

MULTI-MEDIA PERSONAL IDENTITY VERIFICATION

By

ALAN LAWRENCE HIGGINS

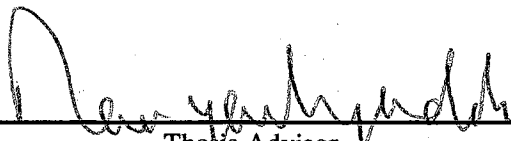
Bachelor of Arts  
University of California, San Diego  
1974

Master of Science  
University of California, San Diego  
1977

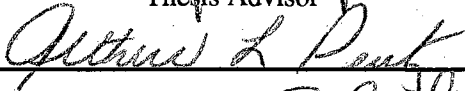
Submitted to the Faculty of the  
Graduate College of the  
Oklahoma State University  
in partial fulfillment of  
the requirements for  
the Degree of  
DOCTOR OF PHILOSOPHY  
July, 1996

MULTI-MEDIA PERSONAL IDENTITY VERIFICATION

Thesis Approved:



Thesis Advisor









Dean of the Graduate College

## PREFACE

The past eleven years of my involvement in the field of voice recognition have seen enormous progress. As voice recognition systems have moved from the laboratory to the real world, my own activities in this area have become ever more market focused. To sustain the market for voice recognition and other biometric security products, further technological advances must address the difficult problems presented by future security applications.

Voice verification is rapidly being incorporated into applications involving telephone transactions. The security provided by these systems, and the trust that is placed in them, is limited by their accuracy and by their resistance to counterfeiting. Overcoming these problems will be a giant step toward enabling secure electronic commerce. The confluence of computer telephony, video teleconferencing, and network computing will provide resources that can be used to this end. This dissertation, and the concept of multi-media personal identity verification, was conceived with these ideas in mind.

This work represents an extension of several years of research and development that was conducted in collaboration with my colleagues and friends, Larry Bahler and Jack Porter. Larry invented the SQNN algorithm for voice recognition, and Jack developed a mathematical rationale to understand its excellent performance. Jack also contributed to the Appendix of this dissertation. Without Jack and Larry, it is unlikely that this work would have been started, much less completed.

The inspiration to undertake a doctoral program after years of work in industry came from Dr. Joe Campbell, an OSU graduate who had been faced with similar circumstances. Joe showed me that it was possible to overcome the hurdles of returning to an academic environment.

The biggest hurdle was in moving to Stillwater, Oklahoma for two semesters while completing my course work. Although my work there was interesting and challenging, I found it quite difficult being separated from my wife, Rebecca, and our two sons, Doug and Alex, who remained in San Diego. The main burden of the move fell on Rebecca, who, in addition to maintaining her dental practice, had to deal single-handedly with the logistics of schools, sporting events, music lessons, Boy Scouts and Cub Scouts, etc. I am grateful to Rebecca for giving me the opportunity to pursue my studies at OSU.

Finally, I thank Professor Yarlagadda, my thesis advisor, for his encouragement, help, and guidance. During my residence in Stillwater and since then, we have had many interesting and useful discussions, and it has been a great pleasure working with Dr. Y.

This work was supported in part by Sandia National Laboratories (Contract #AM-3325) and by the U. S. Army Research Office (Contract #DAAH04-95-1-0463).

## TABLE OF CONTENTS

Section	Page
1. INTRODUCTION	1
1.1 Multi-Media Personal Identity Verification	1
1.2 Overview of Dissertation	2
1.3 Previous Work	5
1.4 Organization of Dissertation	5
2. PIV METHODS AND APPLICATIONS	6
2.1 Passwords and Personal Tokens	6
2.2 Biometric Verification	7
2.3 Law Enforcement	8
2.3.1. Witness Identifications	9
2.3.2. Fingerprints	10
2.3.3. DNA Identification	12
2.4 Local Access Control	13
2.4.1. Retinal Scan	15
2.4.2. Iris Scan	15
2.4.3. Hand Geometry	16
2.5 Network Access Control	16
2.6 Summary	17
3. REVIEW OF RELEVANT LITERATURE	18
3.1 Voice Verification	18
3.1.1. Text-Dependent Approaches	19
3.1.2. Text-Independent Approaches	22
3.2 Face Verification	26
3.2.1. Feature-Based Approaches	27
3.2.2. Template-Based Approaches	29
3.3 Multiple Media	32
3.4 Data Fusion	34
3.5 Biometric Data Protection	37
3.6 Summary	42

Section	Page
4. MULTI-MEDIA PIV SYSTEM DESIGN	44
4.1 System Requirements	44
4.2 Concept of Operation	45
4.3 Counterfeiting	47
4.4 Anti-Counterfeiting	48
4.5 Summary	49
5. PROBABILISTIC MODELING OF INDIVIDUALS	50
5.1 Introduction	50
5.2 Acceptance Criterion	51
5.3 An Example: Height as Evidence of Identity	53
5.4 Probability Density Estimation for Densely Sampled Populations	57
5.5 Probability Density Estimation for Sparsely Sampled Populations	59
5.6 Dimensionality Estimation	65
5.6.1. An Extension of the Method of Pettis, et. al	66
5.7 Likelihood Ratio Estimation	68
5.8 Multiple Models Per Individual	69
5.9 Summary	70
6. EXPERIMENTAL DATA	73
6.1 Introduction	73
6.2 Equipment Setup	73
6.3 Subjects	75
6.4 Prompting	75
6.5 Initial Data Processing	75
6.6 Inventory of Sessions	77
6.7 Subjective Observations	78
6.8 Summary	79
7. VOICE DATA FEATURE EXTRACTION	80
7.1 Introduction	80
7.2 Voice-Only PIV Algorithm	80
7.3 Signal Processing	81
7.3.1. Spectral Analysis	81
7.3.2. Silence Frame Pruning	84
7.3.3. Blind Deconvolution	84
7.3.4. Frequency Differencing	85
7.4 Voice Comparison	85
7.5 Summary	85

Section	Page
8. VIDEO DATA FEATURE EXTRACTION	86
8.1 Introduction	86
8.2 Video-Only PIV Algorithm	86
8.3 Manual Location of Faces	88
8.4 Signal Processing	89
8.4.1. Histogram Stretching	89
8.4.2. Gradient Filtering	90
8.5 Automatic Location of Faces	91
8.6 Face Comparison	97
8.7 Summary	97
9. ANALYSIS AND RESULTS	99
9.1 Introduction	99
9.2 Likelihood Scoring Versus Likelihood Ratio Scoring	99
9.3 Intrinsic Dimensionality	101
9.4 ROC Performance Measurement	101
9.4.1. Integrated Error Measure	104
9.5 Test Procedure	105
9.6 Voice-Only ROC Data	107
9.7 Face-Only ROC Data	109
9.8 Fusion of Voice and Face Data	111
9.9 Summary	115
10. CONCLUSION	118
10.1 Summary of Accomplishments	119
10.2 Suggested Future Research	121
BIBLIOGRAPHY	126
APPENDIX: Density versus NN Distance for Gaussian PDF	135
A.1 Conditional Expectation of $p_x$	136
A.2 Density of $p_x$	137
A.3 Conditional Density of $d_{NN2}$	139
A.4 Numerical Evaluation of $E_{p_x   d_{NN2}}$	140
A.5 Approximation of Median NN Distance	141
A.6 Interpretation	141
VITA	145

## LIST OF TABLES

Table		Page
2.1	Comparison of PIV Systems (from Maxwell 1987)	11
5.1	Cumulative Height Distribution of Adult Males	45
6.1	Inventory of Sessions by Subject	65
7.1	Filterbank Design Data	70
9.1	Example of ROC Computation	88
9.2	Values of $\alpha$ and $\beta$ for Voice and Video Data	96
9.3	Summary of Key Results	99



## LIST OF FIGURES

Figure		Page
2.1	Example of a Biometric Measurement	6
3.1	Illustration of Data Fusion	29
3.2	System for Secure Transmission of Biometric Models	34
4.1	User's View of PIV System	37
4.2	System Interaction with Network	38
5.1	Likelihood Functions for $H_X = 71$	46
5.2	Log Likelihood Ratio Function for $H_X = 71$	46
5.3	Log Probability Density versus Nearest-Neighbor Distance for $N(0, I_2)$	49
5.4	Log Probability Density versus Nearest-Neighbor Distance for $N(0, I_{13})$	51
5.5	Comparison of Log Likelihood Estimators for $N(0, I_{13})$	53
5.6	Approximate Linear Relationship of $\ln r_{ki}$ Versus $\ln k$ , with Slope Equal Reciprocal of Local Dimensionality	57
6.1	Illustration of Experimental Setup	62
6.2	Script Used for Prompting Subjects	63
6.3	Initial Data Processing	63
6.4	Video and Audio Digitization Parameters	64
6.5	Example Image Frame from MA-1	66
7.1	Data Flow Diagram of Voice Data Processing	68
8.1	Data Flow Diagram of Video Data Processing	74
8.2	3 x 3 Neighborhood Used by Sobel Operator	76
8.3	Example Pre-Processed Frame from MA-1	77
8.4	Estimation of Face Box Position	78
8.5	Average Head Box Image after 1, 2, and 3 Iterations	79
8.6	The 12 Most Significant Eigenfaces of the OSU Data	81

Figure		Page
9.1	Log Likelihood Scores for Each Test Session	85
9.2	Example ROC Curve	87
9.3	Test Procedure for False Acceptance Measurement	90
9.4	ROC Performance of SQNN Voice-Only PIV Algorithm	91
9.5	Comparison of INN Versus SQNN Voice-Only PIV Algorithms	92
9.6	ROC Performance of Face-Only PIV Algorithm	93
9.7	ROC Performance of Face-Only PIV Algorithm Using One Test Frame and Various Numbers of Training Frames	94
9.8	Scatter Plot of $LLR_{voice}$ Versus $LLR_{face}$	96
9.9	Scatter Plot Showing Decision Boundary	98
A.1	Illustration of Density Estimation Using NN Distances	114
A.2	Sample Density and Density of $p_x$	117
A.3	Density Estimators for $v = 3$	121
A.3	Density Estimators for $v = 13$	122
A.3	Density Estimators for $v = 13, N = 1000$	123

## ACRONYMS

EER	equal-error rate
FA	false acceptance
FR	false rejection
GMM	gaussian mixture model
HMM	hidden Markov model
IEM	integrated error measure
INN	interpolated nearest neighbor
LLR	log likelihood ratio
LR	likelihood ratio
NN	nearest neighbor
PDF	probability density function
PIV	personal identity verification
ROC	receiver operating characteristic
SQNN	squared nearest neighbor

## CHAPTER 1

### INTRODUCTION

The ongoing de-centralization of computer resources and information raises concerns about information privacy and security. Data that was once physically protected in locked and guarded buildings is, in many cases, now accessible at remote sites through computer networks. It is commonplace, for example, for business people to access proprietary corporate data from airports, hotel rooms, or customer's facilities. The protections that are typically in place are vulnerable to circumvention through stolen passwords or other means. Personal Identity Verification (PIV) systems offer a possible solution. PIV systems use measured physical, or biometric, evidence to establish the authenticity of a person's claimed identity. We refer to the person making the claim as the user, and to the person whose identity is claimed as the claimant. PIV tests the hypothesis that the user and the claimant are one in the same.

