observed probes into the database, as opposed to using (i.e., assuming) a fixed subset of enrolled reference samples. In Chapter 5 an analysis is performed on the data represented by the ROC and CMC curves, where the former is not typically presented in identification-based evaluations. A case is presented suggesting that *both* curves should be reported in an identification-based evaluation, as the CMC curve may not depict the same information as the ROC curve. Chapter 6 presents an investigative study concerning the de-duplication problem, and whether traditional error measures (Section 1.1.4) can be used to predict error in a de-duplication process.

Finally, Chapter 7 summarizes the findings of this dissertation and presents suggestions for future work and recommendations to researchers in the field. A visual overview of this organization is provided in Figure 1.7.

Chapter 2

Methods for Recognition of Human Gait

2.1 Introduction

2.1.1 Design of Gait Recognition Algorithms and Datasets

The early literature surrounding gait recognition served primarily to demonstrate that automated gait recognition may be possible [83, 86, 88, 90, 101, 122]. However, many of these early works demonstrated performance on small datasets, typically consisting of few individual identities and or constraints. Such conditions are not representative of an operational biometric system (traditional or at-a-distance). Thus, in order to advance gait recognition as a candidate biometric modality for a biometric surveillance system it is necessary to (a) design recognition algorithms capable some degree of robustness to environmental (e.g., distance from camera, resolution of captured individuals) and individual variances (e.g., presence of objects, cadence, walking direction, etc.) while remaining computationally efficient and (b) develop and design datasets that more closely represent an operational environment (Chapter 1, Section 1.3.5).

With respect to algorithms for automated recognition of human gait, an ideal gait recognition algorithm would be defined by the following properties:

- *Matching Accuracy*: The algorithm should demonstrate success in identifying large

numbers of individuals.

- *Robustness to Covariates*: The algorithm should be able to perform recognition in the presence of challenging matching situations. Examples of such situations include matching to differing walking directions, speeds, and the presence of objects.

Although trivial, any algorithm developed must first show some degree of baseline matching accuracy. In particular, demonstration of high matching accuracies from an "established" test database with a large number of identities. Here, "established" refers to a database utilized and accepted by researchers in the field. In addition, a recognition algorithm must show some degree of robustness to challenging matching situations (i.e., covariates), which can confound the matching accuracy. Examples of such situations include matching individuals walking in different directions (i.e., walking paths), individuals wearing different clothes and clothing styles, and individuals carrying objects.

With respect to datasets, the following criteria are desirable for operational data:

- *Walking Protocol*: Walking paths should include multiple directions at varying distances to the camera.

- *Environment*: Data collection should occur in an uncontrolled environment; preferably one that is outdoors.

- *Image modality*: The camera hardware should represent the type of data used in an operational setting.

Regarding the walking protocol, in an operational setting, it is likely that an individual will be observed at walking in any direction, at varying distances to the camera. This is important, as a matching algorithm must be able to compare feature vectors obtained from sequences acquired at differing spatial resolutions. Regarding the collection environment, it is likely an operational gait recognition system will encounter multiple individuals and spurious motion artifacts within a scene. It may also be possible that environmental properties such as illumination (via cloud cover) could vary in the short term. Although these issues are more directly related to segmentation and silhouette extraction, they are nonetheless critical

for an operational system. Finally, the image modality (e.g., RGB, Infrared, etc.) should be similar to the type of data used operationally. Although this may appear trivial, images in different spectrums have different challenges. Therefore it is important to target those challenges which are most pertinent for operational data.

2.1.2 Chapter Motivation

In this chapter, both (a) an algorithm for performing gait recognition and (b) a challenging dataset for investigating gait recognition are described. The matching algorithm that will be discussed matches what is to be defined as "gait curve" matching; or more formally, the "Gait Curves" matching algorithm. In general, a gait curve denotes a mathematical representation of ones gait, acquired from a set of frames in which an individual is observed. Compared to existing methods in the literature, matching gait curves is computationally efficient and does not require a training component. In addition, the properties of a "gait curve" can also be exploited for object detection and restoration of distorted silhouettes. The matching performance of the algorithm is compared against two additional algorithms from the gait recognition literature on standard datasets. Finally, the matching performance of the Gait Curves algorithm is evaluated on a new and challenging dataset, which is defined as the WVU Outdoor Gait SWIR (WOSG) dataset. The dataset is designed to mimic a number of the challenges faced in operational data that are not present in existing gait datasets. One feature of the WVU Outdoor SWIR Gait dataset denotes data acquired from the short-wave infrared (SWIR) spectrum, which may be operationally advantageous over other image spectrums (e.g., RGB).

2.2 Recognition by Matching Gait Curves

2.2.1 Static Feature Extraction

As a silhouette traverses across the viewing plane, static parameters can be collected at each frame. For example, raw silhouette height, can be extracted by calculating the

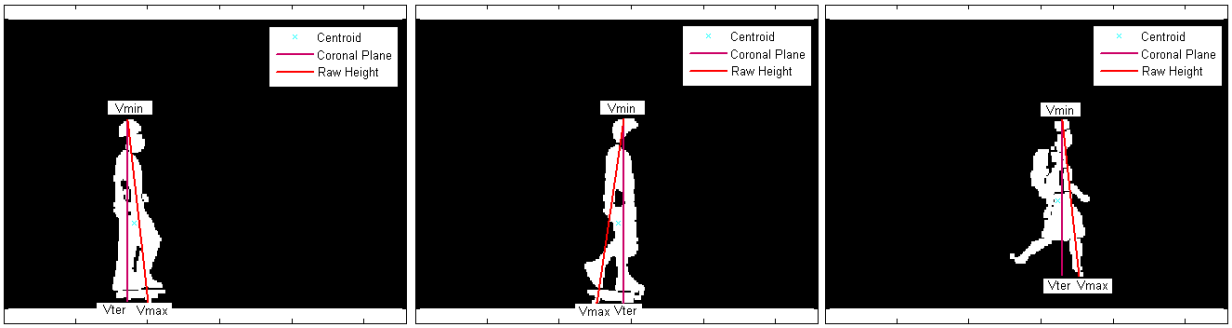

Figure 2.1: Labeled silhouettes. Note the difference in marking the coronal plane and centroid.

maximum difference in vertical silhouette coordinates.[1] Let $v_{min} = \{i_{vmin}, j_{vmin}\}$ and $v_{max} = \{i_{vmax}, j_{vmax}\}$ denote the pixels corresponding to the minimum and maximum vertical coordinates, respectively, for which the binary silhouette image, $B(i, j) = 1$. The silhouette height, h, can then be simply computed as $h = i_{vmax} - i_{vmin}$.

Since this measure is relative to the distance of an individual from the camera, it cannot be used as a unique feature without knowledge of local markers or depth [123]. In lieu of this shortcoming, raw silhouette height can be used in the design of additional features. The first such use is in isolating the coronal plane of the silhouette. In a matrix coordinate system, v_{min} represents the top of the head. This location is stable for any gait sequence, regardless of the direction of walk and therefore, can be treated as the peak of the coronal plane. v_{max} however, will shift to either the left or right foot, depending on variations in stance. The terminus of the coronal plane is then located at $v_{ter} = \{i_{vmax}, j_{vmin}\}$. An alternative method to determine the coordinates of the coronal plane would involve computing the centroid, which is the center of mass of the silhouette. However, presence of carried objects, arm sway, or holes in the silhouette (via segmentation errors) can greatly distort the horizontal position of the centroid. Thus, identification of the coronal plane using v_{min} and v_{ter} is favored as it is less susceptible to these effects. Refer to Figure 2.1 for a fully labeled silhouette.

[1] Refer to Chapter 1 for a definition of a silhouette.

2.2.2 Spatiotemporal Feature Extraction

Following calculation of the coronal plane coordinates, the left and rightmost pixel locations of the outermost contour are obtained for each row in the silhouette. That is, for the *pth* row where $p \in [i_{vmin}, i_{vmax}]$ these pixels are denoted as g_p^{left} and g_p^{right}, respectively. Subtraction of the horizontal position of the coronal plane from these point sets yields a space normalized contour, denoted as the gait curve, G_k, for the k^{th} frame (e.g., image) in the sequence.

Evolution of a set of gait curves across several frame can therefore be regarded as a spatiotemporal feature for shape based analysis. In other words, the output of an arbitrary function $F(G_1, G_2, \ldots, G_K)$ is a single gait curve encompassing the shape dynamics captured in a video sequence of K frames. For example, the output of function F could be the mean of the G_k's.

$$F(G_1, G_2, \ldots, G_K) = \frac{1}{K} \sum_{k=1}^{K} G_k. \tag{2.1}$$

While Equation (2.1) represents one potential method for representing a set of gait curves, alternative solutions exist as well. The Procrustes Meanshape [124, 125] is a mathematically elegant measure of representing and evaluating 2-D shape sets. An advantage of this measure is that differences in translation, rotation, or scale do not negatively impact the resulting match score between shapes transformed into the procrustes shape space. In the context of gait recognition, this is particularly advantageous as it is likely individuals will be observed at varying distances and at varying spatial resolutions. Conversely, computation of Equation (2.1) cannot guarantee these properties.

To derive the procrustes meanshape of K individual gait curves, the number of elements in each individual gait curve must first be normalized (i.e., interpolated) to the same size. Denote this number of elements as T. The spatial coordinates of each element in each gait curve are then encoded as a complex number. Next, the sample mean of each gait curve is computed and subtracted from each individual gait curve, effectively aligning each gait curve at the origin. Finally, a scatter matrix, $\mathbf{S_u}$, is computed. These operations are summarized in Equations (2.2)-(2.6).

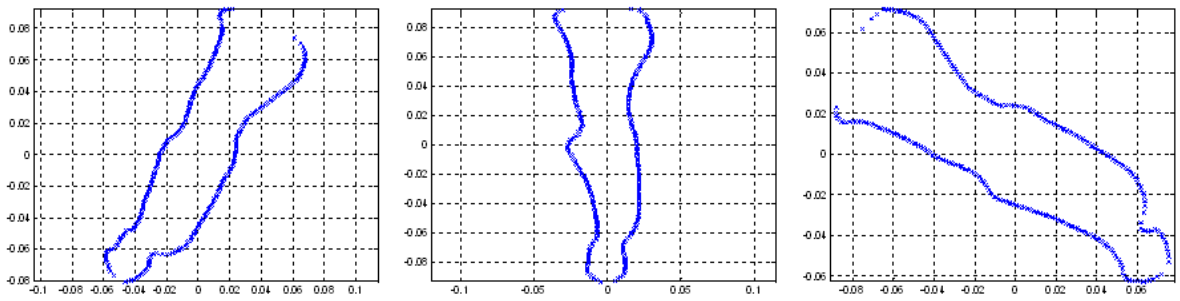

Figure 2.2: The procrustes meanshape computed from gait curves corresponding to three different individuals.

$$\mathbf{z_k} = Re(G_k) + jIm(G_k); \tag{2.2}$$

$$\bar{\mathbf{z}} = \sum_{i=1}^{k} \frac{\mathbf{z_i}}{k}; \tag{2.3}$$

$$\mathbf{u_k} = \mathbf{z_k} - \bar{\mathbf{z}}; \tag{2.4}$$

$$\mathbf{u} = [\mathbf{u_1}, \mathbf{u_2}, ..., \mathbf{u_K}]; \tag{2.5}$$

$$\mathbf{S_u} = \sum_{j=1}^{K} \frac{(\mathbf{u_j u_j^T})}{(\mathbf{u_j^T u_j})}. \tag{2.6}$$

The first eigenvector of the scatter matrix, $\mathbf{S_u}$, is used to denote the procrustes meanshape, $\bar{G}$, from K gait curves. Note that at least one full gait cycle and a minimum of 10 gait curves should be used to generate $\bar{G}$. This constraint is necessary to ensure that enough information has been captured to create a distinguishable $\bar{G}$. A visual example of the procrustes meanshape on a set of gait curves for three individuals is presented in Figure 2.2.

2.2.3 Matching Gait Curves

To obtain a match score between a pair of procrustes meanshapes (e.g., gait curve representations), the procrustes distance is used. This distance is defined in Equation (2.7). The

range of values produced by Equation (2.7) is limited to $[0, 1]$, where the smaller the value, the more similar the shapes $\left(\bar{G_1}, \bar{G_2}\right)$.

$$d(\bar{G_1}, \bar{G_2}) = 1 - \frac{\left|\bar{G_1}^T \bar{G_2}\right|^2}{\left\|\bar{G_1}\right\|^2 \left\|\bar{G_2}\right\|^2} \tag{2.7}$$

2.2.4 Backpack Detection

The gait curve $\bar{G}$ can be further exploited by treating it as a one dimensional signal, $y[t]$ $(t = 1, 2, \ldots, T)$, with the y-axis denoting the silhouette height and the x-axis denoting the distance from the coronal plane to the silhouette boundary. This is accomplished simply by evaluating the distance between the coronal plane and g_p^{left} and g_p^{right}. The end result is a signal of length T that indicates the horizontal distance between each gait curve point to the coronal plane.The first $t = 1, 2, \ldots, \frac{T}{2}$ points correspond to the "back" half of the gait curve and the remaining $t = \frac{T}{2} + 1, \frac{T}{2} + 2, \ldots, T$ points correspond to the "front" half of the gait curve. For the purpose of backpack detection, the "back region" of $y[t]$ is of particular interest. Here, define the "back region" as the subset t corresponding to the back of an individual's extracted gait curve. If an individual is carrying a backpack, it would be expected that the distance between the gait curve and the coronal plane would be greater in the "back region". Intuitively this is likely since the presence of a backpack should outwardly distort the silhouette shape. Thus, the area under the curve in the back region should be greater given the presence of a bag. Since the signal has also been previously interpolated to exactly T points (as per the formation of a set of gait curves), the subset denoting the back is relatively consistent across individuals. This allows for estimation of a window where the back region likely exists. However, this signal is also a function of silhouette resolution. To account for this, a normalization factor δ is included to scale the "waist region" of $y[t]$ to a distance of 1 unit from the coronal. Here, the "waist region" is defined as the values of t pertaining to the waist in the first $\frac{T}{2}$ points. The waist is chosen for its consistency in the gait cycle. That is, it does not perturb much as an individual is observed walking. Refer to Figure 2.3 for a labeled example of this signal for the "back" half of $y[t]$.

In observing the statistics of Figure 2.3, the intuitive notion about the silhouette shape for

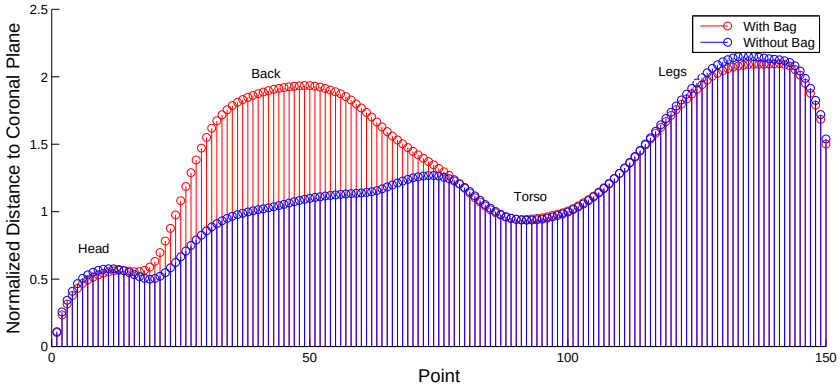

Figure 2.3: Signal representation of half of the gait curve. Note this portion of the gait curve denotes the region encompassing the human back, particularly between values 25 to 75 (where $T/2 = 150$).

individuals with and without bags is verified. This information also provides the necessary window in which to target backpack related features. In this case, a loosely defined back window, w_{back} is the interval $[\frac{T}{6}, \frac{T}{2}]$.

Given a sequence $y[t]$, each of the following features are collected. First, an intuitive feature for backpack detection would be a discrete summation of the values within the window.

$$f_1 = \sum_{t \in w_{back}} y[t] \tag{2.8}$$

Secondly, a threshold-based feature is introduced to observe how often a pre-determined width of the back region, y_θ, is exceeded. The value of y_θ is slightly higher than δ, the normalization threshold, and is empirically defined.

$$f_2 = \sum_{t \in w_{back}} I(y[t] > y_\theta), \text{where } I(.) \text{ is the indicator function} \tag{2.9}$$

The third feature used is the total power of the signal in the back region. The motivation for this feature is to account for high frequency components that may arise as a result of the presence of a backpack.

$$f_3 = \sum_{t \in w_{back}} (\left|\text{FFT}(y[t])^2\right|) \tag{2.10}$$

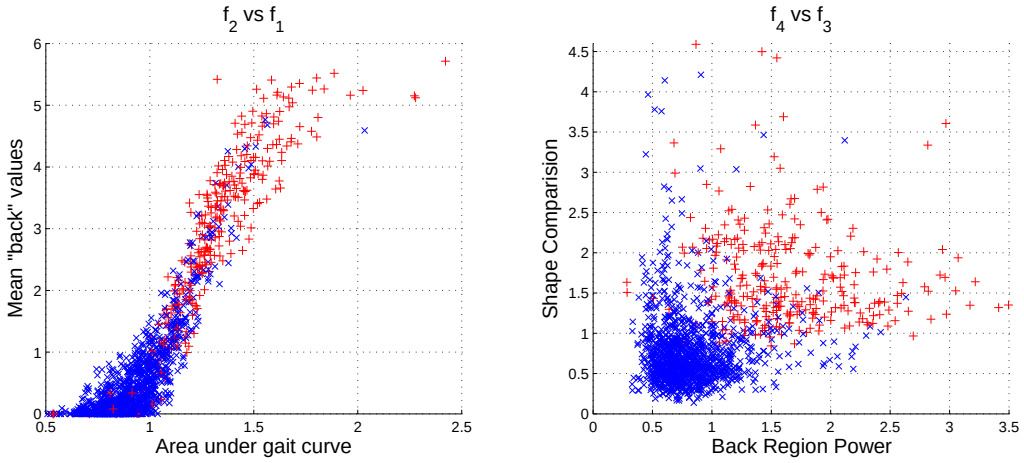

Figure 2.4: Scatter plots of bag detection features. (+) With Bag, (x) Without bag.

Finally, over a sequence of frames, the extracted gait curve eigenvector, $\bar{G}$ can be itself used as a feature for backpack recognition. Using a set of training data, a generalized procrustes shape of an individual, $\hat{G}$ can be extracted using the same procedure as mentioned in Section 2.2.2, and direct comparison between $\bar{G}$ and $\hat{G}$ yields a shape difference statistic where the procrustes distance is used to denote the likelihood an individual is carrying an object.

$$f_4 = 1 - \frac{\left|\hat{\mathbf{G}}^{\mathbf{T}}\bar{\mathbf{G}}\right|^2}{\left\|\hat{\mathbf{G}}\right\|^2 \left\|\bar{\mathbf{G}}\right\|^2} \tag{2.11}$$

Here, f_1, f_2, and f_3 can be computed from each silhouette image, while f_4 requires a sequence of images. To account for this, the full backpack detection feature vector combines f_4 with an average of f_1, f_2, and f_3 computed from K frames. Figure 2.4 provides a visual interpretation of the separability of these features.

2.2.5 Silhouette Rectification

Most model-free approaches to gait recognition generally regress in performance when objects are introduced. This is to be expected for algorithms utilizing shape dynamics, as objects modify the spatial appearance of the silhouette in time. Larger objects, such as a backpack or briefcase fall into this category. However, if the properties of the gait

curve can be used to enable object detection, they can similarly be used in an attempt to correct for object induced distortions. As previously mentioned in Section 2.2.4, the average gait curve extracted from an individual both with and without a backpack are very similar except for the back region, w_{back}. This is evidenced visually in Figure 2.3. Unlike a detection algorithm, where a general window can be targeted to isolate the likely position of a backpack, a *correction* scheme requires additional precision to identify the position of the backpack. This is to ensure that any correction does not erroneously impart non-backpack data points. In order to accomplish this, a refined window of w_{back} is created. This window is based on estimating the values of t corresponding to an increase in $y[t]$ from the "waist" and a decrease in $y[t]$ to the "head". These are denoted as t_{head} and t_{waist}, respectively. Based on the statistics of Figure 2.3, small windows representing t_{head} and t_{waist} are initially implemented to observe their respective divergence and convergence areas. Using minimum and maximum operators, the estimates for t_{head} and t_{waist} are extracted.

$$t_{head} = \min(w_{head}) \tag{2.12}$$

$$t_{waist} = \max(w_{waist}) \tag{2.13}$$

The minimum operator is chosen to identify t_{head} because the head region contains a local minima representing the neck of an individual. From this point, a backpack is likely to rapidly project outwards, as can be seen from Figure 2.3. Conversely, the maximum operator is chosen to isolate t_{waist} since backpack length is variant, and the convergence region trends downward.

Three types of corrections were investigated. Each corrected signal is denoted as $c_i[t]$ $(i = 1, 2, 3)$. The first simply forces any value above the normalization threshold to 1. An advantage of this measure is that the correction is less impacted by erroneous estimation of t_{head} and t_{waist} since the values surrounding these locations are generally less than δ.

$$c_1[t] = \begin{cases} 1, & y[t] > \delta, \quad t_{head} < t < t_{waist} \\ y[t], & \text{otherwise.} \end{cases} \tag{2.14}$$

In the second correction method, $y[t]$ is modified such that the values between $y[t_{head}]$ and $y[t_{waist}]$ are linearly connected. Since the back region of an individual not carrying a

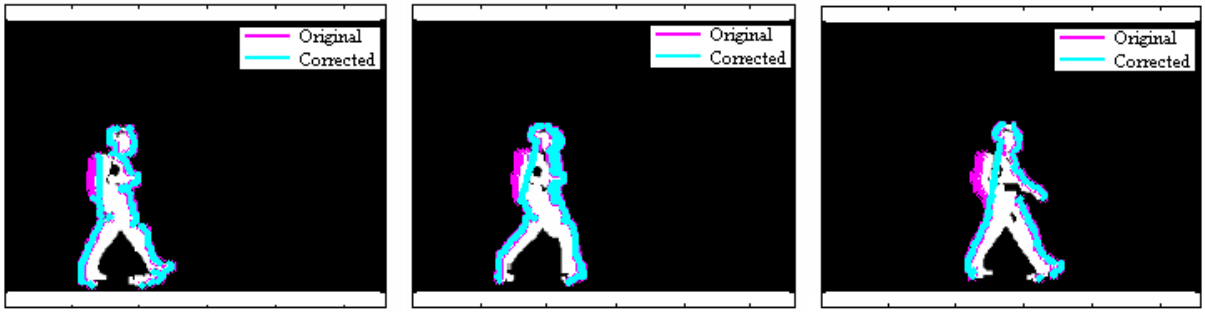

Figure 2.5: Examples for each type of silhouette correction. Left: Threshold. Center: Linear. Right: Interpolated.

backpack is relatively straight, a linear interpolation is a reasonable correction.

$$c_2[t] = \begin{cases} y[t_{head}] + (t - t_{head})\frac{t[t_{waist}] - y[t_{head}]}{(t_{waist} - t_{head})}, & t_{head} < t < t_{waist} \\ y[t], & \text{otherwise.} \end{cases} \tag{2.15}$$

Finally, the third correction method attempts to estimate additional points based on gradient change and linearly interpolates between them. These points include the areas of Figure 2.3 where the gradient changes. Such a point can be denoted as t_{grad}. The numerical value at any particular t_{grad} is given by its expected value, according to Figure 2.3, scaled according to the values of $y[t_{head}]$ and $y[t_{waist}]$ and their expected values.

$$c_3[t_{grad}] = E[y[t_{grad}]] + 0.5 * (E[y[t_{head}]] - y[t_{head}] + E[y[t_{waist}]] - y[t_{waist}]) \tag{2.16}$$

Then, $c_3[t]$ is obtained by repeating Equation (2.15) for each m_{grad}. Refer to Figure 2.5 for examples of each correction type.

2.3 WVU Outdoor SWIR Gait Dataset

2.3.1 Description and Properties

An argument can be made that the advancement of algorithms for human gait recognition (Chapter 1, Section 1.3.1) and the addressing of challenges discovered (Chapter 1, Section 1.3.5) has largely coincided with the release of new and increasingly challenging datasets.

Such datasets are typically characterized by a large number of identities, unconstrained environment with respect to lighting and objects, variations in clothing and footwear, diverse viewpoints with respect to the camera, etc. For example, early gait datsets, such as the CMU Mobo Database [126] and Soton Databases [127], had a limited number of identities and were collected in constrained settings (e.g., indoors, individuals walking on a treadmill). Despite these limitations, these initial datasets were the first publicly available datasets available to the research community and encouraged performance benchmarks and comparison studies.

The UMD Human Identification at a Distance (HID) [72] dataset, collected in 2001, was one of the first datasets to consider multiple viewpoints in an outdoor environment. The dataset included an added challenge of matching low resolution silhouettes. This dataset was later superseded by the USF Gait Challenge dataset [120], which continues to be a benchmark for evaluating and reporting algorithm performance. Published in 2002, the dataset initially consisted of 74 individuals subject to 16 different collection conditions [128], pertaining to viewpoint, walking surface and time. It has since been expanded to include 122 identities and 12 specific experiments. The next major datasets were released by the Chinese Academy of Sciences (CASIA). Referred to as the CASIA B [129] and C Databases [130], these datasets included a larger number of identities (124 and 153, respectively) and exhibited diverse variations, such as viewpoint, clothing, cadence and carrying condition (i.e., with or without a backpack). In particular, the CASIA C database was the first large gait database to expand beyond the visible spectrum, using an infrared (thermal) camera to collect video sequences in a nighttime environment.[2] These datasets contributed towards advancing the state-of-the-art in gait recognition allowing researchers to consider issues such as view invariance [87, 119, 129], object detection [131], clothing [113], time [121] and framerate [132]. A summary of these datasets is provided in Table 2.1.

In order to continue the advancement of automated gait recognition, it is essential for the next generation datasets to have data acquired from less constrained environments. Toward this end, the WVU Outdoor SWIR Gait (WOSG) dataset is introduced. The WOSG dataset denotes a new challenge dataset whose properties are very likely to occur in an operational

[2]The operating wavelength of the image sensor in the CASIA C dataset is not published. It is only reported as "thermal".

Table 2.1: Examples of public datasets available for gait recognition research. The column "Covariates" indicates the types of intra-class variations present in the dataset.

Dataset	**#Identities**	**Environment**	**Spectrum**	**Covariates**
CMU MoBo Dataset[126]	25	Static Indoor	RGB	Viewpoint, Pace, Objects
Georgia Tech Dataset[133]	24	Static Indoor, Static Outdoor	RGB	Pace
UMD HID Database [72]	25-55	Static Outdoor	RGB	Viewpoint
Soton Small Dataset[127]	12	Static Indoor	RGB	Shoe, Clothing, Objects
Soton Large Dataset[127]	115	Static Indoor Static Outdoor	RGB	Viewpoint, Time
USF HumanID Dataset[120]	122	Static Outdoor	RGB	Viewpoint, Shoe, Surface, Objects, Time
Osaka Treadmill (A)[134]	34	Static Indoor	RGB	Pace
Osaka Treadmill (B)[134]	68	Static Indoor	RGB	Clothing
Osaka Treadmill (C)[134]	200	Static Indoor	RGB	Viewpoint
Osaka Treadmill (D)[134]	185	Static Indoor	RGB	Gait Fluctuation
CASIA B Dataset[129]	124	Static Indoor	RGB	Viewpoint, Objects
CASIA C Dataset[130]	153	Static Outdoor	Thermal	Pace, Objects
WVU Outdoor SWIR Gait Dataset	155	Active Outdoor	SWIR (1550nm)	Viewpoint, Illumination

setting. These properties include:

- Data collection occurs in an *active*, *outdoor* environment, wherein environmental factors such as cloud cover (that impacts illumination) and scene factors such as motion artifacts due to trees or additional persons (that impacts segmentation and tracking) exist.
- Multiple walking paths, resulting in video sequences representing different viewing angles.
- The spatial resolution of the observed individual is not the same in every video sequence.

Additionally, the WOSG dataset is assembled using a sensor operating in the short-wave infrared (SWIR) spectrum (900nm-1,700nm), which in an operational setting may be more advantageous than visible light (RGB). For example, in low-light conditions, RGB imagery

requires an *active* illumination source, which can be detected by the human eye. On the contrary, SWIR illumination is for the most part undetectable to the human eye. Thus, a system operating in the SWIR spectrum can operate covertly. Additionally, light emitted from the sun (and reflection from the moon) and stars can act as natural illumination sources, enabling both daytime and nighttime operation. Further, SWIR is tolerant to obscurants such as dense clouds, fog, and smoke. For these reasons, it is worthwhile to evaluate gait recognition in the SWIR domain. It should be noted that the current WOSG dataset does not include nighttime imagery. It is anticipated that future versions of this dataset will incorporate imagery from both daytime and nighttime environments.

In the gait recognition literature, recognition capability in an active outdoor environment has not been adequately tested. In particular, silhouette segmentation is superficially treated and the performance of recognition algorithms on silhouettes extracted from a more operational setting is not well known. On the aspect of multi-directional trajectories, an operational gait recognition system will most certainly encounter individuals walking along arbitrary paths (rather than a simple unidirectional path that is perpendicular to the camera's optical axis), although this issue has been receiving attention as of late [87, 129]. Finally, the problem of dealing with human silhouettes that vary in resolution across video sequences (or even frames) has not been adequately evaluated in the literature. In an operational setting, particularly with cameras capable of performing a pan-tilt-zoom operation, feature extraction may have to be conducted on human objects having variable spatial resolutions. This challenge is evident in the WOSG dataset as the observed field of view may not be the same for all sequences. In summary, the WOSG dataset represents a challenging dataset for the biometrics and computer vision communities in that it embodies the following attributes: matching gait sequences across viewpoints, trajectories and distances; and silhouette extraction in low-resolution (e.g., $\approx$50% reduction in bounding box area (i.e., the rectangular pixel space occupied by a detected human) compared to the CASIA datasets), SWIR imagery.

2.3.2 Hardware Description

The camera used to acquire this dataset was the Sensors Unlimited Goodrich SU640KTSX-1.7RT High Sensitivity InGaAs SWIR Camera (640x512 pixels) with a 50mm f/1.4 SWIR lens. The Goodrich camera was used to capture video sequences of individuals walking at distances ranging from 20-50m. The resolution of the captured video data is 640x512 pixels and the framerate is 30 frames per second.

2.3.3 Collection Protocol

Collection occurred between September and November, 2011. Data collection with the Goodrich camera was performed in an outdoor environment during daylight hours, supplying natural illumination via sunlight. Cloud cover (i.e, ambient illumination) varied between clear skies, partly cloudy and overcast. Video sequences were bandpass filtered to 1550nm ($\pm$50nm FWHM). Operational settings such as integration time were adjusted to achieve the best image quality based on daily environmental conditions (e.g., cloudy, sunny, etc.).

During collection, each individual completed one session and during which, was asked to walk in a continuous motion along eight predefined trajectories. The specified walking directions and approximate distances to the camera are denoted in Figure 2.6. The distances and walking paths were defined such that a minimum of three complete gait cycles could be completed in each direction. The length of each video sequence (in time) varied between 90 and 110 seconds, depending on the walking speed of each individual. Sample video frames from the dataset are shown in Figure 2.7. In total, data was collected from 155 individuals. In terms of gender, there were 93 males and 62 females. In terms of ethnicity, 46% of identities identified themselves as Caucasian, 25% as Asian Indian, 13% as Asian, 8% as African, 5% as Middle Eastern, 2% as African American and 1% as Unknown. In terms of age, 71% of identities were between 20-29 years old, 12% between 18-19, 8% between 30-39, 4% between 40-49, 3% between 60-69 and 2% between 50-59.

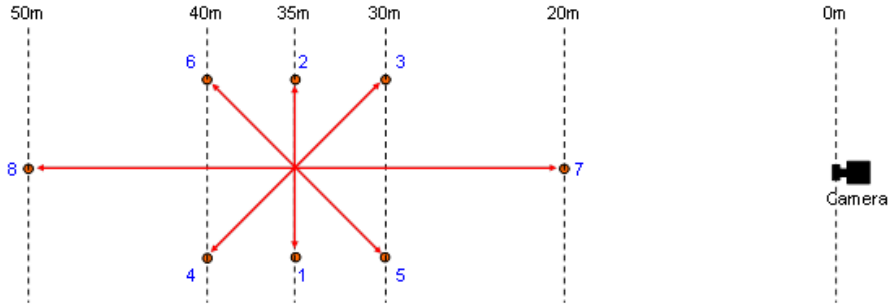

Figure 2.6: Gait collection layout. Each individual is captured walking in the numbered direction one time.

Figure 2.7: Sample frames from the WOSG dataset. Note the variance in contrast and brightness in each frame, occurring as a result of varying environmental conditions.

2.4 Comparison Algorithms

2.4.1 Gait Energy Image

The Gait Energy Image (GEI) is a model-free algorithm for human gait recognition proposed by Han and Bhanu [91]. In contrast to gait curve matching, the GEI algorithm utilizes a weighted combination of pixel values for in set of silhouette images. In other words, the GEI algorithm attempts to reduce the motion dynamics of an individual represented in multiple frames into a single image. As stated in the Chapter 1 (Section 1.3.2), the GEI image is a popular benchmark for comparing performance, owing to its ease of computation. The algorithm computes $\bar{B}$, which is defined as the average of K space-normalized human silhouette images, B_k, $k = 1, 2, \ldots, K$. As in Section 2.2, K denotes the number of frames in one gait cycle. Mathematically, this is described in Equation (2.17).

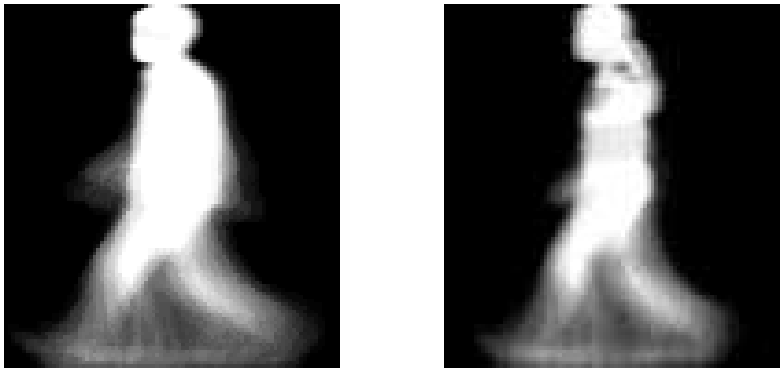

Figure 2.8: Examples of Gait Energy Images (GEI) extracted from two individuals. Note the pixel intensities in each GEI image correspond to the occupied foreground space as a person walks.

$$\bar{B} = \frac{1}{K}\sum_{t=1}^{K} B_k \tag{2.17}$$

In computing the "average" silhouette, the features used for matching correspond to the "moving" pixel intensities as a human silhouette moves in time. In order to accurately compute a GEI image, each silhouette image (B_k) must be normalized to the same number of pixels and appropriately aligned. This is particularly necessary when an individual is observed walking towards or away from the camera. Alignment can be performed using the centroid of the image data. Examples of GEI images are presented in Figure 2.8.

Since the GEI algorithm denotes a weighted combination of K silhouette images, raw GEI images can have a very large dimensionality, which can cause difficulty in matching. In the original work by Han and Bhanu [91] this rectified through Principal Component Analysis (PCA), though other methods such as Linear Discriminant Analysis (LDA) [90], tensor discriminants [94], and 2-D PCA [135] have also been explored. In this implementation of the GEI algorithm, subspace optimization (i.e., dimensionality reduction) is performed using Principal Component Analysis, wherein a principal component is retained if the associated eigenvalue is greater than 0.001. The euclidian distance metric is used to generate match scores between a pair of GEI feature vectors. A brief description of PCA is presented in the following paragraphs.

Principal Component Analysis

The aim of Principal Component Analysis (PCA) is to find an orthogonal subspace ψ that reduces the dimensionality of the original feature space (i.e., the theoretical range of values spanned by each element in a feature vector) while preserving a majority of the data variance. In other words, PCA aims to reduce the number of dimensions, d, in feature vector $\mathbf{v}$, to d' where $d' \leq d$. The output, $\mathbf{v}'$, is a vector whose elements will have a maximal variance with respect to the transformed feature space.

PCA is accomplished by performing an eigenvalue decomposition on the covariance matrix computed from a set of training data. Here, the training data defines the feature space which is to be optimized. Given n samples, $\mathbf{x}_i \in \Re^{d,1}$, $i = 1, 2, \ldots, n$, where $\mathbf{x}_i$ denotes some feature vector, the first step is to compute the sample mean μ ($\mu = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$). Denote $\mathbf{X} \in \Re^{d,n}$ as a matrix containing each feature vector in the training data normalized by the sample mean (i.e., $\mathbf{X} = \{\mathbf{x}_1 - \mu, \mathbf{x}_2 - \mu, \ldots, \mathbf{x}_n - \mu\}$. A scatter matrix, $\mathbf{S}$ is computed from $\mathbf{X}$ as $\mathbf{S} = \mathbf{X}\mathbf{X}^T$, where T denotes the transpose operator.

The optimal subspace ψ is computed by performing an eigenvalue decomposition on $\mathbf{S}$, which yields a matrix of eigenvectors ψ and eigenvalues $\mathbf{\Lambda}$. Note the eigenvector in the i^{th} column of ψ corresponds to the i^{th} diagonal of $\mathbf{\Lambda}$ (i.e., $\mathbf{S}\psi_i = \Lambda_i\psi_i$).

Typically, $d' < d$ eigenvectors are retained, such that

$$d' = \arg\min_{\bar{d}} \frac{sum_{i=1}^{\bar{d'}}\Lambda_i}{\sum i = 1^d \Lambda_i} > V_e \tag{2.18}$$

where $V_e \in [0, 1]$ is the fraction of data variation to be retained.

In general, most of the variance in $\mathbf{X}$ is stored in the largest eigenvector in ψ. By discarding eigenvectors associated with small eigenvalues, the feature dimensionality can be greatly reduced without losing the discriminatory information in the original feature vector.

Note, this description was adapted from the Ph.D dissertation of Klare [136].

2.4.2 Frieze Pattern Matching

The second comparison algorithm is referred to as "Frieze Pattern Matching". In the mathematical context, a frieze pattern is defined as a two-dimensional pattern that repeats

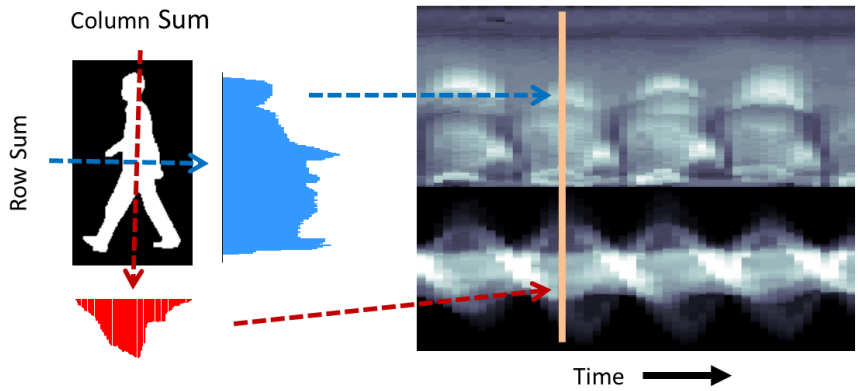

Figure 2.9: Visual example illustrating how the horizontal (row-sum) and vertical projections (column-sum) of the silhouette can be combined to create a spatiotemporal pattern for human gait recognition.

itself infinitely. Liu et al. [101] was the first to exploit the periodic nature of human gait as a frieze pattern. A frieze pattern denoting human gait is defined as the concatenation of the x- and y- projections of a silhouette moving in time. As with GEI, this method for denoting human gait is also classified as a model-free recognition algorithm. A mathematical description of such a pattern is described in the following paragraphs.

Consider a set of K silhouette images, denoted as B_k, $k = 1, 2, \ldots, K$. Define a 2-d frieze image, $F(j, k)$ as the horizontal projection (row sum) of each of K silhouette images. Mathematically, this is described in Equation (2.19).

$$F(j,k) = \sum_{i} B_k(i,j) \tag{2.19}$$

$F(j, k)$, denotes the width of the k^{th} silhouette at the vertical pixel coordinate j. Similarly, $F(i, k)$ can be defined as the vertical projection (column sum) of the silhouette, capturing the height of the k^{th} image at the horizontal pixel coordinate i. In a set of K silhouette images encompassing at least one gait cycle, the silhouette width varies periodically due to factors such as stride and arm sway. By observing $F(j, k)$ as an image, the row projections combine to form a spatiotemporal pattern. A graphical example of how these patterns are constructed is illustrated in Figure 2.9.

In the original work by Liu et al., matching of frieze patterns was accomplished by comparing the central moments of a probe and gallery sequence [101]. Alternatively, Dynamic

Time Warping (DTW) or Hidden Markov Models (HMM's) can be used to perform matching, as described by Kale et al. [100, 103]. In this implementation of the Frieze Pattern matching algorithm, match scores between two gait sequences are generated using 2-D Dynamic Time Warping, [103]. The following paragraphs summarize how two Frieze patterns can be matched using 2-D Dynamic Time Warping.

Dynamic Time Warping

In general, Dynamic Time Warping is a method for evaluating similarity between two sequences that vary in time. For instance, in the context of a frieze pattern for gait recognition, the periodicity varies depending on the speed at which a person is walking. Similarity between patterns is evaluated by generating cost matrix, $\mathbf{C}$, which stores the difference between each pattern for each unit of time (in this case, each frame). Thus, for a pair of patterns, A and B, with length T_A and T_B, $C(1,1)$ denotes the difference between them at $t = 1$ and $C(1, T_B)$ denotes the difference between A at $t = 1$ and B at $t = T_B$. The resulting distance score is obtained by summing the smallest valued path from $C(1,1)$ to $C(T_A, T_B)$ while moving strictly in increments of one in an 8-connected neighborhood.

2.5 Experimental Results

Experiments are designed to convey the following information:

- A baseline analysis demonstrating the identification performance of the Gait Curves matching algorithm on benchmark datasets. For reference, performance is reported with respect to the comparison algorithms (GEI and Frieze Pattern matching) defined in Section 2.4. This experiment provides a quantitative comparison between the Gait Curves matching algorithm to other algorithms and datasets in the gait recognition literature.

- A comprehensive analysis demonstrating the recognition capability of the Gait Curves, GEI, and Frieze patterm algorithms on the WVU Outdoor SWIR Gait dataset. This experiment demonstrates the recognition performance in a less constrained gait dataset.

- An analysis measuring the ability of a background subtraction scheme for silhouette extraction on the WVU Outdoor SWIR Gait dataset compared to data in benchmark gait datasets. This experiment provides a qualitative analysis on the performance of background subtraction in constrained and unconstrained data.

- An evaluation demonstrating whether the features listed for backpack detection (Section 2.2.4) can successfully detect the presence of a backpack on a silhouette.

- An evaluation probing the ability of the defined silhouette rectification schemes (Section 2.2.5) to improve the identification performance of a silhouette distorted by a backpack.

2.5.1 Datasets

The primary dataset in this evaluation is the WVU Outdoor SWIR Gait (WOSG) dataset, a large, outdoor gait dataset collected in the SWIR spectrum. Though the dataset consists of 155 identities, segmentation was not possible for 41 identities. Segmentation failure was often attributed to extreme environmental factors (e.g., rapid illumination variance) or difficulty processing the data (e.g., very low native contrast). Thus, only $N = 114$ identities are used for experimental analysis. In addition, segmentation was not possible in instance when individuals were observed walking directly towards or away from the camera (labels 7 and 8 from Figure 2.6). To be consistent with related gait literature, unless otherwise stated, one feature vector was extracted from each of the remaining walking directions for each individual. In other words, the number of samples per identity, N_G, is six. This results in a total number of $N_T = 684$ gait sequences.

To aid in analysis, in addition to the WVU Outdoor SWIR Gait dataset, experiments are also conducted on a a pair of benchmark datasets from the gait recognition literature. The datasets chosen for comparative study are the CASIA B and C datasets. A brief review of these datasets is presented in the following paragraphs.

CASIA B Gait Dataset

The CASIA B Gait dataset [129] is a large *indoor*, *multi-camera* gait database, which was collected by the Chinese Academy of Sciences (CASIA). The dataset consists of $N = 124$

Figure 2.10: Sample frames from the CASIA B dataset, with an individual captured walking towards (left) and perpendicular to the camera (right)

.

individual identities, wherein each individual is captured walking 10 times (i.e., $N_G = 10$ with one gait feature vector extracted per video). Gait sequences were captured at 11 different viewing angles, linearly spaced such that the individual is viewed walking towards, diagonal to, perpendicular to, and away from the optical axis (i.e., camera viewpoint). In addition, the dataset consists of person-based covariates (Chapter 1, Section 1.3.5) such as carrying a backpack (two sequences per identity) and wearing a coat (two sequences per identity). The dataset was collected in the visible (RGB) spectrum in a controlled indoor environment. The native camera resolution for each image is 320x240 pixels and the framerate is 25 frames per second. A sample of the image data from this dataset is provided in Figure 2.10.

CASIA C Gait Dataset

The second dataset for baseline evaluation is the CASIA C Gait dataset (alternately defined as the CASIA Night Gait dataset) [130], which was also collected by the Chinese Academy of Sciences. Unlike the B dataset, the C dataset was collected from an *infrared* camera in an *outdoor*, *nighttime* environment. In addition, the image data is stored in a single channel.[3]

[3]The explicit wavelength the image sensor operates at is not provided by CASIA and is therefore unknown.

Figure 2.11: Example frames from the CASIA C dataset. Note the reduced contrast and brightness in comparison to the CASIA B dataset.

The CASIA C dataset consists of $N = 153$ unique identities, wherein 10 video sequences are captured for each identity (i.e., $N_G = 10$). In this dataset, the only viewpoint provided denotes the individual walking perpendicular to the optical axis. Covariates in this dataset are person-based, concentrating on cadence (i.e., the rate at which a person walks) and carrying condition. Thus, two sequences are collected for both slow and fast cadence, to contrast against four sequences of normal walk. The remaining two sequences denote normal walk with a backpack. The native resolution of the image data is 279x200 pixels and the framerate is 25 frames per second. A sample illustration of frames from this dataset is provided in Figure 2.11.

Provided the CASIA C dataset was collected outdoors and in a nighttime environment, and that the image data contains one channel of information, this dataset is less constrained than the CASIA B dataset and is arguably more challenging.

2.5.2 Evaluation of Silhouette Quality

Prior to evaluating the matching accuracy of the Gait Curves matching algorithm and the comparison matching algorithms (GEI and Frieze Pattern matching), a good precursor is to evaluate the quality of the silhouettes produced in the silhouette extraction process. Since the image data in the WVU Outdoor SWIR Gait dataset was collected in a significantly less constrained environment, it is useful to compare how well the typical method for silhouette extraction (i.e., background subtraction) performs on more challenging data. The quality

Table 2.2: Median silhouette quality metric for the gait sequences in the WOSG dataset, CASIA C dataset, and CASIA B dataset.

Dataset	**Median ψ** (all frames)	**Median ψ** (50 frame moving window)
WOSG Dataset	0.389	0.443
CASIA C Dataset	0.609	0.654
CASIA B Dataset	1.000	0.971

metric used is the noise of the foreground sum signal as defined in Chapter 1, Section 1.3.4.

Here, ψ (the quality value) is computed using (a) *all* of the frames in a video sequence and (b) the mean value of ψ from a 50-frame moving window. That is, computing the average of ψ from the frame sets: $\{1, 2, \ldots, 51\}$, $\{2, 3, \ldots, 52\}$, ..., $\{K-50, K-49, \ldots, K\}$. For comparison, ψ is also computed on the CASIA B and C datasets. To mitigate the effect of outliers (i.e., exceptionally high quality sequences or exceptionally poor quality sequences), the median value of ψ from all sequences us reported to indicate the global quality of the silhouettes extracted in the whole dataset. These results are tabulated in Table 2.2. Note that the quality value produced for the WOSG dataset is less than the CASIA C dataset, which is less than the CASIA B dataset. This outcome suggests that background subtraction may not be a sufficient method for silhouette extraction on more challenging image data.

2.5.3 Protocol for Measuring Matching Performance

The matching performance of the Gait Curves, GEI, and Frieze Pattern matching algorithms are evaluated using ROC and CMC curves (Chapter 1, Section 1.1.4), which comprise a traditional biometric verification and identification analysis. The specifics by which match scores are extracted for each matching algorithm are described in the following paragraph.

For gait curve matching, extracted gait curves are normalized to a size of 300 elements (i.e., $T = 300$), aligned and then warped to the procrustes space as defined in Section 2.2.2. The procrustes distance measure (Equation 2.7) is used to generate match scores. GEI features are extracted using a 91-pixel horizontal window, which is centered at the centroid of each silhouette image. Extracted silhouette images are normalized to a vertical height of 100 pixels. The resulting GEI image (of size 100x91) is then downsampled using bicubic

interpolation to a resolution of 50x46. Further dimensionality reduction and subspace optimization is performed using principle component analysis. A principle component is retained if its associated eigenvalue has a value larger than 0.001. The match score between two GEI features is evaluated using the Euclidean distance metric. Similar to GEI, silhouettes for Frieze pattern matching are normalized to a height of 100 pixels. The x- and y- projections for each silhouetted image are then computed and concatenated. Match scores are defined as the distance value obtained using Dynamic Time Warping. For each of the matching algorithms, the computed distance score is subtracted from a value of one to convert it into a similarity score.[4]

Since the GEI algorithm requires a training subset for computing the subspace projection matrices, samples for 15% of identities are randomly selected for this purpose (and are not present in either the probe or reference sets). To remove selection bias, all experiments involving GEI are repeated 10 times (e.g., trials). All reported ROC and CMC curves denote the average performance from each trial.

2.5.4 Baseline

In the baseline evaluation, ROC and CMC curves are computed for differing combinations of probe (test) and reference (training) samples. In particular, using the subsets of match scores that can be generated when comparing "normal" walking sequences to each covariate (e.g., with bag, with coat, fast walk, slow walk) and against similarly labeled "normal" sequences. When comparing "normal" to "normal" gait sequences, half of the sequences are designated as probe and reference, respectively. In addition, an additional experiment, denoting comparing each feature vector against the other $N_T - 1$ feature vectors is also performed. Here, this experiment is referred to as "all-to-all" matching. In the interest of being concise, viewpoint (CASIA B) is not considered in the baseline evaluation. Table 2.3 and Table 2.4 summarize the list of experiments for both datasets and the corresponding probe and reference sets for each experiment. ROC and CMC curves for the CASIA B and C datasets are presented in Figures 2.12 and 2.13, respectively.

[4]This computation is cosmetic. However this dissertation assumes all match scores are similarity scores.

Table 2.3: List of experiments on the CASIA B dataset.

# of Identities	Gallery Size	Probe Size	Gallery Sequence	Probe Set
105	315	315	Normal	Normal
105	630	210	Normal	Bag
105	630	210	Normal	Coat
105	210	210	Bag	Bag
105	210	210	Coat	Coat
105	1050	1050	All	All

Table 2.4: List of experiments on the CASIA C dataset.

# of Identities	Gallery Size	Probe Size	Gallery Sequence	Probe Set
130	260	260	Normal	Normal
130	520	260	Normal	Bag
130	612	260	Normal	Slow
130	612	260	Normal	Fast
130	1300	1300	All	All

2.5.5 Identification Performance on the WVU Outdoor SWIR Gait Dataset

In order to perform a comprehensive evaluation of the Gait Curve, GEI, Frieze pattern algorithms, a number of experiments were developed. The experiments performed quantify (a) general matching performance from all gait sequences; (b) the performance when matching sequences of differing viewpoint; and (c) sequences of the same viewpoint. To be consistent with related gait literature, unless otherwise stated, one feature vector per matching algorithm is extracted from each of the $N_T = 684$ gait sequences.

General Matching Performance

In the first experiment, matching is performed using a leave-one-out cross validation scheme. That is, each of $N_T = 684$ gait sequences are compared against the remaining $N_T - 1$ sequences, regardless of walking direction. Establishing the reference data in this way illustrates the matching performance when the constraint of viewing angle is reduced, but not eliminated. These results are illustrated in Figure 2.14.

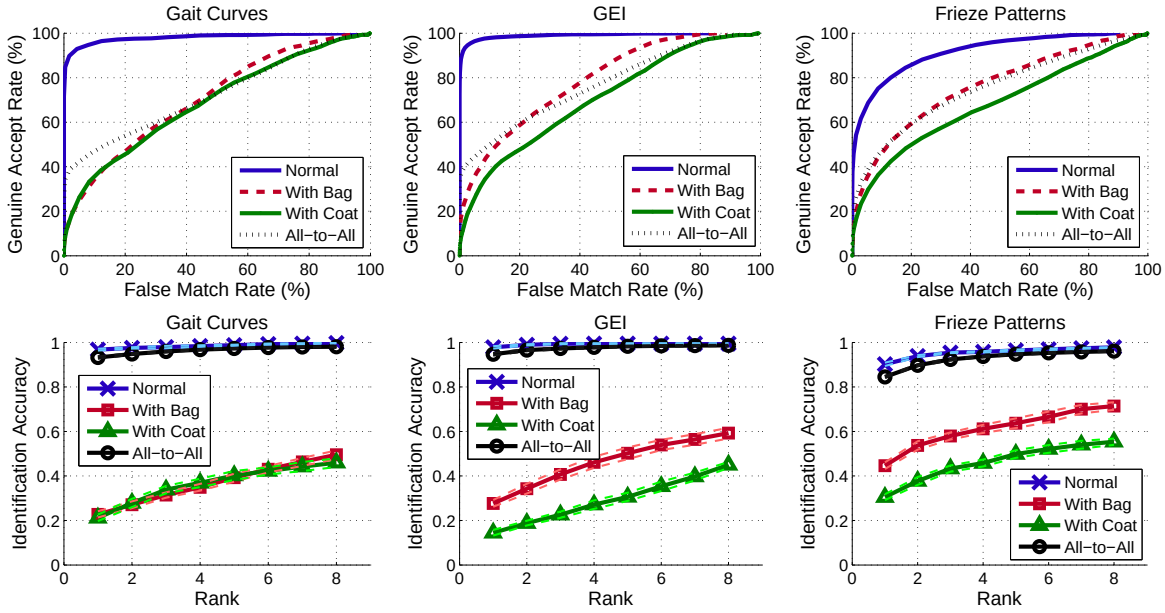

Figure 2.12: Baseline matching performance of the Gait Curve, GEI, and Frieze pattern algorithms on the CASIA B dataset. Dashed lines in the CMC curves (bottom) indicate one standard deviation above or below the mean for ten trials.

Matching Major Differences in Walking Direction

In the second experiment, the ability to match gait sequences corresponding to different viewing angles is evaluated. To avoid redundancies in experimental data, sequences of "leftward" walk are compared against sequences of "rightward" walk. Here, "leftward" walk is defined as those sequences wherein an individual is walking from the left to the right (from the camera's perspective). This includes sequences with direction labels "1", "4", and "5" from Figure 2.6. Similarly, "rightward" walk is defined as those sequences wherein an individual is walking from the right to the left. This includes sequences with direction labels "2", "3", and "6" in Figure 2.6. In particular, the experiment is designed such that each of the three leftward sequences are matched against the three rightward sequences. This results in a total of nine probe and gallery combinations, which are defined in Table 2.5. ROC and CMC curves for this experiment are illustrated in Figure 2.15 and Figure 2.16, respectively.

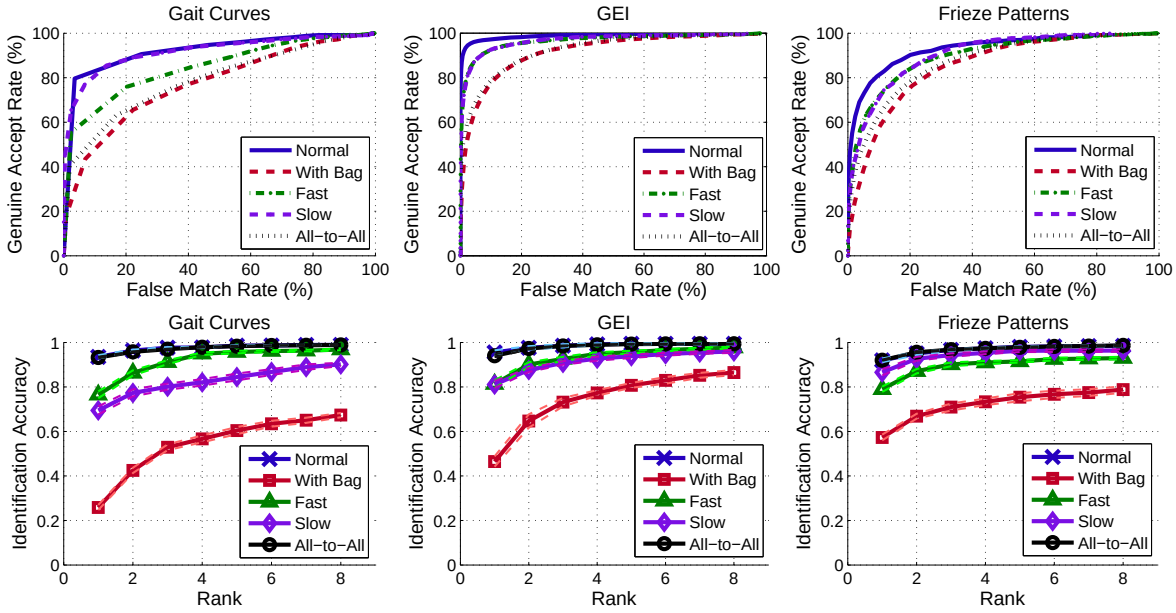

Figure 2.13: Baseline matching performance of the Gait Curve, GEI, and Frieze pattern algorithms on the CASIA C dataset. Dashed lines in the CMC curves (bottom) indicate one standard deviation above or below the mean for ten trials.

Matching Same Walking Direction

In the third experiment, the probe and reference sequences correspond to the same direction. To accommodate this, *two* feature vectors are extracted per video sequence, where frames belonging to the first $\frac{K}{2}$ images are used to generate the first feature vector, while the remaining $\frac{K}{2}$ images are used to generate the second feature vector. Extraction of two feature vectors from a single video sequence is necessary since each individual in the database provided data in only a single session. Since two feature vectors for recognition are extracted sequentially, the similarity between the two vectors should be quite high. As such, this experiment also acts as a measure of local feature variance. If the variance is low, the matching performance should be very good, as both feature vectors for each identity would be approximately equal to one another. If the variance is high, matching performance will be degraded. These results are illustrated in Figure 2.17.

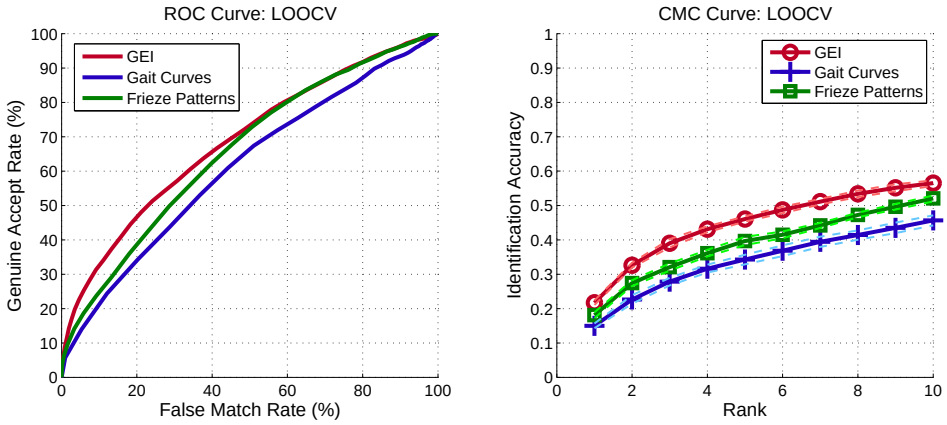

Figure 2.14: General matching performance of the Gait Curves, GEI, and Frieze pattern algorithms on the WVU Outdoor SWIR Gait dataset. Here, ROC (left) and CMC curves (right) illustrate the combined matching performance across all walking directions using leave-one-out cross validation (LOOCV). Dashed lines indicate one standard deviation above or below the mean for ten trials.

2.5.6 Backpack Detection

In this experiment, the ability to predict the presence of a backpack using Equations (2.8-2.11) is evaluated. Here, the experiment is conducted on the CASIA C dataset, wherein each individual is classified as either "carrying a backpack" or "not carrying a backpack". Note that although the CASIA B dataset also contains a "with bag" covariate class, most individuals are not wearing backpacks, but rather, carrying handbags or briefcases. Additionally, one individual in the CASIA C dataset is carrying a briefcase, rather than wearing a backpack. For the purposes of this experiment, this individual is regarded as not carrying a backpack.

To evaluate detection performance, features are compared using a 5-Nearest Neighbor classifier with 10-fold cross validation, where each fold denotes 10% of identities for training (i.e., reference data) and 90% for testing (i.e., evaluation). In addition, the training data is comprised with an equal number of backpack and "non-backpack" samples, in order to mitigate recognition bias towards either class. Here, the "non-backpack" samples denote the first two "normal walk" samples for each identity. Note that this framework suggests that for a given cross validation fold, if an identity is selected to be in the training set, only four of ten samples are used and the remaining six samples are ignored. Alternatively, if an identity

Table 2.5: Probe and reference combinations for matching gait sequences corresponding to different viewing angles. The arrows denote the direction of walk (in the image plane).

Probe (Test data)	**Reference (Training data)**
Direction 1 ($\longleftarrow$)	Direction 6 ($\nearrow$)
Direction 1 ($\longleftarrow$)	Direction 2 ($\longrightarrow$)
Direction 1 ($\longleftarrow$)	Direction 3 ($\searrow$)
Direction 4 ($\nwarrow$)	Direction 6 ($\nearrow$)
Direction 4 ($\nwarrow$)	Direction 2 ($\longrightarrow$)
Direction 4 ($\nwarrow$)	Direction 3 ($\searrow$)
Direction 5 ($\swarrow$)	Direction 6 ($\nearrow$)
Direction 5 ($\swarrow$)	Direction 2 ($\longrightarrow$)
Direction 5 ($\swarrow$)	Direction 3 ($\searrow$)

Table 2.6: Confusion matrix for backpack detection in the CASIA C dataset.

K = 5	**Without Bag**	**With Bag**
Without Bag	9739 (88.5%)	1268 (11.5%)
With Bag	314 (11.4%)	2440 (88.6%)

is selected to be in the test set, all samples are used. Results are summarized in Table 2.6 in the form of a confusion matrix. The confusion matrix is a table whose row entries denote the actual class and the column entries denote the predicted class.

2.5.7 Silhouette Rectification

In this experiment, the silhouette correction module discussed in section presented in Section 2.2.5 is activated. Here, each of the described silhouette correction methods are tested both with and without the detection module. That is, silhouette correction is first performed on all video sequences where the individual is carrying a backpack. Then, the experiment is repeated such that the rectification scheme is applied only when a backpack is detected, either correctly or incorrectly. The results of the previous experiment regarding backpack detection are used to determine whether a backpack was detected. A backpack is determined as having been detected if a probe sample tests positive in five of nine (55.55%) cross validation folds.[5] Since the silhouette rectification module is concerned with estimating

[5]Note, under 10-fold cross validation, a sample participates in nine test sets and one training set.

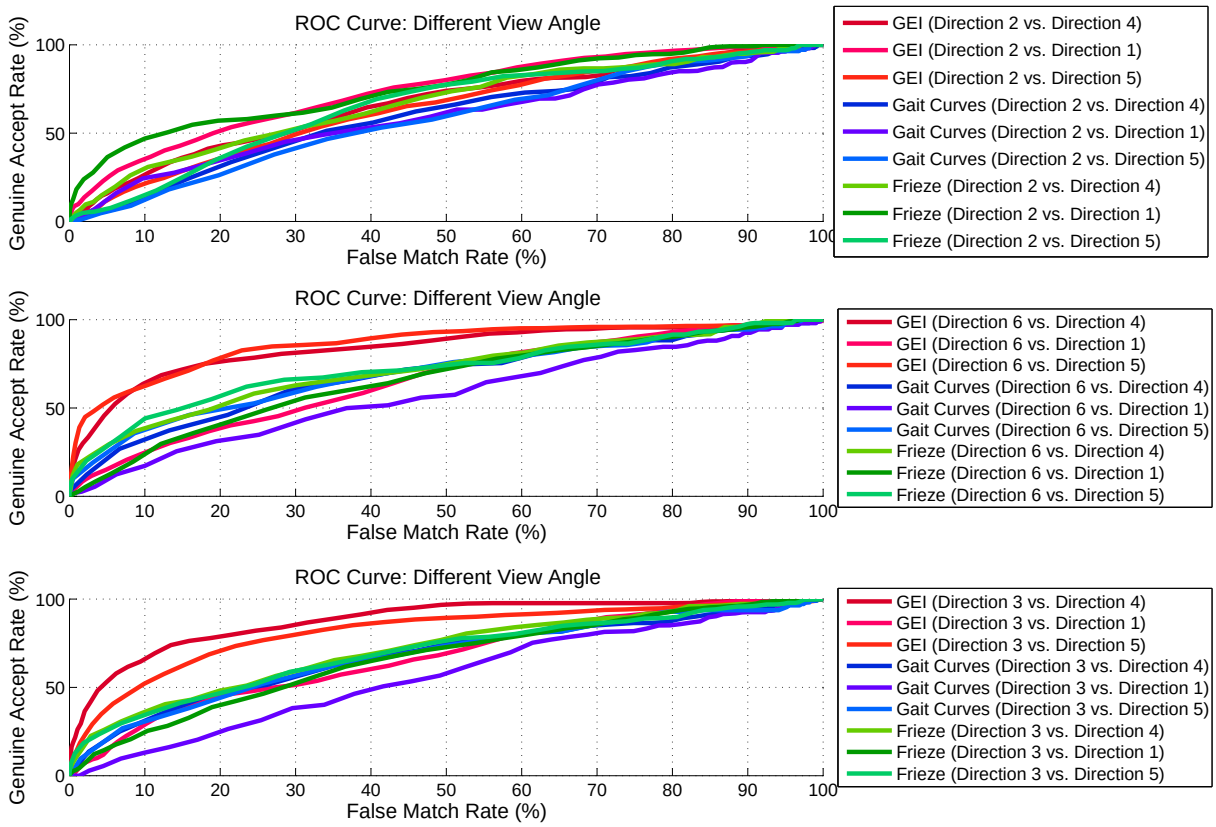

Figure 2.15: ROC curves generated from comparing gait sequences of different viewing angle.

the shape of the silhouette without a backpack, this experiment is also only conducted on the CASIA C dataset. For this experiment, ROC and CMC curves are presented for a "with bag" probe set and "normal walk" reference set (Row two from Table 2.4). Results are presented in Figures 2.18-2.19.

2.5.8 Discussion

Evaluating Silhouette Quality

Prior to performing a performance test of the Gait Curves and baseline algorithms (GEI and Frieze Patterns), an experiment was performed regarding the *quality* of silhouettes (Section 2.5.2) produced for each of the test databases (CASIA B, CASIA C, WVU Outdoor SWIR). In particular, the analysis concentrated on the performance of using simple back-

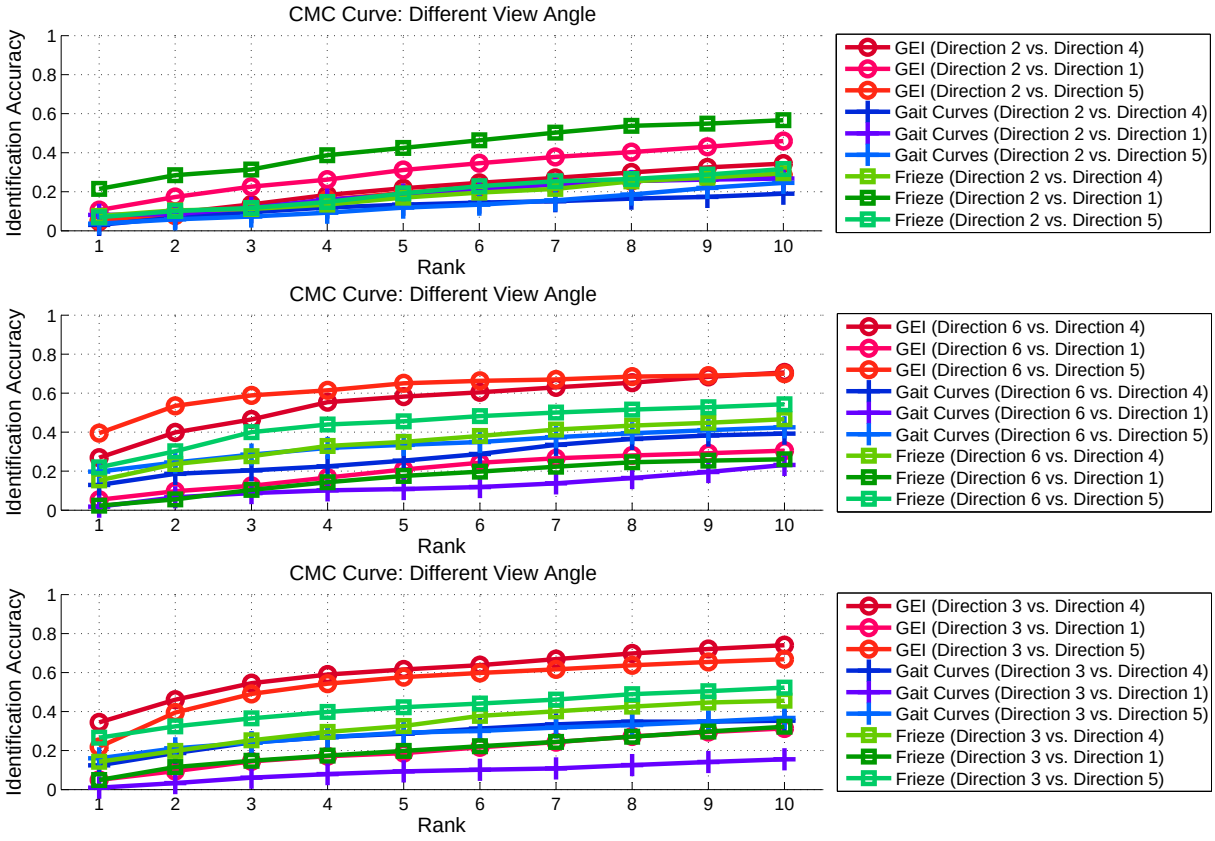

Figure 2.16: CMC curves generated from comparing gait sequences of different viewing angle.

ground subtraction to perform the segmentation and silhouette extraction process. The results demonstrated that produced silhouettes contained a larger concentration of noise as the degree of "challenge" in the image data increased. Numerically, the measure of silhouette quality for the WOSG data was found to be $\approx 39\%$ and $\approx 64\%$ of the values for the CASIA B and C datasets, respectively. Generally, the performance degradation can be attributed to single channel images often being of lower contrast than RGB, which increases the difficulty of properly identifying foreground pixels (i.e., pixels that denote a human body). To further convey this, Figure 2.20 depicts a set of intensity histograms from an image in the CASIA B dataset (RGB) and the WOSG dataset. Note that in the RGB image, regions depicting the human (intensities between 25-75) and background (intensities between 75-150) can be inferred. In the SWIR image, the range of intensities is smaller and no such distinction

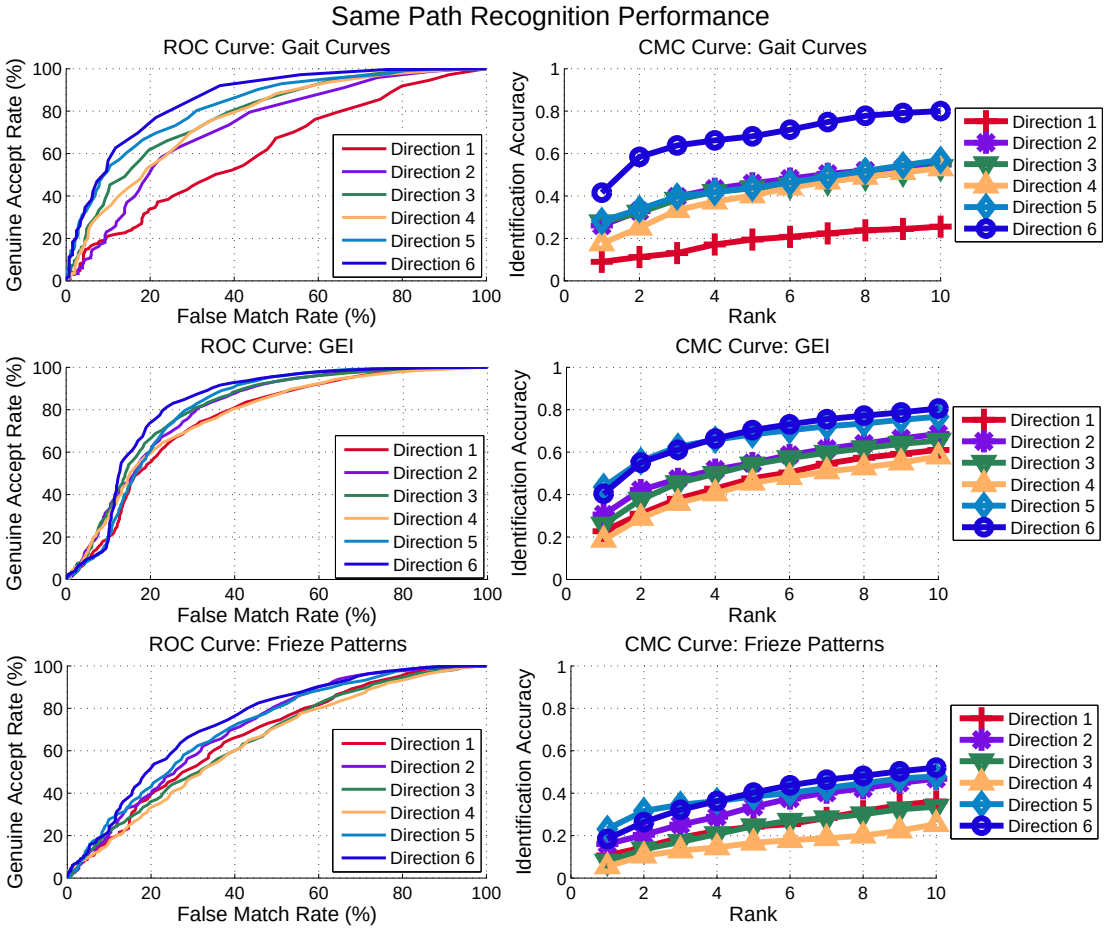

Figure 2.17: Matching performance when comparing feature vectors extracted from the same walking direction Top) Gait Curves. Middle) Gait Energy Image (GEI). Bottom) Frieze Pattern matching.

between person and background can be made.

Thus, as potentially confounding environmental factors are introduced (e.g., illumination variance) in conjunction with a reduction of the number of channels in the image data (i.e., grayscale imagery), simple methods such as background subtraction become less adept at performing silhouette extraction. Figure 2.21 highlights a sequence of image data with extreme short-term illumination variance, which increases the difficulty of identifying foreground pixels. Additionally, Figure 2.22 illustrates a comparison of silhouette images extracted from the WOSG dataset, CASIA B dataset and CASIA C dataset. Note the gradual degradation of silhouette quality, as constraints such as a fixed background (CASIA B and C) are re-

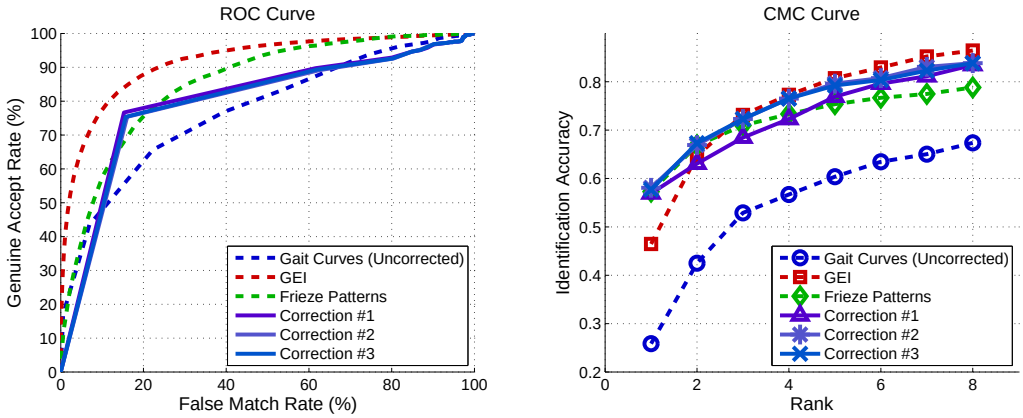

Figure 2.18: Recreation of the "With Bag" vs. "Normal Walk" matching experiment on the CASIA C dataset when the silhouette correction module of the Gait Curves algorithm is activated. Note: This figure assumes perfect backpack detection.

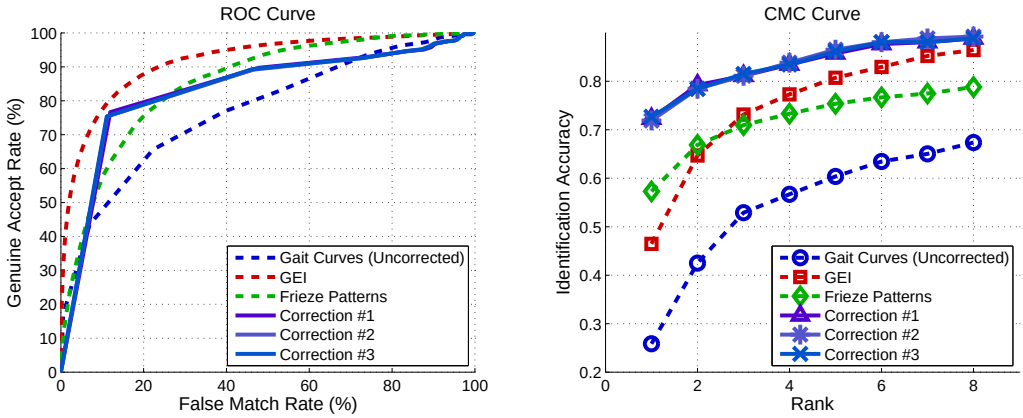

Figure 2.19: Recreation of the "With Bag" vs. "Normal Walk" matching experiment on the CASIA C dataset when the backpack detection and silhouette correction modules of the Gait Curves algorithm is activated.

moved and the number of channels in image data is reduced from three (CASIA B) to one (CASIA C and WOSG).

Baseline Analysis

The baseline analysis compares the performance of the Gait Curves algorithm (Section 2.2) to the GEI and Frieze pattern matching algorithms (Section 2.4 on the CASIA B and CASIA C gait datasets, where the comparison algorithms and datasets denote established algorithms and data in the gait recognition literature. The analysis compared ROC and CMC curves for differing combinations of probe and reference data. The purpose of the baseline

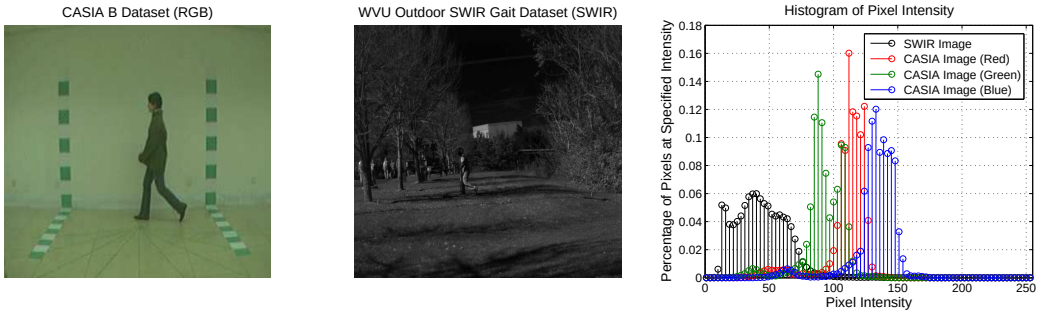

Figure 2.20: Sample images from the CASIA B (RGB) and WVU Outdoor Datasets and their associated intensity histograms. Note the dynamic range for the SWIR image is less than that of the RGB image.

Figure 2.21: Example of a challenging video sequence. Here, changing cloud cover results in a significant change in background pixel intensity over a short period of time, resulting in difficulty in identifying foreground (silhouette) pixels.

analysis was to determine whether the Gait Curves matching algorithm was comparable to existing gait recognition techniques from the literature.

Regarding the results on the CASIA B dataset (Figure 2.12, the matching performance of the Gait Curves algorithm was generally comparable to the performance of the GEI algorithm. For example, rank-1 recognition rates for "normal vs. normal" walk for the Gait Curves and GEI algorithms were 0.972 and 0.978, respectively. The GEI algorithm performed slightly better on the probe set with a backpack (0.226 vs. 0.277), while the Gait Curves algorithm performed slightly better on the probe set with a coat (0.213 vs. 0.145). Frieze pattern matching performed worse than the Gait Curves algorithm at evaluating "normal vs. normal" walk, achieving a rank-1 recognition rate of 0.894. However, for the instances comparing walk with a bag and with a coat, Frieze pattern matching significantly outperformed matching by Gait Curves. Note in all instances when the probe and reference

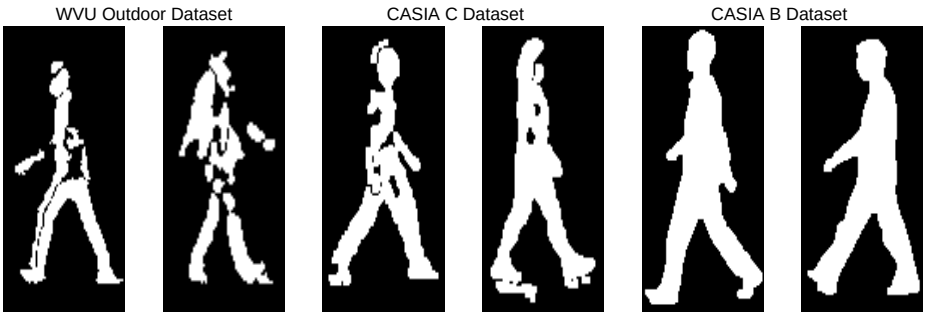

Figure 2.22: Comparison of silhouette quality using simple background subtraction (as described in Chapter 1, Section 1.3.3. Note that the silhouettes produced in the WOSG dataset are of lower quality (i.e., silhouette holes, distorted shape).

sets comprised of differing covariates (e.g., "normal vs. coat"), the matching performance degraded considerably. Since the features generated from each of the three tested algorithms are based on the properties of the silhouette (i.e., model-free algorithms), the performance loss is due to the induced silhouette shape distortions induced from wearing a backpack (expansion of the back region) and wearing a coat (thickening of the upper body).

Regarding the results on the CASIA C dataset (Figure 2.13, the matching performance for each of the algorithms was generally comparable for normal vs. normal walk, as rank-1 recognition rates of 0.936, 0.955, and 0.920 were achieved for the Gait Curves, GEI and Frieze pattern algorithms, respectively. Recognition performance was also comparable when comparing "fast vs. normal" walk. Similar to the CASIA B data, a performance loss was evident in all algorithms when comparing different covariates. Interestingly, a degradation is evident when comparing differing walking speeds. Intuitively, this should be expected, as the dynamics of walk are not the same for all speeds. For example, one might take shorter strides when walking at slower paces. This may reflect why the Gait Curves algorithm was the least robust to changes in walking speed, as it is largely concentrated on changes in shape. Of particular interest is that the ROC curves for the GEI algorithm visually depict a larger area underneath the curve (AUC) than the Frieze pattern algorithm, despite having lower rank-accuracies for the speed and carrying condition covariates. In general, these results suggest that the Gait Curves algorithm performs comparably to the described baseline algorithms.

Here, it should be noted there are other advantages of gait curve matching aside from

matching performance numbers. Namely, the complexity (i.e., computational expense) of gait curve matching is much less than the comparison algorithms. For example, the length of the feature vector used for each gait curve was 300 elements. By comparison, the GEI algorithm was 2300 (50x46) and the Frieze pattern algorithm was 100xT, where T is the number of silhouette images extracted in a video sequence. Since the length of the feature vector is relatively low, the Gait Curves algorithm does not require training for subspace optimization and dimensionality reduction (as with GEI).

Recognition Performance on the WOSG Dataset

To establish a baseline of performance on the WVU Outdoor SWIR Gait dataset, three specific experiments were developed. In the first experiment the relative matching accuracy is determined, without specific regard to any covariate (i.e., all-to-all matching). In this case, the matching performance achieved for each algorithm is significantly less than the accuracies obtained in the baseline experiments (Figure 2.14).

The second experiment expands the analysis to specifically account for matching sequences of differing walking directions. These results, expressed in Figure 2.15, also yielded poor performance numbers, but with some interesting caveats. As with Figure 2.14, the initial observation is that the GEI algorithm generally outperforms the Gait Curves and Frieze pattern algorithms. A closer look into the results (as they pertain to direction) indicates that each algorithm tended to yield reduced performance numbers when matching against sequences of horizontal walk (i.e., directions one and two from Table 2.5). This result is particularly interesting as it is commonly believed that gait recognition is optimally performed when an individual is viewed moving perpendicular to the field of view [72]. In other words, the matching performance is expected to be optimal when an individual is viewed moving horizontally in the image plane. In some instances, this reduction of performance is noticeably significant, as in the instances when the reference data consists of directions three and six.

The third experiment investigates the ability of each recognition algorithm to match the same direction of walk. Provided the WOSG dataset only has one gait sequence per identity in a given walking direction, this was accomplished by extracting two feature vectors

sequentially from one gait sequence. In addition to a measure of performance, this experiment indirectly doubles as an additional measure of silhouette quality. In theory, a pair of gait features extracted sequentially (seconds apart) and in the same walking direction should match very well to one another as the walking dynamics are (presumably) consistent in the short-term. Poor recognition performance, therefore, is an indication that the shape of the silhouette is not consistent (i.e., is of poor quality). Noting this, though the results in Figure 2.17 show an increase in rank- accuracy, the increase is not substantial and is direction dependent. The dependence on direction is also noteworthy, as a uniform increase in performance would suggest issues in the matching algorithms, rather than the data. The results across all algorithms indicate that, in general, the "highest" quality silhouettes were extracted from direction six (the observed individual walks from bottom left to upper right of image plane), while the "lowest" quality silhouettes were extracted from direction one (the observed individual walks from right to left in the image plane). Additional evidence of this can be found in Figure 2.15, where matching with direction one consistently yielded among the lowest rank-accuracies.

In summary, of the three evaluated algorithms, *none* exhibited a high matching performance on the WOSG dataset. However, based on the experimental results (Figures 2.14-2.17), an argument can be made that the likely source of error was the lower quality silhouettes generated from the segmentation process (Table 2.2). Recall that the segmentation process used to generate silhouettes on the WOSG dataset is consistent with related gait literature.

Backpack Detection

With regard to the experiments on bag detection (Section 2.5.6), the features derived from a gait curve are generally able to successfully detect the presence of a backpack, with successful detection (or non-detection) rates approaching 90%. Visual evidence of feature separability from Figure 2.4 supports this claim.

Errors in detection generally occur as a result of error in normalizing $y[t]$ (e.g., gait curve) to a common spatial domain. In the case of false positives (involving the detection of a backpack when one is not present), segmentation errors in the silhouette can greatly cause

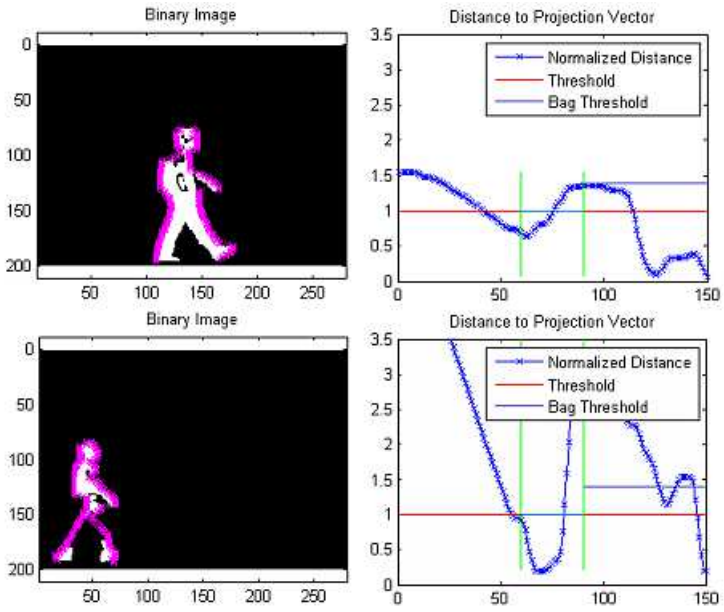

Figure 2.23: Examples of backpack misclassification. Top: False bag rejection. Bottom: False bag positive.

the normalization threshold to be lower than expected, resulting in signal characteristics of a bag. Conversely, missed detection of a backpack can occur if part of a bag exists in the expected waist region. This will result in a higher normalization threshold and reduce the efficacy of the signal based features. Visual examples illustrating the causes of detection errors are found in Figure 2.23. However, although the stated features for backpack detection performed quite well, three of the four are strictly application limited, as they examine only a portion of silhouette shape based on a targeted location of the gait curve. The exception to this is the feature corresponding to expected silhouette shape. This feature is perhaps the most robust because it has no dependence on scale and observes the general silhouette shape. As a result, its usefulness may be extended beyond backpack detection. In spite of the robustness of this feature, it is derived from local database information. Therefore, in an operational setting, each system would need to compute this feature individually, and one feature representation may not work well in a different system or setting.

Silhouette Rectification

Implementation of the silhouette rectification scheme (Section 2.2.5) yielded interesting results. Assuming perfect backpack detection, the rank-1 identification accuracy of the Gait Curves algorithm when comparing "With Bag" samples to "Normal" samples increased by a factor exceeding 100% (0.259 to 0.561, Figure 2.18).[6] The increase in rank-1 identification accuracy exceeded the rank-1 accuracy of the GEI algorithm for each of the three correction schemes and exceeded the rank-1 accuracy of the Frieze pattern algorithm for one of the correction schemes (linear correction, Equation (2.15)). However, the area under the ROC for these corrections is less than the GEI and Frieze algorithms. In general, this result suggests that the silhouette rectification schemes can successfully mitigate the distortions induced from a backpack and improve the recognition accuracy of the Gait Curves algorithm.

Changing the experiment such that samples detected to include a backpack are subject to the rectification schemes (including the reference samples) resulted in an even larger improvement in rank-1 accuracy (Figure 2.19). A similar increase in AUC is also present in the ROC data. This result is particularly interesting, as it might be expected that a *reduced* recognition accuracy would follow due to missed backpack detections. However, in this case, it turns out that some samples incorrectly not recognized as having a backpack are now matching correctly. In other words, the rectification scheme was less beneficial to samples near the decision boundary of "backpack" or "no backpack". Alternatively, there were a few instances where a sample was correctly detected as having a backpack matched correctly to its corresponding sample in the reference set, but the reference sample was incorrectly detected as having a backpack. Such a situation occurs for identities with a large body area, whose shape, even in normal situations might appear as a backpack. Thus, when the correction scheme is applied to both the probe and reference samples, it results in increased similarity. This outcome was in the minority though, found only in three instances. In general, this result suggests that the silhouette rectification schemes are helpful in improving recognition, but not in all cases.

One interesting caveat of the three rectification schemes is that each was found to boost

[6] 0.561 denotes the lowest performing of the three correction schemes.

recognition performance by the same amount (within ±2%). This is likely due to the fact that each correction generally performs the same task, except with increasing complexity. Although each correction method results in a noticable increase in matching performance, the increase is still less than matching "normal" to "normal" sequences. The results of Figure 2.19 suggests some of this discrepancy is due to adverse effects of the correction method on some samples, but it is also likely that other factors, such as cadence account for differences in the gait biometric between "normal" and "with bag" samples.

2.6 Summary

This chapter introduces (a) a novel algorithm for performing human gait recognition and (b) a new challenge gait database to evaluate recognition performance. Defined as gait curve matching, the recognition algorithm denotes a shape-based approach to characterize human gait. The feature representation is also advantageous in that it can be used to detect for abnormalities in the silhouette (e.g., object detection). If objects can be detected, then subsequent silhouette restoration methods can be used to increase recognition performance. Although the Gait Curves algorithm did not outperform the comparison algorithms, the results were generally comparable and the feature representation is less expensive and does not require training or dimensionality reduction.