#### **1.1 Multi-Media Personal Identity Verification**

The focus of this research is on the use of voice and facial image as biometrics for personal identity verification. An advantage of these biometrics over others is that they can be applied conveniently in the type of application described above. The feasibility of measuring and processing voice and facial images is increased by the recent availability of low-cost multi-media computers incorporating audio, video, and digital signal processing capabilities. We refer to the concept of using multiple biometric

attributes to authenticate a user's claimed identity as *Multi-Media Personal Identity Verification*.

## 1.2 Overview of Dissertation

To provide a motivation for the work reported here, the history of PIV methods and applications was investigated. The results of this investigation support the notion that providing information security over networks is a logical direction in which to extend PIV technology. For reasons alluded to above, voice and facial images are among the most promising biometrics for this purpose. Considerable progress may be needed, however to develop PIV algorithms that are sufficiently robust with respect to variations in equipment and environmental conditions. Feature extraction and statistical decision algorithms for voice and face data are critical elements of the envisioned PIV system.

A literature survey was conducted to assess the state of the art in these areas, among others. Numerous studies are cited involving either voice verification or face verification, separately. The approaches applied to voice verification are generally quite different from those applied to face verification. In part, this is due to the fact that all reported work to date on face verification has been based on still images, which lack the temporal dimension associated with voice.

The recent success of Brunelli and Falavigna [1] in combining voice and still facial images proves that voice and facial appearance carry separate information about the identity of the subject that can be mutually reinforcing. Despite this success, facial movements and expressions remain a source of errors, rather than a source of information, to systems restricted to processing still images. It should therefore be possible to achieve better verification

accuracy using *image sequences* than still images. This conclusion is supported by studies of human performance [2].

Verification using voice and image sequences poses interesting theoretical problems. Both voice and facial appearance are influenced by factors other than the identity of the subject. To this extent, they are reasonably regarded as random, as opposed to deterministic, observations of underlying attributes that characterize the subject. The decision to accept or reject the claimed identity is made with minimum error probability according to Bayes' decision rule. Implementation of Bayes' decision rule requires estimates for the likelihood of the observed feature vector sequence assuming that the claim is true, and assuming that it is false. The likelihoods depend in turn on the local probability densities at the observed feature vector points under the same assumptions. Estimation of these densities is made difficult by the high dimensionalities of the feature spaces and by the limited availability of training data. In some cases (e.g., facial images), feature space dimensionality may actually exceed the number of available training samples.

The nearest-neighbor (NN) method is a well-known approach to nonparametric density estimation. According to the NN method, the log of local probability density is related through an affine transformation to the log of Euclidean distance between the test sample and the nearest training sample. It is shown that the NN method fails in high-dimensional spaces. Empirical investigations in cases of interest indicate that local log probability density is more closely related to the *square* of nearest-neighbor distance than to its logarithm. Estimation of density based on this conjectured relation is referred to here as *SQNN estimation*. Analytic support for SQNN estimation is provided for the special case of multivariate Gaussian densities.

PIV algorithms were developed for voice and facial image sequences, separately. The algorithms differ in their methods of feature extraction, but are otherwise identical. Both algorithms implement an approximation to Bayes' decision rule incorporating SQNN estimation.

An experimental database was collected in which subjects were filmed using a camcorder while reading a set of short, prompted phrases. Each subject was filmed on multiple occasions, providing sufficient data for modeling and simulated verification. Both the audio and video data were digitized and stored in computer files. This data was used for development and evaluation of the PIV algorithms.

Performance of the PIV algorithms was evaluated by simulating a large number of verification trials in which the claimed identity was either valid or invalid. Accuracy of the voice and face PIV algorithms was measured as a function of the duration of the input data. These results demonstrate the superiority of image sequences over still images. Given 10-second segments of input data, both algorithms accept at least 90% of valid claims while rejecting all invalid claims. Analysis of the false-rejection errors reveals that in most cases, they occur when the test conditions differ in some obvious way from the training conditions used to model the subject. For example, the subject may wear glasses during a test session but not during training sessions. Similar phenomena affect the voice data. It is argued that differences of this sort invalidate the subject's model, and therefore the decision derived from it. No examples were found in the experimental database of sessions that were simultaneously anomalous with respect to both the voice and face models. Therefore, perfect verification performance is obtained by accepting the identity claim if the data passes *either* the voice test *or* the face test, and rejecting it otherwise.

### **1.3 Previous Work**

This work extends previous studies by the author in collaboration with others [3, 4] (see Acknowledgements). The voice PIV algorithm based on "SQNN" estimation (to be defined later) is substantially the same as that described in [4]. "INN" estimation was developed as an extension of the work of Pettis, et. al [5]. One important step in the face PIV algorithm involves the use of a subject-independent face model proposed by Turk and Pentland [6].

### **1.4 Organization of Dissertation**

Chapter 2 examines methods and applications of personal identity verification. A historical perspective is taken, and the focus is on methods that employ biometrics other than voice and facial image. Chapter 3 surveys the literature in the areas of voice and face verification, as well as data fusion and protection of biometric data, two other key components of a multi-media PIV system. In Chapter 4, the concept of operation of a multi-media PIV system is presented. The approach to probabilistic modeling of individuals, including definition and rationalization of the SQNN estimation method, is presented in Chapter 5. The experimental database is described in Chapter 6. Chapters 7 and 8 provide details of the voice and face PIV algorithms, respectively. Results of applying the simulated PIV algorithms to the experimental database are presented in Chapter 9. Conclusions of this work are given in Chapter 10.



## CHAPTER 2

### PIV METHODS AND APPLICATIONS

It is said that there are three approaches to authenticating an individual's claimed identity: "What you know, What you have, and What you are". The first two approaches refer to the use of passwords and personal tokens as discussed in Section 2.1. The third approach refers to biometric verification, described in Section 2.2. In Sections 2.3, 2.4, and 2.5, applications of PIV systems are divided into three broad categories: law enforcement, physical access, and network access. Examples of each category are given and the differences between categories are presented. No attempt is made to exhaustively catalog or differentiate specific applications. PIV technologies appropriate for each category of application are described. Conclusions are presented in Section 2.6.

#### **2.1 Passwords and Personal Tokens**

Passwords and personal tokens such as credit cards are commonly used as evidence of personal identity. Both are vulnerable to circumvention by unauthorized individuals because their security hinges on the restricted knowledge or possession of some item. Computer break-ins to sensitive DOD, DOE and NASA facilities involving penetration of password PIV systems have been widely publicized [7]. When passwords are computer-generated, they tend to be difficult to remember, adversely affecting user acceptability. When users select their own passwords, they favor words that are easy to remember (and for hackers to guess). The pervasive use of fraudulent identification

items (including credit cards, driver's licenses, etc.) is also well documented [8]. Annual credit card losses attributed to fraud are estimated to be over one percent of total credit card sales [9]. Because of these problems, there is a need for biometric verification methods to be used in place of or in addition to passwords and personal tokens. This research focuses on biometric verification.

## 2.2 Biometric Verification

Password- and token-based PIV systems associate with each user a unique, discrete pattern that must be matched exactly for the claim to be accepted. Biometric PIV systems, on the other hand, involve measurements of human physical attributes that vary continuously over multiple dimensions. Uncontrolled factors, including the subject's behavior, may influence these attributes. In addition, measurement errors of various types may be present.

Bertillon's system of "Anthropometric Indications", published in 1889, consists of a set of length measurements of the head and limbs that was used for positive identification of known criminals. An illustration of measurement of the right ear, as prescribed by Bertillon [10], is shown in Figure 1.2.



**Figure 2.1: Example of a Biometric Measurement**

Biometric verification requires each user to participate in one or more *enrollment sessions*, in which normative measurements of the relevant attributes are established. These measurements are used to form a *model* of the user. Similar measurements, made later during *verification sessions* or *test sessions*, are evaluated using the claimant's model to determine the validity of the claim.

### **2.3 Law Enforcement**

Law enforcement applications of PIV involve use of evidence at a crime scene to identify the perpetrator. The perpetrator is non-cooperative in providing evidence and may actively seek to hide his identity. Evidence of various types may be discovered through the efforts of a human investigator. The initial stages of the investigation may use criteria such as motive, access, etc. to reduce the search to a small number of suspects. For each suspect, the hypothesis that the evidence was produced by that suspect is tested by the PIV

system. Processing time is not critically important. Ultimately, the strength of evidence connecting the crime with the suspect is decided by a jury.

Types of evidence commonly involved in law enforcement applications include witness identifications, fingerprints, and DNA present in human cells. These are described in the following sections.

### **2.3.1. Witness Identifications**

Witness identification of suspects is routinely admitted as evidence in court. The constitutionality of legal proceedings involving witness testimony was established in three landmark cases: *United States v. Wade*, *Gilbert v. California*, and *Stovall v. Denno*. In all three cases, both visual and voice information was used in identification [11]. Under ideal conditions, visual identification is considered more reliable than voice identification. Although the information content of visual and auditory media depends on factors such as lighting and background noise levels, observers apparently do not modify their decision strategy based on these factors [12]. In simulated police lineups, subjects are about 9 times more likely to be correct than incorrect [13], and they tend to be more accurate in rejecting non-targets than in affirmatively identifying target individuals. It is generally reported that observers' subjective confidence in their decisions has little or no correlation with objective accuracy. Zavala [2] reports that the accuracy of witnesses in identifying suspects is improved by the use of movie clips as opposed to still photographs.

Face recognition systems are currently in use and under continuing development for mug-shot retrieval. The purpose of these systems is to rapidly access individuals in a database whose faces match a verbal description. In FACES (Facial Analysis Comparison and Elimination System) [14], used by the

British police force, a digitizing pad and stylus is used to locate 37 "cardinal points" on a photograph of each known face. The locations of the points are converted to a set of 21 linear and area parameters. The values of these parameters are used as coordinates into a space upon which each known face is represented as one point. Verbal descriptions of the perpetrator are converted to numerical coordinates in this space. Goldstein, et al. [15] estimate that about 5.4 features must be accurately known to specify one individual out of a population of 256, and that the number of features required increases with the log of the population size. The main difficulty in these systems is that witnesses rarely remember more than 3-4 features [15], and the features they remember tend to be poor discriminators [16].

### **2.3.2. Fingerprints**

Papillary ridges, or fingerprints, on the surface of the fingertips have been used to identify individuals since the late 19th century. Fingerprints are unique to each individual, including identical twins, because they depend on the chaotic initial conditions of embryonic development. In 1901, fingerprint matching was officially introduced at Scotland Yard using a classification system developed by Sir Edward Richard Henry. The Henry System was adopted by the FBI and other organizations, and remains in use throughout the world today.