The second component of this chapter included a performance evaluation on the WVU Outdoor SWIR Gait (WOSG) dataset. The dataset is unique as the data was collected in an unconstrained, outdoor environment, with a SWIR image sensor. This is in contrast to the majority of gait datasets (including those with an infrared image sensor), where the background environment is very controlled. Consequently, one of the major challenges in the data is the segmentation component. Experimental results confirm this result by (a) the use of a quality metric, which suggests the silhouette data is of much lower quality than the silhouettes produced from baseline datasets (using the same segmentation algorithm) and (b) the matching performance observed, which is significantly reduced in comparison to the performance observed on the baseline datasets.

These results indicate that recognition of human gait is an appropriate modality for

biometric surveillance systems, while also highlighting that more research is required in the segmentation component, in order to further advance the state of the art.

Chapter 3

Clustering Human Gait

3.1 Introduction

Traditionally, human gait recognition implies deducing the *identity* based on how a person walks. Chapter 1, Section 1.3.2 summarizes a number of gait recognition algorithms from the literature. Though a number of recognition algorithms have been proposed, research has not explicitly captured what these algorithms are actually capturing to infer a decision regarding identity. In particular, this is true for model-free approaches as the matcher is based on the differences in a moving shape, rather than an explicit set of measured parameters. It might be possible that static physical characteristics are embedded within ones gait pattern. Examples of such physical properties might include body size, walking speed (i.e., cadence), and gender.[1] If so, it may further be the case that these characteristics factor into the assessment of similarity (or difference) between gait patterns from a gait matching algorithm. In the context of a biometric surveillance system, this becomes a very interesting property and further advantage of utilizing gait. One of the primary advantages pertains to the instances when an observed individual is not recognized by the system (i.e., no matching data is found in the database). Arguably, this scenario is likely to occur using gait biometrics in a surveillance context. Assuming an individual is not recognized, it may still be possible to generate a *semantic profile* based on which gait patterns the probe data is most similar to. Semantic profiles can be generated by a *cluster analysis* of gait data.

[1] While gender is primarily a genetic property, it impacts the physical aspects of an individual.

In addition to the generation of semantic profiles as an advantage for clustering gait data, additional benefits exist as well. If the reference data is very large (arguably a property of an operational biometric surveillance system), the search for an identity can be confined to specific clusters based on the input gait data, thereby reducing search time (e.g., biometric indexing [137]). Indexing or semantic profiling can also be used to narrow down the search space for a more sophisticated matching algorithm (e.g., a face matcher, which may be more computationally expensive), if present. Further,

3.1.1 Classification of Physical Attributes from Gait Patterns

As stated previously in Chapter 1, in the early psychological studies by Kozlowski and Cutting, *gender* was concluded to be embedded within human gait [81]. Recent studies using force places using force plates have also concluded that men and women have differing gait patterns [138]. Some researchers have attempted to apply these conclusions to test whether gait data can be used to classify gender. Lee and Grimson, published the first study attempting to classify gender from gait data [139]. Their study attempted to model the side-profile of the human body by fitting ellipses. A classification rate of 84% was achieved on a dataset of 24 identities. Later, Huang and Wang performed a similar study fitting ellipses on front-view and side-view gait data in the CASIA B dataset [140]. Using a fusion approach, a 89.5% classification accuracy was achieved with a bootstrapped test set of 50 identities (25 male, 25 female) identities. In another study by Li et al. [141], the Gait Energy Image (GEI) image [91] was divided into six components denoting different regions of the human body (side-view). Using a SVM classifier, a gender classification rate of 98% was reported on a test set of 122 identities (USF Human ID dataset [120]). Li et al. [141] further remark that the torso and movement of the leading leg contribute the most discriminative power for gender recognition. The success of these gender classification schemes suggests gender may be embedded in a gait feature descriptor. However, it is unclear how strong of an influence gender has, or if there are any other attributes that may be stronger.

Beyond gender, Samangooei and Nixon presented a study attempting to ascertain which physical attributes were related to the matching reference of a probe gait sequence [142]

using GEI features. In their study, the physical attributes tested included those that might be used in a surveillance or law enforcement scenario. A sample of these attributes includes gender, age, height, ethnicity, hair length, and hair color. Each physical attribute was categorized by human annotators into simplified classes (e.g., "Male", "Female", "Tall", "Blonde", "Middle-aged", etc.). Physical attributes and gait features were extracted from the University of Southampton "Large" dataset (115 identities). Their results concluded that gender, hair length, height, and age were often similar returned categories between a probe and the top ranked match. However, only gender and hair length were found to be statistically significant.

3.1.2 Clustering Gait Patterns to Measure Groups of Identities

While research indicated that physical attributes such as gender are likely to be embedded in gait patterns, there has been limited research investigating the clustering of gait patterns, and whether gait patterns can be described semantically. In a study by Watelain et al., a *cluster analysis* was performed to discover whether age ("young" or "elderly") could be inferred from gait [143]. In their study, force-plate and 3-D muscle power data was extracted from walking individuals in "young" and "elderly" age groups. Using hierarchical clustering and analysis of variance (ANOVA) hypothesis testing, the muscle power data was found to be significant among age groups.

3.1.3 Chapter Motivation

In a surveillance context, the studies by Samangooei and Nixon [142] and Watelain et al. [143] are interesting as they show that (a) surveillance-based physical attributes are likely to be embedded in gait patterns [142] and (b) identities can be clustered based on physical attributes [143]. The motivation of this chapter is to build upon both of these studies by performing a cluster analysis of gait features from multiple matching and clustering algorithms, and further analyze whether clusters can be described by semantic physical attributes. This is accommodated by subjecting the feature set extracted by a gait matcher to a clustering scheme. The cluster analysis is well-suited to this problem as it depicts

how matching identities group identities. This analysis will effectively demonstrate whether semantic profiles can be generated from gait features.

Since multiple gait matching algorithms are evaluated, additional post-analysis will examine whether there are any differences between clusters generated from different matchers. From an academic standpoint, such an analysis will demonstrate whether matchers perceive gait features differently, and as such, whether certain physical attributes emphasized uniquely in assessing similarity. An operational caveat of this an analysis is that (a) it is possible to identify the matchers that best profile specific physical attributes and (b) fusion of matching algorithms that assess similarity differently are likely to result in increased matching performance.

In this chapter, the physical properties that will be investigated consist of gender, body area, height, stride, and cadence. The matching algorithms used in the cluster analysis consist of the gait curves algorithm from Chapter 2, and the comparison algorithms: GEI [91] and frieze pattern [101] matching, which are reviewed in Chapter 2, Section 2.4. The clustering analysis is performed using k-means and hierarchical clustering, which are introduced in Section 3.2. The cluster analysis is performed in Section 3.3 and an analysis of the results is presented in Section 3.3.7. A brief summary of key findings is presented in Section 3.4.

3.2 Clustering Algorithms

In general, the clustering problem refers to organizing a set of samples into distinct groups such that samples in the same group (i.e., cluster) are most similar to other samples in the same group and less similar to samples in different groups. In the context of pattern recognition, the clustering problem is typically characterized as an *unsupervised learning* problem, as the goal is typically to *discover* structure (i.e., class-labels) for a set of samples. Thus, clustering is popular in areas such as image analysis [144, 145], bioinformatics [146, 147], and data mining [148, 149].

Clustering is not limited to a single method or algorithm. The literature offers many clustering techniques which vary according to how clusters are defined and how samples are assigned. In this chapter, clustering of gait patterns is performed using the k-means and

Hierarchical clustering algorithms, which denote common approaches to the clustering task. These algorithms are chosen as they are well-known and offer a good starting point for a data discovery problem. A brief description of these algorithms are provided in the following subsections.

3.2.1 K-means Clustering

The original k-means clustering algorithm was developed internally within Bell Labs in the 1950's as a technique for pulse-code modulation and was first published in 1982 [150]. A similar version of the algorithm was also developed and published by Forgy in 1965 [151]. The k-means algorithm is defined as a data partitioning algorithm, where each cluster is defined (and modified) simultaneously.

Algorithmically, the k-means algorithm is an iterative process with an initialization followed by two successively alternating steps: *Assignment* and *Update*. The algorithm initializes with c "means" or "cluster centroids". Denote these "means" as $\mathbf{m}_1^{(0)}, \mathbf{m}_2^{(0)}, \ldots, \mathbf{m}_c^{(0)}$. Note the superscript (0) denotes a time step, t (initially $t = 0$). Each cluster centroid, $\mathbf{m}_i^{(t)}$, $(i \in [1, c])$ has a dimensionality of d, equal to the dimensionality of the sample data being clustered. In other words, $\mathbf{m}_i^{(t)}$ is a vector with the same number of elements as the sample data. Assigning initial values to each $\mathbf{m}_i^{(t)}$ can be performed in a number of different ways. For example, each $\mathbf{m}_i^{(t)}$ can be assigned random values or be assigned values equal to one of the samples to be clustered.

In the assignment step, each sample is assigned to a cluster, c_i, that has the smallest distance between the sample, $\mathbf{x}_j$ $(j = 1, 2, \ldots, N_T)$, and the cluster centroid $\mathbf{m}_i^{(t)}$. Denote $D(\mathbf{m}_i^{(t)}, \mathbf{x}_j)$ as the distance between cluster centroid $\mathbf{m}_i^{(t)}$, and sample data $\mathbf{x}_j$. Mathematically, this is described in Equation (3.1).

$$\mathbf{x}_j \cup c_i : \min_i D(\mathbf{m}_i^{(t)}, \mathbf{x}_j),\ i = 1, \ldots, k \tag{3.1}$$

In the update step, the numerical values pertaining to cluster centroid $\mathbf{m}_i^{(t)}$, are updated to reflect the mean of the sample data assigned to cluster c_i. Mathematically, this is described in Equation (3.2).

$$\mathbf{m}_i^{(t+1)} = \frac{1}{|c_i|} \sum_{\mathbf{x_j} \in c_i} \mathbf{x_j} \tag{3.2}$$

The algorithm terminates when the sample data at iteration t is assigned to the same clusters as in the previous iteration.

One of the advantages of the k-means algorithms is that it will always converge to *some* minimum. However, the identified solution is not guaranteed to be the global minimum. Another limitation of the k-means algorithm is that the number of clusters must be pre-specified as an input parameter. Additionally, the range of values specified for clusters is sensitive to the properties of the distance metric used. For example, using the Euclidean distance metric, k-means tends to generate spherical clusters with respect to the data space.

3.2.2 Hierarchical Clustering

Hierarchical Clustering is a clustering technique that builds clusters by "linking" samples together. This can be performed using an *agglomerative* or *divisive* strategy. In an agglomerative strategy, each sample is initialized as its own cluster (i.e., initialize with N_T clusters). Next, samples are linked together (using some distance metric) such that the number of clusters is successively reduced from N_T to 1. A divisive strategy is the opposite of an agglomerative strategy. Initially, each sample is initialized to the same cluster (i.e., initialize with one cluster). Samples are successively removed and placed in new clusters until each sample is its own cluster. The process of linking samples together results in a "tree", which depicts which samples are linked together for c clusters ($1 \leq c \leq N_T$). This is defined as a dendrogram.

Algorithmically, an agglomerative strategy can be defined as follows. Given a samples $s_1, s_2, \ldots, s_{N_T}$, define dendrogram $\mathcal{D}$ as the set of cluster labels $(l_1^{(c)}, l_2^{(c)}, \ldots, l_{N_T}^{(c)})$ for c clusters ($1 \leq c \leq N_T$). Initially (at $c = 1$), each cluster label has the same value (e.g., $l_k^{(1)} = 1$, $k = 1, 2, \ldots, N_T$). For $c = 2$ to N_T, denote sample s_a, with label $l_a^{(c)} = \alpha$ ($\alpha \geq 1$), as the best matching sample to the samples whose label is equal to β ($\alpha \neq \beta$). The label for sample s_a is set to β and c is increased by one.

The primary advantage of hierarchical clustering is that the analysis depicts how samples

are linked as the number of clusters increases or decreases. In addition hierarchical clustering can isolate small clusters, which *may* be beneficial for discovery problems. However, hierarchical clustering can be computationally expensive, where computation is at least $\mathcal{O}(n)^2$ and n is the number of samples. Additionally, the dendrogram can be significantly impacted by the linkage established between the first few samples.

3.3 Experimental Results

3.3.1 Dataset

Cluster analysis is performed on the "normal walk" subset of the CASIA B Gait dataset. Recall from Chapter 2, Section 2.5.1, the "normal walk" subset of the CASIA B dataset consists of $N_G = 6$ video sequences for $N = 124$ individuals and that the CASIA B Gait dataset is a *controlled*, *indoor*, gait dataset with *high quality* silhouette data (Chapter 2, Table 2.2). The CASIA B dataset is chosen for cluster analysis as it (a) has a large number of identities, and (b) yields high quality silhouettes. The latter point is particularly important as the cluster analysis could be degraded by noisy silhouettes [108].

In addition to collecting the gait biometric from each video sequence, ancillary physical features regarding each identity are also extracted. The ancillary features consist of: Gender, Stride , Cadence (viz., walking speed), Height, and Area.[2] Gender is determined by visual observation. Stride (in pixels) is computed from stride vector $\mathbf{s}_k$ $(k = 1, 2, \ldots, K)$, where $\mathbf{s}_k = max_i\{B_k(i,j)\} - min_i\{B_k(i,j)\}$ and $B_k(i,j)$ refers to the k^{th} silhouette image. The estimated stride corresponds to the average of $\mathbf{s}_\tau$, where τ corresponds to the values of k that denote a local maxima in $\mathbf{s}_k$ (i.e., feet apart). Local maxima are identified by counting the *negative* zero-crossings in the difference vector, $\dot{\mathbf{s}}_k$. Cadence is measured by $|\tau|$, the total number of *negative slope* zero-crossings in the stride difference vector, $\dot{\mathbf{s}}_k$. Height (in pixels) is computed from height vector $\mathbf{h}_k$ $(k = 1, 2, \ldots, K)$, where $\mathbf{h}_k = max_j\{B_j(i,j)\} - min_j\{B_k(i,j)\}$. The estimated height corresponds to the average of $\mathbf{h}_\pi$, where π corresponds to the values of k that denote a local maxima in $\mathbf{h}_k$ (i.e., feet together). Local maxima are identified by

[2]Ground truth for this information is not provided by CASIA.

Table 3.1: Extracted physical characteristic data from the CASIA B dataset.

Height (pixels)	**Area** (pixels)	**Stride** (pixels)	**Cadence** (half cycles)	**Gender** (male / female)
137.0 ± 7.33	3335 ± 443.9	69.1 ± 6.48	5.17 ± 1.23	474 (75%) / 156 (25%)

counting the *positive slope* zero-crossings in $\dot{\mathbf{s}}_k$.[3] Area (in pixels) is measured by area vector $\mathbf{a}_k$, where $\mathbf{a}_k = \sum_i \sum_j B_k(i,j)$. The estimated area corresponds to the average of $\mathbf{a}_\tau$. A summary of the extracted physical data is provided in Table 3.1. Note that "Height", "Area", and "Stride" are measured in pixels. "Cadence" is measured in half gait cycles completed, and "Gender" is a binary value.

3.3.2 Matching Algorithms

Cluster analysis is performed using the Gait Curves (Chapter 2, Section 2.2), GEI [91], and Frieze Pattern matching approaches [101]. Refer to Chapter 2 for a review of these matching algorithms.

Feature vectors for the GEI, gait curve, and frieze pattern algorithms were constructed using all video frames where an individual was viewed fully in the image plane. GEI images were extracted using a 90-pixel horizontal window, with the silhouette height normalized to 100 pixels. The resulting GEI image was then downsampled by a factor of two. Subspace optimization was performed using PCA, wherein a principal component was retained if the associated eigenvalue was greater than 0.001. The Euclidean distance metric was used to compare GEI features. For the gait curves algorithm, each gait curve was normalized to contain 300 elements and the procrustes distance metric (Chapter 2, Equation (2.7)) was used to compare feature vectors. As with GEI, frieze patterns were extracted from a 90-pixel horizontal window and normalized to a height of 100 pixels. Frieze patterns are compared using Dynamic Time Warping (Chapter 2, Section 2.4.2). Since the GEI algorithm utilizes PCA, 15% (19 identities) of identities are used for training the subspace optimization. This is consistent with the method for subspace optimization in the experiments evaluating matching performance in Chapter 2, Section 2.5. Clustering is performed on the remaining $N_C = 105$

[3]The stride vector is used in this calculation as it was found to have a larger range in amplitude.

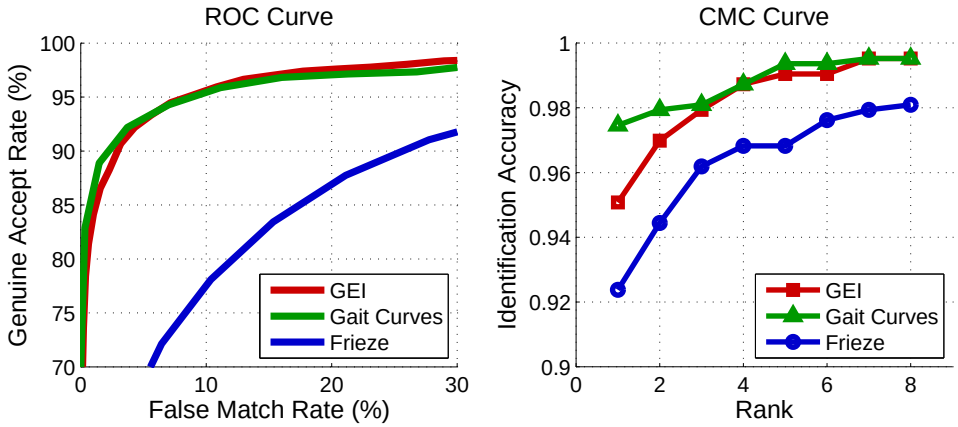

Figure 3.1: Baseline recognition performance of the data used for clustering. Left) ROC Curve. Right) CMC Curve.

identities for all algorithms (N_C: number of identities available for clustering). A baseline recognition performance evaluation consisting of ROC and CMC curves of the cluster data is presented in Figure 3.1. Note the AUC obtained from the samples utilized for clustering is 0.982, 0.981, and 0.922 for the GEI, Gait Curve, and Frieze Pattern algorithms. The corresponding rank-1 identification accuracies are 0.951, 0.975, and 0.924, respectively.

3.3.3 Protocol For Generating Clusters

Clusters are generated using the *k-means* and *hierarchical* clustering algorithms, as described in Section 3.2. Cluster centroids are estimated from a training set based on randomly sampling one of N_G samples for the N_C remaining identities. This set is denoted as C_{train}. The remaining samples are assigned to the test set C_{test}. Each sample in C_{test} is assigned to the cluster whose centroid is closest to the sample. Thus, the maximum number of clusters that can be created is 105 and the minimum is one. Although it is best practice to ensure there are no overlapping identities in C_{train} and C_{test} (to prevent "overtraining"), it is performed here primarily to maximize the amount of data available for post-analysis. Clustering experiments are performed using $c = 5$ and 10 clusters, which are arbitrarily selected.[4] Though arbitrarily selected, the number of clusters must be in a range that will not diminish the physical attributes that we seek to find in individual clusters. The distance metrics are

[4]Note the selection of $c = 5$ and $c = 10$ is arbitrary and may not be the best number of clusters for the analyzed gait data

Sub-Experiment A: Generating and Assigning Clusters (k-means)
Step 1: Draw one sample per identity to comprise C_{train}.
Step 2: Assign the remaining samples to C_{test}.
Step 3: Initialize k-means with an input of c clusters and randomly choose c samples from C_{train} as the initial cluster centers.
Step 4: Learn cluster centroids using C_{train}.
Step 5: Note the identified cluster centers and the total cluster-member distance.
Step 6: Repeat Steps 4-5 2,000 times.
Step 7: Keep the cluster centers with the smallest cluster-member distance.
Step 8: Assign samples in C_{test} to the closest cluster.

Sub-Experiment B: Generating and Assigning Clusters (Hierarchical)
Step 1: Draw one sample per identity to comprise C_{train}.
Step 2: Assign the remaining samples to C_{test}.
Step 3: Initialize hierarchical clustering with each sample as its own cluster.
Step 4: Merge the two clusters that minimize median distance.
Step 5: Repeat Step 4 until c clusters remain.
Step 6: Assign samples in C_{test} to the closest cluster.

Euclidean (GEI), Procrustes (Gait Curves) and Dynamic Time Warping (Frieze Patterns). Regarding k-means clustering, c of N_C samples in C_{train} are randomly selected and initialized as the cluster centroids. The convergence criteria is met when the change in summed distance between each cluster centroid and its members is less than 1e-6 between iterations. This process of random initialization and convergence is repeated 2,000 times (with different initial centroids) and the centroids with the smallest summed cluster-member distance are kept for further evaluation. This is necessary to increase the likelihood of converging to the globally optimal solution. Regarding hierarchical clustering, median-linkage is used to assign clusters. In median-linkage, a sample is added to the cluster whose sample members have the minimum-median distance. The process for generating clusters using k-means is summarized in Sub-Experiment A and the process for generating clusters using hierarchical clustering is summarized in Sub-Experiment B.

3.3.4 Basic Cluster Analysis

First, a basic analysis is performed investigating the composition of the generated clusters. This includes a visual analysis (i.e., a visual look at the clusters produced), an analysis

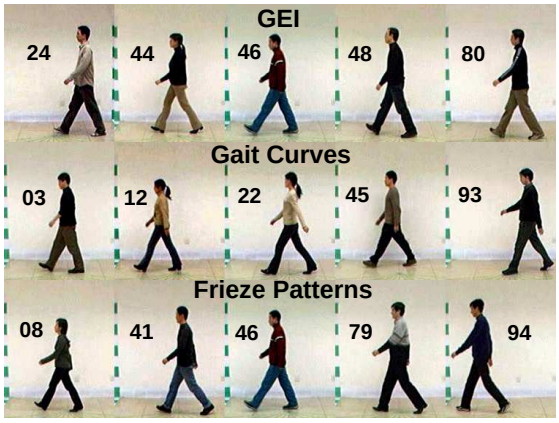

Figure 3.2: Samples nearest to the identified cluster centroid for each matching algorithm for k-means clustering with $c = 5$ clusters. Note that with exception to #46, nearest-centroid samples are different across matchers and appear reasonably distinct.

of *cluster membership* (i.e., samples per cluster), *cluster matching* (i.e., the proportion of samples in C_{test} assigned to the same cluster as their corresponding sample in C_{train}), and *cluster purity* (i.e., identities per cluster) for both the k-means and hierarchical clustering algorithms. The motivation of the basic analysis is to ascertain whether identities are, in fact, being clustered into different groups.

Visual Analysis

In the visual analysis, the samples nearest to each cluster centroid as they pertain to each matcher are visualized for one execution of the k-means and hierarchical clustering algorithms (with $c = 5$).[5] The intent of the visual analysis is to visualize the samples identified near the cluster centroids. In a simple context, if the nearest cluster samples appear distinct visually, it may be an indication that meaningful clusters are being generated. The nearest centroid samples for each matcher are visualized in Figure 3.2 for the k-means clustering algorithm and Figure 3.3 for the hierarchical clustering algorithm. Note the physical variations between identities in Figures 3.2-3.3.

[5]For hierarchical clustering, the sample nearest to the cluster centroid is defined as the sample that has the maximum similarity to all other samples in the same cluster.

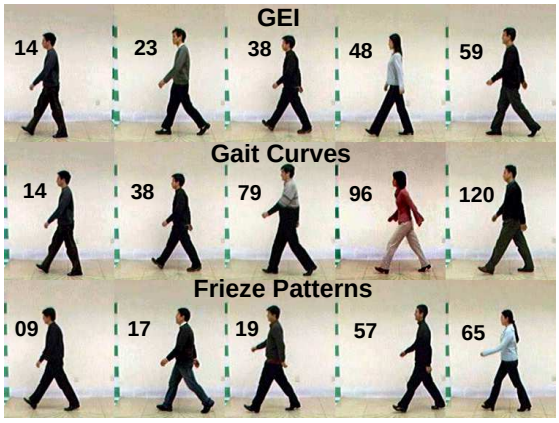

Figure 3.3: Samples nearest to the identified cluster centroid for each matching algorithm for hierarchical clustering with $c = 5$ clusters. Note that with exception to #14 and #38, nearest-centroid samples are different across matchers and appear reasonably distinct.

Cluster Membership

Here, an analysis is performed to investigate the distribution of cluster members. In other words, an analysis of the number of samples per cluster. Though it is difficult to ascertain what would comprise an ideal distribution, it may be more prudent to investigate whether cluster membership is not dominated by one (or two) clusters. Figure 3.4 and Figure 3.5 depict a histogram of member size for one trial of cluster generation for $c = 5$ and $c = 10$ clusters for k-means and hierarchical clustering, respectively. This includes the C_{train} cluster training samples and the C_{test} assigned samples. For aesthetic purposes, cluster membership size is sorted from largest to smallest for each feature representation scheme (e.g., Gait Curves, GEI, Frieze Patterns). Note that for both k-means and hierarchical clustering, one cluster is much larger than the others. This attribute is much more pronounced in hierarchical clustering (Figure 3.5).

Evaluating Cluster Matching

Here, an analysis is performed to investigate whether samples in C_{test} are being assigned to the same cluster as their corresponding sample in C_{train}. This is analogous to the *hit rate* in the biometric indexing problem in the literature. The indexing problem denotes a runtime optimization problem, wherein the reference data in the database is assigned a

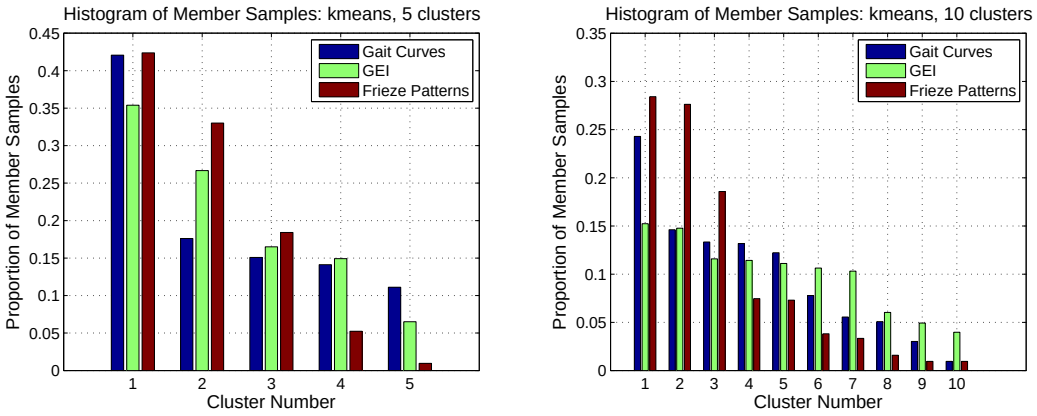

Figure 3.4: Histogram of samples per cluster for k-means clustering.

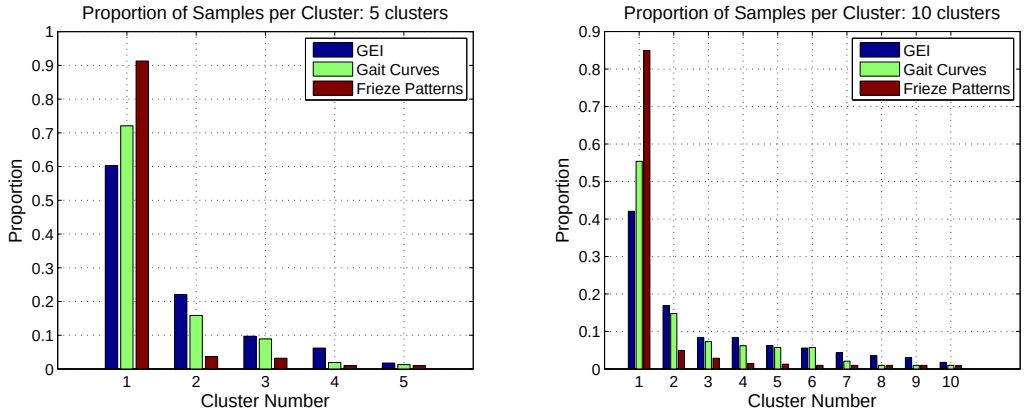

Figure 3.5: Histogram samples per cluster for hierarchical clustering. Note that most samples are placed in a single cluster.

second identifier (i.e., a cluster label), which is used to reduce the search space (i.e., the number of match scores computed) for a given probe [137]. In the indexing literature, the *hit rate* is defined as the probability that a probe sample matches to a cluster that contains the identity of the probe. Mathematically, this is described in Equation 3.3 where N_{hit} denotes the number of times a probe matches to a cluster containing the identity belonging to the probe and N_{probe} denotes the number of probe samples (i.e., the number of samples in C_{test}).

$$\text{Hit Rate } = \frac{N_{hit}}{N_{probe}} \tag{3.3}$$

The proportion of correctly assigned test samples (e.g., hit rate) is tabulated in Table 3.2

Table 3.2: Computed hit rates for the GEI, Gait Curve, and Frieze Pattern algorithms with $c = 5$ and $c = 10$ clusters.

k-means	**Hit Rate** ($c = 5$)	**Hit Rate** ($c = 10$)
GEI	0.798	0.685
Gait Curves	0.745	0.741
Frieze Patterns	0.598	0.512
Hierarchical	**Hit Rate** ($c = 5$)	**Hit Rate** ($c = 10$)
GEI	0.878	0.754
Gait Curves	0.836	0.739
Frieze Patterns	0.966	0.973

for $c = 5$ and $c = 10$ clusters. Ideally, the hit rate would be 1.0, indicating that samples are being matched to the same cluster as its corresponding training sample. Values significantly less than 1.0 may be an indication that clustering of identities is not occurring.

Note a complete description of the biometric indexing problem and analysis is treated as supplementary material and can be found in Appendix A.

Evaluating Cluster Purity

Here, an analysis is performed to investigate the "purity" of clusters, where "purity" is defined by the proportion of clusters that contain a single identity. Ideally, samples for each identity would be assigned to the same cluster. This would indicate a "high" cluster purity. If samples for each identity are assigned across several clusters, this might suggest that the clusters do not convey any meaningful information. This analysis is performed by noting the cluster labels for both the C_{train} and the C_{test} assigned samples. Figure 3.6 and Figure 3.7 conveys this information in the form of a histogram of the number of clusters each identity is represented in. Note that for both clustering algorithms, most identities are represented by one or two clusters.

Conclusions of the Basic Analysis

Summarizing the results of the basic analysis, k-means clustering was generally found to assign samples to multiple clusters, which is a trait indicating groups are being formed via the clustering process. However, the hit rates produced are lower than might be desired.

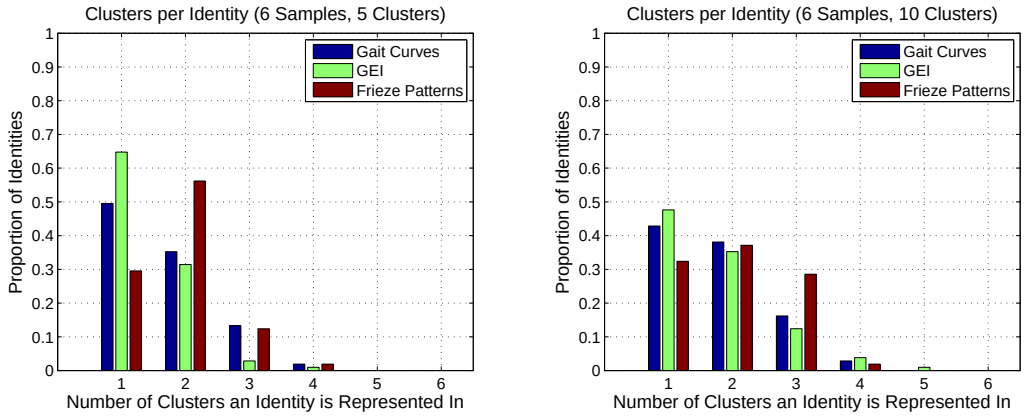

Figure 3.6: Histogram of identities per cluster for k-means clustering. Generally, most identities are represented by no more than two clusters.

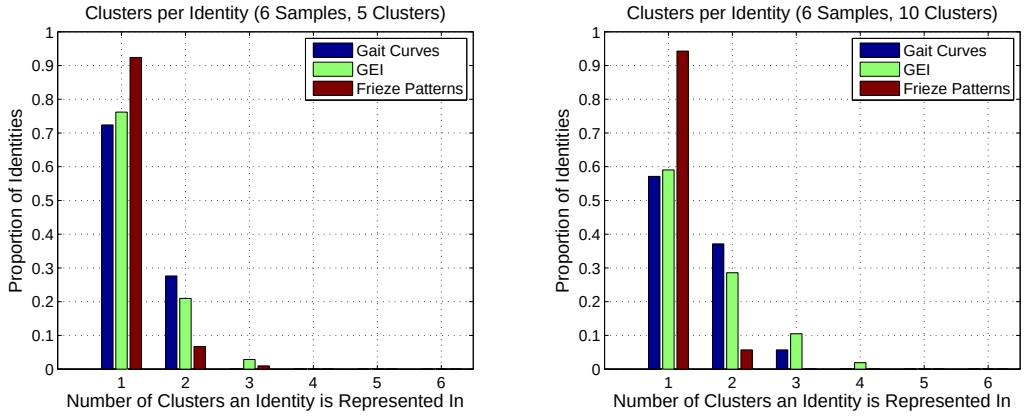

Figure 3.7: Histogram of identities per cluster for hierarchical clustering. Generally, most identities are represented by no more than two clusters.

Despite this, most identities are generally observed in no more than two clusters. Hierarchical clustering differed from k-means clustering in that samples were more likely to be distributed to the same cluster. In other words, one "large" cluster was generated accompanied by "smaller" clusters. As such, the corresponding hit rates and cluster purity were artificially higher. Given the differences between the two clustering algorithms, it is likely that k-means clustering is "clustering" samples while hierarchical clustering is not. Generally, this is because the k-means clustering algorithm is generating more diverse clusters (e.g., clusters of proportional size). Though hierarchical clustering produced a higher hit rate (Table 3.2) and more "pure" clusters (Figure 3.5), this is largely due to most samples being assigned to the same cluster (Figure 3.7). In detail, the small number of member samples for clusters

4 to 5 in Figure 3.5 (left) and clusters 7 to 10 in Figure 3.5 (right) are concerning as they involve a small proportion of samples. This is particularly evident for the clusters generated from Frieze pattern features, where over 80% of the sample data is assigned to a single cluster for both $c = 5$ and $c = 10$. As a result of these findings, hierarchical clustering may not be adequately clustering the sample data. Consequently, **the remainder of the experimental analysis will only utilize k-means clustering**.

3.3.5 Evaluating Identity Pairs in Clusters

In this experiment, an analysis is performed identifying whether *pairs of identities* are clustered similarly among different feature representation schemes. The results of such an experiment demonstrate whether different gait matchers group identities similarly. This is accomplished by tabulating which of the C_{train} samples (corresponding to one identity per sample) appear in the same cluster following cluster generation. Here, the test samples (C_{test}) are not considered. Define Ω as a binary matrix (e.g., table) of identity pairs, of size N_CxN_C. Let $\alpha(i,j)$ $(i,j = 1,2,\ldots,N_C)$ be the row and column entries of Ω. Each $\alpha(i,j)$ is assigned a value of "true" if the i^{th} and j^{th} identity is assigned to the same cluster. Otherwise $\alpha(i,j)$ is "false". The proportion of identity correspondences between two matrices, Ω_1 and Ω_2 (e.g., GEI and gait curves) is described in Equation (3.4).

$$\omega = \mid \frac{\Omega_1 \text{ AND } \Omega_2}{\Omega_1 \text{ OR } \Omega_2} \mid \tag{3.4}$$

Here, $\mid \Omega \mid$ counts the number of "true" entries in Ω. High values of ω denote denote increased cluster similarity between matchers. Equation (3.4) can be expanded to show common identity pairs for $\mathcal{M}$ matching algorithms ($\mathcal{M} = 1,2,\ldots$). This expansion is defined in Equation (3.5).

$$\omega = \mid \frac{\Omega_1 \text{ AND } \Omega_2 \text{ AND } \ldots \text{ AND } \Omega_{\mathcal{M}}}{\Omega_1 \text{ OR } \Omega_2 \text{ OR } \ldots \text{ OR } \Omega_{\mathcal{M}}} \mid \tag{3.5}$$

Using Equation (3.5), ω is computed using the following combinations: Gait Curves (Ω_1), GEI (Ω_2), and Frieze Patterns (Ω_3), $\mathcal{M} = 3$; Gait Curves (Ω_1) and GEI (Ω_2), $\mathcal{M} = 2$; Gait Curves (Ω_1) and Frieze Patterns (Ω_2), $\mathcal{M} = 2$; GEI (Ω_1) and Frieze Patterns (Ω_2), $\mathcal{M} = 2$.

Table 3.3: Proportion of similarly paired identities when clustering GEI, gait curve, and frieze pattern data with $c = 5$ and $c = 10$ clusters.

Combination	ω $(c = 5)$	ω $(c = 10)$
Gait Curves, GEI, Frieze Patterns	0.068 ± 0	0.045 ± 0
Gait Curves, GEI	0.333 ± 0	0.183 ± 0
Gait Curves, Frieze Patterns	0.155 ± 0	0.172 ± 0
GEI, Frieze Patterns	0.161 ± 0	0.106 ± 0

These values are computed for 1,000 different trials of cluster generation. The mean of these values ($\pm$ one standard deviation) are tabulated in Table 3.3 for $c = 5$ and $c = 10$ clusters, respectively. Note that in general, the values of ω produced between matchers are low, suggesting each matcher assesses similarity between gait patterns differently.

3.3.6 Measuring Significance of Physical Characteristics

In this experiment, an analysis is performed in order to determine if any of the physical attributes (e.g., Gender, Stride, Cadence, Height, Area) are distinct across clusters. The intent of this experiment is to discover whether identities in the same cluster share certain physical characteristics. This experiment also doubles as an indirect evaluation of the clustering tendency of the gait data and also utilizes the test C_{test} samples.

Significance of a physical variable is measured using two statistical tests. First, a *chi-square test of independence* is performed using the cluster labels and each physical variable. The chi-square test of independence evaluates whether a physical variable is independently distributed among cluster labels. If the physical variable is independently distributed (i.e., the null hypothesis is accepted), it does not have any impact on the clustered data. Second (assuming the null hypothesis is rejected), the *Spearman Rank Correlation Coefficient* (r_s) is computed for each physical variable and cluster label. The strength of correlation is measured by how close r_s is to a value of 1.0. In computing r_s, the labels for each cluster are assigned in *increasing* order of the cluster-mean corresponding to each physical variable. This ensures that r_s is valued between [0,1] and is not adversely impacted by the arbitrary cluster label assignment.

These experiments are performed 1,000 times, with $c = 5$, and $c = 10$ clusters, for the

Table 3.4: Cluster dependence on physical attributes; $c = 5$ clusters. The proportion of accepted null hypothesis tests is displayed (where appropriate).

Variable	**Matcher**	**Independent?**	p-**value**
Gender	GEI	No	0 ± 0
Gender	Gait Curve	No	0 ± 0
Gender	Frieze Patterns	No	0 ± 0
Stride	GEI	No	0 ± 0
Stride	Gait Curve	No	0 ± 0
Stride	Frieze Patterns	No	0 ± 0
Cadence	GEI	No	0 ± 0
Cadence	Gait Curve	No	0 ± 0
Cadence	Frieze Patterns	Yes (41%)	0.19 ± 0.23
Height	GEI	No	0 ± 0
Height	Gait Curve	No	0 ± 0
Height	Frieze Patterns	No	0 ± 0
Area	GEI	No	0 ± 0
Area	Gait Curve	No	0 ± 0
Area	Frieze Patterns	No	0 ± 0

GEI, Gait Curves, and Frieze Pattern matchers. Tables 3.4-3.5 tabulate the assertion of dependence and average p-value ($\pm$ one standard deviation) for each variable and matcher. Dependence is asserted if exactly *zero* hypothesis tests result in acceptance of the null hypothesis. Tables 3.6-3.7 tabulate the average Spearman r_s and average p-value. Both values are reported $\pm$ one standard deviation. Low p-values denote the likelihood the asserting that the test outcomes were not due to random chance.[6]

3.3.7 Discussion

Again looking first at the output of the clustering process, Figure 3.4 illustrates the distribution of samples per cluster. Note that in general, neither the k-means clustering algorithm or the hierarchical clustering algorithm resulted in an equal distribution of samples per cluster. However, the clusters generated via k-means clustering did not contain a large majority of the sample data. The largest clusters via k-means contained $\approx$40% of the sample

[6]In Tables 3.4-3.7 p-values smaller than 0.001 are rounded to 0.

Table 3.5: Cluster dependence on physical attributes; $c = 10$ clusters. The proportion of accepted null hypothesis tests is displayed (where appropriate).

Variable	**Matcher**	**Independent?**	p-**value**
Gender	GEI	No	0 ± 0
Gender	Gait Curve	No	0 ± 0
Gender	Frieze Patterns	No	0 ± 0
Stride	GEI	No	0 ± 0
Stride	Gait Curve	No	0 ± 0
Stride	Frieze Patterns	No	0 ± 0
Cadence	GEI	No	0 ± 0
Cadence	Gait Curve	No	0 ± 0
Cadence	Frieze Patterns	Yes (3%)	0.007 ± 0.054
Height	GEI	No	0 ± 0
Height	Gait Curve	No	0 ± 0
Height	Frieze Patterns	No	0 ± 0
Area	GEI	No	0 ± 0
Area	Gait Curve	No	0 ± 0
Area	Frieze Patterns	No	0 ± 0

data, whereas the largest clusters generated via hierarchical clustering contained $\approx 40\% - 90\%$ of the sample data. These results indicate that the k-means algorithm is likely clustering identities into discernable groups. The large bias in samples per cluster for hierarchical clustering makes it difficult to ascertain whether discernable groups are being created. It may be possible that the sample data was not of sufficient size or variability for hierarchical clustering to perform well.

From the first experiment, Table 3.3 demonstrates the proportion of similarly clustered identities (ω, Equation (3.4)) between matchers. The motivation for such an experiment is to understand if gait matching algorithms tend to group the same identities together. The low values of ω in Table 3.3 indicate that the *gait matchers tended to group identities differently.* This suggests that different gait matchers assess similarity between gait patterns differently. In addition, the values of ω were found to be equal for each trial of cluster generation (with different training samples). This suggests that the clustering process is converging to the same local minimia, which could be the global minimum.

Provided that the matching algorithms cluster identities differently, it is worthwhile to

Table 3.6: Spearman's correlation coefficient for the clustered physical attributes; $c = 5$ clusters.

Variable	**Matcher**	**Spearman** r_s	p**-value**
Gender	GEI	0.691 ± 0	0 ± 0
Gender	Gait Curve	0.501 ± 0.002	0 ± 0
Gender	Frieze	0.234 ± 0.006	0 ± 0
Stride	GEI	0.457 ± 0	0 ± 0
Stride	Gait Curve	0.450 ± 0	0 ± 0
Stride	Frieze	0.153 ± 0.010	0 ± 0
Cadence	GEI	0.192 ± 0	0 ± 0
Cadence	Gait Curve	0.293 ± 0	0 ± 0
Cadence	Frieze	0.096 ± 0.012	0.022 ± 0.024
Height	GEI	0.691 ± 0	0 ± 0
Height	Gait Curve	0.323 ± 0.003	0 ± 0
Height	Frieze	0.227 ± 0.010	0 ± 0
Area	GEI	0.804 ± 0	0 ± 0
Area	Gait Curve	0.495 ± 0.001	0 ± 0
Area	Frieze	0.378 ± 0.012	0 ± 0

evaluate whether fusion of the matchers that are the most distinct from one another results in the highest gains in recognition performance. In Table 3.3, the lowest value of ω produced between a pair of matching algorithms denotes the pair with the least similarity between clustering results. For $c = 5$, this corresponds to Gait Curves and Frieze Patterns. For $c = 10$, this corresponds to GEI and Frieze Patterns. One might argue that fusion of either the GEI or Gait Curves matcher with the Frieze Pattern matcher should achieve the highest recognition performance, as the pair of algorithms have minimum information overlap. To test this, ROC and CMC curves of the fused data are created. These results are presented in Figure 3.8. The AUC values produced are 0.984, 0.984, and 0.974 for fused GEI and Gait Curves, GEI and Frieze Patterns, and Gait Curves and Frieze Patterns, respectively. The rank-1 accuracies are 0.967, 0.968, and 0.975 for fused GEI and Gait Curves, GEI and Frieze Patterns, and Gait Curves and Frieze Patterns, respectively. As hypothesized, the fusion results with the best recognition performance involved the Frieze Pattern algorithm.

The second experiment attempts to ascertain whether the generated clusters can be described by semantic terms (i.e., physical attributes). This is facilitated by extracting physical

Table 3.7: Spearman's correlation coefficient for the clustered physical attributes; $c = 10$ clusters.