The Henry System classifies ten-print records into one of approximately 1000 types [17]. When searching for a matching ten-print record, the search is restricted to consider only file prints of the same type. Within each fingerprint, a number of points are located (about 12 on average) where a ridge either ends or bifurcates, becoming two ridges. These points are called

*minutiae* points. Matches within prints of the same type are established from the relative positions of the minutiae points.

The most reliable method of locating minutiae points is by a human examiner. A method in current practice for latent prints is the following [18]. The prints are photographically enlarged using a 5:1 scale. Tracing paper is placed over the photograph, and the ridges are traced by hand. The origin of a 2-dimensional coordinate system is established at the center of the outermost joint at the end of the finger. A digitized image of the tracing is created and stored as a computer file. The examiner then enters the coordinates of the minutiae points using a digitizing tablet.

Automated methods have been developed for location and matching of minutia points in ten-print records [19-21]. This technology was incorporated into the FBI's Automated Fingerprint Identification System (AFIS), which became operational in 1983 [22]. The use of AFIS is growing rapidly. It is estimated [23] that by the year 2000, the FBI will process 61,000 fingerprint checks per day against a file of 34 million prints. Manual editing is still required for latent prints [24].

The results of a search include a list of up to ten individuals whose file prints best match the unknown, and scores associated with the matches. Scores range from 0 to 9999, higher scores generally indicating better matches. Scores are influenced by factors including the number of minutia points and how tightly they are clustered. In the case of latent prints, scores are also influenced by the quality of the print and the alignment of the axes used as the reference for the minutia coordinates. Because of these factors, score values themselves are less indicative of the likelihood of a match than differences between score values. The best quantitative evidence of a match is a large gap between the scores of the first and second candidates [18, 22].

In recent work, fingerprint technology has been applied to local access control applications (see Section 2.4). A fully-automatic fingerprint based PIV system using an optical reader is described by Takeda et al. [25]. Other recent work includes improved algorithms for minutia detection [26] and improved measurement devices [27].

### **2.3.3. DNA Identification**

The hereditary characteristics individuals are transmitted from generation to generation by means of *genes*. Genes are represented physically by *chromosomes*, which are present within the nucleus of every cell of the human body. Chromosomes are made up of deoxyribonucleic acid, or DNA, a complex organic molecule that contains a sequence of simpler molecules called *nucleotides*. There are four different nucleotides, usually abbreviated A, G, T, and C, which act as symbols in the *genetic code*. Parts of the sequence of nucleotides on a chromosome, called *exons*, are used to encode the genes. Other parts, called *introns*, are long repeating sequences that may occur at any point on the chromosome. Although introns are genetically insignificant, their lengths (number of nucleotide sequence repetitions) are unique to each individual, and are the basis of DNA identification [28].

DNA identification uses a process known as *gel electrophoresis*. In this process, a set of specific introns is extracted from the DNA sample and added to a gel placed between two electrodes. When electrical current flows between the electrodes, the negatively charged DNA fragments migrate slowly toward the positive electrode. Shorter molecules move more quickly through the viscous gel than longer ones. The current is stopped before the migration is complete, and a radioactive probe is used to expose the positions of the molecules on a photographic plate. The resulting photograph shows a series

of bands along a "lane" between the two electrodes. Two or more DNA samples can be compared by forming multiple lanes in a single gel electrophoresis run.

The accuracy of DNA identification depends on the particular introns used in the process, and on the probability distributions of the lengths of these introns in the general population. As an example, if ten introns are used and each possible intron length is shared by ten percent of the population, the probability of two randomly selected people having the same DNA signature (assuming each intron is independent of the others) is  $1 \times 10^{-10}$ . DNA identification requires a minimum of several weeks and is quite costly. For this reason, it is only used in connection with very serious crimes.

#### **2.4 Local Access Control**

Local access control applications involve transactions that take place at public places such as commercial offices, kiosks, or point-of-sale terminals. The PIV equipment at each site is shared by many users. The equipment is physically protected to minimize to the possibility of tampering. Verification must be completed within a few seconds. Subjects are cooperative and there is little or no human supervision. Examples include:

- Access control to rooms or buildings
- ATM Machines
- Credit Card Verification
- Electronic Benefits Transfer
- Ticketless Travel

Biometric based PIV systems for local access control were first developed and tested in the 1970s [29, 30]. These stand-alone systems incorporated microprocessors and were fully automated. A comparison of



various PIV systems for use in a local access control application was conducted by Sandia Laboratories [31]. The results are shown in Table 2.1.

	Hand Geometry	Voice (AT&T)	Voice (Voxtron)	Retinal Scan	Finger Print
Verify Attempts	1491	3206	2564	3082	3384
Imposter Attempts	4055	3415	3795	5027	4849
Enrollment Time	54 sec	18 sec	144 sec	126 sec	114 sec
Verification Time	4.4 sec	8.8 sec	10.1 sec	7.5 sec	9.8 sec
False Rejections	0.9%	12.5%	17.0%	10.8%	9.1%
Fales Acceptances	0.4%	0.1%	0.6%	0%	0%

**Table 2.1: Comparison of PIV Systems (from Maxwell 1987)**

The development of optically-based fingerprint readers (as opposed to ink pads) has enabled fingerprint technology (See Section 2.3.2) to be applied to physical access control applications. Electronic benefits transfer (EBT), whereby government benefits such as food stamps are dispensed electronically at point-of-sale terminals, is an important current application. A 1995 study by the General Accounting Office concluded that "electronic fingerprinting may be the most viable option for deterring fraud in an EBT environment". The Federal EBT Task Force recommended that EBT with fingerprinting be used for all disbursements of social security, military pensions, civil service retirement, food stamps, and Aid to Families with Dependent Childeren by 1999. The total of these payments in 1994 was \$433 billion, of which it is estimated that 10% was fradulent [32].

#### **2.4.1. Retinal Scan**

Retinal identification of individuals is based on the pattern of blood vessels in the subject's eye. According to a 1935 medical finding, no two retinal patterns are identical. The patterns are disrupted only by eye surgery or serious eye injuries such as detached retina. Either or both eyes can be used, depending on the required accuracy. Retinal scanners have been commercially available since 1985. To use a retinal scanner, the subject looks into the eyepiece, which is similar to the eyepiece of a microscope. The distance between the eye and the scanner must be very small. A weak infrared light is directed through the pupil to the retina, and the reflected pattern is observed using a CCD camera. The pattern is stored digitally using about 35 bytes. The retinal scan is one of the most accurate available biometrics [33]. A scanner instrument, excluding supporting computer equipment, costs about \$3500.

#### **2.4.2. Iris Scan**

The iris, or colored part of the eye, has also been found to be a reliable identifier of individuals. Viewed closely, the iris contains numerous features such as pits and striations. Like fingerprints, these details depend on the initial conditions of embryonic development, and are therefore unique even among identical twins. The principles and technology of personal identification using iris features is patented by two ophthalmologists, L. Flom and A. Safir [34], and described in [35]. The iris scan device is available commercially, and can be used at a distance of up to 45 cm.

### **2.4.3. Hand Geometry**

Hand geometry identification is based on the three-dimensional shape of the subject's hand. Hand shape is a stable individual characteristic over time, and the performance of hand geometry devices is reported to be unaffected by dirt or cuts on the hand. A hand geometry reader includes a base plate with metal alignment pins which guide the placement of the hand. Simultaneous top- and side-view pictures are taken using a digital camera. Features are extracted automatically, including hand height, finger length and width, and distance between knuckles. These features are transformed and compressed to form a 9-byte digital pattern. Enrollment for an individual is accomplished by averaging the 9-byte patterns from three readings. Equal-error rates for hand geometry are now reportedly about 0.2 percent [33]. The cost of a hand-geometry unit is about \$2150.

## **2.5 Network Access Control**

Network access control applications involve access to networked computer, data, or telecommunications resources by users throughout the network. The defining characteristic that differentiates network access control from local access control applications is that the number of sites from which users may be allowed access is unlimited. Users may provide their own equipment, including the PIV measurement equipment. Economic considerations dictate that the cost of this equipment be minimized. PIV systems for network access control must be robust with respect to variations in measurement equipment and user environments.

The use of biometric PIV methods for network access control is in the early stages of development. An early example was the Sprint FONCARD, in which voice verification was used to control access to Sprint's long distance

telephone network. Other potential applications include home banking and home shopping.

The need for network access control is increasing rapidly, as networks themselves increase in size. It is estimated that the number of internet addresses doubles every nine months. At the same time, audio and video capabilities are becoming available at moderate cost for use in video conferencing. As these capabilities become "standard equipment" on future personal computers, it will be possible to obtain voice and facial image data without added cost. The growing need for network access control, combined with affordable measurement hardware, create a powerful motivation for development PIV algorithms based on voice and facial images.

## 2.6 Summary

Biometric technology has evolved since the time of Bertillon toward becoming partially or fully automated. This evolution has enabled new applications such as automated local access control. The biometrics most often used for local access control - hand geometry, fingerprints, iris scans, and retinal scans - all require specialized (and expensive) measurement equipment. The cost of providing this equipment at every node of a network makes these biometrics impractical for use in network access control. Audio and video capabilities required for voice and facial-image PIV systems, on the other hand, are becoming available at modest cost for use in video conferencing. Normal human experience provides proof of the concept of identifying individuals from their voices and faces. These observations suggest that voice and facial image biometrics will become increasingly important in the evolution of PIV systems toward network access control.

## CHAPTER 3

### REVIEW OF RELEVANT LITERATURE

This chapter reviews the literature in areas of specific relevance to multi-media PIV: voice verification, face verification, multi-media verification, data fusion, and biometric data protection.

#### 3.1 Voice Verification

The speech signal from a microphone is segmented into a series of contiguous 10-20 ms *frames*, and features related to the spectral shape within each frame are extracted. The measured spectral shape is determined primarily by the instantaneous size and shape of the vocal tract, and is therefore characteristic of the speaker. Types of features include normalized power spectra [36], cepstra [37], and various transforms of the impulse response of linear prediction filters [38]. Typically, each frame is represented by 10 to 20 features.

A sequence of spectral features over time can be visualized by means of a *spectrogram*, in which the x axis represents time, the y axis represents frequency, and the gray scale level represents power or intensity. Spectrograms (also called "voice prints") have been used for forensic identification since about 1962 [39].

It is common for voice verification systems to prompt the user to speak one or more phrases containing words from a small vocabulary. Data collected in enrollment sessions is used to create models for the vocabulary words. During verification, the similarity of the observed speech data to the word

models is evaluated by means of algorithms similar to those commonly used for speech recognition. The word models allow temporal stretching and compression to accommodate variability of speaking rate. These systems are referred to as *text dependent*.

Voice verification using unconstrained speech material is also possible. In this case, word modeling or other forms of linguistic modeling have relatively little benefit, compared with purely acoustical modeling of the data. Voice verification systems that handle unconstrained speech material are called *text independent*. Text dependent systems are generally capable of higher verification accuracy than text independent systems. The advantage of text dependent systems derives from comparing test and training data frames which are time aligned with respect to articulation, or vocal tract configuration.

### 3.1.1. Text-Dependent Approaches

Suppose the test speech signal,  $x(t)$ , is known to contain a string of words,  $W$ . Word *templates* may be derived from the enrollment data and concatenated together in the sequence  $W$  with restricted contraction or dilation of the time scale to accommodate a range of speaking rates. Let  $F(t)$  be a time scale warping function from a set  $\Phi$  of allowable functions, and let  $C(W, F(t))$  be the concatenation of templates in the sequence  $W$  using warping function  $F(t)$ . Define a measure of distortion,  $D(x, C(W, F(t)))$  between the input data and the concatenated templates. The best time warping function is

$$F^*(t) = \operatorname{argmin}_{F(t) \in \Phi} D(x, C(W, F(t))) \quad (3.1)$$

and the minimum distortion is given by  $D^* = D(x, C(W, F^*(t)))$ .  $F^*(t)$  and  $D^*$  can be computed by *template matching* methods, which employ an efficient dynamic programming algorithm [40, 41].

The above matching method was used by Doddington [29] in one of the first working speaker verification systems. The identity claim was accepted if the distortion  $D^*$  was less than a threshold value. The concatenation of word templates  $C(W, F^*(t))$  is equivalent to a phrase template. A weakness of Doddington's method is that optimization of  $F(t)$  within the set  $\Phi$  does not allow an adequate diversity of pronunciations, particularly with respect to the length of inter-word pauses. This leads to false rejections in cases where the system cannot distinguish pronunciation deviations from voice deviations.

Rosenberg [42], Furui [43] and others developed improved speaker verification algorithms based on a more flexible syntax-driven dynamic programming procedure incorporating word templates [44, 45] as opposed to phrase templates. The use of separate templates for the various words and silence made these algorithms more robust with respect to pronunciation. Another important advantage was that the amount of enrollment data required was determined by the vocabulary size, as opposed to the number of possible phrases.

Furui also introduced the method of cepstral subtraction [37] to compensate for linear channel distortions such as the frequency response of a non-ideal microphone. He showed that convolution of the input signal is equivalent to addition of a constant bias in the cepstral domain (assuming the convolutive distortion does not contain spectral zeros). Therefore, if two signals differ from one another only by a fixed linear distortion, their cepstral sequences can be made comparable by subtracting the long-term average

cepstrum from each one. Cepstral subtraction is analogous to blind deconvolution in the power spectrum domain [46].

Further improvements in text-dependent verification resulted from the use of hidden Markov models (HMMs) [47, 48] rather than dynamic programming as the basis for alignment and matching of the input speech with voice models. Although the theory of hidden Markov modelling is out of the scope of this work, there are several excellent tutorials on the subject [49, 50]. Template matching and HMMs have been compared in several studies [51-53], and consistently better results have been reported for HMMs. A likely explanation is that the probabilistic HMM training procedure produces more stable and robust word models than the deterministic procedure used for template training.

The HMM algorithm provides an estimate of the speaker likelihood, or the conditional probability of the observations given the speaker model. Prior to 1991, the speaker likelihood was commonly used as the numerical criterion for acceptance or rejection of the identity claim. Higgins et al [3] described a system in which verification decisions were based on a likelihood-ratio test of the form shown in Equation 2.2. The denominator of the likelihood ratio,  $p(X | \sim C)$ , was approximated using a group of enrolled users other than the claimant. This approximation requires matching the input speech with the voice models of these other users as well as the claimant, thereby multiplying the required computation. The added computation is justified by reductions in error rates of 2-5 times, leading to widespread adoption of likelihood-ratio scoring methods by other researchers [54-56].



### **3.1.2. Text-Independent Approaches**

One of the first automatic algorithms for text-independent speaker recognition was reported by Pruzansky in 1963 [57]. In this study, the information contained in digital spectrograms of the reference and test utterances was used as the basis of comparison between two speakers. The most important finding was that recognition accuracy remained essentially unchanged when spectrograms were averaged over time to form a single long-term power spectrum per utterance. A similar result was reported by Pfeifer [58], in a study in which speakers were identified from handmarked samples of five vowels. Performance was found to improve by pooling the vowel samples, as opposed to maintaining separate statistical models of each vowel. A likely explanation of this finding is that the vowel samples were highly influenced by the phonetic contexts in which they occurred.

This explanation is supported by the findings of Paul, et al. [59], involving a database of 250 speakers. Acoustic features were extracted from thirteen phonetic categories (10 vowels and 3 nasals). Three methods of selecting reference and test samples were compared. The first method, called "context independent", compared any two events that were of the same phonetic category. The second method, called "context dependent", compared two events of the same phonetic category only if the second formants of the adjacent phonemes were at similar frequencies. The third method, called "text dependent", deemed two events comparable only if they occurred in the identical phonetic contexts. Speaker separability was found to increase monotonically from context independent to text dependent across all phonetic categories.

A recurring question in text-independent speaker verification is whether (and if so, how) linguistic modeling can be used to advantage. The

answer appears to be that linguistic modeling is useful to the extent that it restricts comparisons between enrollment and test data to like phonetic events within the same context. A vast quantity of enrollment material must be available to enable such comparisons. Phonetically-based approaches have little if any advantage over purely acoustic approaches when the text is not known [56, 60], except on very short test utterances. In this case, language constraints provided by a large-vocabulary speech recognizer [61] have proven to be useful.

For longer unprompted utterances (30 seconds or more), good performance can be obtained using purely acoustical modelling. Several studies have investigated the use of acoustic models spanning several time frames [62]. These models capture feature *trajectories*, as opposed to instantaneous features. They are capable of representing vocal gestures or coarticulations that may be speaker specific. Like context-dependent phonetic models, multiple-frame acoustic models require a very large amount of enrollment data. A way of incorporating trajectory information with less impact on enrollment requirements is to estimate the time derivative of the spectrum at each frame, and augment the feature vector to include this information [37].

In the studies cited above by Pruzanski and Pfeifer, speakers were compared with one another based on Euclidean distances between their mean vectors. Markel and Davis [63] extended this approach to use a Mahalanobis, or inverse-covariance weighted Euclidean distance. Another method of measuring distance between speakers [64] is based on the observation that vector quantizers (VQs) can be strongly speaker dependent. In vector quantization, each speech frame to be quantized is replaced with the nearest frame in a "codebook" consisting of multiple frames. The frames belonging to

a codebook are typically derived using a clustering procedure in which the mean squared error in quantization of a set of training data is minimized. When new speech data is processed, the mean squared quantization error tends to be lowest for the speaker whose training speech was used to create the VQ codebook. Speaker identification methods that rely on measures of distance between speakers are known as *minimum-distance* methods.

In a study of various approaches to text-independent speaker identification, Schwartz et al [65] concluded that probabilistic approaches (both parametric and nonparametric) are capable of superior performance to minimum-distance approaches. The premise of probabilistic approaches is that the sounds produced by speakers can be statistically described by stable probability density functions (PDFs), and that these PDFs provide a basis for classifying or testing hypotheses concerning new speech data.

One of the most successful probabilistic approaches is based on the multivariate Gaussian model. Gish and Schmidt [66] derive simple expressions for the likelihood of speech data, given a Gaussian model PDF. Likelihoods are computed for a set of speakers including the claimant, and the likelihood ratio of the claimant versus other speakers is estimated. A virtue of Gaussian models is that relatively few independent parameters are employed, minimizing the required size of the enrollment data. However, evidence that speech PDFs are well approximated as Gaussian has not been reported.

Another parametric model used successfully for speaker verification is the Gaussian mixture model (GMM). A GMM is a weighted sum of multivariate Gaussian densities in which the weights sum to unity, ensuring that the mixture is a proper density function. The parameters (means, covariance matrices, and weights) of a GMM are estimated using the Estimate-Maximize (EM) algorithm [67]. GMMs provide greater flexibility than Gaussian models to

match arbitrary PDFs, at the expense of a larger number of free parameters. Like ordinary Gaussian models, Gaussian mixture models may be used to evaluate the likelihood of an observation being produced by a particular speaker. This approach to speaker verification is described by Reynolds and Rose [68].

Nonparametric methods enable estimation of speaker likelihoods from a given body of enrollment data without recourse to any assumed parametric family of PDFs. The *nearest-neighbor* method is a well known technique of estimating density from a collection of sample points. Given  $N$  samples, the nearest neighbor estimate of density  $p(\mathbf{x})$  at test point  $\mathbf{x}$  is:

$$p(\mathbf{x}) = \frac{1}{NV} \quad (3.2)$$

where  $V$  is the volume of a spherical ball centered on  $\mathbf{x}$ , and just enclosing the frames of the model speaker's enrollment data as the  $N$  samples, and treating each frame of the test data as an independent test point. One of the conclusions of a speaker recognition study that considered this approach [65] was that its effectiveness compared with parametric methods decreases with the dimensionality of the feature space. Higgins et al. [4] reported that a modified form of nearest neighbor estimation gave better speaker recognition performance than the conventional method. This finding and the reasons for it will be further investigated herein.

Text-independent systems are inherently more robust than text-dependent systems with respect to non-linguistic or paralinguistic behaviors such as stuttering or hesitations and to non-speech sounds such as breathing and background noises. This robustness, in addition to the high accuracy that can be obtained with small vocabularies, make text-independent approaches attractive for the intended PIV application.

### 3.2 Face Verification

Until recently, most quantitative investigations of facial recognition has been based on profile measurements [69]. Sir Francis Galton measured the relative positions of five cardinal points. These points were defined in terms of facial features in a manner that would be considered unambiguous to most observers. One point, for example, was defined as "the notch between the nose and the upper lip". Individuals from the training set were selected as similar if all five measurements were within a pre-specified tolerance of the test measurements [70]. A similar procedure using more measurements and an improved decision procedure was developed by Harmon et al. [71]. More recently, Wu and Huang [72] developed a fully automatic system using back-lit photography that correctly recognized 17 out of 18 people. The approach of characterizing individuals by the geometric relationships between a set of cardinal points continues to be used for both profile and frontal recognition, and is generally referred to as a *feature-based* approach.

In 1965, Preston reported using an optical computer to recognize faces [73]. A coherent light source was directed into the device through a photographic film upon which the input image was printed. Within the optical computer, a second photographic film, called a "matched filter", contained the faces of six kings arranged side by side in two rows of three. When the input image matched one of the kings' faces, a bright spot of light appeared on the device's "output plane" at one of six locations corresponding to the arrangement of faces on the matched filter. It was shown that the image on the output plane was the cross-correlation of the input image with the matched filter. The approach of characterizing individuals by a complete image of the face is referred to as a wholistic or *template-based* approach.

Feature based approaches have the advantage of representing faces in a simple and compact form, enabling rapid searching over large databases. Template based approaches, on the other hand, have been found to be more accurate [74], probably because they preserve facial details such as wrinkles, scars, and unusual markings that are known to carry personal identity information [2]. Feature-based and template-based approaches are discussed in the following sections.