Variable	**Matcher**	**Spearman** r_s	p**-value**
Gender	GEI	0.707 ± 0.008	0 ± 0
Gender	Gait Curve	0.604 ± 0.032	0 ± 0
Gender	Frieze	0.427 ± 0.078	0 ± 0
Stride	GEI	0.526 ± 0.027	0 ± 0
Stride	Gait Curve	0.484 ± 0.031	0 ± 0
Stride	Frieze	0.321 ± 0.054	0 ± 0
Cadence	GEI	0.278 ± 0.036	0 ± 0
Cadence	Gait Curve	0.337 ± 0.034	0 ± 0
Cadence	Frieze	0.258 ± 0.070	0 ± 0
Height	GEI	0.673 ± 0.019	0 ± 0
Height	Gait Curve	0.483 ± 0.048	0 ± 0
Height	Frieze	0.374 ± 0.052	0 ± 0
Area	GEI	0.818 ± 0.011	0 ± 0
Area	Gait Curve	0.617 ± 0.028	0 ± 0
Area	Frieze	0.484 ± 0.051	0 ± 0

attributes of the identities in the CASIA B dataset, and attempting to find correlation in the cluster data. Such an analysis also provides insight into which physical attributes are captured by a gait matcher and their relative "weight" as it pertains to matching. Tables 3.4-3.5 first demonstrate that each of the physical variables are *dependent* with respect to their assigned cluster, with exception to cadence in the frieze pattern matcher, which was found to be independent in some trials (41% and 3% for $c = 5$ and $c = 10$, respectively). From Tables 3.6-3.7, each of the three matchers exhibited the highest correlation in body area and gender. Gender is particularly interesting because it was not evenly distributed in the dataset. Of the 105 identities available for clustering, only 26 ($\approx$25%) were female. In Figure 3.9, a histogram of gender and cluster number is presented for each of the matchers. Note that each cluster is largely dominated by one gender. Given the small number of females in the test data, this likely describes the why the cluster sizes were not the same from the k-means clustering algorithm (Figure 3.4) and presents evidence that *meaningful* clusters are being generated. Additionally, the "male" and "female" clusters are further subdivided primarily based on body area. This is evidenced in Figure 3.10. As the number of clusters

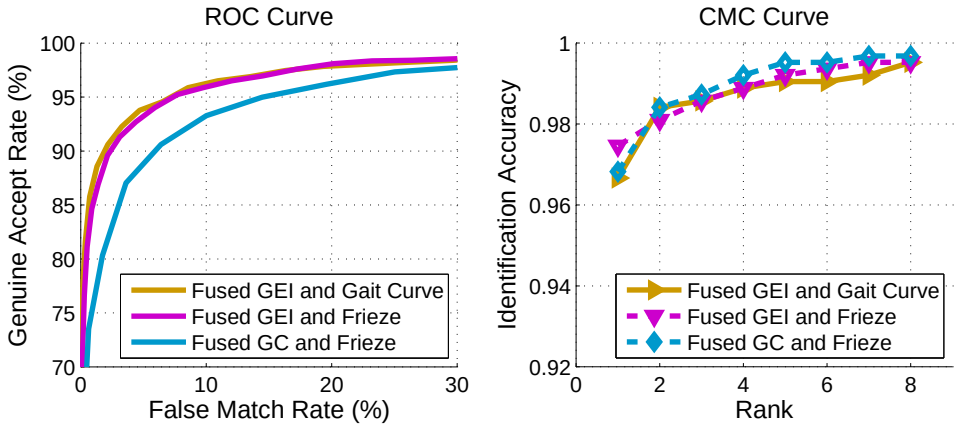

Figure 3.8: Matching performance when the matchers used in the cluster analysis are fused (score-level). Note the best fusion results involve the algorithms most distinct from one another (viz. Table 3.3). Left) ROC Curve. Right) CMC Curve.

increased from 5 to 10, the r_s values for height, stride, and cadence increased slightly. This suggests that these properties play a secondary role when separating identities into a larger number of groups.

Between matchers, the strength of body area and gender in cluster generation was the most pronounced in the GEI algorithm. This is likely attributed to the fact that the GEI matcher is an appearance-based recognition scheme, where the pixel content is the primary feature vector. The gait curves matcher is unique in that stride and cadence play a larger role than in any of the other matchers. This suggests that the gait curves matcher captures more information regarding the dynamics of the lower limbs. The frieze pattern matcher presented comparatively lower values of r_s for each physical variable. In particular, the low value of r_s for cadence is expected as it was found to show a weak dependence on cluster label. These values, as well as the lower values of ω when comparing identity pairs (Table 3.3) suggest either (a) that there is an unknown latent variable that is driving how the Frieze Pattern matcher assess similarity between gait patterns or (b) there is a higher proportion of identities whose samples are distributed across multiple clusters.

Overall, these results suggest that (a) gait patterns can be clustered, (b) the clusters can be described by certain physical properties, and (c) the description of clusters *is* dependent on the matcher. These results are very important in the context of an operational surveillance system utilizing human gait. As mentioned in Section 3.1, a surveillance system utilizing gait

Figure 3.9: Histogram of gender and assigned cluster. Note that most clusters are predominantly characterized by a single gender.

may observe individuals it is not able to recognize as a consequence of not having available reference data. In such a scenario, clustering enables the generation of profiles, which would be beneficial to the operator. Such profiles could be used to direct resources for more computationally expensive matchers (e.g., a face localizer and matcher) while maintaining real-time operation. In another example, suppose the reference database is very large. This could occur naturally if the system stores sample data acquired in real-time. In this case, clustering can be utilized to *reduce* the search space in the reference dataset.

One question that might arise from this study is if clusters can be described via physical characteristics, why not use the physical descriptors as features for recognition. However, it turns out that using the physical characteristics as features for recognition does not result in outstanding matching performance. In Figure 3.11, Receiver Operating Characteristic (ROC) curves are presented for the physical features, as well as the GEI, gait curve, and frieze pattern matchers on the data used in the cluster analysis. Here, match scores from the physical features are generated using a normalized-Euclidean distance and the Mahalanobis distance. Note to use Euclidean distance, the features must be normalized to remove bias, as each has a different mean and variance. On the other hand, the Mahalanobis distance does not require normalization, but rather an estimate of the covariance between features, which

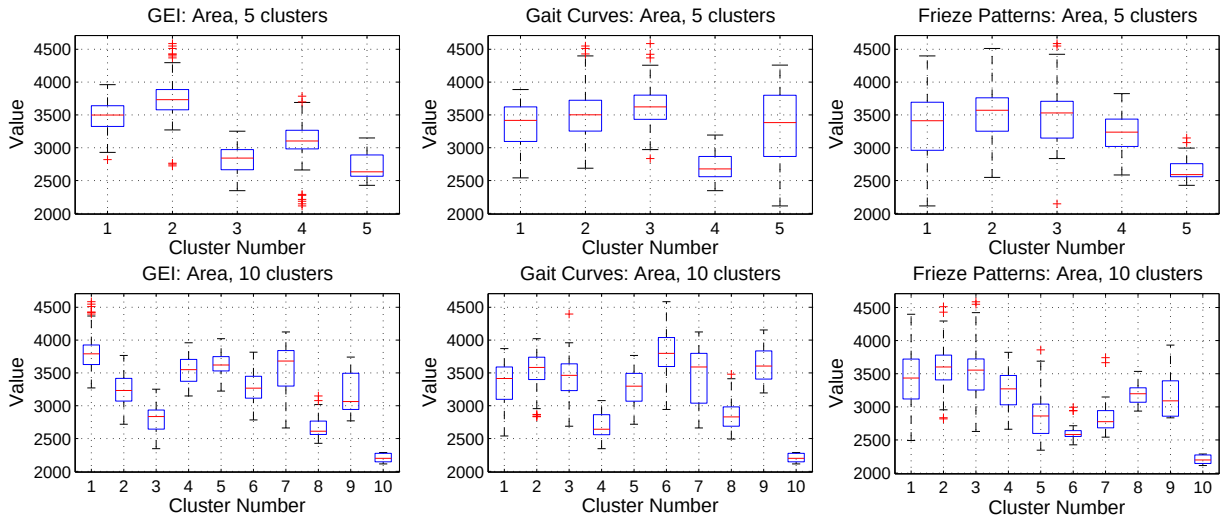

Figure 3.10: Box plot of body area and assigned cluster. Note that the distinct male and female clusters present in Figure 3.9 are separated by differences in body area.

is acquired from the samples used to train subspace optimization for the GEI algorithm (e.g., samples not used in clustering). In Figure 3.11, note that the ROC data for the physical features suggests *decreased* recognition performance when compared to the three matchers. The decrease in recognition performance is likely due to the fact that matchers are utilizing additional physiological properties that were not utilized in this analysis.

One limitation of this study is that the analysis is computed on a single dataset and clustering algorithm, both of which are areas of future work. With respect to clustering algorithms, empirical analysis using hierarchical clustering demonstrated that the distribution of samples per cluster was too biased to a single cluster to reliably conclude whether meaningful clusters were being generated. These results occurred using a number of linkage strategies. Advanced clustering schemes, such as affinity propagation [152], which sets the numbers of clusters intrinsically, tended to converge to a solution with $c = 2$ clusters and most sample data being assigned to a single cluster. Future studies should consider the OU-ISIR gait database [134], which is another high-quality gait dataset that emphasizes cadence. Provided the resources were available, it would be ideal to recreate this type of study on a larger and more demographically diverse dataset. However, such a dataset does not exist in the public domain.

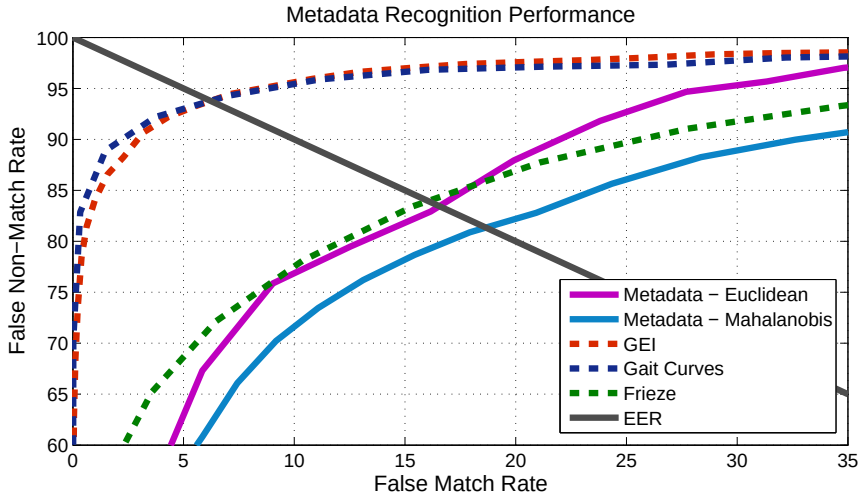

Figure 3.11: ROC curves comparing general recognition performance of the metadata and the GEI, Gait Curve, and Frieze Pattern matching algorithms. Note that the metadata does not perform recognition as well as the matching algorithms.

3.4 Summary

In this chapter, a clustering analysis is performed on three gait recognition matchers from the literature (Gait Energy Image, Gait Curve matching, Frieze Pattern matching) to investigate whether certain physical properties (body area, gender, height, stride, cadence) are utilized by gait matchers to assess similarity between gait patterns. Using k-means clustering, the analysis conveys the following:

- Human gait *can* be clustered into a small number of groups (e.g., 5-10 clusters). Within clusters, identities sharing similar *body area* and *gender* tend to occur together.
- Clustering of gait problems is relevant to a biometric surveillance system in that it may be possible to generate a physical profile of an individual, which can yield some information, or potentially as an indexing scheme for a more traditional biometric matching algorithm.
- The three gait matchers studied in this work assess similarity between gait patterns differently. Thus, fusion of matchers that cluster identities differently is likely to increase recognition performance.

Chapter 4

Anonymous Identification: A Matching Framework for a Biometric Surveillance System

4.1 Introduction

4.1.1 Matching in Traditional (Overt) Biometric Systems

In a typical biometric system [153], the input probe (query) biometric data is compared against the reference samples residing in the reference database. Traditionally, these samples are added into the database during an *overt* collection procedure, which is defined as an enrollment. During enrollment, an individual submits their biometric data, which is paired with an identifier (e.g., name, user-id, etc.). The biometric data and identifier is then added to the database. This process enables the system to either *deduce* the identity of some input biometric data (referred to as identification or 1:N matching) or *verify* the identity of some input data (referred to as verification or 1:1 matching). In the traditional verification and identification problems, the contents of the reference database do not change following a matching decision. In other words, the contents of the reference database are generally *fixed*. For example, a biometric system deployed for access control would contain a database comprised of individuals whom are allowed access to something tangible. Conversely, a biometric system deployed for border throughput might be concerned in preventing a select

number of people from crossing. Again, in this scenario the database would comprise of a list of individuals to be on the lookout of (i.e., a watch-list).

4.1.2 Matching Requirements in a Biometric Surveillance System

In a surveillance application, the primary application might be to identify the presence of suspicious individuals (via a watch-list). Conversely, the application might seek to determine if an unfamiliar individual is present, which could also be indicative of a security threat. However, a traditional biometric identification system is limited to making a decision on whether an observed individual matches to the watch-list, which may be limited in size and only contain a number of high-profile threats. Other suspicious behavior, such as the frequency in which individuals are observed, cannot be ascertained. This reflects an additional surveillance need, as suspicious patterns of observation could be indicative of a malicious act.

Therefore, in a biometric surveillance system, the following properties within the matching scheme would be ideal:

- The system should be able to identify whither or not a specified individual (i.e., an enrolled identity) is present.
- The system should be able to ascertain whither or not a *unknown* individual is present.
- The system should be able to determine whither or not *any* individual (i.e., an enrolled or not enrolled identity) has been encountered before.

A traditional open-set identification can be used to address items 1 and 2 above. However, item 3, which asks "Has this person been encountered before?", cannot be directly ascertained. This is due to the fixed nature of the database. In order to fully address this question, the database must be able to automatically enroll *all* probe samples that do not match to any reference sample in the database. In doing so, the database can grow following the matching decision rendered from the matching algorithm.

4.1.3 Anonymous Identification

To involve the capability of enrolling unrecognized biometric samples to the database, the framework of a classical biometric identification system is modified such that:

- An explicit enrollment process is not required. Instead, biometric data presented to the system for matching is added to the reference database directly following the outcome from the matcher (e.g., after a probe is tested for a match, it is added to the database).
- If a probe sample is matched to an entity in the reference database, it is assigned the same identifier as that of the matching entity.
- If a probe sample does not match to any entity in the reference database, it is assigned a unique identifier from the system.

Thus, the system is strictly determining if a matching sample in the reference database exists. Consequently, without *a priori* identity information (via an enrollment process), the recognition problem is fundamentally changed and the system addresses the question: "Has this person been encountered before?" Therefore, such a system no longer performs classical identity management, but engages in what is defined as *anonymous identification*. The term anonymous identification is used to define this matching process, as the "true" identity of any individual observed is never collected or stored within the system. Rather, the system assigns its own class labels (i.e., identifier) to each entity in the reference database. As a result, neither the identifier from the matcher nor the matching process necessarily deduces the "true" identity of a probe. Note this formulation is distinct from the classical identification problem, since the identification problem assumes (a) samples in the reference database are absolutely associated with a known identity and (b) the reference database is static (i.e., it does not update following each probe observation). Figure 4.1 illustrates the functionality of an anonymous identification system. Note that identity information is not explicitly presented to the system, a separate enrollment process is not necessary and that the matching outcome does not report a specific identity, but rather a label.

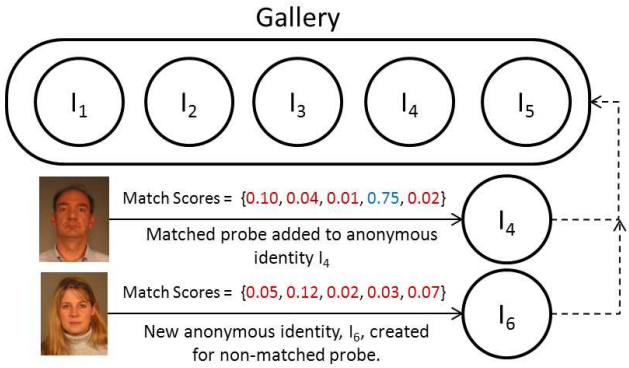

Figure 4.1: Simple flow diagram of an anonymous identification system. Here, the input probe is compared against the reference database in order to determine if there is a match. If a match exists (top), then the probe is labeled with the identifier of the matching reference. If a match does not exist (bottom), then a new identity profile is created. Face images are taken from the FRGC dataset [3].

4.1.4 Benefits of Anonymous Identification

Anonymous identification, as defined in this Chapter, confers a number of benefits specifically pertaining to a biometric surveillance system. For example, as defined in Chapter 1, Section 1.2.2, a biometric surveillance system performs identification-at-a-distance. In applications involving public areas (e.g., shopping centers, airports, etc.), it is likely the system encounters a large number of individuals that might not have corresponding biometric data in the reference database. It may be therefore advantageous to (a) detect the presence of unknown identities and (b) generate "identity profiles" by storing the newly collected biometric data in the reference database. The generation and storage of identity profiles enables the ability for future matches to be made, should the identity appear again. The anonymous identification framework enables this by enrolling observed identities into the reference database as they are encountered in real-time. In doing so, an identity profile is either created (an unseen individual was detected) or updated (a previously stored identity was identified). This dynamic expansion of the reference database is made possible by taking advantage of the fact that an identification-at-a-distance system does not necessarily require an overt collection of biometric data. Thus, an anonymous identification system could also operate covertly, a property that is highly desirable in surveillance applications. Anonymous identification may also be a natural matching framework for incorporating "soft" biometric

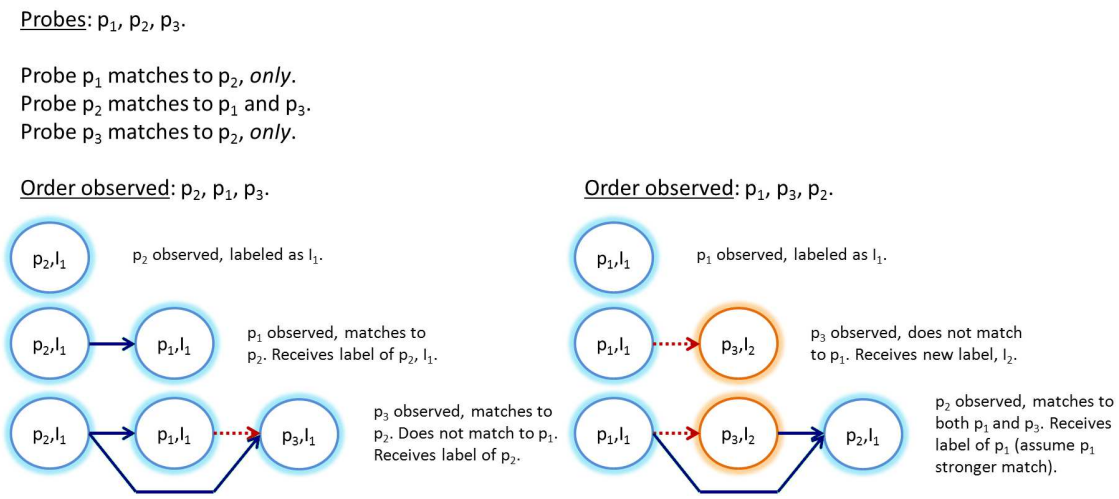

Figure 4.2: Example demonstrating the effect of order of probe encounter in an anonymous identification framework. Here, depending on the order in which probes are observed, either one or two identity profiles are created.

modalities, such as human gait.

In addition to the operational benefits for surveillance systems, the matching and error properties of anonymous identification can be expanded to the biometric de-duplication and re-identification problems [55, 56, 57, 58, 59]. In the context of biometric recognition, de-duplication denotes presenting the system with a probe (query) and *strictly* determining if the probe exists in the reference database (i.e., has been encountered before). This problem has recently gained considerable traction, particularly in the context of national scale ID programs [154]. In the biometric re-identification problem, the general aim is to match biometric data acquired using one camera, to data acquired from another camera in the short-term. The existing literature typically assumes a distinct reference database is available to the system when performing "re-identification" [56, 57, 58, 59]. However, in reality, when an individual enters a scene, a corresponding "track" (i.e., reference biometric data) may not exist, and the system must be able to recognize this and subsequently create a new "track". This process is analogous to the proposed anonymous identification framework (Figure 4.1).

4.1.5 Error in an Anonymous Identification System

Consider a biometric system that encounters N_T probes, denoted as $\{p_1, p_2, \ldots p_{N_T}\}$ in some (arbitrary) sequential order. In an anonymous identification system, in order to determine if an individual has been encountered before, the system assesses if the k^{th} ($k = 1, 2, \ldots, N_T$) probe is similar to any of the preceding $k-1$ probes. As with a traditional biometric system, the probability the system incurs a decision error is critical to understanding the matching accuracy. In general, two types of errors are possible: (a) an encountered probe, p_k is incorrectly matched with one of the previously encountered probes, $p_1, p_2, \ldots, p_{k-1}$ and (b) an encountered probe is incorrectly *not* matched with any of the previously encountered probes, $p_1, p_2, \ldots, p_{k-1}$. Traditionally, the respective probability of these errors is estimated through performance metrics such as FMR (False Match Rate), FNMR (False Non-match Rate), FPIR (False Positive Identification Rate), FNIR (False Negative Identification Rate) and ROC (Receiver Operating Characteristic), each of which has been well studied in the literature [9, 155, 156]. However, these measures do not completely describe the error dynamics of an anonymous identification system for two specific reasons. First, error rates for a traditional biometric system are derived from a *fixed* reference database. In other words, the identifiers associated with an identity are always assumed to be the same. This is appropriate for a traditional biometric system, as the reference database is assembled in a controlled, overt, setting. Second, in a traditional biometric system, the occurrence of an error is a static event that cannot impact future matches. In contrast, in an anonymous identification system, the reference database is dynamically evolving, as new identity profiles are created or old identity profiles are updated following each probe. As a consequence, the sequential order in which probes $p_1, p_2, \ldots, p_{k-1}$ are observed and entered into the reference database can affect the probability the k^{th} probe is incorrectly matched (or not matched). This can lead to two error scenarios. In the first scenario, probes pertaining to a single identity may be erroneously placed in different (multiple) profiles. In the second scenario, probes from different identities may be placed in a single profile. If $\mathcal{P}$ is defined to be the set of all possible permutations $\{p_1, p_2, \ldots p_{N_T}\}$ of probe orders that can be observed by the system, then two such permutations $\Pi \in \mathcal{P}$ and $\Theta \in \mathcal{P}$, can result in different error

probabilities. In Figure 4.2, an example is provided demonstrating how two different probe orders affect the manner in which identity profiles are created.

4.1.6 Chapter Motivation

The motivation for this chapter is to formally introduce, model, and analyze the performance of a biometric system operating with the anonymous identification framework. In particular, this chapter introduces and discusses the following points:

1. Formally introduce the framework and pertinent definitions of an anonymous identification system (Section 4.2.1).

2. Explicitly define decision errors in an anonymous identification system and demonstrate how these errors are different from those encountered in a traditional biometric system (Section 4.2.3).

3. Develop mathematical expressions to model errors in an anonymous identification system (Section 4.2.4).

4. Demonstrate that the sequential order in which probes are observed can have a significant impact on the probability of decision error (Section 4.2.5).

5. Validation of the error model and effect of sequential probe order through a single experiment conducted on two different sets of match scores pertaining to the face and fingerprint modalities (Section 4.3.2).

4.2 Anonymous Identification

4.2.1 Formal Definitions

An anonymous identification system consists of exactly the same architecture as a traditional biometric system. This includes components such as a matching algorithm, decision threshold and a database of reference samples.

Def 1. **Matching Algorithm** - Given two biometric samples s_A and s_B, the matching algorithm computes $S(s_A, s_B)$ and returns a similarity match score, x, between them. The similarity match score, x, is assumed to be normalized in $[0, 1]$.

Def 2. **Decision Threshold** - A pair of biometric samples, s_A and s_B, are said to match if the match score returned by the matching algorithm is above a numerical threshold γ; else, it is a non-match.

Def 3. **Reference Database** - Reference database $\mathcal{G}$, represents a local database where the encountered probes are stored. Initially, the database is a null set.

The fundamental difference between anonymous and traditional biometric systems is in the definition of identity. To highlight the difference, an identity is defined as being either unique or anonymous.

Def 4. **Unique Identity** - The true identity representing a biometric sample (e.g., this sample belongs to "Jason," or "user_123").

Def 5. **Anonymous Identity**: An identifier that is assigned to a probe by the matching algorithm. Anonymous identities are defined in the integer interval $[1, N_T]$ and the list of identifiers are stored in set I. A matched probe receives the identifier corresponding to the matching entity in the reference database. Non-matched probes receive a new identifier which is 1 more than the maximum value in I.

In this definition, it is assumed that the matching algorithm generates similarity scores and that the reference database $\mathcal{G}$ is initialized to the null set. During online operation, a biometric system will observe a set of probes in a particular order. Each observation of an individual probe is defined as an encounter.

Def 6. **Encounter** - The instance when the biometric system observes a probe. Denoted by e_k for $k = 1, 2, \ldots, N_T$ probes received.

When reference database, $\mathcal{G}$, is empty, the very first probe p_1, associated with encounter e_1 is automatically added to the reference database and assigned anonymous identity I_1. For all remaining encounters, probe p_k is matched against the contents of the reference database. A *dynamic match* with previously encountered probe p_i occurs if $S(p_k, p_i) \geq S(p_k, p_j)$ and $S(p_k, p_i) \geq \gamma$, $\forall i \neq j$, $i, j = 1, 2, \ldots k-1$. Following the match, p_k is enrolled into the reference database with matching anonymous identity I_i. Here, I_i is used to indicate the anonymous identity of probe p_i. If a match does not exist, a *dynamic non-match* occurs and

a new anonymous identity is created and added to the reference database. The algorithm describing this procedure is indicated in Alg. 4.1.

Finally, it is necessary to state the relationship between anonymous identification and cluster generation. Here, the term identity cluster is defined to refer to a particular subset of anonymous identity entries stored in set I.

> ***Def 7.*** **Identity Cluster** - Elements in I sharing a common anonymous identity number as designated by the matching algorithm. Each unique identifier represents at least one entry in $\mathcal{G}$.

Note that Alg. 4.1 represents one operational approach towards implementing an anonymous identification system. Other approaches may be adopted in the creation and matching of identity clusters (i.e., profiles) within the reference database. As with any biometric system, the method used by the matching algorithm to select the best matching reference entity (in this case, identity cluster) is a controllable parameter which can affect the performance of the system.

4.2.2 Extension to Multibiometrics

Section 4.2.1 outlined the framework of a single modality anonymous identification system. Next, that foundation is expanded upon to include multiple biometric modalities working collectively to produce a single match outcome. The motivation behind this is that a multibiometric system is less likely to generate a decision error as the number of biometric cues pertaining to an individual increases. This effect has been extensively observed in the literature [157, 158, 159].

Consider a biometric system with r modalities, wherein upon each encounter, r probes pertaining to r different modalities are observed. Thus, a random permutation of N_T probes follows $\{\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_{N_T}}\}$, where $\mathbf{p_k}$ is a vector with elements $\langle p_{k,1}, p_{k,2}, \ldots, p_{k,r} \rangle^T$ and $p_{k,i}$ is the i^{th} modality of the k^{th} probe. Now, the matching algorithm is presented with a *set* of probes at each encounter and the decision is based on the fusion of information pertaining to the r modalities.

> ***Def 8.*** **Fusion Operation** - Given probe vector $\mathbf{p_k}$, fusion operation $F(:)$ fuses the information related to $p_{k,1}$, $p_{k,2}$, ... $p_{k,r}$ and generates a single similarity match score.

Algorithm 4.1: Anonymous Identification

Input: Biometric probes $p_1, p_2, \ldots, p_{N_T}$
Output: Reference database $\mathcal{G}$ comprised of N_T probes with assigned anonymous identity numbers $I = \{I_1, I_2, \ldots, I_{N_T}\}$.
Define: $S(p_k, p_j)$ as similarity score between p_k and p_j.

Initialize:
$I_1 = 1$ \\ assign p_1 anonymous identity number 1.
Reference database $\mathcal{G} = \{(p_1, I_1)\}$ \\ the first probe is placed in the reference database.
$I_2 = I_3 = I_{N_T} = -1$ \\ probes $p_2, \ldots, p_{N_T}$ are yet to be observed.

//Begin algorithm
for $k = 2$ **to** N_T **do** \\ iterate through the rest of the probes.
  for $j = 1$ **to** $k - 1$ **do**
    \\ compare p_k with the previous set of encountered probes.
    $R(j) = S(p_k, p_j)$
    \\ compute similarity between p_k and p_j.
  if $max_j\{R(j)\}_{j=1}^{k-1} \geq \gamma$ **then**
    $I_k = I_m$ where $m = \arg max_j\{R(j)\}_{j=1}^{k-1}$
    \\ there is a match with the m^{th} reference.
  else
    $I_k = max(I) + 1$
    \\ if there is not a match, assign p_k an anonymous identity number one higher than the maximum value in I.
  $\mathcal{G} = \mathcal{G} \cup \{(p_k, I_k)\}$
    \\ add the new probe, along with its anonymous identity number to the reference database.
//End algorithm
Return $\mathcal{G}$

The fusion operation can occur at the feature level, score level, or decision level. In feature level fusion, feature vectors of probes belonging to different modalities are combined according to $F(:)$. The end result is a single probe feature vector for which the matching algorithm can compute a match score. In score level fusion, a matching algorithm is invoked for each of the r modalities. The fusion operation $F(:)$ converts the set of r scores into a single score, which is then compared against a decision threshold γ. In decision level fusion, the matching algorithm is called to report a matching identity for each of r modalities. Fusion operation $F(:)$ uses this information to determine the single best matching identity. In this work, $F(:)$ is defined to be the SUM rule for score level fusion. The SUM rule states that for r modalities, the final match score is the sum of r match scores returned by the matching algorithm. This is defined in Equation (4.1).

$$F(\mathbf{P_A}, \mathbf{P_B}) = \sum_{i=1}^{r} S(p_{A,i}, p_{B,i}) \tag{4.1}$$

4.2.3 Error Analysis

An anonymous identification system incurs error akin to traditional biometric systems. Typically, the matching performance of a traditional biometric system is evaluated through measures such as False Match Rate (FMR), False Non-match Rate (FNMR), False Positive Identification Rate (FPIR), False Negative Identification Rate (FNIR), Receiver Operating Characteristic (ROC) curves, Cumulative Match Characteristic (CMC) curves, d-prime statistic, etc. Classical CMC analysis, for example, illustrates the (closed-set) probability that when presented a probe (with a corresponding entity in the reference database) the matching algorithm will return the correct identity within N ranks (e.g., estimations from the matcher), where N is the number of unique identities in the reference database. However, CMC analysis assumes that the identifier associated with a biometric sample is always the same. That is, probe p_k is always associated with a specific subset of entities in the reference database. In the anonymous framework, this condition does not hold. Here, depending on which probes were observed prior to p_k, the actual identity pertaining to probe p_k may or may not have been encountered previously and subsequently, may not exist in the reference

database (i.e., anonymous identification is open-set). Further, if multiple true matches exist in the reference database, they may each exist in separate identity clusters as a result of error induced by the matching algorithm. As a result, decision errors and the order probes are encountered can alter the (a) composition, and (b) number of identity clusters within the reference database. Decision errors can be classified into one of two distinct types. Let N denote the number of unique identities encountered and M denote the number of anonymous identities. The first type of error occurs when probe p_k incorrectly matches to an anonymous identity I_m, $m = 1, 2, \ldots, M$. This is defined as a *false dynamic match* (FDM). As a consequence, the single identity I_m is then associated with two or more (of N) unique individuals. The second type of error occurs when probe p_k, which in fact belongs to some identity in I, is not matched with any identity in I. This error is defined as a *false dynamic non-match* (FDNM). Note by definition, a genuine (true) match for p_k must exist within the reference database for a false dynamic non-match to occur. On the other hand, a false dynamic match is not bound by this constraint. Further, a false dynamic match does not occur when a probe correctly matches to an identity cluster consisting of the true identity *in addition* to other identities.

The consequences of these errors can impact system performance in different ways. For example, a large incidence of false dynamic matches can potentially bias the matcher to repeatedly match multiple probes to the same anonymous identity in I. The extreme representation of this error occurs at a decision threshold of $\gamma = 0$, where all probe encounters are deemed to have a "match" in the reference database. Refer to Figure 4.3 for a visual representation of this error.

The result of a false dynamic non-match is different from that of a false dynamic match. Instead of multiple unique identities being represented in one identity cluster, here a single unique identity appears in several identity clusters. Identity clusters created as a result of a false dynamic non-match will typically consist of few members. Such clusters often remain small in size, as members have low similarity scores with respect to the reference database and range of candidate probes. This effect occurs as a result of a classifier decision threshold being set towards a high degree of similarity. Again, the extreme case of this error occurs at a decision threshold of $\gamma = 1.0$, where the decision outcome is a "non-match,". Figure 4.4

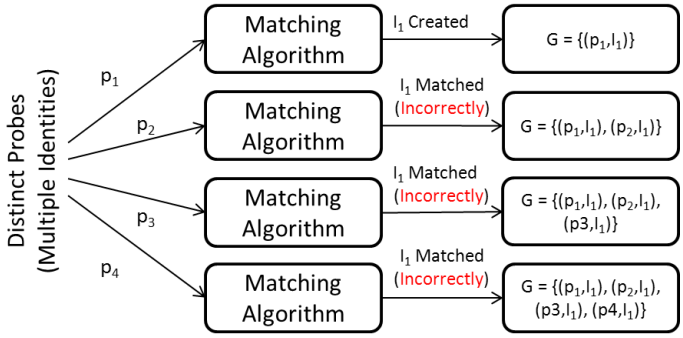

Figure 4.3: Flowchart of a false dynamic match. Here, probes belonging to multiple (unique) identities are incorrectly matched, resulting in multiple (unique) identities being merged into a single anonymous identity profile.

presents a simple flowchart illustrating false dynamic non-matches.

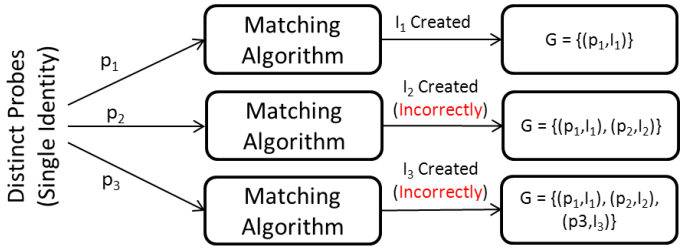

Figure 4.4: Flowchart of a false dynamic non-match. Here, probes belonging to a single (unique) identity are incorrectly not matched, resulting in a single (unique) identity appearing in multiple anonymous identity profiles.

4.2.4 Error Modeling

Although the performance of an anonymous identification system is dependent on the sequential order in which probe data is observed, prediction of expected error rates can still be accomplished. Suppose a set of N_T probes pertaining to several different identities is available and each identity is represented by multiple probes. Assuming that the probability of encountering any one of N_T probes is uniform, an analytical approach using combinatorics can be used for error prediction. In order to derive a model for error analysis, it is necessary to understand the "events" that contribute to the occurrence of a false dynamic match or false dynamic non-match. Once this is accomplished, the events can be modeled using probabilistic expressions. In deriving a model for error prediction, an assumption is made

Table 4.1: Summary of assumptions for FDMR and FDNMR estimation.

Assumption	**Description**
"Closed" Sample Space	The model assumes sample data is available for a population of identities.
Uniform Sampling	The probability of observing any sample belonging to any identity is uniform.
Matching Algorithm	The rank-1 matching identity returned from the matching algorithm corresponds to the identifier associated with the reference sample from which the maximum match score is generated.
Model Training	Match scores used to train the model can be designated as genuine or impostor.

that the best matching reference entity selected by the matching algorithm corresponds to the entity with the maximum match score. In addition, it is assumed that the match scores used to train the model can be designated as genuine or impostor. Meaning, a genuine match score represents the similarity between two biometric probes sharing the same (unique) identity, while an impostor match score represents the similarity between two biometric probes of two different (unique) identities. Finally, it is assumed that the probability of observing any one of N_T probes is uniform. A summary of these assumptions is provided in Table 4.1.

False Dynamic Match

By definition, a false dynamic match occurs when a probe is incorrectly matched to an identity cluster whose entries do not contain the true identity of the probe. This occurs if one of the following events occur.

> ***Event A***: When probe p_k (observed during encounter e_j, $j = 1, 2, \ldots, N_T$) is matched against $\mathcal{G}$, there are no genuine scores generated and at least one impostor score is greater than γ.
>
> ***Event B***: When probe p_k (observed during encounter e_j) is matched against $\mathcal{G}$, both genuine and impostor scores are generated, and there is at least one impostor score that (a) exceeds γ and (b) is greater than all the genuine scores.

Mathematically, the union of Events A and B therefore denote the probability of observing a False Dynamic Match. This is expressed in Equation (4.2). Visual examples of Events A and B are also illustrated in Figure 4.5.

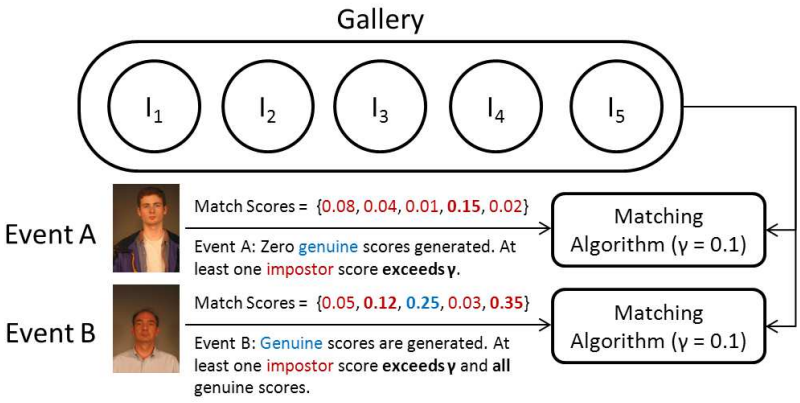

Figure 4.5: Visual example of Events A and B, where the occurrence of either event results in a false dynamic match. Note that these events denote the generation of impostor scores exceeding γ and in the case of Event B, exceeding the maximum generated genuine scores. Face images are taken from the FRGC dataset [3].

$$P(FDM|p_k, e_j) = P(A|p_k, e_j) \cup P(B|p_k, e_j) \tag{4.2}$$

Individually, the probabilities of Events A and B can be elicited through combinatorics. In this case, the goal is to objectively determine the probability that a specific probe p_k, observed at encounter e_j, has been preceded by some *combination* of $j-1$ probes from a *set* of N_T possible probes, which result in Event A or B. In the context of Event A, denote N_G as the total number of genuine probes (i.e., the number of probes having the same true identity as p_k) and N_G^γ as the total number of probes which, when matched against p_k result in a genuine score exceeding γ. Therefore, $N_T - N_G$ denotes the number of impostor probes with respect to p_k. For Event A to occur, none of the $j-1$ probes in the reference database should have the same true identity as that of p_k. The number of combinations (in this case, hypothetical reference databases) that satisfies this is denoted by $\binom{N_G}{0}\binom{N_T-N_G}{j-1}$. This number is divided by the total number of all possible combinations of $j-1$ probes, denoted by $\binom{N_T}{j-1}$, yielding the probability that a reference database of $j-1$ probes has no genuine matches. Second it is necessary to generate at least one impostor score exceeding γ. Define N_I^γ as the number of impostor probes, that when matched against p_k, generate a match score exceeding γ. The number of combinations (i.e., hypothetical reference databases) that satisfy this is denoted by $\sum \binom{N_I^\gamma}{z}\binom{N_T-N_I^\gamma}{j-z-1}$ for $z = 1, 2, \ldots, N_I^\gamma$. The summation is necessary

since it may be the case that multiple impostor probes which could result in a match with p_k could have been observed in the previous $j-1$ encounters. Again, division by $\binom{N_T}{j-1}$ yields the probability a reference database satisfying this condition occurs. Multiplication of these two probabilities yields the probability of Event A, for probe p_k, observed at the j^{th} encounter. This probability is expressed in Equation (4.3).

$$P(A|p_k,e_j)=\sum_{z=1}^{N_I^{\gamma}}\left(\frac{\binom{N_I^{\gamma}}{z}\binom{N_T-N_I^{\gamma}}{j-z-1}}{\binom{N_T}{j-1}}\right)\cdot\frac{\binom{N_G}{0}\binom{N_T-N_G}{j-1}}{\binom{N_T}{j-1}} \tag{4.3}$$

Deriving Event B is slightly more complicated, as in this case, the objective is to identify a combination of $j-1$ probes wherein at least one impostor score exceeds γ *and* any genuine scores that are generated. Here, denote $N_I^{\gamma G}$ as the number of impostor scores above both γ and the maximum genuine score. In addition, define C as a set of genuine probes (with 1 to N_G elements), representing the genuine probes which could have been observed in the previous $j-1$ encounters. For example, suppose there are two probes, p_α and p_β that share the same true identity as p_k (i.e., $N_G=2$). For a genuine score to be generated at the j^{th} encounter, either (a) p_α was observed, (b) p_β was observed, or (c) p_α and p_β were observed. Therefore, C is defined as $\{p_\alpha\}$,$\{p_\beta\}$,$\{p_\alpha,p_\beta\}$. Finally, define C_ℓ as the number of elements in a particular realization of C ($C_\ell=1,1,2$) in the aforementioned example. The number of combinations (databases) satisfying the presence of an impostor score exceeding γ and the maximum genuine score is given by $\sum\binom{N_I^{\gamma G}}{z}\binom{N_T-N_I^{\gamma G}}{j-z-1}$ for $z=1,2,\ldots,N_I^{\gamma}$. This term is multiplied by $\sum\binom{N_T-N_G}{j-C_\ell-1}$, the number of combinations enabling C_ℓ genuine scores to be generated, for all possible realizations of C. Again, division by $\binom{N_T}{j-1}$ converts the number of combinations for each term into probabilities and multiplication of the two terms denotes the probability of event B. This is expressed in Equation (4.4).

$$P(B|p_k,e_j)=\sum_{\forall C}\sum_{z=1}^{N_I^{\gamma G}}\frac{\binom{N_I^{\gamma G}}{z}\binom{N_T-N_I^{\gamma G}}{j-z-1}}{\binom{N_T}{j-1}}\cdot\frac{\binom{N_T-N_G}{j-C_\ell-1}}{\binom{N_T}{j-1}} \tag{4.4}$$

False Dynamic Non-Match

Conversely, a false dynamic non-match occurs when a probe does not match to a genuine reference and all impostor probes which potentially could match have not been observed as

yet. This can be described by the simultaneous occurrence of the following events.

> ***Event C***: When p_k (observed during encounter e_j) is matched against $\mathcal{G}$, all genuine scores generated are below γ.
>
> ***Event D***: When p_k (observed during encounter e_j) is matched against $\mathcal{G}$, all impostor scores generated are below γ.

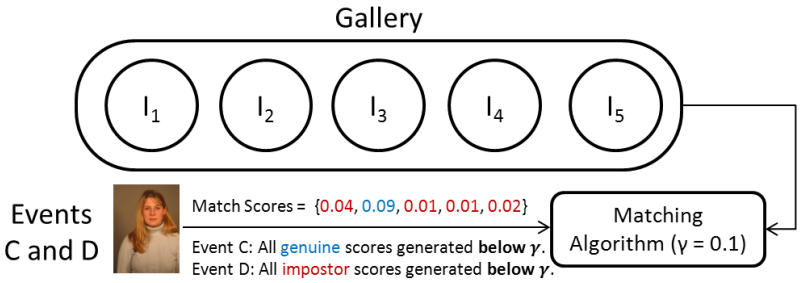

Figure 4.6: Visual example of Events C and D, which, when occurring simultaneously, results in a false dynamic non-match. Note that Event C denotes the instance when *all* genuine scores are less than γ and Event D denotes the instance when *all* impostor scores are less than γ. Face images are taken from the FRGC dataset [3].

Thus, the probability of observing a false dynamic non-match for probe p_k at e_j is dependent on the intersection of Events C and D, and is given by Equation (4.5). A visual example of these events is also described in Figure 4.6.

$$P(FDNM|p_k, e_j) = P(C \cap D|, p_k, e_j) \tag{4.5}$$

Event C represents the first condition for a false dynamic non-match by choosing a reference database whose entries do *not* produce a genuine match score greater than γ. To describe this, first denote N_G^γ as the number of genuine probes, which when matched with p_k, result in a match score exceeding γ. Let N_G retain its previous definition. Should N_G^γ be nonzero, the term $\binom{N_G - N_G^\gamma}{z}\binom{N_G^\gamma}{0}\binom{N_T - N_G}{j-z-1}$ denotes the number of reference databases that would result in the generation of z genuine scores below γ and 0 genuine scores exceeding γ, for $z = 1, 2, \ldots, N_G - N_G^\gamma$. Division by $\binom{N_T}{j-1}$ converts this number into a probability for Event C, and is given by Equation (4.6).

$$P(C|p_k, e_j) = \sum_{z=1}^{N_G - N_G^\gamma} \frac{\binom{N_G - N_G^\gamma}{z}\binom{N_G^\gamma}{0}\binom{N_T - N_G}{j-z-1}}{\binom{N_T}{j-1}} \tag{4.6}$$

Event D satisfies the second condition of a false dynamic non-match, wherein the probe does not (incorrectly) match to a reference entity originating from a different identity. This component is essential, otherwise a false dynamic match would occur, and can be described fairly easily. Recall N_I^γ is defined as the total number of probes which, when matched against p_k, result in a genuine score exceeding γ. For Event D to occur, the database must consist of a combination of probes that generate impostor scores with values less than γ. That is, the reference database contains none of the N_I^γ probes which could result in a match. Mathematically, this combination is expressed as $\binom{N_I^\gamma}{0}\binom{N_T-N_I^\gamma}{j-1}$ and the corresponding probability is given in Equation (4.7).

$$P(D|p_k, e_j) = \frac{\binom{N_I^\gamma}{0}\binom{N_T-N_I^\gamma}{j-1}}{\binom{N_T}{j-1}} \tag{4.7}$$

As is, Equations (4.2) and (4.5) define the probability of a specific probe, p_k, observing an error during encounter e_j. Computing the mean probability across all probes yields the general probability of error at e_j. Further, summation of this probability yields an estimation of observed errors for N_T encounters. Appropriate scaling establishes an expected value for each of the two rates of error, resulting in:

$$E(FDMR) = \frac{100}{N_T}\sum_{e_j}\sum_{p_k} P(FDM|p_k, e_j) \tag{4.8}$$

$$E(FDNMR) = \frac{100}{N_T}\sum_{e_j}\sum_{p_k} P(FDNM|p_k, e_j) \tag{4.9}$$

A summary of the parameters used in Equations (4.2)-(4.9) and their interpretation is provided in Table 4.2.

4.2.5 Probing for Worst-Case Error

Although the previous section states that it is possible to estimate the false dynamic match rate and false dynamic match rate, it is possible that certain permutations of $\mathcal{P}$ (probe orders) could result in widely different error rates than what might be expected. Thus, it is necessary to identify such permutations in $\mathcal{P}$ that contribute to exceptionally

Table 4.2: Summary of estimated parameters for FDMR and FDNMR estimation.

Parameter	Interpretation
N_G	Number of genuine scores per identity
N_I^{γ}	Number of impostor scores exceeding γ
$N_I^{\gamma G}$	Number of impostor scores exceeding both γ and the maximum genuine score
N_G^{γ}	Number of genuine scores exceeding γ
C	Set of hypothetical genuine scores that may exist in the reference database
C_ℓ	Number of elements in C
N_T	Number of samples to observe

poor performance. Identification of these permutations can yield an approximation of a "worst-case" estimation of FDMR and FDNMR, which could serve as a secondary measure for understanding the performance of an anonymous identification system.