### **3.2.1. Feature-Based Approaches**

One of the first automated systems for recognizing facial front views was developed by Kanade [75]. The method is similar in principal to the mug-shot retrieval method described previously in Section 2.3.1. Facial features (analogous to "cardinal points") are located on the image, and parameters derived from these features serve as indexes into a "face space". To locate features on the image, edge detection is first performed using the Laplacian operator, and the image is quantized to binary intensity values. A set of "integral projections" are computed on narrow horizontal slits at various positions in the image. At each pixel along the length of the slit, the integral projection is the number of 1's along the width of the slit. Heuristic procedures are applied to locate facial features in the following sequential order: (1) top of head; (2) sides of face; (3) nose, mouth and bottom of chin; (4) chin contour; and (5) eyes. The algorithm is iterative in the sense that feature locations can be revised based on later computations. The final locations are converted to a set of 13 ratios and angles that are invariant with respect to the image scale. The face space is therefore 13 dimensional. Fifteen out of 20 people were correctly identified in a test of the system.

More recently, explicit geometric models have been used to locate facial features. Govindaraju et al. reported an algorithm to automatically locate faces in newspaper photographs [76]. The outline of the face was modeled as a closed contour consisting of arcs for the hair-line and chin-line with connecting straight lines for the face sides. Candidate arcs and lines were detected using the generalized Hough transform [77]. These segments were then grouped together to form candidate faces using an algorithm that is also based on the Hough transform. Candidate faces are pruned using spatial constraints derived from the caption of the photo and heuristics of photo journalism. A slightly different approach was taken by Yuille et al. [78] in locating eyes and mouths. They used simple geometric models involving 11 parameters of the eye model and the 10 parameters of the mouth model. These parameters specified the location, size, shape and tilt of the model features with respect to the image. The parameters were adjusted by a steepest descent algorithm to minimize an error function [79] of the image and the models. Similar approaches have been reported by others [80] [81].

Empirically-determined shape models were used by Lanitis et al. [82]. These models were created from a set of training examples, which consisted of manually traced contours of eyes, nose, mouth, ears, and chin. The principal modes of variation of the shapes encountered in the training data were determined statistically [83], and used to define a family of allowable shapes, controlled by a small number of parameters. An algorithm similar to that of Yuille, et al. was used to align the shapes with faces present in test images.

Having located a set of facial features using methods such as those above, the problem remains of how this information may be used to recognize individuals. Kamel et al. [84] developed a transformation and matching technique in which nine original feature locations were converted to a set of

five invariant features to make the system robust with respect to viewing angle. The extracted information was represented using a data structure designed to enable efficient matching. In a test involving 84 test images with various viewing angles, 66% were correctly recognized. A test of a similar system, which did not involve multiple viewing angles [85], resulted in 89% correct recognition.

### **3.2.2. Template-Based Approaches**

Baron [86] postulated that a cross-correlation mechanism like that of Preston's optical computer exists within the human neuroanatomy and is used in face recognition. According to this theory, input images are rapidly (nearly instantaneously) compared with remembered facial images, or templates. He developed a computer simulation of the mechanism, which included cross-correlation together with related "control networks" that performed functions such as input scaling (to normalize the distance between the eyes) and illumination level normalization.

Nakamura, et al. [87] applied a template-matching approach to a set of two-dimensional contours, termed "iso-density maps", which were derived from the original image. An iso-density map is a set of one or more closed contours connecting points in a monochrome image that have the same level of brightness. Each face was represented using eight iso-density maps corresponding to different brightness levels. The maps for high brightness levels were found to be most useful for discriminating between people, but also were most affected by variations in viewing angle. The maps for low brightness levels were more robust but less person specific. In a test involving ten subjects, all faces were correctly recognized, but it was noted



that the algorithm is very sensitive to camera positioning and to the intensity, position, and color of the light source.

A limitation of conventional template matching approaches is that they do not accommodate the kind of facial movements that may be expected during speech, for example. Lades et al. [88] applied a rectangular grid to input images, and modeled the local sub-images at the vertices of the grid. In recognizing a test image, local template matching was performed at each grid point, and the grid was allowed to distort elastically in order to minimize a global cost function. In a set of face verification tests involving 87 people with variations in viewing angle and facial expressions, equal-error rates ranging from 12% to 21% were reported.

Another approach to handling facial image variability is to model the modes of variability with respect to a long-term average image. Sirovich and Kirby [89] applied principal component analysis to a set training faces. A set of training images was analyzed to determine the mean image,  $\Psi$ , and the principal components,  $\mathbf{u}_i, i \leq 0 < L$ , of the covariance of the training images about  $\Psi$ . They showed that an arbitrary facial image,  $\Gamma$ , not belonging to the training set could be approximated in terms of  $\Psi$  and a relatively small number (e.g., 10) of principal components. To do this, the deviation of  $\Gamma$  from  $\Psi$  is first computed as  $\phi = \Gamma - \Psi$ . The projection,  $\phi_f$ , of  $\phi$  on the subspace spanned by the  $\mathbf{u}_i$  is given by  $\phi_f = \sum_{i=1}^L \omega_i \mathbf{u}_i$ , where  $\omega_i = \phi_f^T \mathbf{u}_i$ . The synthesized approximation is:  $\Gamma_f = \Psi + \phi_f$ . Turk and Pentland [6] described a method of locating a face within a larger image by sliding a window over the original image, and selecting the position of the window to minimize the error function:  $\epsilon^2(x,y) = |\phi - \phi_f|^2$ . In the following,  $\Gamma(x,y)$  denotes a subimage of  $\Gamma$  of dimensions equal to those of  $\Psi$  and  $\mathbf{u}_i$ , with upper-left corner at  $(x, y)$ . The

derivation of  $\epsilon(x,y)$  in terms of  $\Gamma(x,y)$ ,  $\Psi$ , and  $\mathbf{u}_i$  is repeated here because several errors were contained in [6]. Dependence on  $(x, y)$  is suppressed.

$$\epsilon^2 = \|\phi - \phi_f\|^2 \quad (3.3)$$

$$= (\phi - \phi_f)^T (\phi - \phi_f) \quad (3.4)$$

$$= \phi^T \phi - \phi^T \phi_f - \phi_f^T (\phi - \phi_f) \quad (3.5)$$

$$= \phi^T \phi - \phi^T \phi_f \quad (3.6)$$

$$= \phi^T \phi - \left( \sum \omega_i \mathbf{u}_i^T \right) \left( \sum \omega_i \mathbf{u}_i \right) \quad (3.7)$$

$$= \phi^T \phi - \sum \omega_i^2 \quad (3.8)$$

Expanding the first term,

$$\phi^T \phi = (\Gamma - \Psi)^T (\Gamma - \Psi) \quad (3.9)$$

$$= \Gamma^T \Gamma - 2\Psi^T \Gamma + \Psi^T \Psi \quad (3.10)$$

Expanding the second term,

$$\sum \omega_i^2 = \sum (\phi^T \mathbf{u}_i)^2 \quad (3.11)$$

$$= \sum ((\Gamma - \Psi)^T \mathbf{u}_i)^2 \quad (3.12)$$

$$= \sum (\Gamma \mathbf{u}_i - \Psi^T \mathbf{u}_i)^2 \quad (3.13)$$

Combining the two terms, and making explicit the dependence on spacial position:

$$\epsilon^2(x,y) = \Gamma^T(x,y)\Gamma(x,y) - 2\Psi^T \Gamma(x,y) + \Psi^T \Psi + \sum_{i=1}^L [\Gamma(x,y)\mathbf{u}_i - \Psi^T \mathbf{u}_i]^2 \quad (3.14)$$

Principal components analysis efficiently represents variability among faces, without regard for discrimination between individuals. In an effort to capture the most useful possible information for face recognition, Cheng, et al. [90] proposed using linear discriminant functions similar to Fisher's discriminant function [91]. In an experiment involving 40 subjects, correct

recognition rates ranged from 87.8% using 3 training images per subject to 96.3% using 27 training images per subject.

Template matching approaches are inherently computation intensive. Burt asserts that the cost of searching for a target over a range of scale factors and orientations is proportional to the 6th power of the target dimensions [92]. To reduce the computational cost, he proposes the use of *image pyramids*. The low-pass, or Gaussian pyramid is generated from the original image through a sequence of steps, each involving low-pass filtering followed by sub-sampling. The band-pass, or Laplacian pyramid is formed as difference images between successive levels of the Gaussian [93]. Laplacian pyramids are potentially useful in face verification because they enable each step of facial location and matching to occur at the appropriate level of spacial resolution.

### 3.3 Multiple Media

Facial movements during speech are known to convey information that can substitute for voice. In recognizing noisy speech, observers gain the equivalent of 8-10 dB of signal-to-noise ratio by seeing the talker's face [94]. Petajan [95] developed a speaker-dependent visual word recognition algorithm. Testing the algorithm on digits pronounced in isolation, a correct recognition rate of about 95 percent was reported. The algorithm cannot distinguish between words that differ in articulations that are not visible. Further evidence of the redundancy of voice and facial movements with respect speech message content is provided by the "McGurk effect" [96]. As an example of this phenomenon, observers presented with a videotape showing articulation of the syllable "ba" together with the sound of the syllable "ga" tend to perceive "da". This is explained by the fact that articulation of /d/ occurs at a point in the vocal tract that is physically between that of /b/ and

/g/, representing a compromise between the conflicting sound and visual cues.

The high correlation between speech sounds and mouth movements has led to recent efforts in joint audio/visual speech recognition [97] and joint audio/visual encoding for teleconferencing [98].

A PIV system using both voice and facial image data was recently reported by Brunelli, et al [1, 99]. The voice component of the system is based on vector quantization, and is similar to the system of Rosenberg and Soong [100]. The face component is similar to the feature-based system of Brunelli and Poggio [74]. The voice and face scores for all known (modelled) individuals were processed using a HyperBF neural network [101] to produce the score of the integrated system. On a database of 33 subjects, the integrated system achieved 100% accuracy, whereas the voice and image components separately each achieved less than 95% accuracy.