A simple metric for measuring how prone a given permutation of probes might be for error is to observe the ratio of genuine to impostor scores computed for each encounter. Intuitively, a probe encounter that results in a decreased proportion of genuine scores and an increased proportion of impostor scores may correspond to an increased probability of decision error. Therefore, if probes $p_1, p_2, \ldots, p_{N_T}$ are ordered such that the observed error (Equations (4.8)-(4.9)) is abnormally high, this may be the result of encounters consistently occurring with a low genuine to impostor score ratio. To demonstrate this effect, two hypothetical permutations of probe orders which result in distinctly different ratios of genuine to impostor match scores generated are defined. First, define permutation *increment subject* (IS), which orders N unique identities with N_G probes successively (Equation (4.10)). The motivation behind "increment subject" is as each additional unique identity is introduced, every probe must be compared against an increasing number of impostor entities. Note in Equation (4.10), the subscripts n and t pertain to the n^{th} and t^{th} unique identity and genuine sample, respectively.

$$\begin{aligned} \text{IS} = \quad & \{p_{n_1 t_1}, p_{n_1 t_2}, \ldots, p_{n_1 t_{N_G}}, \\ & p_{n_2 t_1}, p_{n_2 t_2}, \ldots, p_{n_2 t_{N_G}}, \\ & \ldots, p_{n_N t_{N_G - 1}}, p_{n_N t_{N_G}}\}, \quad \text{IS} \in \mathcal{P} \end{aligned} \tag{4.10}$$

By contrast, define permutation *increment probe* (IP) as a probe order such that probes corresponding to a unique identity occur after every N^{th} encounter, i.e., the first set of N probes correspond to one sample of N different identities, the second set of N probes correspond to another sample of N different identities, and so on. This is summarized in Equation (4.11). Here, the ratio of genuine to impostor reference entities is approximately the same value for every encounter, and each of N unique identities is observed in the minimal number of encounterers.

$$\begin{aligned} \text{IP} = \quad & \{p_{n_1 t_1}, p_{n_2 t_1}, \ldots, p_{n_N t_1}, \\ & p_{n_1 t_2}, p_{n_2 t_2}, \ldots, p_{n_N t_{N_G}}\}, \quad \text{IP} \in \mathcal{P} \end{aligned} \tag{4.11}$$

These permutations are expressed visually in Figure 4.7, where $N = 75$ and $N_G = 5$. Note that for any combination of N and N_G, the genuine to impostor ratio for "increment subject" rapidly declines to values similar to, or less than "increment probe".

In addition to establishing a permutation which is favorable to observing an error, the probability of error is also impacted by the between-class variance (similarity of impostor entities) and within-class variance (similarity of genuine entities) of existing identity clusters. Identity clusters with above average between- and within-class variance are increasingly likely to result in an error. By specifically ordering probes of a test set according to the between- and within-class variance of each unique identity, "increment subject" can be further enhanced to act as an example of "worst-case" error. Here, ordering is designed by assigning a class label to each unique identity according to the framework by Doddington et al., referred to as Doddington's Zoo [7].

Doddington et al. devised a scheme for classifying users of a biometric system according to their contributions to the FMR (false match rate) and FNMR (false non-match rate).

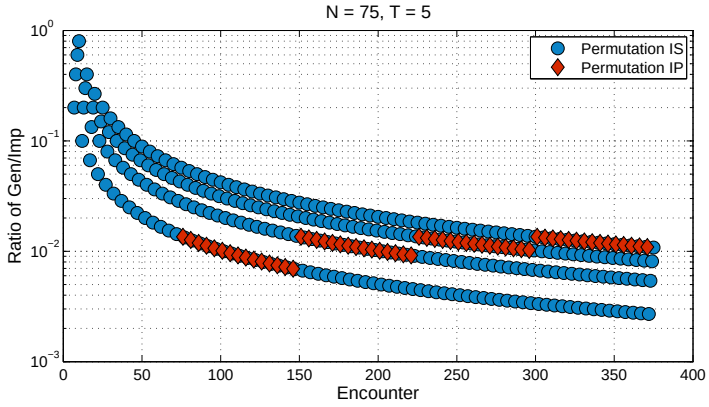

Figure 4.7: Permutations "increment subject" (IS) and "increment probe" (IP). In general, a lower ratio of genuine to imposter comparisons, increases the probability of decision error. Note that as the number of encounters increases, the ratio of genuine to impostor comparisons made for "increment subject" declines steadily. Conversely, for "increment probe", the ratio is relatively stable (i.e., similar in value).

Their classification scheme presents four separate classes: Sheep, Goats, Lambs and Wolves. Sheep are defined as ordinary users, who do not significantly contribute to adverse system performance. Goats are users who are difficult to recognize, thus contributing to the FNMR. Such users will typically have lower genuine similarity scores. By contrast, Lambs are users who are easily imitated by others. These users will commonly exhibit above average impostor similarity scores. Finally, wolves are users who are capable of imitating others. Here, an assumption is made that users are not willfully attempting to spoof the system, so the contribution of wolves is ignored. In this case, a sequence of probe encounters that may result in increased decision error would follow: Lambs, Sheep and Goats. The reasoning for this is fairly straightforward. Since goats are difficult to recognize, they are placed last, when the conditions for error are more favorable. Lambs are placed first, as they are most likely to falsely match with a goat (or even a sheep) in a future encounter.

User categorization is based on the definitions supplied by Ross et al. [160]. A user is labeled as a goat if the mean genuine score from all of their probes is below the 30^{th} percentile. Lambs are identified as users who have mean-maximum impostor scores above the 90^{th} percentile. The mean-maximum operation is defined as the average maximum impostor score for a set of probes pertaining to a single identity. When this information is used to generate a sequential order of encounter, the subset of identities that are observed first are

Lambs. Within this subset, the ordering is based on the mean-maximum impostor score generated for each identity (from highest to lowest). The next subset of identities are those identities classified as lambs. These users are also sorted based on their mean-maximum impostor score. Finally, the identities classified as goats are observed, again based on mean-maximum impostor score.

4.3 Experimental Results

4.3.1 Datasets

Experiments are conducted using similarity scores generated from the face and fingerprint subsets of the WVU Multimodal Dataset [6]. The face subset contains 5 frontal face images for each of 240 unique individuals. The fingerprint subset consists of 5 fingerprint images for each of 240 unique individuals. Fingerprints captured include the right index (R1), right middle (R2), left index (L1), and left middle (L2) fingers. In the interest of being concise, analysis is restricted to R1 scores. Match scores for face and fingerprint were obtained from the commercial software VeriLook and VeriFinger, respectively. The face and fingerprint datasets were used to create multimodal sets of scores as well, comprising a set of fused face and fingerprint (R1) scores. Fusion of scores was performed using the SUM rule, given by Equation (4.1). Scores corresponding to individual modalities were normalized between [0,1] using min-max normalization [161].

In total, 2,400 genuine and 717,000 impostor scores are generated from each test set for each modality. The aforementioned datasets were chosen as they represent commonly used biometric modalities where previous studies have demonstrated acceptable results. DET Curves for the face and fingerprint (R1) subsets are provided in Figure 4.8 along with fused face and fingerprint (R1) scores. The intent of Figure 4.8 is to provide a reference to the separability of the match scores, rather than precise performance numbers.

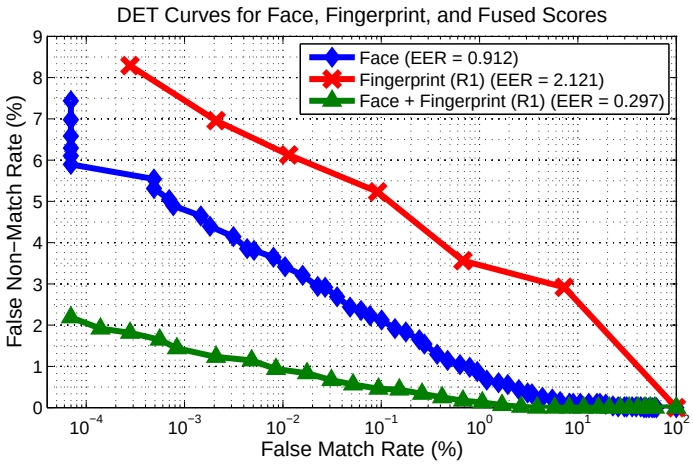

Figure 4.8: DET curves for face; fingerprint (R1); and fused face and fingerprint.

4.3.2 Experimental Protocol

Here, an experiment is presented to highlight (a) the impact of probe order on an anonymous identification system; (b) the ability of the error model presented in Section: *Error Modeling* to estimate error in an anonymous identification system (i.e., FDMR and FDNMR); and (c) the inability of traditional error metrics (e.g., FMR, FNMR, FPIR, FNIR) to appropriately measure error in an anonymous identification system. To accomplish these goals, an analysis is performed comparing the average *observed* (i.e., empirical) FDMR and FDNMR to the *expected* FDMR and FDNMR (Equations (4.8) and (4.9)). These rates are also compared against the traditional error measures of the verification (FMR and FNMR) and open-set identification (FPIR and FNIR) recognition tasks. **Note, the traditional analysis is included as a means to assess their ability to describe the error dynamics of anonymous identification**.

To distinguish between the *observed* and *expected* error rates, the match score data is divided into two random partitions of 120 identities. These partitions are denoted by "testing" and "training", respectively. To reduce the effect of selection bias, 100 partition pairs are sampled and the results from each pair are averaged together. Each "testing" and "training" partition is mutually exclusive. That is, identities in a particular "testing" partition are not in the corresponding "training" partition.

Observed values of FDMR and FDNMR are computed by implementing Alg. 4.1, while

setting γ between $(0, 1)$ in increments of 0.001. Using the actual (unique) identity of each individual in the "testing" partition as ground truth, the average observed FDNMR and FDNMR is obtained by noting the percentage of encounters where a decision error (Section 4.2.3) occurred. To demonstrate the impact of probe order, observed error rates are computed for three distinct ordering schemes. The first scheme is defined as *random draw*. In "random draw", probes are sampled at random without replacement. The second and third schemes are "increment subject" and "increment probe", as defined in Equations (4.10) and (4.11), respectively. Visual examples of these probe orders are provided in Figure 4.9. For each sampled "testing" partition, the observed FDMR and FDNMR is computed and averaged for $P = 10,000$ instances of "random draw", "increment subject", and "increment probe". Additionally, a single instance of "increment subject" is structured with specific ordering according to Doddington's Zoo assignment (as defined in Section 4.2.5) is included (denoted as Increment Subject + Zoo), which is hypothesized to be an estimate of "worst-case" error. This process for generating the observed false dynamic match rate and observed false dynamic non-match rate is summarized under the label Sub-Experiment C.

Sub-Experiment C: Obtaining Observed FDMR and FDNMR
Step 1: Sample N_T probes.
Step 2: Set γ (Between [0,1]) for normalized similarity scores.
Step 3: Implement Alg. 4.1. Maintain a record, Err_{FDM} and Err_{FDNM}, the number of false dynamic matches and false dynamic non-matches incurred for a specified γ.
Step 4: Repeat steps 1-3 P times.
Step 5: Division of E_{FDM} and E_{FDNM} by P yields the observed FDMR and FDNMR for a specified γ.

Predicted (expected) rates of FDMR and FDNMR are obtained by implementing Equations (4.2) and (4.5) on the "training" partition. This is accomplished by selecting a value for $\gamma \in (0, 1)$ and obtaining values for N_G, N_I^{γ}, $N_I^{\gamma G}$, N_G^{γ}, C, C_{ℓ}, and K for each sample in the "training" partition. Refer to Table 4.2 for a summary of these parameters and their interpretation. This enables the implementation of Equations (4.2)-(4.7). This process for generating the predicted false dynamic match rate and false dynamic non-match rate is also provided under the label Sub-Experiment D.

To provide a contrast against the observed and predicted rates of FDMR and FDNMR, a

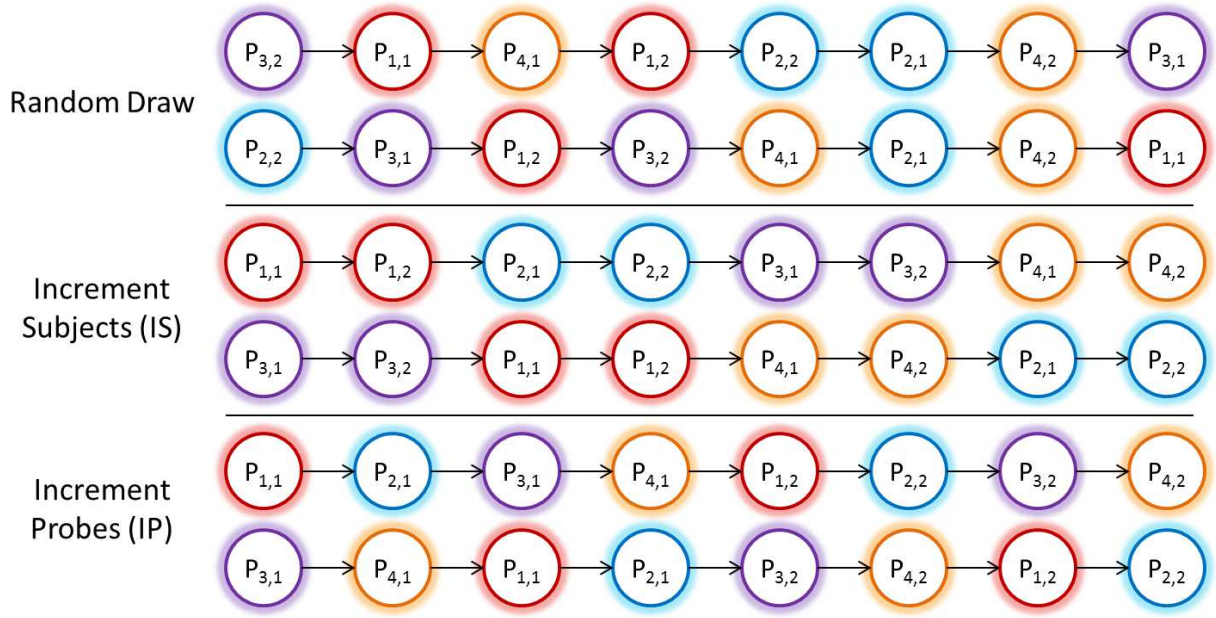

Figure 4.9: Potential sequences in which probes are observed for permutations "random draw", "increment subject" (IS), and "increment probe" (IP), where $N = 4$ and $N_G = 2$. For each permutation, the first subscript denotes the identity number and the second subscript denotes the probe number. Note the first subscript does not necessarily follow $1, 2, \ldots, N$, but rather any combination of $1, 2, \ldots, N$ (e.g., $2, 1, 3, 4$, or $3, 2, 4, 1$).

Sub-Experiment D: Obtaining Predicted FDMR and FDNMR
Step 1: Compute match scores for N identities, denoting a total of N_T probes and N_G probes per identity ($N_T = N_G \cdot N$).
Step 2: Set γ (Between [0,1]) for normalized similarity scores.
Step 3: Obtain N_I^{γ}, $N_I^{\gamma_G}$, N_G^{γ}, C, and C_ℓ
Step 4: Using the match score data, apply Equations (4.2)-(4.7) to obtain the predicted FDMR and FDNMR for a specified γ.

traditional analysis comprised of standard verification (FMR and FNMR) and identification (FPIR, FNIR) is also conducted from the "training" partition. Values for the FMR and FNMR are obtained according to the definitions supplied in Equation (1.5) and Equation (1.6) in Chapter 1, Section 1.1.4. Values for the FPIR and FNIR are obtained according to the definitions supplied in Chapter 1, Section 1.1.4.

Results of this experiment are presented in Figures 4.10-4.15 in the form of a bargraph. Each bargraph highlights values of average observed FDMR and FDNMR, expected FDMR and FDNMR, and the traditional analysis comprising FMR, FNMR, average FPIR, and average FNIR for four specific values of γ. The selected values for γ denote the approximate

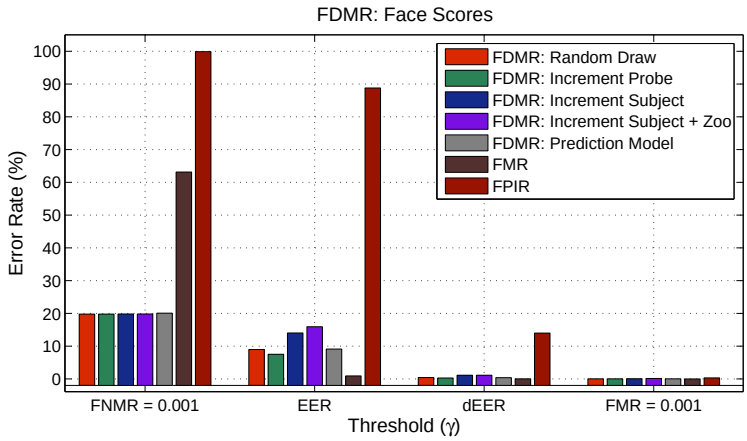

Figure 4.10: Bar graphs of observed FDMR, predicted FDMR, FMR and FPIR for face scores at selected values of γ. Note (a) the observed FDMR is different for each probe order ("random draw", "increment probe", "increment subject"); (b) the predicted FDMR is very close to the observed value; and (c) FMR and FPIR are not accurate models of anonymous identification error. To observe the differences in error rates between the proposed model and the traditional metrics for the full range of thresholds see Figure B.2 and Figure B.3 in Appendix B.1 and Appendix B.2, respectively.

values for which FNMR = 0.001, FMR = 0.001, the Equal Error Rate (EER), and what is defined as the dynamic Equal Error Rate (dEER), the threshold where FDMR is equal to FDNMR.

Supplemental experimentation demonstrating (a) observed FDMR and FDNMR for the full range of γ, (b) the variance in observed error rates within each type of probe order, and (c) further evaluation of the prediction model can be found in Appendix B.1 and B.2.

4.3.3 Discussion

Based on the experimental results, it is immediately apparent that the general shape of the curves in Figures 4.10-4.15 is similar to performance curves for traditional biometric recognition. That is, there exists a trade-off between false dynamic match rate and false dynamic non-match rate as the decision threshold is varied. However, such similarities are strictly visual. With regard to comparing FDMR and FDNMR to the error measures from traditional biometric verification (FMR and FNMR), and identification (FPIR and FNIR), Figures 4.10-4.15 show that classical metrics poorly describe the errors of an anonymous identification system. In general, FMR and FNMR are not appropriate measures as they

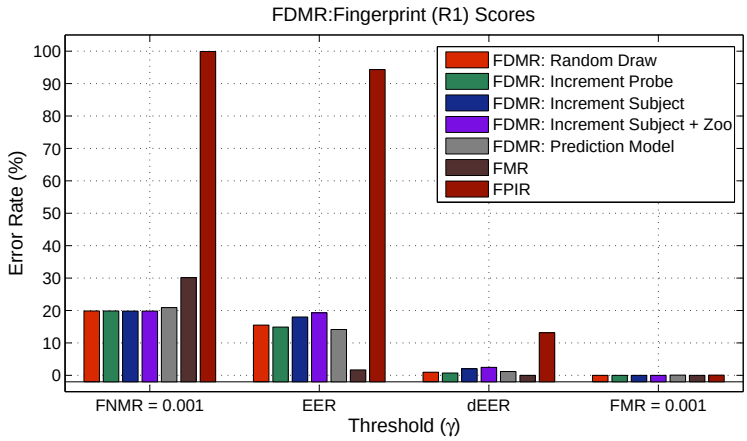

Figure 4.11: Bar graphs of observed FDMR, predicted FDMR, FMR and FPIR for fingerprint (R1) scores at selected values of γ. Note (a) the observed FDMR is different for each probe order ("random draw", "increment probe", "increment subject"); (b) the predicted FDMR is very close to the observed value; and (c) FMR and FPIR are not accurate models of anonymous identification error.

denote the general probability a *single* impostor and genuine score are incorrectly classified, respectively. Here, most matching outcomes are based on the comparison of multiple match scores. Regarding the false positive identification, one of the reasons the FPIR does not accurately describe the FDMR is that the FPIR only considers instances where the probe does not have a corresponding match in the reference database, while the FDMR is valid both when a genuine match in the reference database is and is not present. Similarly, for the false negative identification, while both the FNIR and FDNMR require a genuine match in the reference database to be present, the FDNMR also requires all generated genuine and impostor scores to be less than γ, a condition not necessary to procure a "non-match" identification error. In addition, the FPIR and FNIR assume (regardless of the number of references) that the reference elements are "correctly" labeled, a condition that cannot be presumed in an anonymous identification system. For these reasons, metrics such as FMR, FNMR, FPIR, and FNIR cannot be used to describe anonymous identification performance.

Although traditional metrics failed to describe anonymous identification performance, the proposed prediction performance model (Section 4.2.4) proved to be very good, as illustrated in Figures 4.10-4.15. In general, the model successfully predicted FDMR for the probe order "random draw", and FDNMR for each type of probe order. The model was

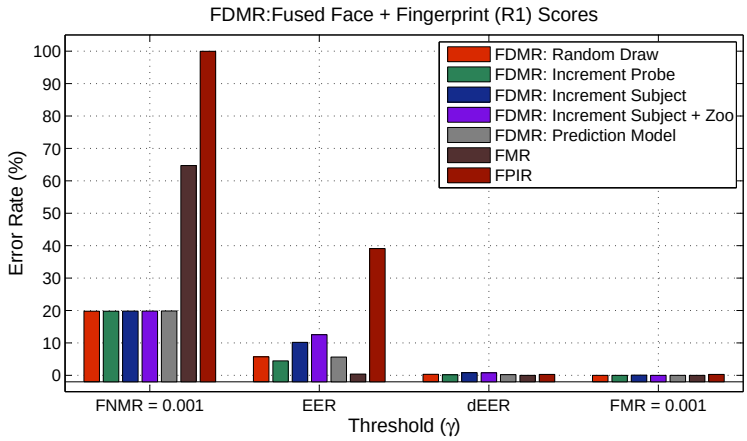

Figure 4.12: Bar graphs of observed FDMR, predicted FDMR, FMR and FPIR for fused face and fingerprint (R1) scores at selected values of γ. Note (a) the observed FDMR is different for each probe order ("random draw", "increment probe", "increment subject"); (b) the predicted FDMR is very close to the observed value; and (c) FMR and FPIR are not accurate models of anonymous identification error.

less adept at predicting FDMR for the probe orders "increment probe" (overestimate) and "increment subject" (underestimate), but was generally within 5% of the observed FDMR and substantially better than traditional metrics. Excluding the effects of uncertainty in the database [162], these results suggest that the prediction model is able to reasonably approximate error rates. To appropriately estimate the expected FDMR and FDNMR for operational data, as with classical verification or identification, a training set of reasonable size is necessary [163].

Regarding the effect of probe order on observed error, Figures 4.10-4.12 demonstrate that the probability of observing a false dynamic match can be significantly impacted by the sequential order in which probes are encountered. This is evidenced from the different values of FDMR for the probe orders: "random draw", "increment probe", and "increment subject". In Figures 4.10-4.12, these observations are the most evident when γ is set to the equal error rate (FMR = FNMR) and dynamic equal error rate (FDMR = FDNMR). As predicted in Section 4.2.5, the order "increment subject" yielded larger error rates than both "random draw" and "increment probe". Further, by explicitly structuring "increment subject" such that the identities that would be classified as "lambs" (via the Doddington's Zoo classification scheme) are observed first, the observed FDMR can be significantly increased.

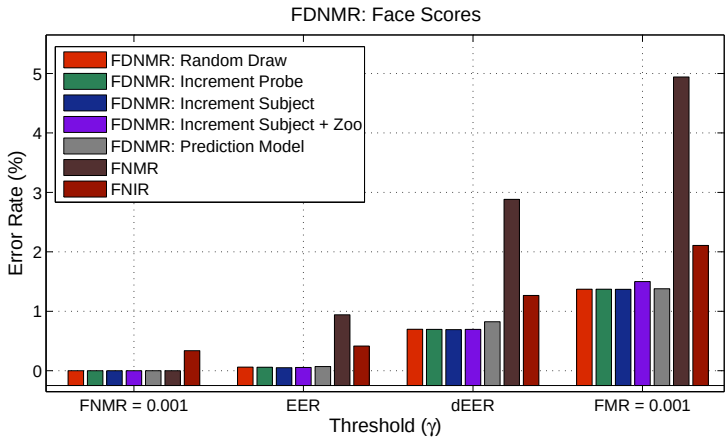

Figure 4.13: Bar graphs of observed FDNMR, predicted FDNMR, FNMR and FNIR for face scores at selected values of γ. Note (a) the predicted FDNMR is very close to the observed value; and (b) FNMR and FNIR are not accurate models of anonymous identification error. To observe the differences in error rates between the proposed model and the traditional metrics for the full range of thresholds see Figure B.2 and Figure B.3 in Appendix B.1 and Appendix B.2, respectively.

This demonstrates how the intra- and inter-class variation between identities contribute to error rates that vary as a result of probe order. Additionally, establishing probe orders this way may demonstrate a possible "worst-case" error. On the other hand, permutation "increment probe", which was designed to mitigate conditions resulting in decision error, yielded the lowest FDMR and FDNMR rates. This implies a relationship between the number of unique individuals encountered by the system in its early operating life and future performance. Interestingly, the observed FDNMR rates from Figures 4.13-4.15 for "random draw", "increment probe", and "increment subject" (randomized and Doddington-based) were approximately equal for all sets of match scores. This suggests that although FDNMR is a dynamic quality, it appears to be much less likely to be influenced by probe order. However, it may be the case that factors that increase or decrease the probability of a false dynamic non-match are not the same as factors affecting a false dynamic match, which was the primary aim in establishing the orders "increment probe" and "increment subject".

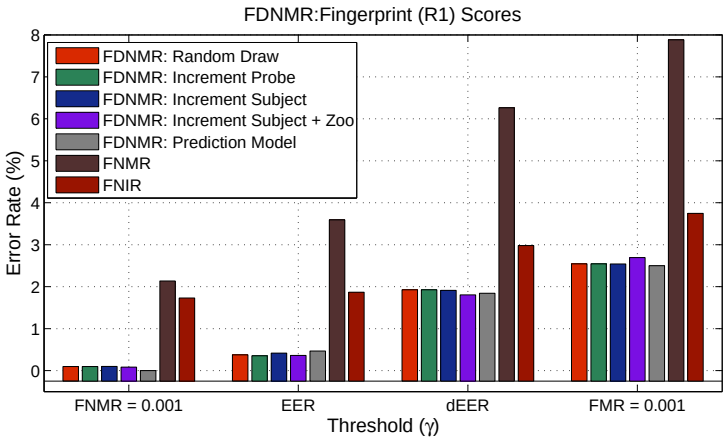

Figure 4.14: Bar graphs of observed FDNMR, predicted FDNMR, FNMR and FNIR for fingerprint (R1) scores at selected values of γ. Note (a) the predicted FDNMR is very close to the observed value; and (b) FNMR and FNIR are not accurate models of anonymous identification error.

4.4 Summary

In a traditional biometric identification system, one of the limitations that has not been addressed in the literature is how to detect whether an individual has been observed before. This is especially important for surveillance applications. Additionally, in the gait recognition literature, one limitation that has not been addressed is how a database of gait features can be dynamically assembled. This chapter introduces a variant of the traditional open-set identification system, which enables the enrollment of biometric data as it is observed and matched by the system and addresses the aforementioned limitations. Defined as *Anonymous Identification*, this approach goes beyond deducing unique identity information, or verifying a claimed identity. Rather, the system observes a probe and asserts "*Has this person been encountered before?*". Therefore, the probe is either (a) merged into an existing identity profile, or (b) emplaced within a new identity profile, depending on whether a matching entity in the reference database is found or not.

A consequence of this matching framework is that the matching outcome becomes a dynamic process, as the reference database can potentially change following each probe observation. As such, the probability of observing a decision error is now dependent on both (a) the contents of the reference database (as in the traditional open-set identification prob-

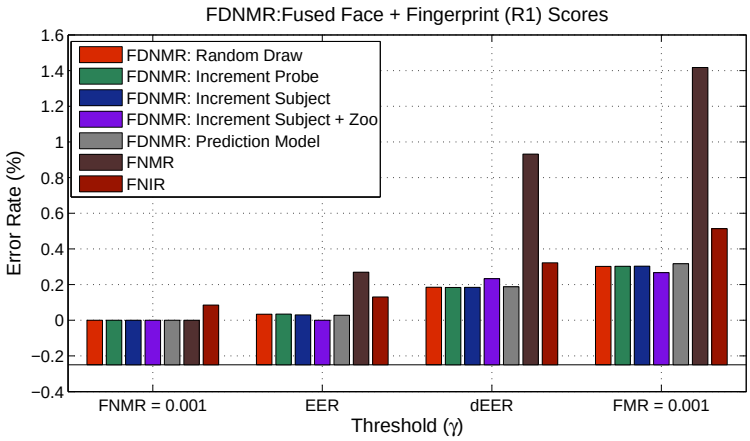

Figure 4.15: Bar graphs of observed FDNMR, predicted FDNMR, FNMR and FNIR for fused face and fingerprint (R1) scores at selected values of γ. Note (a) the predicted FDNMR is very close to the observed value; and (b) FNMR and FNIR are not accurate models of anonymous identification error.

lem) and (b) the explicit order in which prior probes were observed. The latter point is particularly notable as the identifiers assigned by the system to probes of the same identity (individual) may not be the same. Since the probability of decision error differs from that of a traditional biometric system, new terminology is introduced to define anonymous identification error. These errors are defined as either a false dynamic match (FDM), or false dynamic non-match (FNDM).

To confirm that these errors are different than those in traditional biometric recognition, an experiment is presented to demonstrate that the rates of these errors cannot be represented by traditional metrics that measure biometric performance (e.g., FMR, FNMR, FPIR, FNIR) (Section 4.3.2). Additionally, the impact of sequential probe order on anonymous identification error (FDMR and FDNMR), is verified by comparing rates of FDMR and FDNMR for three distinct "classes" of probe orders (Figure 4.9). Further, a method to organize the sequential probe order based on the Doddington's Zoo user-classification scheme is described, which leads to an estimate of "worst-case" error.

Since the FDMR and FDNMR (a) varies as a function of the observed probe order, and (b) cannot be characterized by traditional measures, a model capable of estimating FDMR and FDNMR is presented using combinatorial analysis. Experimental analysis demonstrates that the error model is a significantly better representation of the error dynamics of an

anonymous identification system, in comparison to FMR, FNMR, FPIR and FNIR (Section 4.3.2).

The anonymous identification framework, and in particular, the impact a dynamic matching framework has on error is applicable to researchers studying the re-identification and de-duplication problems. Researchers in these fields should consider the consequences of a dynamic matching process and at a minimum be cautious when reporting traditional error rates to describe the matching accuracy of algorithms in these fields.

Chapter 5

Relating the ROC and CMC Curves via the Biometric Menagerie

5.1 Introduction

5.1.1 Academic Performance Evaluations

In the academic literature, the matching accuracy of a biometric system is typically quantified through measures such as the Receiver Operating Characteristic (ROC) curve and Cumulative Match Characteristic (CMC) curve (Figure 1.2).[1] The ROC curve, measuring verification performance (Chapter 1, Section 1.1.4), is based on aggregate statistics of match scores corresponding to all identities, while the CMC curve, measuring closed-set identification performance (Chapter 1, Section 1.1.4), is based on the relative ordering of match scores corresponding to each identity.

In general, most *operational* biometric identification applications, including a biometric surveillance system operate in the open-set mode [13, 14]. However, in the *academic literature*, most performance evaluations are conducted in the closed-set mode [8, 23, 24, 38, 164]. One such reason this may be the case is that academic researchers are often limited by the amount of data available for testing, or resources available to procure data. As such, a closed-set evaluation offers maximum utility. Additionally, it enhances the reproducibility

[1] In this chapter, the terms "CMC curve" and "ROC curve" will be interchangeably used with the terms "CMC" and "ROC", respectively.

of a study, as no bias can be claimed by omitting difficult or problematic samples. Often, researchers studying distance-based biometric matching algorithms (such as human gait recognition) concentrate explicitly on the identification problem. In such studies it is common for matching performance to be reported via the CMC curve, without the associated ROC curve [92, 93, 94, 97, 98, 102, 104, 112, 121]. This is also common in studies in biometric "re-identification" [58, 59, 61]. If the data expressed in the ROC curve is an extension of the data of the CMC curve (i.e., the curves are "correlated"), then there is no issue with reporting a single curve. However, if the data in the CMC curve is not associated to a particular ROC curve (i.e., the curves are not "correlated"), then reporting only identification accuracy (via the CMC curve) may not denote a comprehensive evaluation of the matcher.

5.1.2 The Relationship Between the ROC and CMC

In Chapter 1, Section 1.1.4, it is stated that the ROC (aggregate-based) and CMC (rank-based) curves are estimated from the same set of match scores. Thus, it is not unreasonable to expect some degree of "correlation" between the two curves. This topic has received some attention in the literature, yielding mixed conclusions.

Phillips et al. [19], first developed a measure for estimating the CMC curve directly from the ROC curve. The measure was found to consistently underestimate the values of an experimentally derived CMC [5]. Later, Bolle et al., argued that the CMC is directly related to the ROC and can be used to deduce the performance of a 1:1 verification system [4]. Additionally, Bolle et al. developed a mathematical model for estimating the CMC based on the ROC when the database consists of one reference entity per identity. This measure is given in Equation (5.1). Note in Equation (5.1), K denotes the number of ranks in the CMC curve, N denotes the number of identities, $FMR(x)$ refers to the False Match Rate evaluated at a decision threshold of x, and $f_G(x)$ denotes the genuine match score distribution.

$$CMC(K) = \sum_{k=1}^{K} \binom{N-1}{k-1} \int_0^{\infty} [FMR(x)]^{k-1} f_G(x) [1 - FMR(x)]^{N-k} ds \; K = 1, 2, \ldots, N \quad (5.1)$$

Similarly, Hube also argued in favor of a direct relationship between the ROC and CMC, developing a different model for estimating the CMC from the ROC [5], also with the as-

sumption of one reference entity per identity. This measure is provided in Equation (5.2). In Equation (5.2), the variables K, N, and $FMR(:)$, retain their previous definitions, and $FNMR(:)$ refers to the False Non-match Rate evaluated at some decision threshold.

$$CMC(K) = 1 - FNMR(FMR = \frac{k}{N})\ K = 1, 2, \ldots, N \tag{5.2}$$

In the recent past, however, the notion that the ROC and CMC are directly related has been challenged. Gorodnichy first presented an argument stating that aggregate-based metrics such as the FMR, FNMR, and ROC fail to appropriately evaluate operational systems characterized by large sample size and non-static populations, or systems performing identification at a distance (e.g., systems without a controlled biometric acquisition protocol) [165, 166]. Further, Gorodnichy argues that verification systems should be evaluated (and developed) as 1:N identification systems [166], stating that measures for identification (i.e., *ranked* statistics) reveal more information regarding the relationships between users involved in a biometric system.

Based on the conclusions drawn from Bolle et al. [4], Hube [5], Gordnichy [165, 166], it is clear that support in the literature for a *direct* relationship between the ROC and CMC curves is *mixed*. In Figure 5.1, the CMC prediction models of Bolle et al. [4]. and Hube [5] are compared on two different sets of match scores generated by two different matching algorithms. The first set of match scores represents gait scores generated using the Gait Curves matching algorithm (Chapter 2, Section 2.2) on a subset of the CASIA B dataset [129]. The CASIA B dataset consists of $N = 124$ identities and $N_G = 10$ videos per identity, pertaining to "normal walk" (six videos), "with bag" (two videos) and "with coat" (two videos). Here the subset used denotes the first two samples of "normal walk".[2] The second set of match scores are fingerprint (left-index) scores from the WVU Multimodal Dataset [6]. These scores were generated using VeriFinger, a commercial fingerprint algorithm. Here, $N = 240$ and $N_G = 2$. Note that the intent of Figure 5.1 is *not* to show the performance of the matchers, but rather to analyze the ability of the two models to predict the experimentally obtained CMC curve. The data in Figure 5.1 suggests the prediction models of Bolle et al.

[2]For more information on the CASIA B dataset, refer to Chapter Chapter 2, Section 2.5.1.

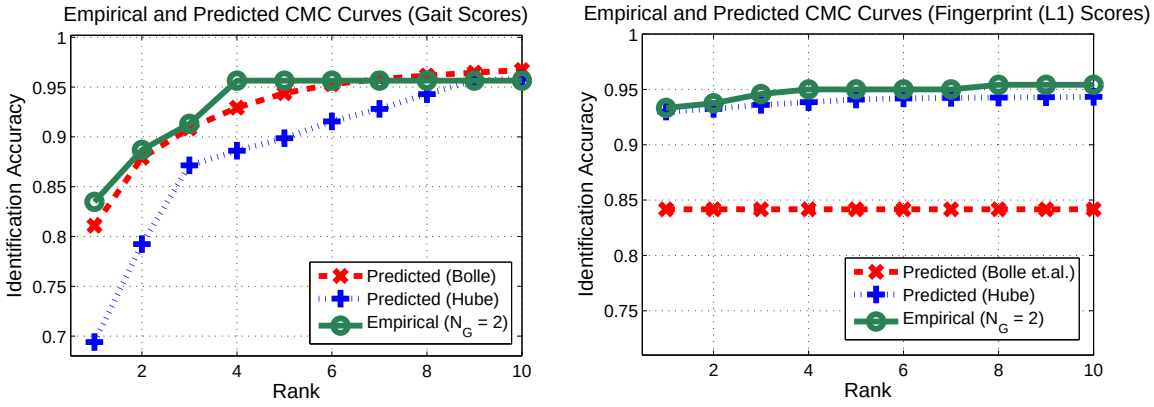

Figure 5.1: Output of the CMC prediction models (from ROC curve data) by Bolle et al. [4] and Hube [5] on match scores obtained from the Gait Curves algorithm in Chapter 2 (left), and match scores obtained from VeriFinger, a fingerprint matcher (right). Note that neither model perfectly predicts the CMC curve for both sets of match scores.

or Hube do not accurately estimate the CMC curve in all cases.

Although the data in Figure 5.1 demonstrates that there may be some degree of "correlation" between the ROC curve and CMC curve, it is clear that neither model completely predicted the empirical CMC curve based solely on the ROC data. One reason this might be the case is that aggregate-based statistics do not account for the unique manner in which different individuals contribute towards the overall performance of a biometric system. In other words, the genuine and impostor score distributions pertaining to two different individuals can be significantly different from the overall genuine and impostor match score distributions, $f_G(x)$ and $f_I(x)$. Such differences cannot be captured in aggregate statistics. Visually, this is depicted in Figure 5.2, where a subset of three individual genuine and impostor score distributions are shown using the left-index match scores from the WVU Multimodal Dataset [6]. Note that each of the three genuine and impostor distributions are distinct from one another, and that the accumulation of these subsets result in the aggregate distributions, $f_G(x)$ and $f_I(x)$.

Doddington et al. [7] first discussed the notion that different identities contribute differently towards overall biometric system performance by introducing a scheme to classify identities based on their propensity to generate a false match or false non-match error in speaker recognition [7]. This observation is referred to as the *Biometric Menagerie* in the

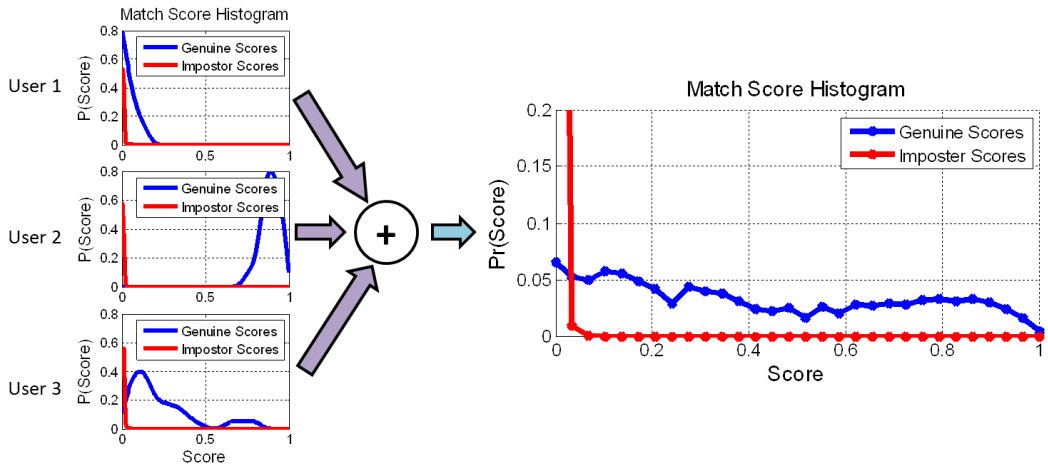

Figure 5.2: Visual example depicting the contribution of *individual identities* towards the overall genuine and impostor match score distributions, $f_G(x)$ and $f_I(x)$. Note that genuine and impostor score distributions corresponding to an identity may be *unique* (left) and the *aggregation* of these individual distributions comprises the global genuine and impostor match score distributions (right). Here, the individual match score distributions are based on fingerprint scores (L1) computed on the WVU Multimodal Dataset [6].

literature [167]. If each identity contributes to the performance of a biometric system differently, it may be possible that for a *single* pair of genuine and impostor match score distributions, *multiple* rank-based statistics (e.g., CMC curves) can be generated. Further, these differences in rank-based statistics may result in multiple CMC curves with large differences in cumulative rank-K accuracy. In general, it may be theoretically possible for a set of match scores to exhibit "good" or "poor" performances via the ROC and CMC curves, depending on how the match scores are distributed among each identity.

5.1.3 Chapter Motivation

Thus, in this chapter, the aim is to demonstrate that the ROC and CMC curves may not be directly related while also accounting for the unique per-identity statistics of individuals (i.e., the role of the Biometric Menagerie). This can be accomplished in two ways: (a) Empirically, with a sufficiently large number of match scores collected experimentally from multiple datasets and matching algorithms, or (b) analytically, via modeling the inter- and intra-class relationships in match scores, such that "faux identities" can be created from an input set of empirical match scores. This chapter focuses on the latter (e.g., modeling

per-identity statistics), as it may not be feasible to procure the amount of empirical data necessary to derive meaningful conclusions regarding the differences in aggregate-based and rank-based statistics. On the other hand, by modeling the inter- and intra-class relationships in match scores, it is possible to demonstrate that a *fixed* set of match scores can be reassigned *differently* among N identities. This reassignment of existing match scores to faux identities is accomplished by utilizing the "Doddington Zoo" user classification scheme. As such, it is possible to demonstrate that a variety of *distinct* ranked statistics (i.e., CMC curves) can be accompanied by *the same* aggregate statistics (i.e., ROC curves).

Thus, the contributions of this chapter are as follows:

- Introduce a framework for categorizing a biometric evaluation into one of four outcomes, based on the performance exhibited by the associated ROC and CMC curves (Section 5.2.1).

- Given a set of real match scores pertaining to multiple identities, a method for reassigning the scores to faux identities is described. Faux identities are created to represent varying types of intra-class and inter-class statistics, based on the Doddington Zoo phenomenon (Section 5.3).

- The validity of the proposed score reassignment model is asserted by recreating the intra- and inter-class statistics present in empirically obtained match scores (Section 5.4.2).

- Experimentally demonstrate that the score reassignment model can be used to generate a set of faux identities whereby a "good" ROC curve can be accompanied by a "poor" CMC curve (and vice-versa), while maintaining distinct per-identity statistics (Section 5.4.3).

- Experimentally validate that match scores sharing common aggregate statistics (ROC curves) can have differing ranked statistics (CMC curves) (Section 5.4.4).

5.2 Outcomes of A Performance Test

5.2.1 Performance Outcomes

Assume for the sake of argument that the matching performance depicted by the ROC curve and the CMC curve can be treated as two independent outcomes. That is, provided a hypothetical set of match scores, derived from some genuine and impostor score distribution, suppose it is possible to generate an ROC curve exhibiting "good" or "poor" performance and a CMC curve exhibiting "good" or "poor" performance, and vice-versa. The basis for this assumption is based on the data in Figure 5.1, where neither of the CMC prediction models were able to correctly predict the empirical CMC from ROC data. Although this is not necessarily evidence of statistical independence, it is reasonable to suggest that the two curves are not absolutely correlated. Therefore, if the information expressed within the ROC and CMC curves can be summarized as being "good" or "poor", the following outcomes may occur:

Good Verification Good Identification (GVGI): A performance test is classified as good verification good identification (GVGI) when the properties of the ROC and CMC indicate excellent performance. Here, the system is adequately able to perform both verification and identification tasks. Such an outcome is perhaps the most desirable for a biometric system.

Good Verification Poor Identification (GVPI): A performance test is classified as good verification poor identification (GVPI) when the properties of the ROC indicate good performance, while the CMC demonstrates poor identification accuracy. Such a system is adept at verification tasks, but is generally unreliable at performing identification.

Poor Verification Good Identification (PVGI): A performance test is classified as poor verification good identification (PVGI) when the properties of the ROC indicate poor performance, while the CMC demonstrates good identification accuracy. Such a system is capable of performing identification, but not verification.

Poor Verification Poor Identification (PVPI): A performance test is classified as poor verification poor identification (PVPI) when the properties of both the ROC and CMC indicate poor performance. Such a system is not capable of performing verification or identification tasks adequately.

5.3 Modeling Match Score Relationships

In an operational biometric system, or in empirically collected biometric data, it is likely that each identity contributes uniquely towards the overall performance of the system [7, 160]. Evidence of this can be found in empirically derived match scores (Figure 5.2). In this section, a model is developed for characterizing inter- and intra-class relationships between match scores. The model functions by assigning match scores from to faux identities, whereby the per-identity statistics for each faux identity (i.e., the local level) are distinct from the statistics of the population (i.e., the global level). As such, the manner in which match scores are distributed among each faux identity can impact the manner in which they are sorted, which in turn can affect the outcome expressed by the CMC curve.

5.3.1 Model Framework

As previously stated, the model for characterizing match score relationships functions by assigning match scores to faux identities. Here, a *faux identity* is defined as an identity, n, whose individual genuine and impostor match score distributions, $f_G^n(x)$ and $f_I^n(x)$, have been sampled (without replacement) from $\mathbf{x}_{Gen}$ and $\mathbf{x}_{Imp}$. Note that $\mathbf{x}_{Gen}$ and $\mathbf{x}_{Imp}$ denote sets of genuine and impostor scores generated by a biometric matcher on a dataset of N *real* identities. For example, $\mathbf{x}_{Gen}$ and $\mathbf{x}_{Imp}$ may be the fingerprint match scores illustrated in the bottom of Figure 5.2. Thus, faux identities can be created from *real* and empirically obtained match score data. By creating a set of faux identities, it is possible to generate *multiple* sets of N faux identities, with each set sharing the same aggregate $f_G(x)$ and $f_I(x)$ but with differing rank statistics.

In defining each faux identity, an assumption is made that the range of genuine and impostor scores for each faux identity is smaller than the range of the overall distributions, $f_G(x)$ and $f_I(x)$. In other words, if a genuine match score distribution, $f_G(x)$, is nonzero in the interval $[\alpha, \beta]$, the nonzero range of an individual genuine match score distribution, $f_G^n(x)$, is in $[\alpha_n, \beta_n]$ where $(\beta_n - \alpha_n) < (\beta - \alpha)$. The "tightness" of $[\alpha_n, \beta_n]$ can be defined by the variance in match scores on a per-identity basis. Define these per-identity variances as σ_{n-n}^2 and σ_{n-m}, where σ_{n-n} denotes the average variance in genuine scores for each identity

Table 5.1: Summary of assumptions for FDMR and FDNMR estimation.

Assumption	**Description**
Per-identity Match Score Distributions	The range of values for per-identity genuine and match scores is smaller than the range of genuine and impostor scores for all identities. Mathematically, this is described by: $max(f_G^n(x)) - min(f_G^n(x)) \ll max(f_G(x)) - min(f_G(x))$ and $max(f_I^n(x)) - min(f_I^n(x)) \ll max(f_I(x)) - min(f_I(x)) \; \forall n$.
Model Integrity	Exceptions to the above condition are permitted when no match scores can be assigned to a faux identity within its initial range.

and σ_{n-m} denotes the average variance in impostor scores for each pair of identities. Note that the intent of this assumption is to ensure created faux identities *do not* share the same individual genuine and impostor match score distribution as the aggregate genuine and impostor score distributions. In addition, this assumption allows for a more plausible representation of inter- and intra-class relationships in match scores (in contrast to sampling match scores randomly). A summary of these assumptions is provided in Table 5.1.

The output following the creation of each faux identity is $\mathcal{S}$, which denotes a table of size N_TxN_T, wherein each column (or row) of $\mathcal{S}$ contains match score information for one "faux" biometric sample, matched against $N_G - 1$ samples from the same "faux" identity and $N_T - N_G$ samples from the remaining $N - 1$ "faux" identities. Note that this exercise preserves the aggregate score statistics; what changes is the set of match scores pertaining to every identity.

5.3.2 Modeling Inter- and Intra-class Variations

The model for reassigning match scores to faux identities is inspired by the "Doddington's Zoo" user-classification scheme, which characterizes identities based on their contribution towards the FMR and FNMR [7]. The Doddington's Zoo classification scheme consists of four classes: Sheep, Goats, Lambs, and Wolves. Sheep are defined as "well behaved" individuals who are easily recognized and do not incorrectly match with others. Goats are individuals who are intrinsically difficult to recognize and contribute to false non-match errors. Lambs are individuals whose biometric data can often be confused with other identities, resulting in false match errors. Finally, wolves are defined as individuals who willfully and successfully

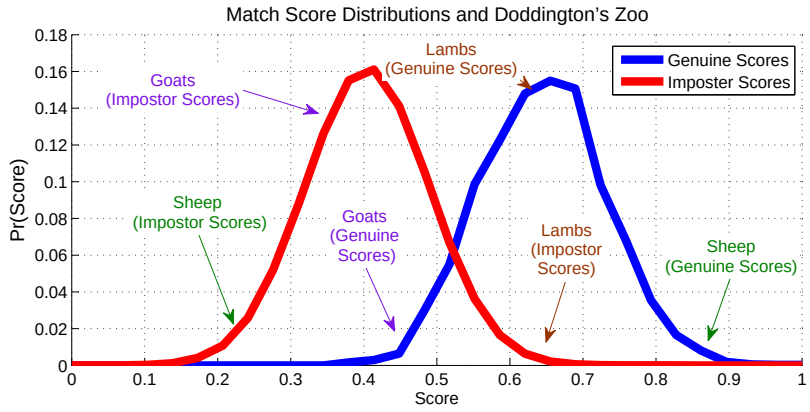

Figure 5.3: Visual illustrating the general concept of the proposed model for defining inter- and intra- class relationships in match scores, which creates faux identities based on the "Doddington's Zoo" framework [7].

spoof the biometric data of other individuals, increasing the rate of false match errors. Studies have also demonstrated that a single identity can share multiple classes, an example being an individual exhibiting both goat and lamb characteristics [160].

In terms of match scores, sheep can be loosely characterized as having "high" genuine scores and "low" impostor scores. Meanwhile, goats can be loosely characterized as having "low" genuine scores. Finally, lambs (and wolves) can be loosely characterized as having "high" impostor scores. These simple characterizations formulate the basis of the model for reassigning scores to faux identities, and is visually depicted in Figure 5.3.

The score reassignment model consists of two stages: initialization and sampling. During initialization, each of N identities are assigned a label, χ_n $(n = 1, 2, \ldots, N)$, $\chi_n \in \{Sheep, Goat, Lamb\}$. The number of faux identities corresponding to each label is pre-specified (see Section 5.4). Next, each identity is *assigned* match scores (from the original score set) based on the properties of a "Sheep", "Goat", or "Lamb". Sampled match scores are drawn (without replacement) from the original scores $\mathbf{x}_{Gen}$ and $\mathbf{x}_{Imp}$, and stored in $\hat{\mathbf{x}}^n_{Gen}$ and $\hat{\mathbf{x}}^n_{Imp}$, which are the reassigned genuine and impostor scores for the n^{th} faux identity. Finally, a table of match scores of size $N_T \text{x} N_T$ is created (denoted by $\mathcal{S}$). Each row in $\mathcal{S}$ stores the $N_G - 1$ assigned genuine scores and $N_T - N_G$ assigned impostor scores for each sample of a given faux identity.

Assignment of genuine scores to each faux identity is a relatively straightforward process.

For each faux identity, $\binom{N_G}{2}$ genuine scores are drawn without replacement[3] from $\mathbf{x}_{Gen}$ and stored in $\mathcal{S}$. Depending on the label of the faux identity, a target range from which scores will be sampled, is first defined. This range is assumed to be between $[\mu_{Gen} + \sigma_{Gen}, 1]$, $[0, \mu_{Gen} - \sigma_{Gen}]$, and $[0, \mu_{Gen} + \sigma_{Gen}]$ for "Sheep", "Goats", and "Lambs", respectively. Denote the subset of genuine scores within this range as $\mathbf{x}_{rng}$. If $\mathbf{x}_{rng}$ is a null set, the target range is opened (i.e., increased) by scaling the lower and upper bounds of $\mathbf{x}_{rng}$ by a factor of δ $(0 < \delta < 1.0)$ until $\mathbf{x}_{rng}$ contains at least one element. Next, one element (i.e., score) from $\mathbf{x}_{rng}$ is sampled and stored in $\mathcal{S}$. Denote the *value* of this score as x. The remaining $\binom{N_G}{2} - 1$ scores are sampled from the range $x \pm \epsilon_{Gen}$, where ϵ_{Gen} is a scaling parameter. As with the range used to sample x, if no match scores are found within $x \pm \epsilon_{Gen}$, the range is opened by scaling ϵ_{Gen} by δ. This process for sampling genuine scores is summarized in Alg. 5.1. Note that this sampling method ensures that (a) sampled genuine scores for each identity are consistent, and (b) the genuine scores for a "Sheep" are distinct from those of a "Goat", and a "Lamb" (when possible).

Assignment of impostor scores to each faux identity captures the inter-class relationships between identities. As such, assignment of impostor scores is viewed as being between pairs of identities (and therefore labels), rather than for a single identity. This results in one of six possible scenarios, viz. "Sheep-Sheep", "Sheep-Goat", "Sheep-Lamb", "Goat-Goat", "Goat-Lamb", and "Lamb-Lamb".

When sampling impostor scores between a pair of identities, ${N_G}^2$ impostor scores are sampled from $\mathbf{x}_{Imp}$, of which a single score, x, is first drawn from a target range, $\mathbf{x}_{rng}$. $\mathbf{x}_{rng}$ is dependent on the labels denoting the pair of identities. Denote $\hat{\mathbf{x}}^n_{Gen}$ and $\hat{\mathbf{x}}^m_{Gen}$ as the set of *assigned* genuine scores for the n^{th} and m^{th} identities (i.e., the genuine scores assigned following implementation of Alg. 5.1). When both of the labels are a "Sheep" or "Goat", $\mathbf{x}_{rng}$ is limited to $[0, min\{max\{\hat{\mathbf{x}}^n_{Gen}\}, max\{\hat{\mathbf{x}}^m_{Gen}\}\})$, the *minimum* of the maximum genuine score observed for both identities. This constraint attempts to ensure that sampled impostor scores for a "Sheep" or a "Goat" will always be less than their corresponding genuine scores, preventing the occurrence of a false match error.

When one of the labels is a "Lamb", the only constraint emplaced is that $\mathbf{x}_{rng}$ is below

[3] equiprobable sampling

Algorithm 5.1: Reassigning Genuine Scores

Input: Vector $\mathbf{x}_{Gen}$, containing the genuine scores.
Table $\mathcal{S}$, where sampled genuine scores are stored.
Vector χ, a set containing the labels of each identity (e.g., "Sheep", "Goat", "Lamb").
Define: δ, ϵ_{Gen}: Scaling parameters.
Output: Table $\mathcal{S}$ populated with genuine scores.
\\ *begin algorithm*
Step 1: For each identity, note the assigned label.
Step 2a: Draw a genuine score (without replacement), x, $\mathbf{x}_{Gen}$, from within subset $\mathbf{x}_{rng}$, where
$\mathbf{x}_{rng} = [\mu_{Gen} + \sigma_{Gen}, 1]$, if $\chi_n = Sheep$.
$\mathbf{x}_{rng} = [0, \mu_{Gen} - \sigma_{Gen}]$, if $\chi_n = Goat$.
$\mathbf{x}_{rng} = [0, \mu_{Gen} + \sigma_{Gen}]$, if $\chi_n = Lamb$.
Step 2b: If $\mathbf{x}_{rng}$ is a null set, and $\mathbf{x}_{rng} = [a, b]$, set $a = \delta \cdot a$, $b = \frac{b}{\delta}$ and repeat Step 2a.
Step 3a: Draw $\binom{N_G}{2} - 1$ scores (without replacement) from $\mathbf{x}_{Gen}$ within $x \pm \epsilon_{Gen}$.
Step 3b: If less than $\binom{N_G}{2} - 1$ scores can be drawn set $\epsilon_{Gen} = \frac{\epsilon_{Gen}}{\delta}$ and repeat Step 3a.
Step 4: Store the sampled genuine scores in $\mathcal{S}$.
return $\mathcal{S}$
\\ *end algorithm*

the maximum genuine score for the paired identity. That is, if $\chi_n = Lamb$, and $\chi_m = Sheep$, $\mathbf{x}_{rng} = [0, max\{\hat{\mathbf{x}}^m_{Gen}\})$. When $max\{\hat{\mathbf{x}}^m_{Gen}\} > max\{\hat{\mathbf{x}}^n_{Gen}\}$, this enables (but does not guarantee) the possibility of drawing an impostor score which can generate a false match (at rank-1) for the identity denoted as a "Lamb", but not the "Sheep". Should $\chi_n = \chi_m = Lamb$, no constraints are emplaced on $\mathbf{x}_{rng}$, enabling (but not guaranteeing) the possibility of a false match (at rank-1) to occur for both faux identities.

As with the sampling of genuine scores, if $\mathbf{x}_{rng}$ is a null set, $\mathbf{x}_{rng}$ is opened fully to $[0, 1]$. Once a valid range of $\mathbf{x}_{rng}$ is identified, one impostor score is drawn from $\mathbf{x}_{Imp}$ and stored in $\mathcal{S}$. The remaining ${N_G}^2 - 1$ impostor scores are sampled from a range of $x \pm \epsilon_{Imp}$, where ϵ_{Imp} is a scaling parameter. If no match scores are found within $x \pm \epsilon_{Imp}$, the range is opened by scaling ϵ_{Imp} by δ. This process for drawing impostor scores is summarized in Alg. 5.2. Note that this sampling method ensures that (a) the impostor scores between pairs of identities are consistent, and (b) the error dynamics for a "Sheep", "Goat", and "Lamb" are upheld

Algorithm 5.2: Reassigning Impostor Scores

Input: Vector $\mathbf{x}_{Imp}$, containing the impostor scores.
Table $\mathcal{S}$, where sampled genuine scores are stored (from Alg. 5.1) and sampled impostor scores will be stored.
Vector χ, containing the labels of each identity (e.g., "Sheep", "Goat", "Lamb").
$\hat{\mathbf{x}}^n_{Gen}$, $\hat{\mathbf{x}}^m_{Gen}$, *Assigned* genuine scores for identities n, m.
Define: δ, ϵ_{Imp}: Scaling parameters.
Output: Table $\mathcal{S}$ populated with genuine and impostor scores.
\\ *begin algorithm*
Step 1: For all combinations of n and m ($n = 1, \ldots, N$, $m = n + 1, \ldots, N$), note χ_n and χ_m.
Step 2: Draw an impostor score, x from $\mathbf{x}_{Imp}$, within interval $\mathbf{x}_{rng}$, where
$\mathbf{x}_{rng} = [0, min\{max\{\hat{\mathbf{x}}^n_{Gen}\}, max\{\hat{\mathbf{x}}^m_{Gen}\}\})$, if $\chi_n = Sheep$ or $Goat, \chi_m = Sheep$ or $Goat$.
$\mathbf{x}_{rng} = [0, max\{\hat{\mathbf{x}}^n_{Gen}\})$, if $\chi_n = Sheep$ or $Goat, \chi_m = Lamb$.
$\mathbf{x}_{rng} = [0, max\{\hat{\mathbf{x}}^m_{Gen}\})$, if $\chi_n = Lamb, \chi_m = Sheep$ or $Goat$.
$\mathbf{x}_{rng} = [0, 1]$, if $\chi_n = \chi_m = Lamb$.
Step 3: If $\mathbf{x}_{rng}$ is a null set, $\mathbf{x}_{rng} = [0, 1]$.
Step 4a: Draw ${N_G}^2 - 1$ scores from $\mathbf{x}_{Imp}$ within $x \pm \epsilon_{Imp}$.
Step 4b: If less than ${N_G}^2 - 1$ scores can be drawn set $\epsilon_{Imp} = \frac{\epsilon_{Imp}}{\delta}$, and repeat Step 4a.
Step 5: Store the sampled impostor scores in $\mathcal{S}$.
return $\mathcal{S}$
\\ *end algorithm*

(when possible).

5.4 Experimental Results

5.4.1 Datasets and Experimental Design

Experiments are conducted to demonstrate the following:

- Demonstrate the model for reassigning match scores (Section 5.3.2) is able to create *viable* representations of how match scores could be distributed among identities. This is accomplished by using the model to "recreate" the per-identity statistics of empirical match scores.

- Using the model for reassigning match scores, demonstrate that it is at least theoretically possible to observe the performance outcomes: GVGI, GVPI, and PVGI (Section 5.2.1) using synthetic match scores.

- Using the model for reassigning match scores, explore whether empirically obtained match scores denoting a GVGI outcome, could otherwise be reassigned to generate a GVPI outcome. Similarly, explore whether a PVGI outcome could be reassigned to generate a PVPI outcome.

Experiments in this chapter are computed using match scores pertaining to the face and gait modalities, as well as a set of synthetic match scores. Face scores were extracted from the WVU Multimodal Dataset [6] using the commercial software VeriLook, and are the same match scores used in Chapter 4. Recall from Chapter 4, Section 4.3.1, the face subset of the WVU Multimodal Dataset consists of $N_G = 5$ frontal face images for $N = 240$ unique individuals. Gait match scores were collected using the Gait Curves algorithm (Chapter 2, Section 2.2) on the CASIA B Dataset [129]. The CASIA B dataset (previously described in Chapter 2, Section 2.5.1) is a multi-camera dataset for human gait recognition, containing $N = 124$ individuals walking "normally" (six sequences), "with a coat" (two sequences), and "with a backpack" (two sequences) from 11 different viewpoints. For the purposes of this experiment, only the instances where an individual is walking "normally" and whose viewpoint is perpendicular to the optical axis of the camera are considered (i.e., $N_G = 6$, individual is viewed as traversing the horizontal axis of the image plane).

Synthetic match scores are sampled from a parametric normal distribution with parameters μ_{Gen}, σ^2_{Gen}, μ_{Imp}, σ^2_{Imp}, and normalized between [0,1]. Note, the assumption that match scores can be sampled from a normal distribution is made *strictly* to define a means for which hypothetical distributions of match scores can be understood. In general, the probability distributions of $f_G(x)$ and $f_I(x)$ can take on any function and can vary on an algorithmic basis [168]. **Although the usage of synthetic match scores may not be ideal, the usage of such data is done *strictly* to support a *theoretical* analysis regarding the relationship between the ROC and CMC curves**. To be consistent with the empirical data, synthetic scores denote $N = 240$ identities with $N_G = 5$ samples per identity.

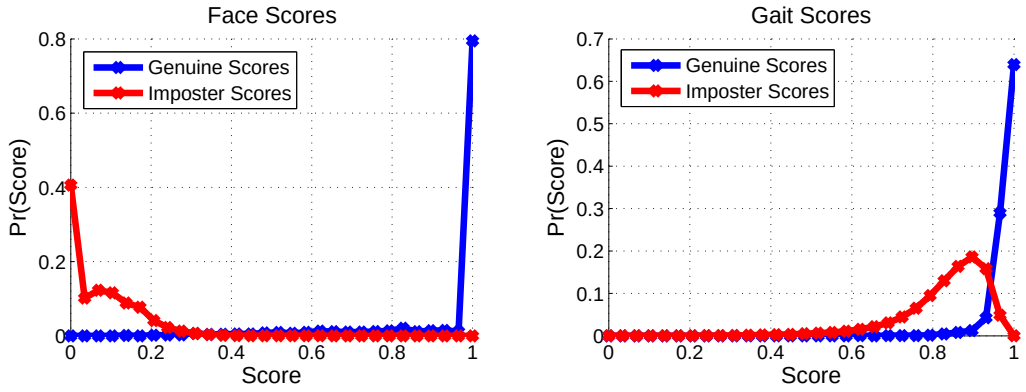

Figure 5.4: Genuine and impostor score distributions, $f_G(x)$ and $f_I(x)$, for the face and gait scores used in this evaluation.

For this analysis, *aggregate* statistics (i.e., the performance expressed in the ROC curve) are expressed by the area underneath the ROC curve (denoted by AUC). *Rank* statistics (i.e., the performance expressed in the CMC curve) are expressed via the Weighted Rank-M identification accuracy, which is a weighted sum of the identification accuracies corresponding to the first M ranks in the CMC curve. Here, M is defined as 5% of the number of identities, N. The weight of the i^{th} rank, w_i, $i = 1, 2, \ldots, M$, is defined by $\frac{1}{i}$, and normalized such that $\|\mathbf{w}\|_2 = 1$. A weighted strategy is chosen such that performance can be quantified relative to N.

In Figure 5.4, a visualization of $f_G(x)$ and $f_I(x)$ is presented for the face and gait scores. A baseline evaluation consisting of AUC, weighted rank-M, predicted weighted rank-M (via the models of Bolle et al. [4] and Hube [5]) and the empirically obtained proportions of "Sheep", "Goats", and "Lambs" is provided in Table 5.2. The strategy used to obtain empirical proportions of "Sheep", "Goats", and "Lambs" is the same as the one defined by Ross et al. [160].[4] Note that in Figure 5.4 and Table 5.2, the performance values for both modalities are similar, but the genuine and impostor distributions, $f_G(x)$ and $f_I(x)$, for the gait scores share a larger range of nonzero values.

[4]It should be noted that this scheme will always classify at least 30% of identities as having properties of a "Goat", or "Lamb", regardless of whether these identities contribute to adverse recognition performance.

Table 5.2: Baseline AUC, Weighted Rank-M, estimated Weighted Rank-M, and the empirically obtained proportion of "Sheep", "Goats", and "Lambs" for the face and gait datasets.

Measure	**Face Scores**	**Gait Scores**
AUC (Empirical)	0.999	0.980
Weighted Rank-M (Empirical)	1.0	0.978
Weighted Rank-M (Bolle et al.[4])	0.991	0.895
Weighted Rank-M (Hube [5])	0.991	0.878
Proportion of {Sheep, Goat, Lamb} (%) (Ross et al. [160])	{62, 28, 10}	{66, 24, 10}

5.4.2 Model Viability

Provided the basis of the match score reassignment model (Section 5.3.2) is to create *plausible* representations of the intra- and inter-class match scores. As such, it is necessary to verify whether the score reassignment model can accomplish this. Arguably, one criteria for success is whether the model can *recreate* or mimic the inter- and intra-class relationships present in empirically collected match scores.

To evaluate this, the model (i.e., score reassignment process) is used to recreate the inter- and intra-class relationships of the face and gait scores defined in Section 5.4.1. When creating faux identities, the proportion of identities labeled as "Sheep", "Goats", and "Lambs" is set to {90%, 05%, 05%}, for face scores and {75%, 10%, 15%} gait scores, respectively. This information is used to generate χ_n.[5] The parameters δ, ϵ_{Gen}, and ϵ_{Imp} are set to 0.98, $3.2\sigma_{Gen}$, and $1.7\sigma_{Imp}$, for face scores and 0.98, $3.5\sigma_{Gen}$, and $0.77\sigma_{Imp}$ for gait scores.

For both the actual (empirical) and created faux identities, the mean variance in match scores within a single identity (σ^2_{n-n}) and between pairs of identities (σ^2_{n-m}) is computed. If the respective variances are the same for both the original data and the recreated data, it is reasonable to conclude that the relationships have been successfully recreated. Since the created faux identities are generated from the match scores of actual identities, the aggregate statistics (i.e., ROC curves) will be equal.

In addition, it is necessary to demonstrate the score reassignment process is capable of producing consistent outputs (i.e., the same input parameters produce a similar output). As

[5]The basis for choosing these proportions is strictly towards the design of the faux identities and should not be interpreted as the ratio of "Sheep", "Goats", and "Lambs" identified in the respective datasets via a match score classification scheme.

Table 5.3: Evaluating the viability of the reassigning model on the face scores in the WVU Multimodal Dataset.

Measure	**Original Face Scores**	**Reassigned Scores**
AUC	0.999	0.999
Weighted Rank-M	1.0	0.999
σ^2_{Gen}	2.1e-2	2.1e-2
Mean σ^2_{n-n}	1.3e-2	1.5e-2
Rejects Null Hypothesis	N/A	Yes
p-value	N/A	0
σ^2_{Imp}	6.2e-3	6.2e-3
Mean σ^2_{n-m}	3.9e-3	3.8e-3
Rejects Null Hypothesis	N/A	Yes
p-value	N/A	0

such, the score reassignment process is repeated 250 times, and the average value of σ^2_{n-n} and σ^2_{n-m} is computed. Further, a T-test is performed to evaluate whether the output the score reassignment process generates values of σ^2_{n-n} and σ^2_{n-m} in some distribution whose mean is the actual values of σ^2_{n-n} and σ^2_{n-m}, respectively. This denotes a rejection of the null hypothesis. Also computed is the corresponding p value, which denotes the probability the observed output of the faux data is outside the actual (empirical) values of σ^2_{n-n} and σ^2_{n-m}. These results are tabulated in Tables 5.3 and 5.4 for face and gait scores, respectively. Note that the data for AUC, Rank-M, σ^2_{n-n}, and σ^2_{n-m}, are approximately the same in both columns (empirical and faux). In addition, note that the result of the T-test suggests a rejection of the null hypothesis and that the output of the score reassignment process (at the stated inputs), results in an output whose expected values of σ^2_{n-n}, and σ^2_{n-m} denote the actual values obtained empirically.

5.4.3 Evaluating Theoretical Performance Outcomes

In this experiment, an evaluation is performed to ascertain whether the outcomes defined in Section 5.2.1 (e.g., GVGI, GVPI, PVGI) can, on a theoretical level, occur on match score data that could denote empirical match score data. This experiment assumes (in some sense) that the score reassignment process is capable of generating plausible representations of empirical match scores (via Tables 5.3 and 5.4 in the prior experiment) in comparison to

Table 5.4: Evaluating the viability of the reassigning model on the gait scores in the CASIA B Dataset.

Measure	**Original Gait Scores**	**Reassigned Scores**
AUC	0.980	0.980
Weighted Rank-M	0.978	0.994
σ^2_{Gen}	3.8e-4	3.8e-4
Mean σ^2_{n-n}	4.2e-4	4.4e-4
Rejects Null Hypothesis	N/A	Yes
p-value	N/A	0
σ^2_{Imp}	4.4e-3	4.4e-3
Mean σ^2_{n-m}	1.7e-3	1.9e-3
Rejects Null Hypothesis	N/A	Yes
p-value	N/A	0

random assignment of scores to faux identities. With this assumption, the score reassignment process can be used to either validate or deny whether it is theoretically possible to observe a set of match scores whose respective ROC and CMC curves denote a GVGI, GVPI, or PVGI outcome.

To facilitate this, faux identities are created from synthetically generated match scores. Synthetic match scores are sampled from a parametric normal distribution with parameters μ_{Gen}, σ^2_{Gen}, μ_{Imp}, and σ^2_{Imp} (as defined in Section 5.4.1).

For the purposes of this experiment, the performance expressed by the ROC curve (e.g., verification performance) is defined to be "good" if the AUC is above 98% and "poor" if the AUC is below 75%. The performance expressed by the CMC curve (e.g., identification performance) is then defined as "good" if the weighted rank-M identification accuracy is greater than 90% and "poor" if the rank-M accuracy is below 50%. Note this definition neglects the situation where the AUC or rank-M accuracy is between the ranges specified as "good" or "poor". This two tiered threshold is necessary to suggest a "poor" outcome is sufficiently poor and not "almost good" (or vice-versa). A summary of these outcomes is provided in Table 5.5.

Here, the parameters used to generate the genuine and impostor match score distributions, in conjunction with the score reassignment parameters used, are as follows:

- GVGI: $N = 240$, $N_G = 5$, $\epsilon_{Gen} = 0.25\sigma_{Gen}$, $\epsilon_{Imp} = 0.25\sigma_{Imp}$, $\delta = 0.98$, $\mu_{Gen} = 0.500$,

Table 5.5: Range of AUC (row) and rank-M (column) identification rate resulting in a PVPI, PVGI, GVPI and GVGI outcome. Outcomes outside these definitions are denoted by "***".

AUC / Rank-M	**0.00-0.50**	**0.50-0.90**	**0.90-1.00**
0.00-0.75	PVPI	***	PVGI
0.75-0.98	***	***	***
0.98-1.00	GVPI	***	GVGI

$\sigma^2_{Gen} = 3.4e-3$, $\mu_{Imp} = 0.120$, $\sigma^2_{Imp} = 0.011$. Proportion of "Sheep", "Goat", and "Lamb" = {96% ,2% ,2%}, respectively.

- GVPI: $N = 240$, $N_G = 5$, $\epsilon_{Gen} = 0.25\sigma_{Gen}$, $\epsilon_{Imp} = 0.25\sigma_{Imp}$, $\delta = 0.98$, $\mu_{Gen} = 0.500$, $\sigma^2_{Gen} = 3.4e-3$, $\mu_{Imp} = 0.120$, $\sigma^2_{Imp} = 2.42e-2$. Proportion of "Sheep", "Goat", and "Lamb" = {15% ,35% ,50%}, respectively.

- GVPI: $N = 240$, $N_G = 5$, $\epsilon_{Gen} = 0.9\sigma_{Gen}$, $\epsilon_{Imp} = 0.25\sigma_{Imp}$, $\delta = 0.98$, $\mu_{Gen} = 0.500$, $\sigma^2_{Gen} = 0.190$, $\mu_{Imp} = 0.230$, $\sigma^2_{Imp} = 6.60e-3$. Proportion of "Sheep", "Goat", and "Lamb" = {96%, 2%, 2%}, respectively.

Results are presented in Figures 5.5-5.7, which depict (a) a visual of the genuine and impostor score distribution and the associated AUC value, and (b) a scatter of the maximum genuine score and impostor score assigned to each "sample" of each faux identity when match scores are assigned with, and without regard to inter- and intra-class variations, along with the corresponding weighted rank-M value. Visualization in this way aids in depicting the rank statistics of match scores.[6] The data in Figures 5.5-5.7 illustrate that each of the stated performance outcomes are theoretically possible, even when intra- and inter-class match score statistics are considered.

5.4.4 Evaluating Empirical Score Distributions

Whereas the previous experiment utilized synthetic match scores to justify the theoretical existence of a GVGI, GVPI, or PVGI outcome, in this experiment, the score reassignment model for creating faux identities is implemented on *empirical* match score distributions to

[6]For a biometric sample, when its maximum impostor score exceeds its maximum genuine score, then a rank-1 identification error will occur.

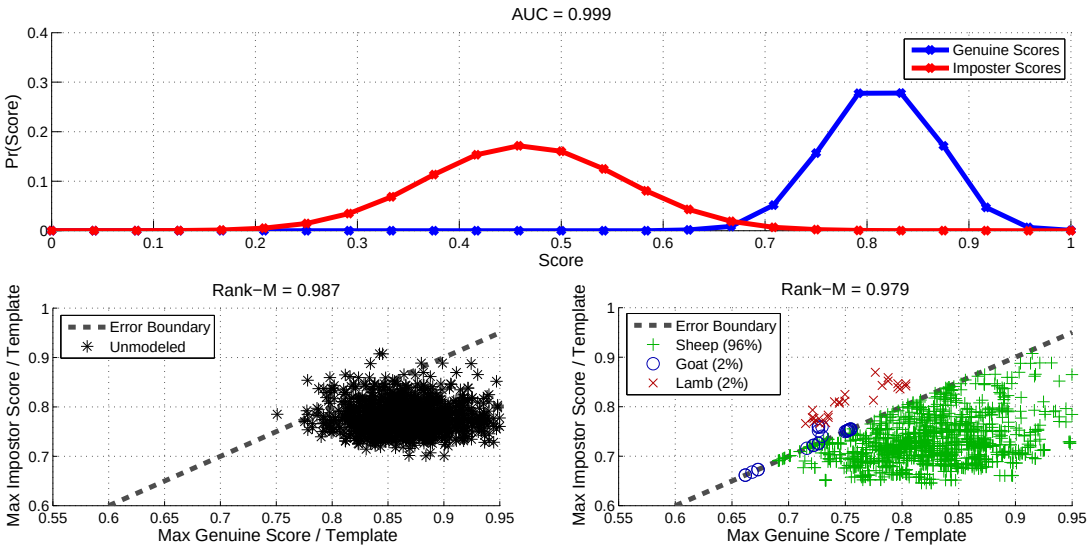

Figure 5.5: Example of a synthesized GVGI result (AUC > 0.98, Rank-$M > 0.90$), where intra- and inter-class relationships are not considered (left) and modeled (right). Note that the model is able to reproduce the intended result (i.e., a high Rank-M accuracy).

Table 5.6: AUC and Weighted Rank-M values after match score reassignment for different proportions of "Sheep", "Goats", and "Lambs" using face scores. Note that in this case, the weighted rank-M accuracy does not change much.

Sheep (%)	**Goats** (%)	**Lambs** (%)	**AUC**	**Weighted Rank-M**
100	0	0	0.999	1.0
82	10	8	0.999	1.0
50	26	24	0.999	0.997
15	10	75	0.999	0.997

generate *alternative* realizations of intra- and inter-class relationships from the same set of scores. The intent of this experiment is to demonstrate that two sets of match scores sharing the same aggregate statistics (e.g., CMC curves) can result in different ranked statistics (i.e., CMC curves) on empirical data. To enable this, the model is run with multiple proportions of "Sheep", "Goats", and "Lambs". Parameters for δ, ϵ_{Gen}, and ϵ_{Imp}, are set to 0.98, $0.25\sigma_{Gen}$, and $0.25\sigma_{Imp}$, respectively, for both face and gait scores. These results are tabulated in Tables 5.6 and 5.7.

In addition, one proportion of "Sheep", "Goats", and "Lambs" that might result in a *decrease* in rank-M performance is highlighted. Ideally, the decrease would be significant enough such that both GVGI and GVPI outcomes could be observed from the same match

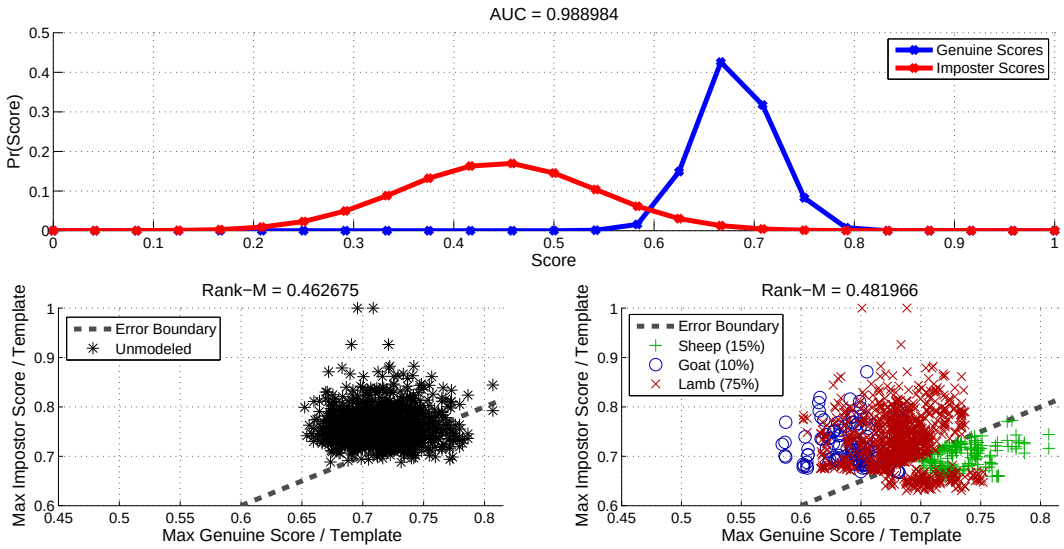

Figure 5.6: Example of a synthesized GVPI result (AUC > 0.98, Rank-$M < 0.50$), where intra- and inter-class relationships are not considered (left) and modeled (right). Note that the model is able to reproduce the intended result (i.e., a low Rank-M accuracy).

Table 5.7: AUC and Weighted Rank-M values after match score reassignment for different proportions of "Sheep", "Goats", and "Lambs" using gait scores. Note that in this case, the rank-M accuracy changes significantly.

Sheep (%)	**Goats** (%)	**Lambs** (%)	**AUC**	**Weighted Rank-M**
100	0	0	0.980	1.0
82	10	8	0.980	0.966
50	26	24	0.980	0.915
15	10	75	0.980	0.800

score data. This is accomplished by reducing the number of "Sheep" or "well-behaved" faux identities in χ_n. The highlighted proportions for face and gait scores are {50%, 26%, 24%} and {15%, 10%, 75%} for "Sheep", "Goats", and "Lambs", respectively and are illustrated visually in Figures 5.8 and 5.9, which, similar to Figures 5.5-5.7 plot the maximum impostor score against the maximum genuine score for each biometric sample for the empirical and reassigned face and gait scores. In addition, Figure 5.10 illustrates the actual ROC and CMC curves generated from the empirical and reassigned match score data for both face and gait scores.

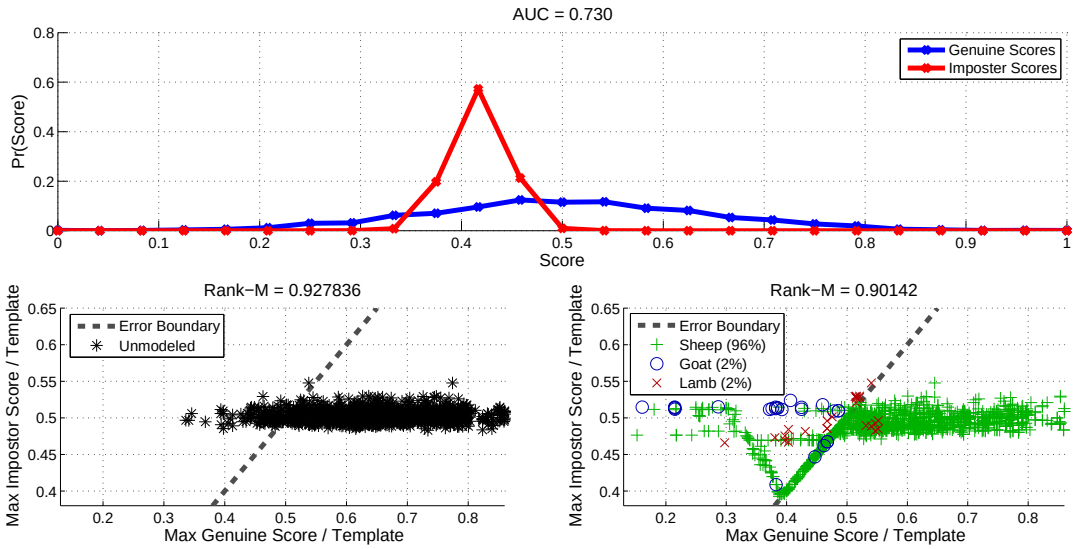

Figure 5.7: Example of a synthesized PVGI result (AUC < 0.75, Rank-$M > 0.90$), where intra- and inter-class relationships are not considered (left) and modeled (right). Note that the model is able to reproduce the intended result (i.e., a high Rank-M accuracy).

5.4.5 Discussion

The first experiment (Section 5.4.2) demonstrates that the proposed score reassignment model is able to generate "viable" representations of intra- and inter-class relationships between identities via their match scores. This is important, as in order to effectively model the per-identity statistics in match scores, it must be demonstrated that such a model is creating meaningful relationships. Further, such a model must demonstrate that the per-identity statistics generated are not equal to the aggregate genuine and impostor score variances. In addition, a measure of "viability" is important to show that the analysis may be relevant to real-world problems. Viability is confirmed via the data in Tables 5.3 and 5.4, as the weighted rank-M accuracy and average intra-class variance per identity (σ^2_{n-n}), and inter-class variance between pairs of identities (σ^2_{n-m}) are approximately equal to the original face and gait scores (with equal values of AUC). In addition, Figures 5.8 and 5.9 demonstrate the model is behaving as stated (Section 5.3.2), since "Goats" and "Lambs" are seen contributing to rank-based error by having at least one of the following properties: (a) low genuine scores and (b) high impostor scores.

In the second experiment (Section 5.4.3), the score reassignment model for creating faux

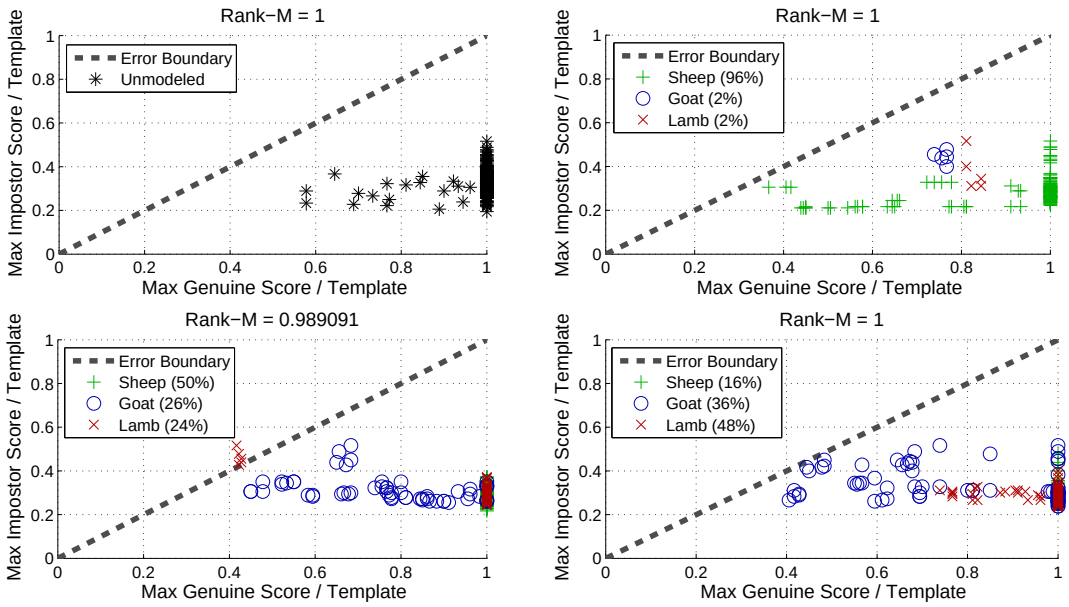

Figure 5.8: Comparing weighted rank-M accuracies before (above) and after (below) the score reassignment process for the face dataset. Note that here, although it is possible to generate a different realization of ranked match scores, the resulting rank-M accuracy does not significantly vary (1 and 0.989091).

identities is implemented on a set of synthetic match scores, in order to evaluate whether on a theoretical level, if it is possible to observe an ROC curve and CMC curve whose respective performances correspond to one of the three "interesting" outcomes defined in Section 5.2.1 (e.g., GVGI, GVPI, PVGI). The results demonstrate that it is theoretically possible to generate each of these outcomes. The primary limitation of this experiment is that it is conducted on synthetic data. However, in this case, the benefit of synthetic data is that it allows for the rapid generation of a large number of hypothetical match score distribution, which is beneficial for a theoretical analysis.

The third experiment (Section 5.4.4) serves to address the primary limitation of the second, by invoking the score reassignment model on empirical data, in order to query whether differing ranked-based statistics (CMC curves) can be observed from match score distributions with similar aggregate-based statistics (ROC curves) and if so, whether the differences can be significant. In Figure 5.8 (involving face scores), when varying the proportion of "Sheep", "Goats", and "Lambs", while a difference in ranked statistics can be observed, the

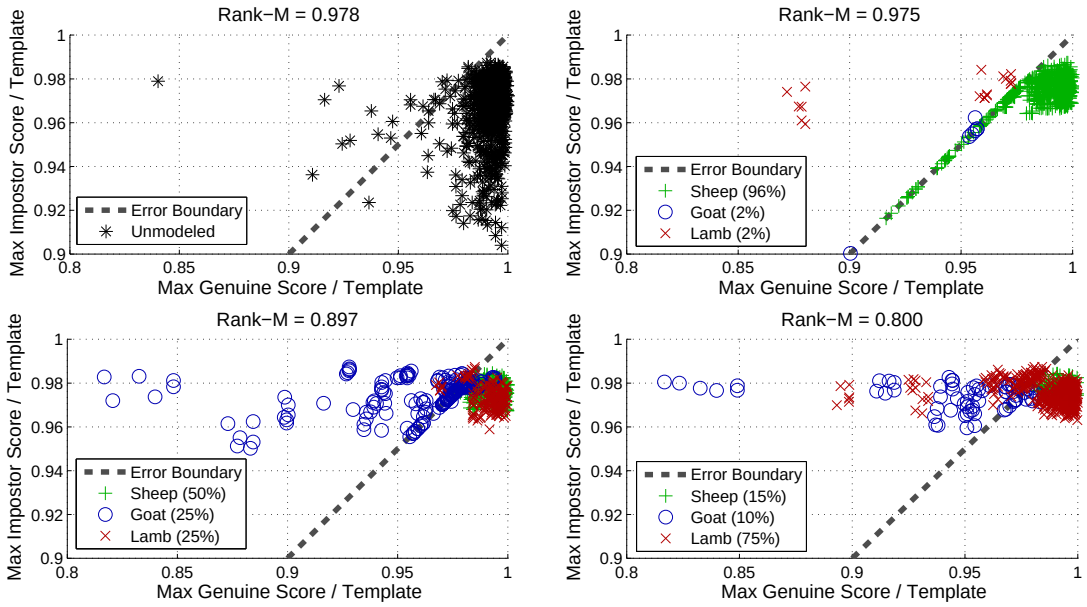

Figure 5.9: Comparing weighted rank-M accuracies before (above) and after (below) the score reassignment process for the gait dataset. Note that here, it is possible to generate a different realization of ranked match scores with a significantly different weighted rank-M accuracy (0.978 and 0.8). This suggests that multiple CMC curves can be accompanied with the same ROC curve.

resulting weighted rank-M accuracy was not significantly different than that of the original data nor the predicted values (Table 5.2). However, in Figure 5.9 (involving gait scores), while varying the proportion of "Sheep", "Goats", and "Lambs", a realization with a significantly lower weighted rank-M accuracy (weighted rank-M $= 0.8$) was discovered, which is enough to categorize the reassigned outcome as a GVPI. These observations are also evident in the CMC curves from Figure 5.10.

The overarching question(s) then become: Why was this phenomena observed with the gait scores and not the face scores? What is unique to a set of match scores that enables the possibility of a GVPI outcome (as in Figure 5.6 and Figure 5.9) or a PVGI outcome (as in Figure 5.7). The answer has to do with the extent of overlap between $f_G(x)$ and $f_I(x)$ (i.e., the range of x for which both $f_G(x)$ and $f_I(x)$ are non-zero). If $f_G(x)$ and $f_I(x)$ have less overlap, although match scores can be arranged differently between identities, this is unlikely to change the ordered ranking of match scores in the CMC curve. However, if $f_G(x)$ and $f_I(x)$ are reasonably overlapped, then **it cannot be guaranteed that aggregate-based**

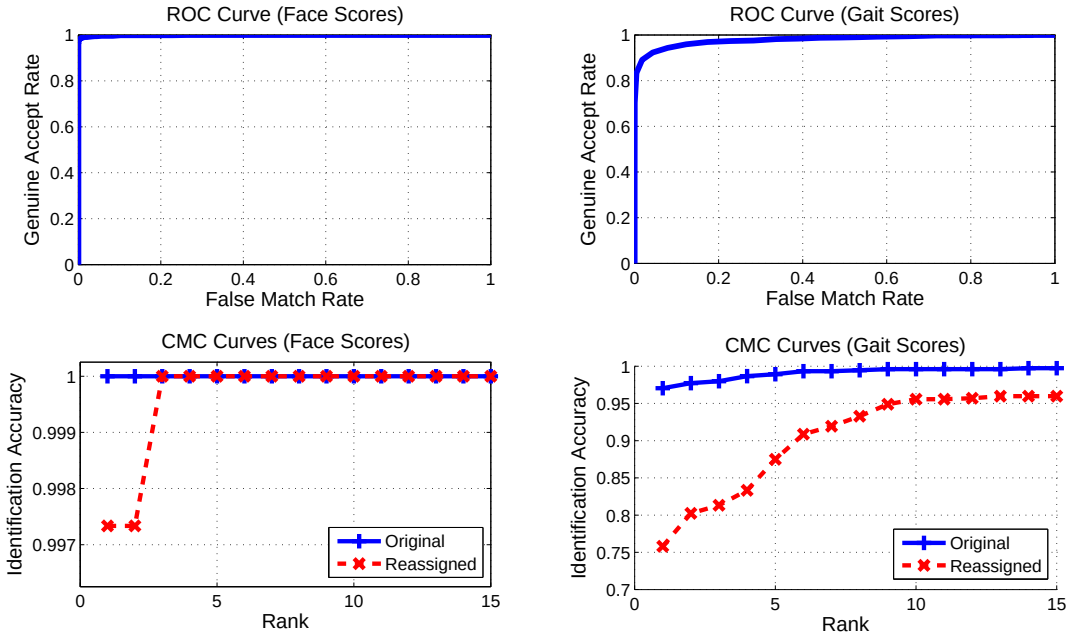

Figure 5.10: ROC and CMC curves for the original and reassigned face (left) and gait (right) match scores. Note that for both sets of match scores, the ROC data is the same, while the CMC data is different for the original and reassigned scores.

statistics will correlate with rank-based statistics. This property becomes particularly important when biometric systems increase in scale, as $f_G(x)$ and $f_I(x)$ cannot be certain to conform to any distribution [168], and also in unconstrained biometric systems, which may yield larger intra- and inter-class variances as a consequence of uncontrolled biometric acquisition. Thus, researchers in the biometric community should report the ROC curve if a CMC curve is used as a measure of reporting matching performance. Similarly, the ROC curve should not be used as a stand-alone performance measure of a biometric matcher operating in identification mode. This recommendation is especially important for academic performance evaluations, as CMC curves are more likely to be reported.