Testing the synchrony between speech sounds and mouth movements would be a useful means of detecting possible counterfeiting in a PIV system. Performing this test would of course require visual processing to be focused on the mouth area. In a study of audio/visual speech recognition by human listeners in a noisy environment, Le Goff et al. [102] found that about half the information carried by viewing the speaker's natural face could be provided by a model of the movement of the lips. About two thirds of the information could be provided either by the natural lips, or by a model of the combined movement of the lips and jaw. Presumably, the remaining visual information is carried by the tongue and teeth. Brooke [94] reported that "articulatory excursions from a neutral facial position, in which the lips and jaw are lightly closed, rarely if ever exceed 25 mm". Therefore, any test of lip

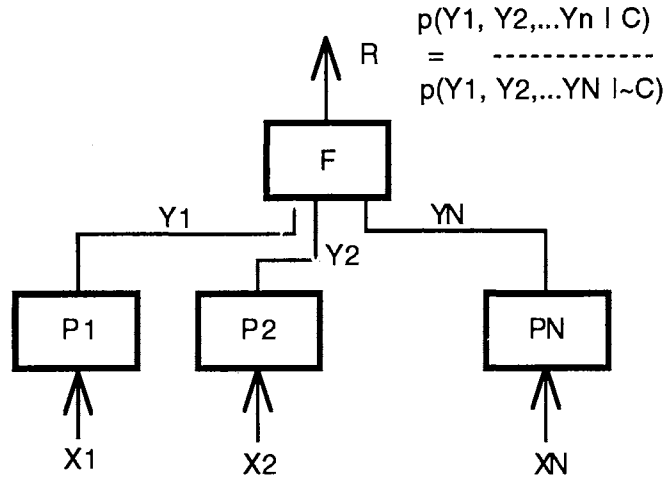
synchronization requires that images be captured with sufficient resolution to observe these small movements.

A substantial body of literature addresses human performance at identifying individuals from voice or face observations [2, 11-13]. Surprisingly, to the author's knowledge there have been no studies of human identification performance using *both* voice and face observations. One interesting study, however, concerned the detection of deceit in interpersonal communications [103]. The best correlate of deceit was derived from a combination of voice and facial features, and achieved 86% correct classification of deceitful versus nondeceitful communications.

### 3.4 Data Fusion

Combining multiple tests of the claimed identity is known to improve security. For example, the procedure used in automated bank teller machines provides greater security than passwords or credit cards [104] because it requires the user to possess a valid card *and* to know the corresponding PIN number. The probability of unauthorized access is reduced by combining independent tests of the claimed identity.

Data fusion of biometric data can be accomplished at various stages of the data processing. Suppose we have  $N$  sources of data:  $X_1, X_2, \dots, X_N$ , which are passed through processors  $P_1, P_2, \dots, P_N$  to produce processed data  $Y_1, Y_2, \dots, Y_N$ , respectively. All processed data are applied as inputs to fusion processor  $F$ , as shown in Figure 1.3. The output of  $F$  is an estimate of the likelihood ratio,  $p(Y_1, Y_2, \dots, Y_N | C) / p(Y_1, Y_2, \dots, Y_N | \sim C)$ .



**Figure 3.1: Illustration of Data Fusion**

In the architecture of Figure 3.1, each source is processed independently of the other sources, deferring possible consideration of correlation between sources to F. The literature refers to various types of data fusion, which differ in the nature of the processed data,  $Y_i$ .

- $Y_i = X_i$ . In this case, no processing is performed by  $P_i$ . The measurements  $X_i$  are passed directly to F, where they are processed jointly. This is referred to as *measurement fusion*.
- $Y_i = p(X_i | C) / p(X_i | \sim C)$ . In this case, each  $P_i$  estimates the likelihood ratio of C relative to  $\sim C$  based on its input data  $X_i$ . We refer to this case as *likelihood-ratio fusion*.
- $Y_i = \begin{cases} 1 & \text{if } p(C | X_i) > p(\sim C | X_i) \\ 0 & \text{otherwise} \end{cases}$ . In this case, each  $P_i$  forms a decision to accept or reject the claim based on its input data  $X_i$ . This case is referred to as *decision fusion*.

In the case of measurement fusion, the data from all sources is fed directly to F. F produces an estimate of the likelihood ratio,  $p(X_1, X_2, \dots, X_N | C) /$

$p(X_1, X_2, \dots, X_N | \sim C)$ , where the dimensionality of the observation space equals the sum of the dimensionalities of the  $N$  sources. This approach preserves all correlation information. It is therefore potentially the best estimator [105] and the one that requires most computation.

Likelihood-ratio fusion is equivalent to measurement fusion if the sources are mutually independent. In this case, by definition,  $p(X_1, X_2, \dots, X_N | C) = P(X_1 | C)P(X_2 | C)\dots P(X_N | C)$ . The fusion processor,  $F$ , simply multiplies together the  $N$  likelihood ratio estimates.

$$R = \prod_{i=1}^N Y_i \quad (3.15)$$

If the sources are not independent, likelihood-ratio fusion is suboptimal because it neglects to account for correlation between sources at the measurement level.

Decision fusion methods were studied and developed in the 1980s to enable military information systems to integrate multiple reports originating from a distributed network of independent sensors [106-109]. A large number of sensors were typically involved, and the bandwidth allowed for communication between the each local sensor and the central "decision post" (equivalent to  $F$ ) was assumed to be highly restricted. Optimal strategies for setting the local decision thresholds and fusing the resulting decisions were derived. Decision fusion is less accurate than likelihood-ratio fusion because each likelihood ratio is presented to  $F$  with a precision of only one bit.

For multi-media PIV, either measurement fusion or likelihood-ratio fusion are appropriate, depending on whether the information sources are considered to be independent. Decision fusion is not appropriate because restrictions on bandwidth or representational precision do not apply.

### 3.5 Biometric Data Protection

A practical requirement of biometric security devices is that some means should be in place to insure that personal models are not forged or modified between enrollment and access time. The opportunity for such alterations may be greatest when personal models are transported from the enrollment site to various access sites using telecommunications media or personal tokens such as smart cards. The requirements for biometric data protection are likely to include the following:

- It must be possible to verify that a model belongs to the person who presents it or claims it (authenticity).
- It must be possible to detect any alterations of a model occurring after enrollment (integrity).
- It may be desirable to keep the model data private to minimize the potential advantage to impersonators (secrecy).

In 1976, Diffie and Hellman proposed the principles of *public-key cryptography*, whereby secure communication can take place without any transfer of secret keys [110]. Their method is based on *one-way functions*, which are easy to compute and which have inverses that are infeasible to compute. One-way functions for which computation of the inverse is made feasible by knowledge of a key are called *trap-door one-way functions*. Consider a family of encoding functions  $E_z$  and their inverse or decoding functions  $D_z$ , indexed by integer  $z$ , for which  $Y = E_z(X)$  and  $X = D_z(Y)$  can be computed easily given  $z$ , but for which  $D_z(Y)$  is infeasible to compute otherwise, even when  $E_z$  is known.

A public-key cryptosystem providing secure communications to a network of users works as follows. Each user,  $A$ , randomly chooses an integer



$z$  and forms algorithms for computing  $E_A$  and  $D_A$ . He then publishes  $E_A$  in a public directory and keeps  $z$  and  $D_A$  secret. User A may send a secret message,  $X$ , to user B by retrieving  $E_B$  from the public directory, forming  $Y = E_B(X)$ , and transmitting  $Y$  to B. User B uses his private decoding algorithm  $D_B$  to decrypt  $Y$ .

*Information authentication* refers to methods of proving the identity of the originator of a message. Cryptographic methods are often used to implement authentication. Secrecy and authentication, however, are independent attributes [111]. Public-key cryptography can be used to create a *digital signature* as follows: User A "signs" his message by applying his secret algorithm,  $D_A$ , to form  $S = D_A(X)$ . Anyone may decode  $S$  by applying  $E_A$ , available in the public directory. The decoded message, if intelligible, could only have been signed by User A because only User A knows  $D_A$ . Also, since  $S$  is a function of message  $X$ , any alteration of  $S$  will render the decoded message unintelligible. It is possible to simultaneously achieve secrecy and authentication by signing with  $D_A$  and then encrypting with  $E_B$  at the transmitting end, and applying the inverse operations at the receiving end.

A theory of information authentication developed by G. Simmons [112] defines an authentication scheme by the set of messages the receiver will accept as authentic and the set of messages the transmitter may transmit. The probability of deception,  $P_d$ , is equal to the ratio of the sizes of these two sets. An essential feature of authentication is the presence of redundant information known to the receiver. Encryption is used to spread the set of acceptable messages in what appears to be a random manner among the set of all possible messages. Simmons cites the example of a common military communications protocol in which the transmitter and receiver have matching secret authenticator codes. The transmitter appends the

authenticator code  $Z$  to the message and encrypts the resulting extended message before transmitting. The receiver accepts the message as authentic if  $Z$  is recovered after decryption. For an authenticator containing  $r$  bits chosen at random, the probability of a random message being accepted is equal to  $2^{-r}$ . This reasoning leads to the following lower bound on  $P_d$ :

$$\log P_d \geq -H(Z), \quad (3.16)$$

where  $H(Z)$  is the entropy of the authenticator code. A tighter bound, derived by Simmons, is:

$$\log P_d \geq -I(Y ; Z) \quad (3.17)$$

where  $I(Y ; Z)$  is the mutual information between the encrypted message  $Y$  and the authenticator code. This surprising result states that the probability of deception can only be small when the encrypted message provides a large amount of information about the key! This can be achieved by using a long key.

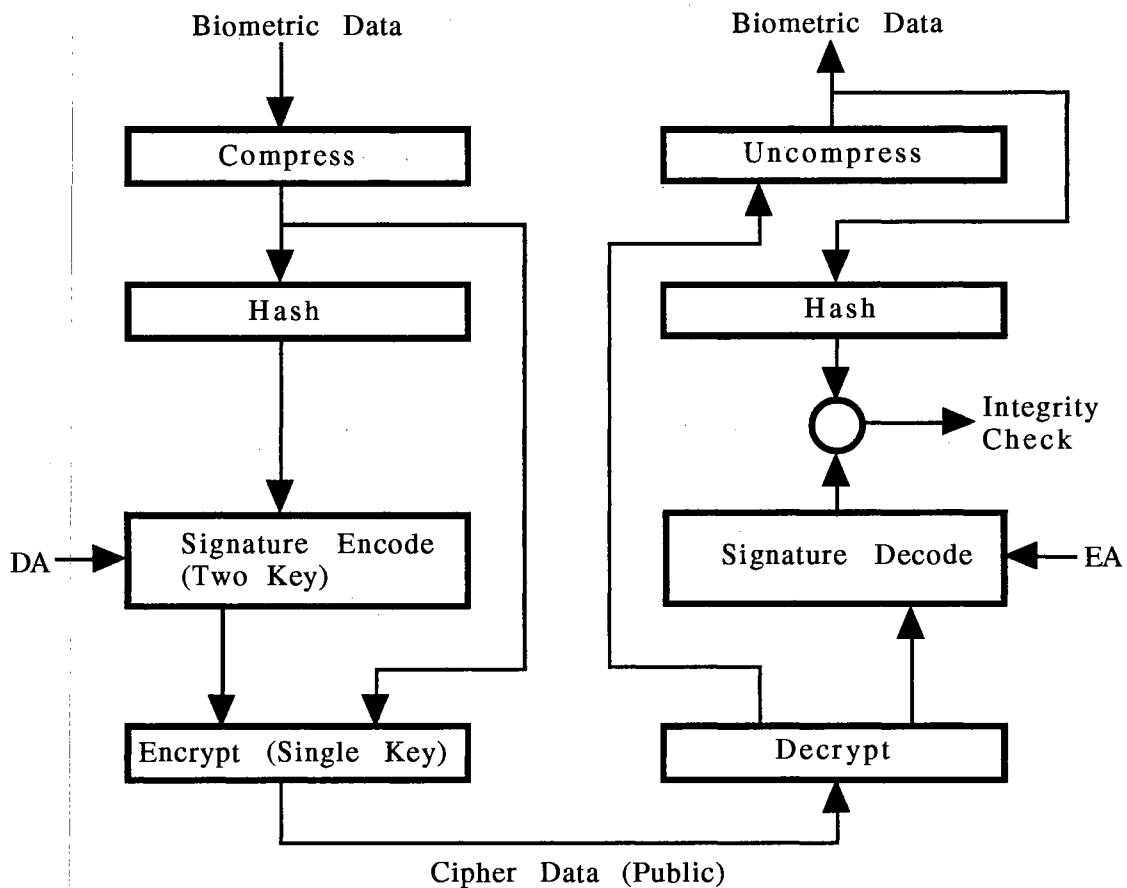
The algorithm proposed by Rivest, Shamir, and Adelman [113] is the most thoroughly studied public-key cryptographic algorithm that remains viable today. It is believed (but not proven) that the security of the RSA algorithm is equivalent to the difficulty of factoring large numbers. The RSA method is described in [114] as "the only well-known system discovered to date which is secure, practical and suitable for both secrecy and authentication". A drawback of the RSA algorithm is that it is patented and subject to a licensing fee by RSA Data Security, Inc. Alternatives to RSA include the public-domain PGP algorithm and the NSA/NIST Digital Signature Standard (DSS) algorithm [115].