5.5 Summary

In this chapter, an analysis was performed regarding the relationship between the ROC curve, which traditionally denotes verification performance, and the CMC curve, which traditionally denotes identification performance (in particular, in academic performance eval-

uations). Our analysis discusses that although there are models to predict the CMC from the ROC data [4, 5], such models may not always be accurate as the data in the ROC curve is based on the match scores generated from *every* identity (i.e aggregate-based), while the CMC data is based on the match scores on a per-identity basis (i.e., rank-based). As such, it is possible that (a) an ROC indicating "good" performance can be accompanied by a CMC curve indicating "poor" performance (and vice-versa), and (b) a single ROC curve can be accompanied by *multiple* CMC curves, where depending on the extent of overlap between the genuine and impostor match score distributions, the per-identity statistics expressed in the CMC curve vary such that large differences of performance can be observed (i.e., both "good" and "poor" CMC performance).

To facilitate this, terminology mapping the performance of the ROC curve and CMC curve to one of four possible outcomes is developed. These terms include: GVGI (Good Verification Good Identification), GVPI (Good Verification Poor Identification), PVGI (Poor Verification Good Identification), and PVPI (Poor Verification Poor Identification).

Next, a model for characterizing the inter- and intra-class relationships found in match scores (e.g., the Biometric Menagerie) is defined. The model is used to generate *faux identities* from an input set of match scores, where the created faux identities can be defined to have differing per-identity statistics (e.g., CMC data) while sharing *the same* aggregate statistics (e.g., ROC data) as the input.

The ability of the model to generate "plausible" representations of match scores is evaluated by recreating the per-identity statistics of empirically collected face and gait scores. Next, the model is used on synthetic data to theoretically validate the occurrence of a GVGI, GVPI, and PVGI outcome. Finally, the model is again implemented on empirically collected face and gait scores to probe whether the match scores can be re-distributed such that a large variance in CMC data can be observed from the same ROC data.

The results of this study suggest that aggregate-based statistics (i.e., the ROC curve) may not be directly related to rank-based statistics (i.e., the CMC curve). In particular, a single ROC curve can be associated with multiple CMC curves. Consequently, when reporting the CMC curve as an indication of identification performance (as is common in academic evaluations), the ROC curve should also be presented, in order to have a more comprehensive

understanding of the matching performance.

Chapter 6

De-duplication Error in Biometric Systems

6.1 Introduction

6.1.1 Identity Duplication in Biometric Systems

In a biometric system, it is possible for a single individual to be associated with multiple identities or labels. This is referred to as identity *duplication*.[1] In a dynamic enrollment framework, such as in the Anonymous Identification framework (Chapter 4), this may occur as a result of decision error by the system, whereby *multiple* identity profiles are generated containing the same identity. In an overt (traditional) enrollment framework, a duplicate identity may be created by a malicious individual who intends to derive multiple benefits from the system (e.g., a welfare disbursement system). Alternatively, a duplication may be a result of unintentional oversight by the system administrator during enrollment.

The process of detecting and managing duplicate entries associated with a single individual is referred to as *de-duplication*. The de-duplication task occurs during the enrollment phase of a biometric system, wherein the input biometric sample is compared against the previously enrolled data by a biometric matcher in order to determine if a duplicate reference entity exists. If a duplicate entity is found, then the current input sample is flagged by the

[1]This phenomenon is independent of whether the system is operating with a traditional enrollment, in the Anonymous Identification framework (Chapter 4), overtly, or covertly.

system.[2] In the simplest case, the input biometric data is not stored in the system. If no duplicates exist, then the input biometric sample is associated with a new label (i.e., identity profile) and stored in the system. A simple illustration of de-duplication is given in Figure 6.1.

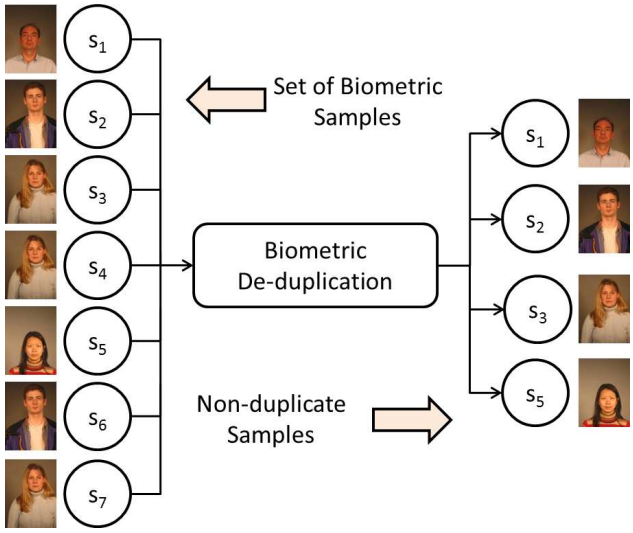

Figure 6.1: Simple illustration of the input (left) and output (right) of a de-duplication task. Note that the output set contains one sample per identity (i.e., no duplicates). Face images are from the FRGC dataset [3].

The de-duplication task has gained considerable attention as of late, particularly in the context of national scale ID programs such as the UID program in India [154], and in maintenance of large scale forensic or government databases [75]. However, the application itself has not been rigorously studied in the literature. In particular, there has not been any work pertaining to the types of errors in a de-duplication task, their potential consequences, and whether they can be appropriately estimated.

As discussed in Chapter 1, errors in classical biometric recognition are quantified using the False Match Rate (FMR), False Non-match Rate (FNMR), and Receiver Operating Characteristic (ROC) curve (in the verification scenario); the False Positive Identification Rate (FPIR) and False Negative Identification Rate (FNIR) (in the open-set identification scenario); or the Cumulative Match Characteristic (CMC) curve (in the closed-set identification scenario). However, these measures may not adequately model de-duplication error.

[2]The response to a flag for a duplicate can vary according to system needs.

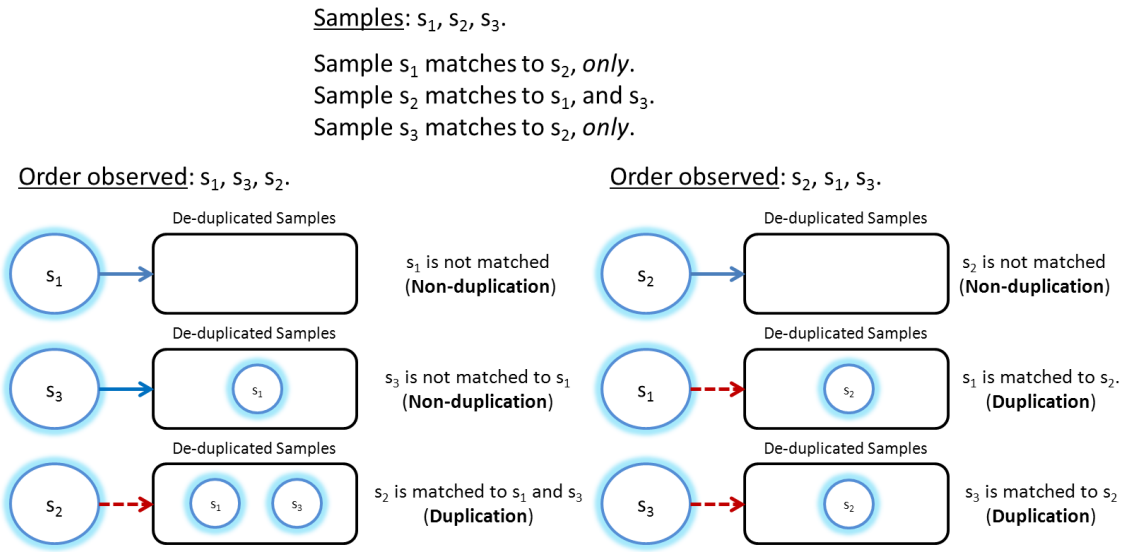

Figure 6.2: Example illustrating the effect of sample order on the outcome of a de-duplication process. Note that the probability of a sample being "de-duplicated" depends on both the "non-duplicate" sample list and its position in the sequence when it is tested for a duplicate.

In the traditional biometric verification and identification problems, the occurrence of a matching error is assumed to be a static event and cannot impact future matches. This concept was previously discussed in Chapter 4. However, in the de-duplication problem, the reference database (with "non-duplicate" entries) has the potential to expand following each test for a duplicate (in particular, when a duplicate is not found). Consequently, the order in which biometric samples are observed during enrollment can impact the error rate of the de-duplication task. This is not unlike the challenges of measuring Anonymous Identification error (Chapter 4).[3] In Figure 6.2, an example is presented demonstrating how two different sample orders can affect which samples are de-duplicated.

6.1.2 Chapter Motivation

The motivation for this chapter is to formally introduce and analyze errors in biometric de-duplication, and determine whether these errors can be reliably estimated via traditional measures (e.g., FPIR, FNIR, etc.). Thus, the contributions of this chapter are the (a) Intro-

[3]The difference between Anonymous Identification and de-duplication is that in de-duplication samples are added to the reference database *only* when a matching reference entity is not found. In Anonymous identification, samples are added to the reference database regardless of whether a match is found.

duction and definition of biometric de-duplication errors (Section 6.2); (b) An investigation regarding whether traditional biometric error measures can first be leveraged to estimate de-duplication errors in a *simplified* problem space (Section 6.2.2); and (c) An evaluation regarding whether the designed measures accurately estimate constrained de-duplication error, and a discussion of confounding factors, if present (Section 6.4).

6.2 Understanding De-duplication

6.2.1 The De-duplication Task

Consider a set of N individuals, where each individual has provided N_G biometric samples. Denote the total number of ordered samples as N_T, where $N_T = N \cdot N_G$. Assuming a perfect biometric matcher, a de-duplication task would reduce the total number of samples to N, such that each individual is represented by *exactly* one sample. Note that this is done as follows:

Suppose a set of N_T biometric samples, $\mathcal{G}_{init} = \{s_1, s_2, \ldots, s_{N_T}\}$, is to be de-duplicated. Additionally, define $\mathcal{G}_{out}$ as the set of non-duplicate samples remaining after the de-duplication task. Let $N_{out} =\mid \mathcal{G}_{out} \mid$. Initially, $\mathcal{G}_{out}$ is initialized to the empty set and $N_{out} = 0$. When the first sample, s_1, is checked against $\mathcal{G}_{out}$ for a duplicate, there are no samples to match against and s_1 is placed in $\mathcal{G}_{out}$. For all remaining samples, the k^{th} sample ($k = 2, 3, \ldots, N_T$) is matched against all the entries in $\mathcal{G}_{out}$. A *de-duplication* occurs if the similarity match score generated between the k^{th} sample and the i^{th} element in $\mathcal{G}_{out}$ ($i = 1, 2, \ldots, N_{out}$) exceeds a value of γ, where γ denotes a decision threshold. In the event of a de-duplication, the k^{th} sample is flagged for further action. Here, assume the sample is discarded (i.e., removed from the sample set). If a de-duplication does not occur, a *non-duplication* occurs and sample s_k is added to $\mathcal{G}_{out}$ and the value of N_{out} is increased by one. This process is summarized in Alg. 6.1.

Note that this scheme represents one approach towards performing the de-duplication task. In particular, the action taken following the flag for a duplicate, which can vary according to operator needs. Other actions taken may involve, for example, consolidating

Algorithm 6.1: Biometric De-duplication

Input: Biometric samples $s_1, s_2, \ldots, s_{N_T}$
Output: Non-duplicate set of samples, $\mathcal{G}_{out}$
Define: $S(s_1, s_2)$ as the similarity score between s_1 and s_2
N_{out} as the number of elements in $\mathcal{G}_{out}$
γ as a decision threshold with a value in the range $[0, 1]$
Initialize:
$\mathcal{G}_{out} = \{s_1\}$ \\ the first sample is placed in $\mathcal{G}_{out}$.

```
//Begin algorithm
for k = 2 to N_T do \\ iterate through the rest of the samples.
  duplicate_found = FALSE
  for i = 1 to N_out do
      \\ compare s_k, to the contents of G_out.
    χ = S(s_k, s_i) \\ compute similarity between s_k and s_j.
    if χ > γ then
      duplicate_found = TRUE
      break \\ A match (i.e., duplicate) was found.
    end if
  end for
  if duplicate_found == FALSE then
  G_out = G_out ∪ s_k
      \\ if no duplicates are found, add the sample to
        the non-duplicate sample set.
  end if
end for

//End algorithm
Return G_out
```

information from the probe and its matching sample to update the stored identity profile.

6.2.2 De-duplication Errors

The de-duplication task incurs type-1 (false match) and type-2 (false non-match) errors. These errors are defined as follows:

> ***False de-duplication*** **(FDD)**: A sample incorrectly matches to an identity in the non-duplicate set, $\mathcal{G}_{out}$. As a result, the input data *is not* added to $\mathcal{G}_{out}$ and the identity of an individual input may not be present in the non-duplicated set.
>
> ***False non-duplication*** **(FND)**: A sample, which has a matching identity in $\mathcal{G}_{out}$, is incorrectly not matched to any sample in $\mathcal{G}_{out}$. As a result, the same individual may have multiple identities.

Table 6.1: Summary of assumptions for FDMR and FDNMR estimation.

Assumption	**Description**
Matching Scenario	The matching scenarios defined for estimating de-duplication denote basic matching scenarios.
Uniform Sampling	The probability of observing any sample belonging to any identity is uniform.
Initial Knowledge	It is not known initially whether any samples are duplicates.
Matching Algorithm	The rank-1 matching identity returned from the matching algorithm corresponds to the identifier associated with the reference sample from which the maximum match score is generated.

The consequences of these errors can impact the outcome of the de-duplication task in different ways. For example, a large incidence of false de-duplication errors will result in a majority of identities not being represented in $\mathcal{G}_{out}$. The operational impact of this error might be that several individuals will be unable to utilize services or receive resources, having been inadvertently deleted from the list of individuals.

The result of a false non-duplication, on the other hand, is that a single individual is represented by *multiple* identities in the system. Thus, a single individual may then be able to "double-dip" and procure services or resources intended for a single person.

6.3 Estimating De-duplication Errors

Given that these errors persist in the de-duplication task, it is necessary to determine whether they can be estimated. In the traditional biometric literature, these errors are often measured through the false match rate (FMR), false non-match rate (FNMR), false positive identification rate (FPIR), and false negative identification rate (FNIR). In this section, two *simplified* de-duplication test scenarios are defined such that on the surface, appear to enable direct usage of the FMR, FNMR, FPIR, and FNIR for estimating de-duplication error rates. A summary of assumptions used to estimate these error rates is provided Table 6.1.

6.3.1 False de-duplication

Suppose N_T samples representing N identities are to undergo a de-duplication test (as defined in Section 6.2.1). Additionally, suppose each identity is represented in the initial set of samples, $\mathcal{G}_{init}$, *exactly* once (i.e., $N_G = 1, N_T = N$). Under these conditions, when each sample is tested for a duplicate, no genuine matches will exist in the non-duplicate set, $\mathcal{G}_{out}$. Further, the probability of error cannot be confounded by a false non-match. Therefore, the probability of observing a false de-duplication error (under these conditions) depends on the probability of generating at least one of N_{out} impostor scores exceeding γ.

FMR-based Estimation

In the classical verification task [10], the FMR can be loosely interpreted as the probability that a generated impostor score exceeds a decision threshold (γ). Thus, an argument can be made that the FMR raised to the power m denotes the probability that m impostor match scores exceed a decision threshold. Conversely (1−FMR) raised to the power m denotes the probability that m impostor match scores are less than a decision threshold [19]. This formulates the basis for estimating false de-duplication error via the FMR. As such, the probability of observing a false de-duplication error when $\mathcal{G}_{out}$ contains N_{out} elements is the complement of the probability that all generated match scores are less than a decision threshold and is defined in Equation (6.1).

$$P(FDD|N_{out}) = 1 - (1 - FMR)^{N_{out}} \tag{6.1}$$

FPIR-based Estimation

In the classical identification task [10], a probe biometric sample is matched against a database of N labeled identities. The system computes match scores for every identity in the database and orders them from highest (similarity) to lowest. The output is a set of L identities whose match scores exceed a certain decision threshold. In open-set identification, the actual identity of the probe may or may not exist in the database (common with the de-duplication problem). Traditionally, the performance of open-set identification is measured

through the FPIR and FNIR as demonstrated in the evaluation tests conducted by NIST [13, 14]. The FPIR is defined as the proportion of probe samples that do not have a matching identity in the database but whose match scores with one or more database entries exceed γ. Thus, the FPIR (in this case) is equal to the probability that at least one generated impostor score exceeds γ, which corresponds to the probability of false de-duplication. This is expressed formally in Equation (6.2).

$$P(FDD|N_{out}) = FPIR(N_{out}) \tag{6.2}$$

6.3.2 False Non-duplication

Again, suppose N_T samples representing N identities is to undergo a de-duplication test. Here, let $N_G = 2$ and constrain γ such that the false match rate is negligible (i.e., FMR ≈ 0). Note these constraints are more stringent than in the previous section, but are introduced in order to mitigate any confounding effects from *prior* errors (i.e., errors made in the de-duplication process prior to encountering the current sample). For example, $N_G > 1$ is required to produce genuine scores (and thus the false non-duplication error), however, in the general case, $N_G - 1$ genuine scores can theoretically be generated for a given test sample, as a result of previous false non-duplication errors. Establishing $N_G = 2$ eliminates this artifact, as each test sample can *at most* generate one genuine score. Similarly, FMR ≈ 0 is introduced such that a *prior* false de-duplication error does not impact whether a genuine matching sample to the *current* test sample was erroneously discarded.[4] Thus, in the interest of simplicity, the problem is constrained to prevent this artifact and isolate the non-duplication error. Thus, the false non-duplication error (in this case) reduces to the probability a matching sample to s_k was previously observed, multiplied by the probability a generated genuine score is less than γ.

The probability that a generated genuine score is less than γ can be approximated using the FNMR and FNIR. By definition, the FNMR is defined as the proportion of genuine scores that are lower than a threshold, γ. Loosely interpreted, the FNMR denotes the probability

[4]The challenge in measuring this probability is that it also depends on whether the erroneous matching samples were also observed and not subject to false de-duplication errors.

that a generated *genuine* match score is less than a decision threshold. The FNIR is defined as the proportion of times a probe that does have a matching entry in the database generates a genuine score less than γ *or* is observed at a rank greater than R ($R = 1, 2, \ldots, N$).[5] When $R = 1$, the FNIR denotes the probability a probe with a genuine matching identity in the database is incorrectly not matched due to one of two conditions: (a) the generation of a genuine match score less than γ *or* (b) a better match was found with another identity in the database. Note the second condition suggests that there are impostor scores greater than γ. However, with the assumption that FMR ≈ 0, this is not likely to bias the probability that a genuine score is less than γ.

FNMR-based Estimation

Regarding whether a genuine matching sample to s_k has been previously observed, if the samples are tested uniformly, this probability will simply be $\frac{k}{N_T}$. However, due to true de-duplication events, $k \neq N_{out}$. Thus, it is necessary to derive an estimate of k, given N_{out}. Let $P(k, m)$ denote the probability N_{out} is equal to m after testing k samples. Then, the expected value of N_{out} after testing k samples, $E[N_{out}|k]$ is the sum of products of m and $P(k, m)$ for $m = 1, 2, \ldots, N_T$:

$$E[N_{out}|k] = \sum_{m=1}^{k} mP(k, m). \tag{6.3}$$

However, the principal interest is in computing $E[k|N_{out}]$, the expected value of k, given N_{out}. Define $\rho^m = \{\rho_1^m, \rho_2^m, \ldots\}$ as the set of values of k for which $P(k, m)$ is non-zero for a specific m. In other words, for $N_{out} = m$, this set denotes the range of potential sample indexes and $\sum_{k \in \rho^m} P(\rho^m, m) = 1.0$. In addition, define $|\rho^m|$ as the number of elements in this set. The expected value of k, given N_{out} can be computed as the average of ρ_i^m, for $i = 1, 2, \ldots, |\rho^m|$. This is given in Equation (6.4).

$$E[k|N_{out}] = \sum_{i=1}^{|\rho^m|} \frac{\rho_i^m}{|\rho^m|}, \;\; N_{out} = 1, 2, \ldots, N_T \tag{6.4}$$

[5]Here, R denotes the length of the candidate list returned by an identification system.

The probability, $P(k, m)$ can be derived iteratively, for $m = 1, 2, \ldots, k$. In a de-duplication test, after the first sample is observed, it is added to $\mathcal{G}_{out}$ and $N_{out} = 1$. Thus, $P(k = 1, m = 1) = 1.0$. After the second sample is tested, one match score is generated, which can be either genuine or impostor. Denote the probability the match score can be classified as genuine as P_{Gen}, and the probability the match score can be classified as impostor as P_{Imp}. Note these probabilities must be "assumed" depending on the problem space. Concerning the outcome following the second sample ($k = 2$), $N_{out} = 1$ occurs if (a) the match score is genuine and correctly de-duplicated, or (b) the match score is impostor and falsely de-duplicated. The latter is assumed not to occur with FMR ≈ 0. Thus, $P(k = 2, m = 1)$ can be estimated using the FNMR, where a correct de-duplication event (i.e., a match was correctly found) is estimated as 1-FNMR (probability a genuine score exceeds γ), scaled by the probability the match score is genuine, P_{Gen}. The other outcome, $N_{out} = 2$ at $k = 2$, occurs if (a) the match score is genuine and falsely non-duplicated, or (b) the match score is impostor and correctly non-duplicated. Here, the probability of the former is defined by P_{Gen} multiplied by the FNMR, while that of the latter is defined by P_{Imp}. This process can be repeated to compute $P(k, m)$ for $k = 1, 2, \ldots, N_T$ and $m = 1, 2, \ldots, k$, enabling implementation of Equation (6.3). Thus, the probability of observing a false non-duplication is the product of the FNMR and $E[k|N_{out}]$, divided by N_T and is summarized in Equation (6.5).

$$P(FND|N_{out}) = \frac{FNMR \cdot E[k|N_{out}]}{N_T} \tag{6.5}$$

FNIR-based Estimation

The FNIR can be substituted for the FNMR in the derivation of $E[k|N_{out}]$ (Equation (6.4)), and $P(k, m)$, as a measure for estimating the false non-duplication rate under the stated assumptions. This is summarized in Equation (6.6).

$$P(FND|N_{out}) = \frac{FNIR(N_{out}) \cdot E[k|N_{out}]}{N_T} \tag{6.6}$$

6.4 Experimental Results

6.4.1 Datasets and Evaluation

Experiments are conducted to (a) demonstrate the effect the sequential testing order has on de-duplication error and (b) evaluate whether traditional error measures can describe de-duplication error in the constrained scenarios presented in Section 6.3.1 and Section 6.3.2. To enable this, similarity match scores were generated from a subset of the Facial Recognition Technology (FERET) database [8]. In particular, the subsets for regular frontal facial expression (code "fa") and alternative frontal facial expression (code "fb") are used. These subsets contain $N = 1009$ identities, with $N_G = 2$ samples per identity. Match scores were obtained using the commercial software VeriLook, similarly used in a study by Gyaourova and Ross [169].

In the presented experiments, two mutually exclusive partitions of 504 identities are randomly selected for training and testing. These partitions are divided into two further subsets, denoted by the labels "A", "B", "C", or "D". In subsets "A" and "B", only one sample per identity is utilized (from FERET code "fa"). In subsets "C" and "D", both samples per identity are utilized. These partitions are summarized in Table 6.2.

Table 6.2: Data partitions from the FERET database [8].

Partition	**# Samples**	**# Identities**	**Code(s)**
Partition A (Test)	504	504	"fa"
Partition B (Train)	504	504	"fa"
Partition C (Test)	1008	504	"fa" and "fb"
Partition D (Train)	1008	504	"fa" and "fb"

Samples in partitions "B" and "D" are used to generate estimates of the false match rate (FMR), the false positive identification rate (FPIR), and where applicable, the false non-match rate (FNMR), and the false negative identification rate (FNIR). Samples in partitions "A" and "C" are used to generate the empirical error rates after executing the de-duplication algorithm specified in Section 6.2.1.

6.4.2 De-duplication Error and Testing Order

In this experiment, the observed false de-duplication error rate is computed for different sequential orders of test data. The intent of this experiment is to demonstrate that de-duplication error is *dynamic*, and can vary depending on the explicit order in which samples are tested.

To demonstrate this, a de-duplication test (as defined in Section 6.2.1) is performed using samples from Partition "C" (Table 6.2). In total, 10,000 tests are performed using the same test data but ordered differently. In each test, the *average* observed false de-duplication rate and false non-duplication rate is computed for a set of five decision thresholds, γ. The values of γ used in this experiment are those which correspond to a false match rate approximately equal to 0.25, 0.1, 0.01, 0.001, and 0.0001.

These results are illustrated in Figures 6.3 and 6.4 in the form of a box plot. The width of each box denotes the upper and lower quartile of observed false de-duplication error. The lines extending beyond each box denote the full range of observed false de-duplication error. Outliers are designated by a "+". In order to reduce redundancy in the results, data from FMR = 0.0001 and FMR = 0.25 are neglected in Figures 6.3 and 6.4, respectively. These figures demonstrate that de-duplication error is *dynamic* and varies (between 3-10% for FDD and 0-0.3% for FND) depending on the order samples are tested for a duplicate.

6.4.3 Estimating De-duplication Error

In this experiment, the ability to estimate false de-duplication and false non-duplication error under the *simplified* conditions described in Section 6.2.2 is evaluated. This is accomplished by comparing the observed false de-duplication error rate to the FMR-based and FPIR-based measures, as presented in Section 6.3.1. Similarly, observed false non-duplication error is compared with the FNMR-based and FNIR-based measures presented in Section 6.3.2. In addition, estimates of FMR, FNMR, and FNIR are included for additional comparison (where appropriate).

Here, parameters for the false de-duplication error models are estimated using Partition "B" and the empirically observed false de-duplication error is computed on Partition "A". In

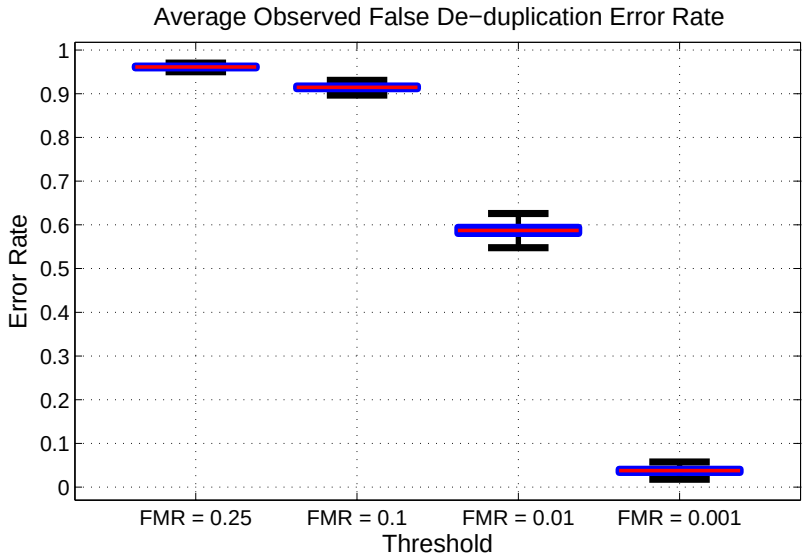

Figure 6.3: Boxplot of the average false de-duplication error for selected values of γ. Note that the error rate *varies* depending on the order samples are tested.

addition, observed and estimated error rates correspond to a decision threshold, γ, resulting in FMR ≈ 0.01. The parameters of the false non-duplication error models are estimated from Partition "D" and the empirically observed false non-duplication error is computed from Partition "C". Here, observed and estimated error rates correspond to a value of γ resulting in FMR $\approx \frac{1}{N_T}$. To remove sampling bias, 100 different combinations of Partitions "A", "B", "C", and "D" are computed and the resulting errors are averaged. These results are illustrated in Figures 6.5 and 6.6, where the false de-duplication error (Figure 6.5) and false non-duplication error (Figure 6.6) is shown in the form of a bargraph for set values of N_{out} at the stated value of γ. Note that since the data in Figures 6.5-6.6 is computed from different values of N_T and γ, the maximum value of N_{out} will be different (due to true and false de-duplication events).

6.4.4 Discussion

The above experiments highlight two major points. First, that de-duplication (and its errors) are *dynamic* and largely influenced by (a) the sequential order in which samples are tested for a duplicate (Figures 6.3-6.4) and (b) the number of elements in the non-duplicated sample set, $\mathcal{G}_{out}$ (Figures 6.5-6.6). This effect is peculiar to recognition tasks where matching

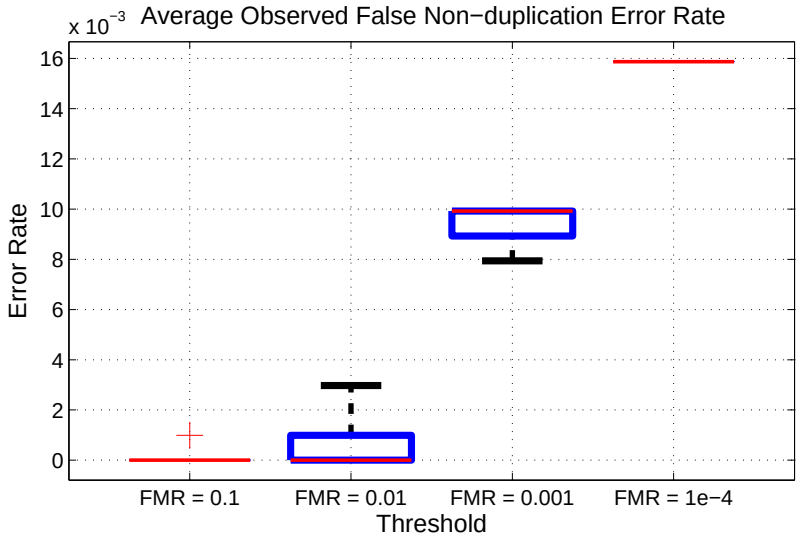

Figure 6.4: Boxplot of the average false non-duplication error for selected values of γ. Note that the error rate *varies* depending on the order samples are tested.

outcomes can influence the future composition of the reference database as discussed previously in Chapter 4. Second, that de-duplication errors are complex and difficult to predict. In the defined "simple case" for the false de-duplication error ($N_T = N$), the FMR and FPIR estimators denote noticeable *bias* (Figure 6.5). This is an interesting result, as the assumptions built into the problem appear to directly correlate with the definition of the FPIR. Therefore the logical question is: "What is the source of the bias?", for which two sources are identified.

The first source is related to the fact that classical error measures (in particular, the FMR and FNMR) denote *aggregated* match score statistics, which may not provide accurate representations of error on a per-identity level. In other words, these measures are based on a *global* analysis of error, while at a per-identity level, the error rate of an individual identity may differ from the FMR, FNMR, FPIR and FNIR. An example of this phenomenon is the Doddington's Zoo classification system of individuals in a biometric system based on their *individual* contributions towards the FMR and FNMR [7]. For example, in the Doddington's Zoo framework, "lambs" denote identities whose biometric feature set overlaps significantly with others. Such identities are likely to generate false de-duplication errors, and depending on *when* such identities are observed and the proportion of them that exist

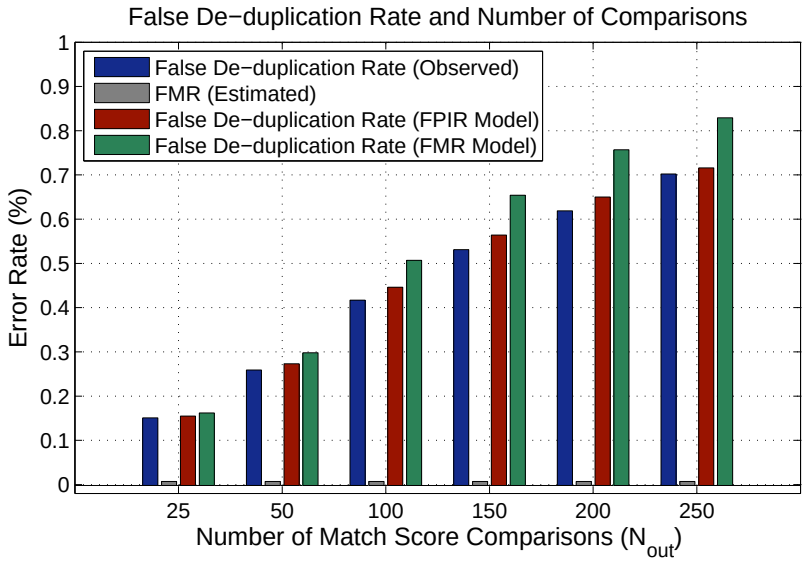

Figure 6.5: Comparison of the FMR-based and FPIR-based error models to the observed false de-duplication rate. Note in this case, the error models denote a biased estimation of the false de-duplication error.

in $\mathcal{G}_{out}$, the observed false de-duplication error rate can vary drastically. Similarly, "goats" denote identities whose biometric feature set does not match well against itself. Such users are likely to generate false non-duplication errors, which can also impact the observed error rate.

The second, and perhaps more significant source concerns a specific assumption built into the definition of FPIR and FNIR (and to a lesser extent, the FMR and FNMR). That being, these measures generally assume a probe can be compared against a database containing any combination of the other $N_T - 1$ samples. For example, the FPIR is computed by selecting some subset of samples (1 to $N_T - 1$) to define the "enrolled database", and the samples that do not have a corresponding match in the database are tested for a matching error. Note that there is *no restriction* on how the "enrolled database" is created. In other words, any possible combination of N_{gal} samples ($1 \leq N_{gal} \leq N_T - 1$) is valid. However, in de-duplication, some combinations of samples to comprise $\mathcal{G}_{out}$ are *outside* the set of possible outcomes (i.e., cannot occur).

To demonstrate this effect, consider the following "toy-example". Let θ denote a set of three biometric samples ($\theta = \{s_1, s_2, s_3\}$), where each sample denotes a different identity.

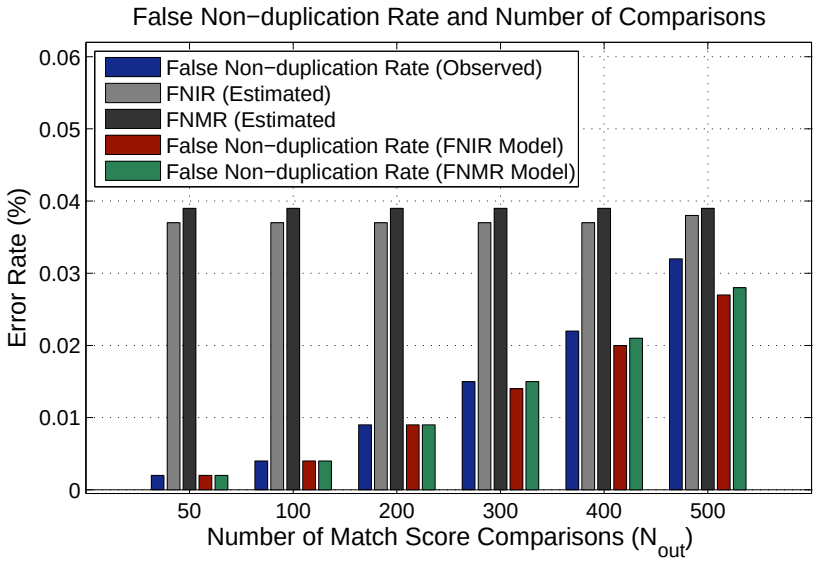

Figure 6.6: Comparison of the FNMR-based and FNIR-based error models to the observed false non-duplication rate. Note in the case (with constraints), the error models appear to accurately estimate false non-duplication error.

Assume that s_1 and s_3 "match" (incorrectly) to s_2 (and vice-versa). To compute the FPIR N_{gal} ($1 \leq N_{gal} \leq 2$) samples are chosen to denote the database and the remaining samples are used to test for an error. Let $\mathcal{G}$ denote the set of hypothetical database combinations, which are: $\mathcal{G} = \big\{\{s_1\}, \{s_2\}, \{s_3\}, \{s_1, s_2\}, \{s_1, s_3\}, \{s_2, s_3\}\big\}$. However, in a de-duplication test, the gallery combinations $\{s_1, s_2\}$ and $\{s_2, s_3\}$ cannot occur, as the pair of samples match to one another and a de-duplication event will prevent these combinations from manifesting. Thus, **the sample space for estimating the FPIR *is not* the same as the sample space for estimating the false de-duplication rate**.

Although there is not an apparent bias in the estimation of the false non-duplication error rate, this should not be interpreted as FNIR and FNMR being ideal estimators. In the general case ($N_G > 2$, FMR > 0), the false de-duplication error (which was effectively mitigated for the data in Figure 6.6) can reduce the probability that a test sample has a matching identity in $\mathcal{G}_{out}$. Consequently, if the stated false non-duplication measures are adopted, a biasing artifact will be induced and the model will fail. Figure 6.7 demonstrates this by repeating the false non-duplication error experiment (Section 6.4.3), where the decision threshold is set such that $FMR > 0$ ($FMR = 0.006$).

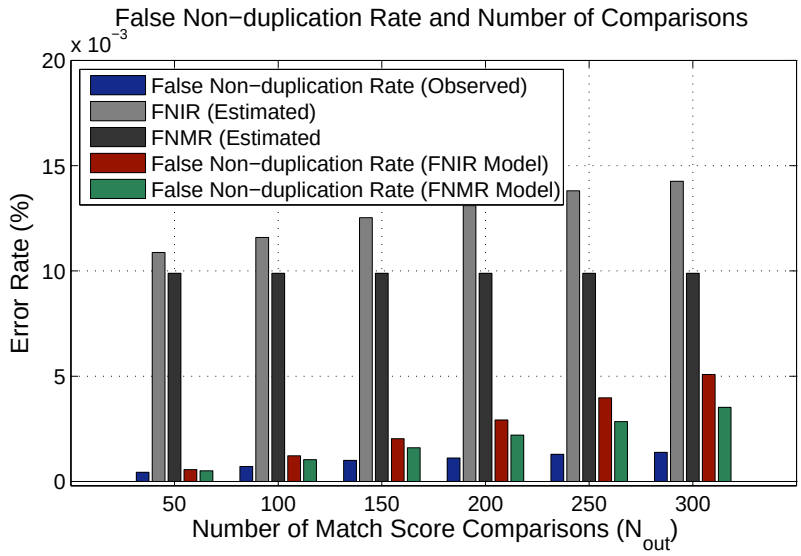

Figure 6.7: Comparison of the FNMR-based and FNIR-based error models to the observed false non-duplication rate. Note in the case (less constrained, $FMR = 0.006$), the error models fail to accurately estimate false non-duplication error.

Similarly, the false de-duplication error can be affected by prior false non-duplication errors. Therefore, given that the false de-duplication error cannot be estimated via traditional measures in the simple case, and that both de-duplication errors influence one another in the general case, it is likely traditional error measures will not provide reliable estimations of generalized de-duplication error. However, if the problem is re-defined such that the interest is in quantifying the error rate *given* N non-duplicate samples (in closed-set), then it is likely the observed error rate would converge to traditional error measures.

6.5 Summary

This chapter formally introduces the errors involved in the biometric de-duplication task, their operational impact, and the conditions required to generate a de-duplication error. Next, a simple experiment is performed that establishes a pair of constrained matching scenarios, which isolate the false match error and false non-match from one another, such that the FMR, FNMR, FPIR, and FNIR can be best leveraged for simple error prediction. The results indicate that under the constrained conditions, the FMR and FPIR result in a biased estimation of false de-duplication error, while the FNMR and FNIR can act as an unbiased

estimation of false non-duplication error. The observed bias in false de-duplication error is due to implicit assumptions present in estimating the FMR, FNMR, FPIR, and FNIR that do not hold for the de-duplication task. In addition, once the matching scenario for false non-duplication was relaxed such that false match errors had a non-zero probability of occurrence (i.e., more realistic case), the error model broke down considerably. Therefore, traditional error measures may not be completely reliable when used to describe de-duplication error in the general case.

Chapter 7

Summary and Conclusions

7.1 Summary

This dissertation discusses two components of an automated biometric surveillance system. The first component, denoted as "Methods and Modalities", refers to the imaging hardware and algorithms for performing human recognition. In this dissertation, an argument is made encouraging the use of the short-wave infrared spectrum (SWIR) for data acquisition and human gait as a biometric trait for recognition. The SWIR spectrum is discussed as an operationally advantageous image-spectrum as it (a) is undetectable to the human eye, (b) has natural illumination sources (outdoor) in both day and night, and (c) has some tolerance to obscurants such as smoke and fog (which impede visible spectrum imaging). Human gait recognition is discussed as a potential candidate for a biometric surveillance system as it (a) is believed to be unique to the individual, (b) can be collected unobtrusively, and (c) can be captured in low resolution video-data. To this end, a novel gait recognition algorithm is proposed and evaluated on a new gait database, which utilizes SWIR image data and an outdoor, unconstrained setting. In addition, a cluster analysis of gait recognition algorithms is performed, demonstrating that gait matching algorithms are capable of grouping individuals into physically distinct groups based on their gait patterns. The established groups were found to be dependent on the matching algorithm utilized. The results of this component suggests that human gait recognition can act as a capable recognition modality in a biometric surveillance system. The biometric can be utilized to infer

identity via recognition or to profile observed identities into groups.

The second component, denoted by "Performance Models and Analysis", refers to the measures used to describe matching error in various recognition tasks. In this dissertation, the recognition tasks denote those applicable to a biometric surveillance system or an identification-at-a-distance system. One such task involves matching of identities that have not been enrolled into the system. This can be accomplished by dynamically enrolling unrecognized sample data to the reference database. This is defined as Anonymous Identification. However, the dynamic creation of new reference data based on a matching outcome introduces errors that are different from classical open-set or closed-set identification and an error model is thus introduced to estimate these errors. Another recognition task relevant to surveillance as well as large-scale systems involves the de-duplication problem, whose matching outcome can also alter the contents of the reference database. Error models using traditional error measures (e.g., False Positive Identification Rate (FPIR), False Negative Identification Rate (FNIR), False Match Rate (FMR), and False Non-match Rate (FNMR)) were developed and were found to provide an inaccurate representation of an empirically derived de-duplication error rate. Finally, a model is developed for understanding the relationship between ROC and CMC curves, which are typically used to denote matching performance in academic literature. The model characterizes the inter- and intra-class relationships between identities and is used to develop sets of faux identities from empirical match scores. At the global level, sets of faux identities share the same match score data, but the values are distributed differently to each identity. The benefit of performing such an analysis is to determine whether large differences in the CMC curve (which reflects a measurement of local match score relationships) can be explained from the same match score data contributing to a single ROC curve (which reflects a measurement of global match score statistics). The results indicate that the interactions between different identities could result in a CMC curve whose suggested performance differs from that of the ROC curve. As such, researchers should present both ROC and CMC curves to better characterize the performance of a matching algorithm.

7.2 Contributions

In summary, the contributions of this dissertation are summarized by the following points.

In Chapter 2, a method for performing automated gait recognition was presented. In addition, a novel dataset for evaluation of gait recognition algorithms was introduced. Defined as the Gait Curves matching algorithm and the WVU Outdoor SWIR Gait (WOSG) dataset, the matching algorithm and dataset offer the following contributions:

- The Gait Curves matching algorithm denotes a unique, shape-based approach to quantifying human gait patterns. The feature data can be further utilized for backpack detection and silhouette restoration.
- The WVU Outdoor SWIR Gait dataset is unique as it is the first gait dataset to utilize the short-wave infrared (SWIR) spectrum and whose collection was performed in an environmentally unconstrained setting, which is likely to mimic surveillance data.

In Chapter 3, a cluster analysis of gait recognition algorithms was performed. The cluster analysis offers the following contributions:

- Three algorithms for gait recognition (Gait Curves, Gait Energy Image [91], and Frieze Patterns [101]) were found to group identities differently. As such, not all matchers assess gait similarly.
- Formed clusters can be described by physical attributes of individuals. In particular, body area and gender were found to contribute significantly towards how a pair of gait patterns are assessed for similarity.

In Chapter 4, a matching scheme for biometric surveillance systems is introduced. Defined as Anonymous identification, the matching scheme offers the following contributions:

- In Anonymous Identification, the system dynamically enrolls unrecognized identities into the reference database, enabling the possibility of future recognition.

- An error model is introduced to provide an estimation of errors in an Anonymous Identification system. The error model provides a better representation of Anonymous Identification error, compared to False Positive Identification Rate and False Negative Identification Rate, which are common error measures for identification systems.

- The matching scheme and error model are useful for understanding biometric recognition problems where the matching outcome can alter the composition of the reference database, as in the re-identification and de-duplication problems.

In Chapter 5, an analysis studying the relationship between the Receiver Operating Characteristic (ROC) curve and Cumulative Match Characteristic (CMC) curve is presented. The analysis offers the following contributions:

- A model for characterizing identity-specific relationships in match scores is developed. The model is capable of generating faux identities from a set of empirical match scores. Created faux identities share the same match scores as in the empirical data (i.e., same global statistics, or ROC curves), but distributed differently (i.e., different per-identity statistics, or CMC curves).

- Utilizing the model, it is possible to show that a range of CMC curves can be associated with a single ROC curve. As such, researchers should utilize both curves when reporting performance of a matching algorithm.

In Chapter 6 an analysis was performed testing whether traditional error measures can be utilized to predict de-duplication errors. The analysis offers the following contributions:

- Formal introduction of de-duplication errors, noting that their occurrence impacts the composition of the reference dataset.

- Development of techniques that utilize the False Match Rate, False Non-match Rate, False Positive Identification Rate, and False Negative Identification Rate to estimate de-duplication error in a simplified matching setting. The results demonstrate that these measures are not adequate representations of de-duplication error.

7.3 Future Work

Although this dissertation presents a number of contributions, the results presented in this dissertation are not without their own challenges, thus offering opportunities for further research. Such opportunities exist within both the "Methods and Modalities" and "Performance Models and Analysis" components.

7.3.1 Segmentation in Unconstrained Video Sequences

Arguably, the primary reason that the Gait Curves (as well as the additional baseline gait matching algorithms) performed poorly on the WVU Outdoor SWIR Gait (WOSG) dataset is that the silhouettes extracted from the video data were of much lower quality than those extracted in existing gait datasets. This argument was validated experimentally in Chapter 2, Section 2.5.2, where the silhouette quality metric developed by Liu et al. [109] was implemented on silhouettes produced on the CASIA B, CASIA C, and WOSG datasets, using a simple background subtraction scheme for silhouette extraction. The results showed that the quality of the silhouettes produced in the WOSG dataset (most challenging) were much lower than those of the CASIA B and CASIA C datasets.

Though it is important to develop methods that are robust to covariates and extraneous variables such as carrying condition, viewpoint, and walking speed, as suggested by Liu and Sarkar [108], it is also naive to presume that robust silhouette extraction is not a critical component in an *operational* gait recognition system. As with any biometric modality, the matching process is certain to be less than ideal if the data passed to a feature extraction algorithm is noisy or otherwise corrupted. Though advanced segmentation methods have been developed in the literature, the majority of these algorithms are developed specifically for image data in the visible spectrum, rather than the infrared spectrum. In particular, many of these algorithms rely on multiple channels of image data (i.e., as in RGB data), an example being the "Codebook" model for background subtraction by Kim et al. [114]. Given the operational advantages of using SWIR image data (natural illumination sources, nighttime operation, etc.), it is essential to develop segmentation methods explicitly for SWIR images, or methods specific to single channel video and images.

7.3.2 Clustering of Gait: Additional Datasets and Evaluation

As stated in Chapter 3, Section 3.3.7, one limitation of the cluster analysis was that it was performed on a single dataset. It would be beneficial to perform the analysis on several datasets to see whether the metadata variables (gender, body area, cadence, stride, and height) contribute similarly to the formation of clusters. Such a study would aid in confirming the results presented in Chapter 3.

However, the cluster analysis is best suited towards *high quality* silhouette data. The reason high quality data is required is that research has demonstrated that lower quality silhouettes can match together based on erroneous foreground pixels (e.g., shadows), and not the actual "gait pattern" [108]. This means that datasets such as the CASIA C gait dataset [130] and the WVU Outdoor SWIR gait dataset (Chapter 2, Section 2.3) would not be appropriate for such an analysis. Other datasets, involving the use of treadmills might be beneficial for a clustering analysis as they denote data collected in a static lab environment. Examples include the Soton Large dataset [127] and the Osaka Treadmill dataset [134]. In particular, the "A" subset of the Osaka Treadmill dataset compares gait at different walking speeds. It would be interesting to perform clustering on this subset to see if cadence and stride play a prominent role in cluster generation. It is likely that the Gait Curves algorithm (Chapter 3, Section 2.2) would result in clusters that show increased correlation with stride and cadence, as this algorithm was found to show a reduced matching performance when comparing samples of varying walking speed (Figure 2.13).

7.3.3 Empirical ROC and CMC Analysis

One of the reasons for the development of a simulation model for characterizing match scores (Chapter 5) was that it is a resource intensive task to generate empirical match scores from multiple recognition schemes and modalities, in order to perform a comprehensive analysis relating the two curves. However, if a large number of empirical ROC and CMC curves could be extracted for a large number of matching algorithms, the analysis would be extremely relevant to the field.

7.4 Conclusions and Recommendations

This dissertation provides a foundation towards the design and development of a biometric surveillance system and should be of interest to researchers across the biometric community. Researchers studying biometric surveillance systems should not dismiss gait recognition as a candidate recognition modality. Though it can be argued that it is not possible to perform a standard enrollment of gait, clustering of unlabeled gait patterns can still aid in providing some information regarding observed individuals. Researchers studying gait recognition problems must consider the silhouette extraction process and advanced segmentation algorithms for gait recognition to proceed as a biometric modality. The WVU Outdoor SWIR Gait dataset illustrates how challenging the localization and segmentation problems can be, and that algorithms noted for acceptable matching performance can fail with poor quality data. Beyond gait recognition, researchers studying biometric surveillance systems, or the biometric "re-identification" problem must be aware of how such systems will function in an operational setting. In particular, questions such as how the system assembles its reference database or how the system discerns whether an observed identity has been seen before, must be considered. Finally, researchers in general should be cautious when reporting or observing reports on de-duplication error that utilize traditional error rates. Similarly, researchers utilizing CMC curves should report the corresponding ROC curve, as the curves cannot be assured to be directly related.

Appendix A

Clustering Extension: Indexing Gait

A.1 The Indexing Problem

In the biometric literature, the indexing problem denotes a runtime optimization problem, wherein the reference data in the database is assigned a second identifier (i.e., a cluster label), which is used to reduce the search space (i.e., the number of match scores computed) for a given probe [137]. In other words, when the system observes a probe, a clustering algorithm identifies a subset of the reference data to match against (i.e., a candidate list), reducing the time required to produce a match report. This process is important for *large* biometric systems that require a fast matching time (e.g., high throughput).

Indexing performance is typically measured via two measures: *Hit Rate* (HR) and *Penetration Rate* (PR). The hit rate reflects the probability that a probe, which has a reference entity in the database, has the reference entity returned in the candidate list. This is summarized in Equation (A.1), where N_{hit} denotes the number of probes for which a correctly matching reference entity is present in the candidate list and N_{probe} is the number of probes.

$$\text{Hit Rate } = \frac{N_{hit}}{N_{probe}} \tag{A.1}$$

The penetration rate denotes the average percentage of the reference database comprising the candidate list returned for each probe. This is summarized in Equation (A.2), where L_i denotes the size of the returned candidate list and N_{ref} is the total number of reference samples. Note, in the context of the previous experiments, $N_{ref} = C_{train}$.

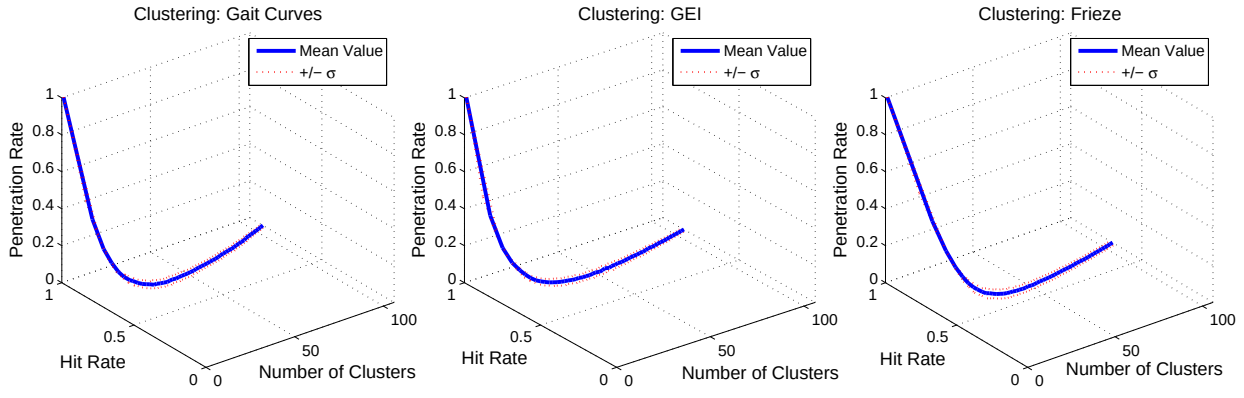

Figure A.1: Indexing performance using k-means clustering. Left) Gait Curve Matching. Center) Gait Energy Image (GEI). Right) Frieze Pattern Matching.

$$\text{Penetration Rate } = \frac{1}{N_{probe}} \sum_{i=1}^{N_{probe}} \frac{L_i}{N_{ref}} \tag{A.2}$$

Similar to the false match rate and false non-match rate, the hit rate and penetration rate are inversely related. For example, with one cluster, the candidate list will always contain the matching reference sample, at the cost of searching the entire database. As the number of clusters increase, the penetration rate and hit rate generally decrease. An effective indexing scheme will aim to maximize hit rate while minimizing penetration rate.

A.2 Indexing Performance

In this experiment, an extension of the cluster analysis in Chapter 3 is presented to denote indexing performance. Here, hit rates and penetration rates are computed for $c = 1, 3, 5, \ldots, 105$ clusters according to the protocol outlined in Chapter 3, Section 3.3.3. Since the observed hit rate and penetration rate is a function of the samples comprising C_{train}, cluster generation is repeated 100 times. The mean hit rate and penetration rate for the k-means clustering algorithm is presented in Figure A.1, respectively. Also included in Figure A.1 is the mean value adjusted $\pm$ one standard deviation.

In addition to hit rate and penetration rate, some researchers have attempted to quantify the hit rate and penetration rate trade-off as a single valued measure. For example, Gadde et. al. define ζ [170], which combines both the hit and penetration rate as follows:

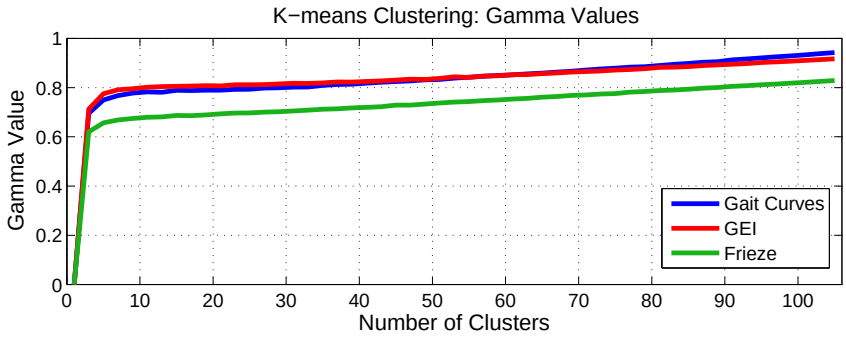

Figure A.2: Computed ζ values from k-means clustering.

$$\zeta = \sqrt{\text{Hit Rate} \cdot (1 - \text{Penetration Rate})}. \tag{A.3}$$

In Figure A.2, ζ is computed for both the k-means clustering algorithm for, $c = 1, 3, 5, \ldots, 105$. These plots aid in visualizing the indexing performance across the different matching algorithms (e.g., Gait Curves, GEI, Frieze Patterns).

Appendix B

Supplemental Anonymous Identification Analysis

B.1 Effect of Sequential Probe Order on Observed FDMR and FDNMR

Here, a more detailed analysis regarding the effect of sequential probe order as it pertains to the false dynamic match rate and false dynamic match rate is provided. Recall in Chapter 4, Section 4.3.2, the observed FDMR and FDNMR were evaluated for the probe orders: random draw, increment probe, increment subject, and a specified version of increment subject, where individuals that are more prone to falsely match to others are encountered first. In Figures B.1 and B.2 the mean observed error rates (denoted by a circle (o)) for these probe orders is shown for the full range of γ for face scores. Additionally, the standard deviation of observed FDMR and FDNMR (denoted by dots ($\cdot$) is also provided, demonstrating that the error for each "class" of probe orders is also dynamic.

B.2 Predicting FDNMR and FDNMR

Here, an experiment is provided demonstrating the ability of the prediction model to perform on sequestered data. That is, estimating the FDMR and FDNMR on data that is not used to generate the observed error rates. To enable this, bootstrapping of the original

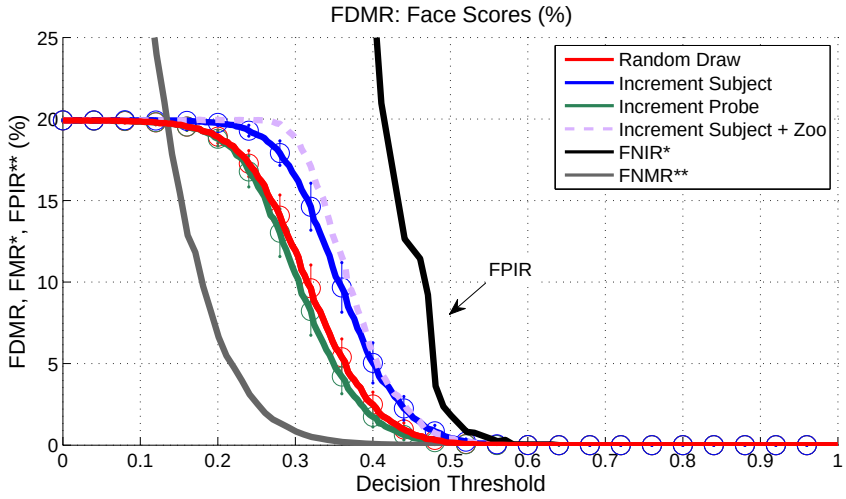

Figure B.1: Comparison of FDMR, FMR, and FPIR for face scores. Each circle (o) denotes the mean FDMR at that threshold. Dots (·) indicate one standard deviation from the mean. Note that each type of probe order exhibits different ranges of error.

test data into smaller subsets is performed. Each bootstrapped subset consists of 300 sampled probes, pertaining to $M = 60$ identities. By performing several predictions and observations on bootstrapped data, an estimation of the model performance on data not explicitly used in training can be established. This is illustrated in Figure B.3 for face scores. Note that in this experiment, predicted error rates were generated using the procedure described in Sub-Experiment B, only with 300 probes. Observed error rates were generated using the procedure described in Sub-Experiment A, using random draw to order the probes and $P = 2,500$.

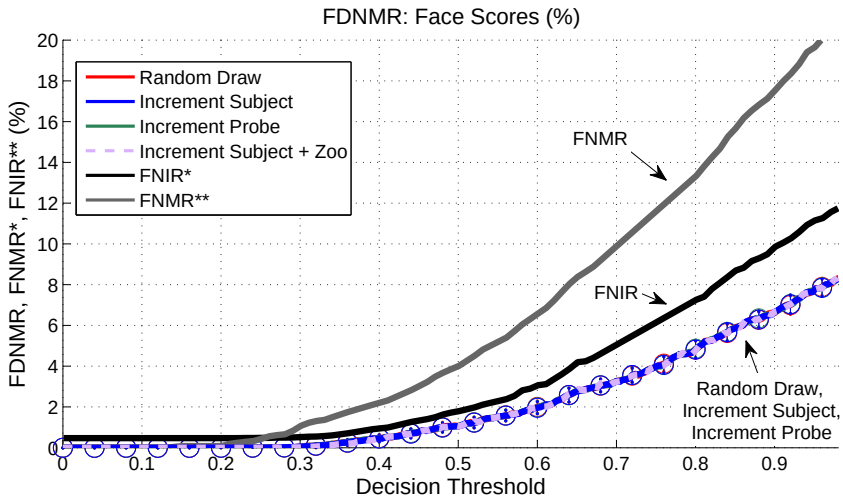

Figure B.2: Comparison of FDNMR, FNMR, and FNIR for face scores. Each circle (o) denotes the mean FDNMR at that threshold. Dots (·) indicate one standard deviation from the mean. Note that the range of error for each probe order is similar to one another.

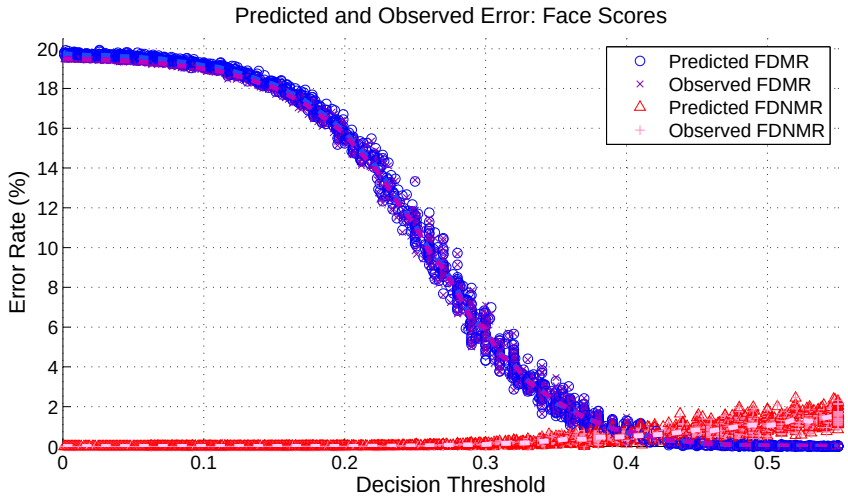

Figure B.3: Predicted and observed error rates for face scores. Each bootstrap is marked with its predicted pair. Note that in general, the predicted FDMR or FDNMR for a given threshold is within ±2% of any observed value.

References

[1] "PETS 2007," Tenth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, 2007.

[2] R. Zhang, C. Vogler, and D. Metaxas, "Human Gait Recognition," *Proceedings of the IEEE International Workshop on Computer Vision and Pattern Recognition*, 2004.

[3] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Cheng, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the Face Recognition Grand Challenge," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[4] R.M Bolle, J.H. Connell, S. Pankanti, N.K. Ratha, and A.W. Senior, "The Relation Between the ROC Curve and the CMC," *Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, pp. 15–20, 2005.

[5] J.P. Hube, "Using Biometric Verification to Estimate Identification Performance," *Biometrics Symposium*, pp. 1–6, September 2006.

[6] S. Crihalmeanu, A. Ross, S. Schuckers, and L. Hornak, "A Protocol for Multibiometric Data Acquisition, Storage and Dissemination," Tech. Rep., West Virginia University, 2007.

[7] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, Goats, Lambs and Wolves: A Statistical Analysis of Speaker Performance," *IEEE International Conference on Language and Speech Processing*, pp. 1351–1354, November 1998, Sydney, Australia.

[8] P. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET Evaluation Methodology for Face-recognition Algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[9] A. Jain, P. Flynn, and A. Ross, *Handbook of Biometrics*, Springer, 2008.

[10] A. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, January 2004.

[11] A. Jain, S. Dass, and K. Nandakumar, "Soft Biometric Traits for Personal Recognition Systems," *Proceedings of the International Conference on Biometric Authentication*, vol. LNCS 3072, pp. 731–738, 2004.

[12] ISO/IEC, *International Standard ISO/IEC 2382-37*, first edition edition, December 2012.

[13] P. Grother, G. Quinn, and P. Phillips, "Report on the evaluation of 2D still-image face recognition algorithms," Interagency/Internal Report (NISTIR) 7709, National Institute of Standards and Technology (NIST), 2010.

[14] P. Grother, G.W. Quinn, J.R. Matey, M. Ngan, W. Salamon, G. Fiumara, and C. Watson, "IREX III Performance of Iris Identification Algorithms," Tech. Rep. NIST Interagency Report 7836, National Institute of Standards and Technology (NIST), 2012.

[15] J.A. Hanley and B.J. McNeil, "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve.," *Radiology*, vol. 143, pp. 29–36, 1982.

[16] A. Martin, G. Doggington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," *EUROSPEECH*, pp. 1895–1898, 1997.

[17] D.M Green and J.A. Swets, *Signal Detection Theory and Psychophysics*, Wiley, 1966.

[18] H. Moon and P.J. Phillips, "Computatioinal and Performance Aspects of PCA-based Face Recognition Algorithms," *Perception*, vol. 30, no. 5, pp. 303–321, 2001.

[19] P. Phillips, P. Grother, R. Michaels, D. Blackburn, T. Elham, and J.M. Bone, "FRVT 2002: Facial Recognition Vendor Test," Tech. Rep., DoD, April 2003.

[20] A.Y. Johnson, J. Sun, and A.F. Bobick, "Using Similarity Scores From a Small Gallery to Estimate Recognition Performance for Larger Galleries," *AMFG*, p. 100, 2003.

[21] P. Grother and P. Phillips, "Models of Large Population Recognition Performance," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 68–75, 2004.

[22] P. Griffin, "Predicting the Performance of Biometric Identification Systems," *Preprint*, 2005.

[23] A. Jain and J. Feng, "Latent Fingerprint Matching," *IEEE Transactions on Patt*, vol. 33, no. 1, pp. 88–100, 2011.

[24] T.J. Cham Y. Huang, D. Xu, "Face and Human Gait Recognition Using Image-to-Class Distance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20(3), pp. 431–438, 2010.

[25] H. Aghajan and A. Cavallaro, *Multi-Camera Networks: Principles and Applications*, National Academic Press, 2009.

[26] P. Chen, P. Ahammad, C. Boyer, S. Huang, L. Lin, E. Lobaton, M. Meingast, O. Songhwai, S. Wang, P. Yan, A.Y. Yang, C. Yeo, L. Chang, J.D. Tygar, and S.S. Sastry, "CITRIC: A Low-Bandwidth Wireless Camera Network Platform," *Second ACM/IEEE International Conference on Distributed Smart Cameras*, pp. 1–10, 2008.

[27] F. Speilman and F. Main, "City Plans Camera Surveillance Web," Chicago Sun-Times, 2004.

[28] H. Dardick, "City Will Keep Eyes Peeled Big Time," Chicago Tribune, 2005.

[29] "iSee Project: Survey and Web-Based Application Charting the Locations of Closed-Circuit Television (CCTV) Surveillance Cameras in Urban Enviornments," Tech. Rep., Institute for Applied Autonomy (IAA), 2003, http://www.appliedautonomy.com/isee/info2.html.

[30] "More Cities Deploy Camera Surveillance Systems With Federal Grant Money," May 2005, http://epic.org/privacy/surveillance/spotlight/0505/.

[31] T. Edwards, "Magistrate Says CCTV Cameras Are Not Monitored Properly," Worcester News, June 2012, http://www.worcesternews.co.uk/news/9741479.Magistrate_says_CCTV_cameras_are_not_monito

[32] F. Manning, "Worcester City Monitor 100 CCTV Cameras With Only One Person," Big Brother Watch, 2011, http://www.bigbrotherwatch.org.uk/home/2011/08/worcester-city-monitor-100-cctv-cameraswith-only-one-person.html.

[33] R. McKinnon, "Big Brother isn't Watching," Evening Times, November 2007, http://www.eveningtimes.co.uk/big-brother-isn-t-watching-1.976256.

[34] Macnish, "Unblinking Eyes: The Ethics of Automating Surveillance," *Ethics and Information Technology*, vol. 2, pp. 151–167, 2012.

[35] A. Mack, "Intentional Blindness: Looking Without Seeing," *Current Directions in Psychological Science*, vol. 12, pp. 180–184, 2003.

[36] F. Bashir, P. Casaverde, D. Usher, and M. Friedman, "Eagle-Eyes: A System for Iris Recognition at a Distance," *IEEE Conference on Technologies for Homeland Security*, pp. 426–431, May 2008.

[37] A. Hampapur, L. Brown, H. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, S. Pankanti, A. Senior, C. Shu, and Y. Tian, "Smart Video Surveillance: Exploring the Concept of Multi-Scale Spatiotemporal Tracking," *IEEE Signal Processing magazine*, pp. 38–51, March 2005.

[38] M. Shah, O. Javed, and K. Shafique, "Automated Visual Surveillance in Realistic Scenarios," *IEEE Multimedia*, vol. 14, no. 1, pp. 30–39, January 2007.

[39] A. Jain, D. Kopell, K. Kakligian, and Y. Wang, "Using Stationary-Dynamic Camera Assemblies for Wide-area Video Surveillance and Selective Attention," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[40] M. Valera and S.A. Velastin, "Intelligent Distributed Surveillance Systems: A Review," *Vision, Image and Signal Processing*, vol. 152, no. 2, pp. 192–204, April 2005.

[41] K. Brown, "Genetec Unveils the AutoVu Sharp: the Latest in License Plate Recognition," Tech. Rep., Genetec, 2009.

[42] M. Haemek, "License Plate Recognition System offers high-speed operation," Tech. Rep., Hi-Tech Solutions Ltd., 2010.

[43] R. Polana and R.C. Nelson, "Detection and Recognition of Periodic Nonrigid Motion," *International Journal of Computer Vision*, vol. 23(7), 1997.

[44] S. Seitz and C. Dyer, "View-Invariant Analysis of Cyclic Motion," *International Journal of Computer Vision*, vol. 25(3), 1997.

[45] S. Carlsson and J. Sullivan, "Action Recognition by Shape Matching to Key Frames," *Workshop Models versus Exemplars in Computer Vision*, 2001.

[46] M.J. Black, "Exploring Optical Flow Events with Parameterized Spatio-Temporal Models," *Computer Vision and Pattern Recognition*, vol. 1, pp. 33–80, 1999.

[47] A.A. Efros, A.C. Berg, G. Mori, and J. Malik, "Recognizing Action at a Distance," *International Conference on Computer Vision*, 2003.

[48] M. Brand, N. Oliver, and A. Pentland, "Coupled Hidden Markov Models for Complex Action Recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 994–999, 1997.

[49] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky, "Behavior Classification by Eigendecomposition of Periodic Motions," *Pattern Recognition*, vol. 38(7), pp. 1033–1043, 2005.

[50] L. Gorelick, M. Blank, E Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," *Transactions on Pattern Analysis and Machine Intelligence*, vol. (29)12, pp. 2247–2253, 2007.

[51] I. Junejo, E. Dexter, I. Laptev, and P. Perez, "View Independent Action Recognition from Temporal Self-Similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 33, pp. 172–185, 2011.

[52] S. Maji, L. Bourdev, and J. Malik, "Action Recognition From a Distributed Representation of Pose and Apperance," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3177–3184, 2011.

[53] C. Roberts, "Muni's New Cameras Detect 'Pre-Crime'," NBC Bay Area, June 2012, http://www.nbcbayarea.com/news/weird/Munis-New-Cameras-Detect-Pre-Crime-158211365.html.

[54] N. Halverson, "RNC Fortified by Behavior-Recognizing Cameras," Discovery News, August 2012, http://news.discovery.com/tech/rnc-security-cameras-120828.html.

[55] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher, "Apperance-based Person Reidentification in Camera Networks: Problem Overview and Current Approaches," *Journal of Ambient Intelligence and Humanized Computing*, vol. 2, pp. 127–157, 2011.

[56] N. Gheissari, T.B. Sebastian, and R. Hartley, "Person Reidentification Using Spatiotemporal Apperance," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1528–1535, 2006.

[57] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux, "Person Re-Identification in Multi-Camera System by Signature Based on Interest Point Descriptors Collected on Short Video Sequences," *IEEE/ACM International Conference on Distributed Smart Cameras (ICSDC)*, pp. 1–6, 2008.

[58] N. Martinel and C. Micheloni, "Re-Identify people in Wide Area Camera Network," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 31–36, 2012.

[59] R. Satta, G. Fumera, and F. Roli, "Fast Person Re-Identification Based on Dissimilarity Representations," *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1838–1848, October 2012.

[60] M. Nappi and H. Wechsler, "Robust Re-Identification Using Randomness and Statistical Learning: Quo Vadis," *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1820–1827, October 2012.

[61] R. Mazzon, S.F. Tahir, and A. Cavallaro, "Person Re-Identification in Crowd," *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1828–1837, October 2012.

[62] DARPA, "HumanID Program," www.darpa.mil/iao/HID.htm.

[63] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893, 2005.

[64] K. Mikolajczyk, C. Schmid, and A. Zisserman, *Computer Vision - ECCV 2004*, chapter Human Detection Based on a Probabilistic Assembly of Robust Part Detectors, pp. 69–82, Springer Berlin / Heidelberg, 2004.

[65] O. Tuzel, F. Porikli, and P. Meer, "Human Detection via Classification on Reiemannian Manifolds," *IEEE Computer Soc*, vol. 1, pp. 1–8, June 2007.

[66] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4: Real-Time Surveillance of People and Their Activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 809–830, 2000.

[67] J. Daugman, "Iris Recognition Border-Crossing System in the UAE," *International Airport Review*, vol. 8, no. 2, 2004.

[68] J. Daugman, "How Iris Recognition Works," *IEEE Transactions on Circuits and Systems for V*, vol. 14(1), pp. 2130, 2004.

[69] P. Viola and M. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[70] H.A. Rowley, S. Baluja, and T. Kanade, "Neural Network-based Face Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, Jan 1998.

[71] R.L. Hsu, M. Abdel-Mottaleb, and A. Jain, "Face Detection in Color Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696–706, May 2002.

[72] M. Nixon, T. Tan, and R. Chellappa, *Human Identification Based on Gait*, Springer, 2006.

[73] BBC News (UK), "Burglar Jailed Over Unusual Walk," April 2008, http://news.bbc.co.uk/2/hi/uk_news/england/lancashire/7343702.stm.

[74] P.K. Larsen, E.B. Simonsen, and N. Lynnerup, "Gait Analysis in Forensic Medicine," *Journal of Forensic Sciences*, vol. 53, pp. 1149–1153, 2008.

[75] A. Jain, B. Klare, and U. Park, "Face Matching and Retrieval in Forensics Applications," *IEEE MultiMedia*, vol. 19, no. 1, 2012.

[76] D.A. Winter, *Biomechanics and Motor Control of Human Gait: Normal, Elderly, and Pathological*, Waterloo, 1991.

[77] S.R. Simon, "Quantification of Human Motion: Gait Analysis - Benefits and Limitations to its Application to Clinical Problems," .

[78] A. Mirelman, B.L. Patritti, P. Bonato, and J.E. Deutsch, "Effects of Virtual Reality Training on Gait Biomechanics of Individuals Post-Stroke," *Gait & posture*, vol. 31, no. 4, pp. 433–437, 2010.

[79] A. Nimbarte and L. Li, "Effect of Added Weights on the Characteristics of Vertical Ground Reaction Force During Walk-to-Run Gait Transition," .

[80] P. Devita, T. Hortobagyi, and J. Barrier, "Gait Biomechanics Are Not Normal After Anterior Cruciate Ligament Reconstruction and Accelerated Rehabilitation," *Medicine and Science in Sports and Exercise*, vol. 30, pp. 1481–1488, 1998.

[81] L. Kozlowski and J. Cutting, "Recognizing the Sex of a Walker from a Dynamic Point Light Display," *Perception and Psycophysics*, vol. 21, pp. 575–580, 1977.

[82] J. Cutting and L. Kozlowski, "Recognizing Friends by Their Walk," *Bulletin of the Psychonomic Society*, vol. 9(5), pp. 353–356, 1977.

[83] D. Cunado, M. Nixon, and J. Carter, "Using Gait as a Biometric via Phase-Weighted Magnitude Spectra," *Proceedings of the First International Conference on Audio Visual Biometric Person Authentication*, vol. 1206, pp. 95–102, 1997.

[84] C. Yam, M. Nixon, and J. Carter, "Automated Person Recognition by Walking and Running via Model-Based Approaches," *Pattern Recognition*, vol. 37(5), pp. 1057–1072, 2004.

[85] D. Wagg and M. Nixon, "On Automated Model-Based Extraction and Analysis of Gait," *Proceedings of the IEEE International Conference on Face and Gesture Recognition*, pp. 11–16, 2004.

[86] C. BenAbdelkader, R. Cutler, and L. Davis, "Stride and Cadence as a Biometric in Automatic Person Identification and Verification," *Proceedings of the IEEE International Conference on Face and Gesture Recognition*, pp. 372–377, 2002.

[87] M. Goffredo, I. Bouchrika, J.N. Carter, and M.S. Nixon, "Self-Calibrating View-Invariant Gait Biometrics," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 40, no. 4, pp. 997–1008, August 2010.

[88] S. Niyogi and E. Adelson, "Analyzing and Recognizing Walking Figures in XYT," *Proceedings of IEEE Computer Vision and Pattern Recognition*, pp. 469–474, 1994.

[89] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette Analysis-Based Gait Recognition for Human Identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1505–1518, 2003.

[90] P. Huang, C. Harris, and M. Nixon, "Recognizing Humans by Gait via parametric Canonical Space," *Artificial Intelligence in Engineering*, vol. 13(4), pp. 359–366, 1999.

[91] J. Han and B. Bhanu, "Human Activity Recognition in Thermal Infrared Imagery," *Proceedings of the IEEE Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, 2005.

[92] J. Han and B. Bhanu, "Individual Recognition Using Gait Energy Image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 316–322, 2006.

[93] Y. Guan, C.T. Li, and Y. Hu, "Random Subspace Method for Gait Recognition," *IEEE Internaional Conference on Multimedia and Expo Workshops*, pp. 284–289, July 2012.

[94] D. Tao, X. Li, X. Wu, and J. Maybank, "General Tensor Discriminant Analysis and Gabor Features for Gait Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1700, 2007.

[95] E. Zhang, Y. Zhao, and W. Xiong, "Active Energy Image plus 2DLPP for Gait Recognition," *Signal Processing*, vol. 90, no. 7, pp. 2295–2302, July 2010.

[96] J. Liu and N. Zheng, "Gait History Image: A Novel Temporal Template for Gait Recognition," *IEEE Internation Conference on Multimedia and Expo*, pp. 663–666, 2007.

[97] C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian, "Frame Difference Energy Image for Gait Recognition With Incomplete Silhouettes," *Pattern Recognition*, vol. 30, no. 11, pp. 977–984, 2009.

[98] D. Tan, K. Huang, S. Yu, and T. Tan, "Efficient Night Gait Recognition Based on Template Matching," *Proceedings of the International Conference on Pattern Recognition*, 2006.

[99] A. Sundaresan, A. Roy-Chowdhury, and R. Chellappa, "A Hidden Markov Model Based Framework for Recognition of Humans from Gait Sequences," *Proceedings of the IEEE International Conference on Image Processing*, vol. 2, pp. 93–96, 2003.

[100] A. Kale, A Sunderesan, A.N. Rajagopalan, N. Cuntoor, A. Roy-Chowdhury, V. Kruger, and R. Chellappa, "Identification of Humans Using Gait," *IEEE Transactions on Image Processing*, vol. 13, pp. 1163–1173, 2004.

[101] Y. Liu, R. Collins, and Y. Tsin, "Gait Sequence Analysis Using Frieze Patterns," in *Computer Vision - ECCV 2002*, vol. 2351, pp. 733–736. Springer Berlin / Heidelberg, 2002.

[102] S. Lee, Y. Liu, and R. Collins, "Shape Variation-Based Frieze Pattern for Robust Gait Recognition," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[103] A. Kale, N. Cuntoor, B. Yegnanarayana, A.N. Rajagopalan, and R. Chellappa, "Gait Analysis for Human Identification," in *Audio- and Video-Based Biometric Person Authentication*, vol. 2688, p. 1058. Springer B, 2003.

[104] D. Tan, K. Huang, S. Yu, and T. Tan, "Recognizing Night Walkers Based on One Pseudoshape Representation of Gait," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[105] L. Wang, H. Ning, W. Hu, and T. Tan, "Gait Recognition Based on Procrustes Shape Analysis," *IEEE International Conference on Image Processing (ICIP)*, pp. 433–436, 2002.

[106] Y. Yang and M. Levine, "The Background Primal Sketch: An Approach for Tracking Moving Objects," *Machine Vision and Applications*, vol. 5, pp. 17–34, 2002.

[107] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. Romney, J. B. Zimmerman, and K. Zuiderveld, "Adaptive Histogram Equalization and its Variations," *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355–368, 1987.

[108] Z. Liu and S. Sarkar, "Effect of Silhouette Quality on Hard Problems in Gait Recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 35, pp. 170–183, 2005.

[109] J. Liu, N. Zheng, and L. Xiong, "Silhouette Quality Quantification for Gait Sequence Analysis and Recognition," *Signal Processing*, vol. 89, no. 7, pp. 1417–1427, July 2009.

[110] D. Maio, D. Maltoni, R. Cappelli, J. Wayman, and A. Jain, "FVC2004: Third Fingerprint Verification Competition," *International Conference on Biometric Authentication (ICBA)*, pp. 1–7, July 2004.

[111] E. Tabassi, C. Wilson, and C. Watson, "NIST Fingerprint Image Quality," Nistir7151, National Institute of Standards and Technology (NIST), 2004.

[112] S. Yu, D. Tan, and T. Tan, "A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition," *Internation Conference on Pattern Recognition (ICPR)*, vol. 4, pp. 441–444, 2006.

[113] Md. A. Hossain, Y. Makihara, J. Wang, and Y. Yagi, "Clothing-invariant Gait Identification Using Part-based Clothing Categorization and Adaptive Weight Control," *Pattern Recognition*, vol. 43, no. 6, pp. 2281–2291, June 2010.

[114] K. Kim, T.H. Chalidabhongse, D. Harwood, and L. Davis, "Background Modeling and Subtraction by Codebook Construction," *IEEE International Conference on Image Processing*, 2004.

[115] C. Stauffer and W. Grimson, "Adaptive Background Mixture Models for Real-time Tracking," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 246–252, 1999.

[116] X. Chen, Z. He, J. Keller, D. Anderson, and M. Skubic, "Adaptive Silhouette Extraction in Dynamic Enviornments Using Fuzzy Logic," *IEEE International Conference on Fuzzy Systems*, pp. 236–243, 2006.

[117] J. Jacques, C. Silveira, C.R. Jung, and S.R. Musse, "Background Subtraction and Shadow Detection in Grayscale Video Sequences," *IEEE Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, pp. 189–296, 2005.

[118] M. Goffredo, J.N. Carter, and M.S. Nixon, "Front-view Gait Recognition," *IEEE Conference on Biometrics: Theory Applications and Systems (BTAS)*, pp. 1–6, 2008.

[119] A. Kale, A.K. Roy-Chowdhury, and R. Chellappa, "Towards a View Invariant Gait Recognition Algorithm," *IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 143–150, July 2003.

[120] S. Sarkar, P.J. Phillips, Z. Liu, and I.R. Vega, "The HumanID Gait Challenge Problem: Datasets, Performance, and Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, pp. 162–177, 2005.

[121] D.S. Matovski, M. Nixon, S. Mahmoodi, and J.N. Carter, "The Effect of Time on the Performance of Gait Biometrics," *IEEE Conference on Biometrics: Theory Applications and Systems (BTAS)*, pp. 1–6, September 2010.

[122] A. Bobick and A Johnson, "Gait Recognition Using Static, Activity-Specific Parameters," *IEEE Transactions on Computer Vision and Pattern Recognition*, pp. 423–430, 2001.

[123] A. Criminisi, A. Zissermann, and L. Van Gool, "A New Approach to Obtain Height Measurement from Video," *Proceedings of SPIE*, vol. 3576, pp. 227–238, 1998.

[124] J. Kent, *New Directions in Shape Analysis*, Wiley, 1992.

[125] M. Stegmann and D. Gomez, "A Brief Introduction to Statistical Shape Analysis," March 2002.

[126] R. Gross and J. Shi, "The CMU Motion of Body (MoBo) Database," Tech. Rep. CMU-RI-TR-01-18, Robotics Institute, Pittsburgh, PA, 2001.

[127] J.D. Shutler, M.G. Grant, M.S. Nixon, and J.N. Carter, "On a Large Sequence-Based Human Gait Database," *4th International Conference on Recent Advances in Soft Computing*, pp. 66–71, 2002.

[128] P. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. Bowyer, "Baseline Results for the Challenge Problem of Human ID Using Gait Analysis," *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR '02)*, 2002.

[129] S. Yu, D. Tan, and T. Tan, "A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition," *Proc. 18th International Conference on Pattern Recognition (ICPR06)*, pp. 441–444, August 2006.

[130] Chinese Academy of Sciences, "CASIA Night Gait Dataset (Dataset C)," http://www.cbsr.ia.ac.cn/english/Gait

[131] B. DeCann and A. Ross, "Gait Curves for Human Identification, Backpack Detection, and Silhouette Correction in a Nighttime Enviornment," *SPIE Conference on Biometric Technology for Human Identification VII*, April 2010.

[132] N. Akae, A. Mansur, Y. Makihara, and Y. Yagi, "Video from Nearly Still: An Application to Low Frame-rate Gait Recognition," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1537–1543, 2012.

[133] R. Tanawongsuwan and A. Bobick, "Modeling the Effects of Walking Speed on Appearance-Based Gait Recognition," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 783–790, 2004.

[134] Y Makihara, A. Tisuji, M.A Hossain, K. Sugiura, A. Mori, and Y. Yagi, "The OU-ISIR Gat Database Comprising the Treadmill Dataset," *IPSJ Transactions on Computer Vision and Applications*, vol. 4, pp. 53–62, 2012.

[135] X. Yang, Y. Zhou, T. Zhang, G. Shu, and J. Yang, "Gait Recognition Based on Dynamic Region Analysis," *Signal Processing*, vol. 88, no. 2, pp. 316–322, 2008.

[136] Brendan F. Klare, *Heterogeneous Face Recognition*, Phd. thesis, Michigan State Univeristy, 2012.

[137] S. Palla, S. Chikkerur, V. Govindaraju, and P. Rudravaram, "Classification and indexing in large biometric databases," *Biometric Consortium Conference*, September 2004.

[138] S.H. Cho, J.M. Park, and O.Y. Kwon, "Gender Differences in Three Dimensional Gait Analysis Data from 98 Healthy Korean Adults," *Clinical Biomechanics*, vol. 19, no. 2, pp. 145–152, 2004.

[139] L. Lee and E. Grimson, "Gait Analysis for Recognition and Classification," *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 148–155, 2002.

[140] G. Huang and T. Wang, "Gender Classification Based on Fusion of Multi-View Gait Sequences," in *Computer Vision*, pp. 468–471. Springer-Berlin-Heidelberg, 2007.

[141] 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*, vol. 38, no. 2, pp. 145–155, March 2008.

[142] S. Samangooei and M. Nixon, "Performing content-based retrieval of humans using gait biometrics," *Multimedia Tools and Applications*, vol. 49, no. 1, pp. 195–212, 2010.

[143] E. Watelain, F. Barbier, P. Allard, A. Thevenon, and J.C. Angu, "Gait Pattern Classification of Healthy Elderly Men Based on Biomechanical Data," *Archives of Physical Medicine and Rehabilitation*, vol. 81, no. 5, pp. 579–586, May 2000.

[144] G.B. Coleman and H.C. Andrews, "Image segmentation by clustering," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 773–785, 1979.

[145] T.N. Pappas, "An adaptive clustering algorithm for image segmentation," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 901–914, 1992.

[146] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery, and W.L. Ruzzo, "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, vol. 17, no. 10, pp. 977–987, 2001.

[147] A.D. King, N. Przulj, and I. Jurisica, "Protein complex prediction via cost-based clustering," *Bioinformatics*, vol. 20, no. 17, pp. 3013–3020, 2004.

[148] R. Agrawal, J. Gehrke, J. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," *ACM*, vol. 27, no. 2, pp. 94–105, 1998.

[149] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping Multidimensional Data*, pp. 25–71. Springer Berlin Hidlberg, 2006.

[150] S.P. Loyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[151] S.P. Loyd, "Cluster analysis of multivariate data: Efficiency versus interpretability of classifications," *Biometrics*, vol. 21, pp. 768–769, 1965.

[152] B. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[153] A. Jain, R. Bolle, and S. Pankanti, *Biometrics: Personal Identification in Networked Society*, Heidelberg, 1999.

[154] UIDAI, "Role of Biometric Technology in Aadhaar Enrollment," Tech. Rep., Government of India (GoI), January 2012.

[155] A.J. Mansfield and J. L. Wayman, "Best Practices in Testing and Reporting Performance of Biometric Devices," Tech. Rep., UK Govt. Biometrics Working Group, 2002.

[156] J. Wayman, A. Jain, D. Maltoni, and D. Maio, *Biometric Systems: Technology Design and Performance Evaluation*, Springer-Verlag, 2005.

[157] L. Hong, A.K. Jain, and S. Pankanti, "Can Multibiometrics Improve Performance?," *Proceedings of AutoID*, pp. 59–64, October 1999, Summit, NJ, USA.

[158] A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics*, Springer, 2006.

[159] A. Jain, A. Ross, and S. Pankanti, "Biometrics: A Tool for Information Security," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 125–143, June 2006.

[160] A. Ross, A. Rattani, and M. Tistarelli, "Exploiting the Doddington Zoo Effect in Biometric Fusion," *3rd IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–7, September 2009, Washington DC, USA.

[161] A.K. Jain, K. Nandakumar, and A. Ross, "Score Normalization in Multimodal Biometric Systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, December 2005.

[162] J. Wayman, A. Possolo, and A. Mansfield, "Fundamental Issues in Biometric Performance Testing: A Modern Statistical and Philosophical Framework for Uncertainty Assessment," *First International Biometric Performance Conference (IBPC)*, March 2010, Gaithersburg, MD, USA.

[163] S.C. Dass, Y. Zhu, and A.K. Jain, "Validating a Biometric Authentication System: Sample Size Requirements," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1902–1913, 2006.

[164] J. Wright, A. Yang, A. Ganesh, S Shankar, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[165] D.O. Gorodnichy, "Multi-order Analysis Framework for Comprehensive Biometric Performance Evaluation," *SPIE Conference on Defense, Security and Sensing. DS108: Biometric Technology for Human Identification*, April 2010.

[166] D.O. Gorodnichy, "Multi-Order Biometric Score Analysis Framework and Its Application to Designing and Evaluating Biometric Systems for Access and Border Control," *IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, pp. 44–53, 2011.

[167] N. Yager and T. Dunstone, "The Biometric Menagerie," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 220–230, 2010.

[168] J. Wu and C. Wilson, "Nonparametric Analysis of Fingerprint Data on Large Data Sets," *Pattern Recognition*, vol. 40, no. 9, pp. 2574–2584, 2007.

[169] A. Gyaourova and A. Ross, "Index Codes for Multibiometric Pattern Retrieval," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 7, no. 2, pp. 518–529, April 2012.

[170] R. Gadde, D. Adjeroh, and A. Ross, "Indexing Iris Images Using the Burrows-Wheeler Transform," *Proc. of IEEE International Workshop on Information Forensics and Security (WIFS)*, December 2010.