Ordinary credit transactions are characterized by mutual distrust between the merchant and the customer, but common trust in a third party (e.g., a bank or credit card company). The third party, or "issuer", may employ the following methods, proposed by Simmons [112], to facilitate transactions among a network of merchants and customers. Each customer reports to the issuer for measurement of biometric attributes such as voice and facial features. The customer is given a credential (possibly in the form of a "smart card") containing his biometric information encrypted using the issuer's secret key. The issuer's public key is made available to all merchants. When the customer initiates a transaction, he presents his encrypted credential to the merchant. The merchant decrypts it using the issuer's public key. He then measures the customer's biometric attributes and compares them with those derived from the credential. If an acceptable match is obtained, he concludes that: (1) the credential is authentic (endorsed by the issuer); and (2) the customer's biometric data are consistent with the credential. This procedure requires no communication between the merchant and the issuer at the time of the transaction, and requires merchants to store only the public keys of the various issuers.

Authentication of long messages requires a prohibitive amount of computation [114]. To reduce computation, *one-way hash functions* of the message can be used. A hash function converts a variable-length message  $X$  to a fixed-length representation,  $H(X)$ , sometimes referred to as a *message digest*. The message digest is then signed, rather than the message itself. The signed message digest,  $D_A(H(X))$ , together with the original message,  $X$ , are jointly encrypted to produce  $Y = E_B(X, D_A(H(X)))$ . At the receiver, the hash function of the recovered message  $X$  is compared with the recovered message digest. If they are the same, the integrity of  $X$  is proved.

The Secure Hash Algorithm (SHA) developed by NIST [116] takes as input messages of length up to  $2^{64}$  bits and produces a 160-bit message digest. Every bit of the message digest is a function of every bit of the input message. The difficulty of finding a message with a given digest is on the order of  $2^{160}$  operations.

Biometric models created from voice and facial image features require a large number of bytes for their representation (on the order of  $10^4$  to  $10^5$  bytes). Coding and decoding these models using RSA or similar algorithms would incur unacceptable delays. Therefore, the use of hash functions is indicated. The system shown below includes both digital signature (with SHA) and encryption functions.



**Figure 3.2: System for Secure Transportation of Biometric Models.**

An operational system based on the principles described above was developed and tested by Sandia Laboratories for controlling access to a plutonium reactor facility [30]. Biometric information consisted of the subject's weight and hand geometry features. The RSA encryption algorithm was used, with a separate key pair for each employee. Encrypted data was stored on a magnetic stripe on the employee's ID badge.

A recent survey of government applications of smart cards [117] includes a large number of experimental programs related to personal identity verification. Many of these programs appear to include or have future plans to include biometric features and/or cryptographic data protection.

### 3.6 Summary

A substantial body of literature exists in various fields relevant to multimedia PIV. Both voice and face PIV algorithms have been under development for nearly 20 years, resulting in numerous and diverse approaches. Face verification algorithms have used only still images, as opposed to image sequences. Very recently, a PIV algorithm was reported that combines voice and still facial images, achieving better accuracy with the combination than with either voice or face information alone. The use of facial image sequences is a logical extension of the current state of the art. Human performance studies provide at least anecdotal evidence of the merit of this approach.

Fusion of multiple data sources can be performed at various levels, from raw measurements to fully processed binary decisions. In general, accuracy increases as data fusion is performed earlier in the processing. For independent sources, fusion can be performed at the likelihood-ratio level without loss of accuracy.

Cryptographic methods of data protection are available to insure the integrity, authenticity, and privacy of biometric data as it is transported from the measurement site to a processing site or between processing sites at different locations. These methods will be needed to detect and prevent subversion of the system by persons wishing to gain unauthorized access.

## CHAPTER 4

### MULTI-MEDIA PIV SYSTEM DESIGN

This chapter considers the design of a biometric PIV system for network access control. The requirements of the system are specified and a concept of operation is presented. Issues related to counterfeiting and methods of preventing counterfeiting are discussed.

#### 4.1 System Requirements

From the point of view of the authorized user, the primary requirements are that verification be performed quickly, unobtrusively, and with low probability of rejection. From the point of view of the system administrator, the primary requirement is low cost and low probability of admitting an unauthorized person. These requirements may be quantified as follows [118, 119]:

- Access time: 10 seconds or less
- False rejection probability: 1 percent or less
- False acceptance probability: application dependent
- Cost: no special-purpose biometric equipment needed
- Counterfeiting: simple attacks must be blocked.

The requirement on false rejection probability is based on the tolerance of valid users to being blocked from their intended activity. The 1% level is comparable with the probability of being blocked for other reasons such as a mis-dialed telephone number or lack of an available communications circuit.

The maximum tolerable false-acceptance probability depends on the value of what is being protected, and on the other (non-biometric) checks that may also be in effect. For long-distance telephone access control, where the cost of false acceptance is relatively low, a 10% level may be tolerable.

The requirement on counterfeiting is also dependent on the application. Counterfeiting should be discouraged by making it sufficiently costly or time consuming that it is not warranted. Forseeable attacks that could be accomplished simply should be prevented.

#### **4.2 Concept of Operation**

The concept of operation is as follows: The user logs in using the normal procedure, establishing an identity claim. A live video picture of the user's face is displayed on the CRT, and the user is asked to adjust his position or that of the camera so that the face appears entirely within the screen. When the system detects that a face is present, it prompts the user to speak a short, randomly selected phrase. A decision to accept or reject the identity claim is then made based on the available measurements.



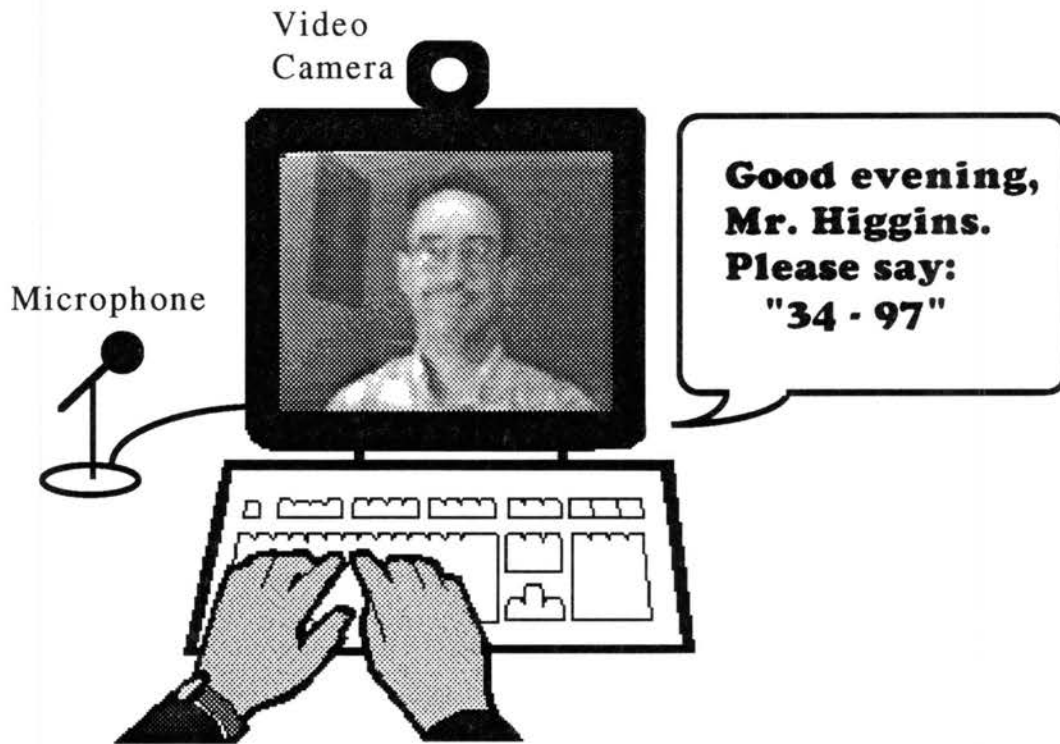
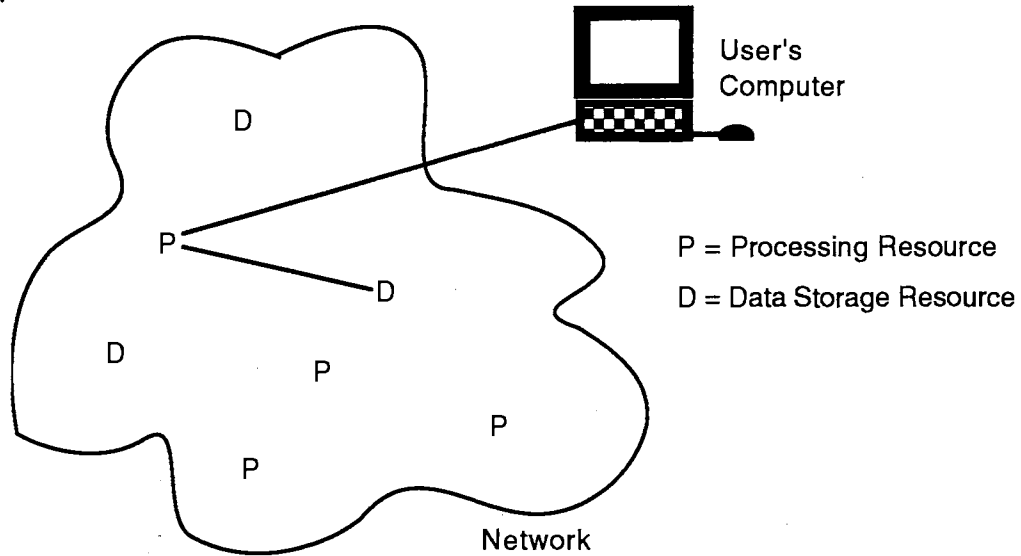


Figure 4.1: User's View of PIV System.

Equipment beyond a normal personal computer required for this procedure consists of a video camera, microphone, and appropriate digitization hardware. It is assumed that the sampled sound waveform and sampled image data are available at resolutions and sampling rates typical of current consumer-quality audio-visual equipment. It is also assumed that valid users are cooperative in responding to prompts and in providing an environment with adequate lighting and reasonably low background noise level.

To minimize the possibility of tampering, the PIV program is executed using processing resources located within the network, as opposed to the user's computer. As shown in Figure 4.2, biometric data collected at the user's computer is transmitted to a processing resource P in the network. P then locates and retrieves the user's model from a data storage resource, D. After

performing the PIV processing, P allows or disallows the requested access to occur.



**Figure 4.2: System Interaction with Network.**

It would be possible to circumvent this PIV procedure in either of two general ways. One approach would be to substitute the imposter's model for the claimed user's model at point D, or to interfere with the communication between P and D in such a way that the imposter's model was received by P. Data protection methods that address this possibility were discussed in Section 3.5. Another approach is to present counterfeit data to the PIV system. This could be done, for example, by substituting pre-recorded audio and video data in place of the microphone and camera "live" inputs. Anti-counterfeiting methods addressing this possibility are presented in the following section.

### **4.3 Counterfeiting**

Conventional approaches to performance measurement employ subjects who behave cooperatively, using the PIV system in the intended manner. Cooperation by authorized users is rational because it increases the likelihood

that the PIV device will recognize their identity and allow them to proceed with their business. Unauthorized users, or impostors, however, might reasonably believe that access could more likely be obtained by employing a *counterfeiting* strategy designed to exploit a perceived vulnerability of the PIV system. For example, a person's voice could be counterfeited using a tape recorder, or a facial image using a mask or photograph. Law enforcement agencies have even reported cases involving specially made rubber globes with fingerprint patterns inscribed on the fingertips [120]. The possibility that an imposter's biometric data might be made available to law enforcement authorities provides further incentive for counterfeiting. Counterfeiting is therefore a rational strategy for imposters. The possibility of counterfeiting has been largely neglected in the literature. PIV systems should include measures to detect and reject counterfeiting attempts.

#### 4.4 Anti-Counterfeiting

The objective of anti-counterfeiting is to verify that the received audio and video signals are "live", as opposed to pre-recorded or synthesized. This can be accomplished by prompting the subject to speak randomly-chosen phrases, and limiting the time permitted for the correct response [3]. Randomized prompting effectively defeats the threat of pre-recording if the a-priori probability of having the needed response (roughly the reciprocal of the number of possible phrases) is less than the false-acceptance rate. The imposed time limit addresses the synthesizer threat. The proposed concept of operation is vulnerable to (future) synthesizers that are sufficiently fast and accurate. Even so, the imposter would have to go to considerable effort to obtain data from the target user with which to train the synthesizer.

## 4.5 Summary

A concept was presented for a multi-media PIV system to be used for network access control applications. The system employs inexpensive audio and video equipment of the type used for desktop video conferencing. It prompts the user to speak randomly selected phrases, while capturing sound and full-motion video of the spoken response. The user is allowed access to the protected network resources if the observations are consistent with the claimed identity. Several system requirements were presented. The possibility of counterfeit evidence being presented was discussed, and approaches to detecting counterfeiting were described.

## CHAPTER 5

### PROBABILISTIC MODELING OF INDIVIDUALS

#### 5.1 Introduction

This chapter describes the rationale for accepting or rejecting an identity claim based on the observed biometric measurements. We wish to minimize the total probability of making an error, which occurs either when a valid claim is rejected or when an invalid claim is accepted. This is accomplished by accepting the claim if and only if it is more likely to be valid than invalid given the observation. Evaluating the likelihood of the claim (or the alternative, that the individual is someone other than the claimant) given the observation requires the use of *individual models*.

Both voice and facial appearance can be consciously influenced by the subject. To this extent, they are reasonably regarded as random, as opposed to deterministic, observations of underlying attributes that characterize the subject. The subject may be considered to "emit" observation vectors according to a multi-dimensional probability density function (PDF). The true PDF associated with an individual is not known in practice, but must be estimated from a set of prior observations known as enrollment data. The enrollment data combined with the PDF estimation algorithm comprise the individual model.

The individual model is the key element of verification by which observations at the frame level are converted to evidence of identity. Evidence

may be accumulated over the length of a verification session under the assumption of mutually independent frames.

The dimensionality (number of measurements per frame) of biometric data is often greater than ten. The number of frames of enrollment data is often on the order of 100 or less. Probability density estimation is problematic under these conditions where the measurement space is sampled sparsely. Two approaches to this problem are presented. A key factor in choosing between the two approaches is the *intrinsic dimensionality*, or the minimum number of independent parameters needed to specify a point in the space.

## 5.2 Acceptance Criterion

Suppose we wish to test the validity of an identity claim,  $C$ , given observation sequence  $X$ . Using Bayes' decision rule,  $C$  is accepted if and only if

$$p(C | X) > p(\bar{C} | X) \quad (5.1)$$

where  $p(C | X)$  is the posterior probability of  $C$  given  $X$ , and  $p(\bar{C} | X)$  is the posterior probability of the alternative (that  $C$  is false) given  $X$ . Bayes' decision rule minimizes the probability of making an incorrect decision [91].

Re-writing the posterior probabilities gives the rule

$$\frac{p(X | C) p(C)}{p(X)} > \frac{p(X | \bar{C}) p(\bar{C})}{p(X)} \quad (5.2)$$

or

$$\frac{p(X | C)}{p(X | \bar{C})} > \frac{p(\bar{C})}{p(C)} \quad (5.3)$$

where  $p(X | C)$  and  $p(X | \bar{C})$  are likelihood functions, and  $p(C)$  and  $p(\bar{C})$  are *a-priori* probabilities of  $C$  and  $\bar{C}$ . The ratio  $p(X | C) / p(X | \bar{C})$  is known as a *likelihood ratio function*, and Equation 5.3 is a *likelihood ratio test*. The

quantity  $p(\bar{C}) / p(C)$  is a constant. The decision rule of Equation 5.3 can also be expressed in terms of the log likelihood ratio (LLR),

$$\ln \frac{p(X | C)}{p(X | \bar{C})} > \ln \frac{p(\bar{C})}{p(C)}. \quad (5.4)$$

In practice, the prior probabilities  $p(\bar{C})$  and  $p(C)$  are unknown. It is therefore common to replace  $\log(p(\bar{C}) / p(C))$  with an experimentally determined threshold value,  $T$ .

Estimation of  $p(X | \bar{C})$  is difficult because of the conditioning on  $\bar{C}$ , the set of all individuals except  $C$ . A reasonable approximation to  $p(X | \bar{C})$  can be derived as follows:

$$p(X | \bar{C}) = \sum_{S_i \in \bar{C}} p(X | S_i) \quad (5.5)$$

$$\approx \sum_{\substack{S_i \in D \\ S_i \neq C}} p(X | S_i) \quad (5.6)$$

$$\approx \max_{\substack{S_i \in D \\ S_i \neq C}} \{p(X | S_i)\} \quad (5.7)$$

where  $S_i$  is a particular individual, and  $D$  is a set of individuals known as a *cohort*, for whom enrollment data is available. Equation 5.6 is valid if the number of individuals included in  $D$  is sufficiently large. Equation 5.7 is valid if the sum over  $S_i \in D$  in Equation 5.6 is dominated by one individual. The LLR decision rule of Equation 5.4 can now be approximated as

$$\ln p(X | C) - \ln \max_{\substack{S_i \in D \\ S_i \neq C}} \{p(X | S_i)\} > T \quad (5.8)$$

or

$$\ln p(X | C) - \max_{\substack{S_i \in D \\ S_i \neq C}} \{\ln p(X | S_i)\} > T. \quad (5.9)$$

Suppose that  $X$  is an observed sequence of feature vectors  $x_i, 0 \leq i < N$ , over the length of an utterance. A further approximation is based on the assumption that the feature vectors  $x_i$  comprising  $X$  are statistically independent:

$$p(X | S_i) \approx \prod_{x_j \in X} p(x_j | S_i), \quad (5.10)$$

or

$$\ln p(X | S_i) \approx \sum_{x_j \in X} \ln p(x_j | S_i). \quad (5.11)$$

Combining Equations 5.9 and 5.11, the LLR decision rule becomes

$$\sum_{x_j \in X} \ln p(x_j | C) - \max_{\substack{S_i \in D \\ S_i \neq C}} \left\{ \sum_{x_j \in X} \ln p(x_j | S_i) \right\} > T. \quad (5.12)$$

The decision to accept or reject the claimed identity based on evidence  $X$  therefore reduces to the problem of accurately estimating probability densities at the test sample points,  $x_j$ . This problem is addressed in the following sections.

### 5.3 An Example: Height as Evidence of Identity

To illustrate estimation of the likelihood ratio, consider a simple PIV system based on the user's measured height. Suppose the height of User  $X$  is determined during enrollment to be  $H_X = 71$  inches. Suppose also that the combined effect of variability of  $X$ 's height, measurement precision, and variability of shoe height introduce errors that are Gaussian distributed with zero mean and a standard deviation of 0.5 inches (a simplification for the purpose of illustration). If a person claims to be User  $X$  and has measured height  $H$ , the likelihood function, assuming the claim,  $C$ , is true, is:



$$p(H | C) = \frac{2}{\sqrt{2}} \exp( -2(H - H_x)^2 )$$

The likelihood function for  $\sim C$  is the PDF of heights for all people *except* X. Assuming that the X has negligible effect on the PDF,  $p(H | \sim C)$  may be approximated by the unconditional likelihood,  $p(H)$ .

$$p(H | \sim C) \approx p(H)$$

The cumulative probability distribution,  $P(H)$ , of heights of males between the ages of 18 and 64 in the US is shown in Table 1.1 [121]. Differentiating to form a density function,

$$p(H | \sim C) \approx \frac{d}{dH} P(H).$$

Height (inches)	Cumulative Probability (%)
60	0.15
61	0.35
62	0.65
63	1.44
64	2.89
65	5.92
66	11.81
67	19.88
68	31.37
69	45.93
70	60.18
71	73.73
72	83.94
73	83.94
74	91.65
75	95.81
76	98.04
77	99.26

**Table 5.1: Cumulative Height Distribution of Adult Males**

The two likelihood functions,  $p(H | C)$  and  $p(H | \sim C)$ , are shown in Figure 5.1. The LLR,  $\log \frac{p(H | C)}{p(H | \sim C)}$ , is shown in Figure 5.2.

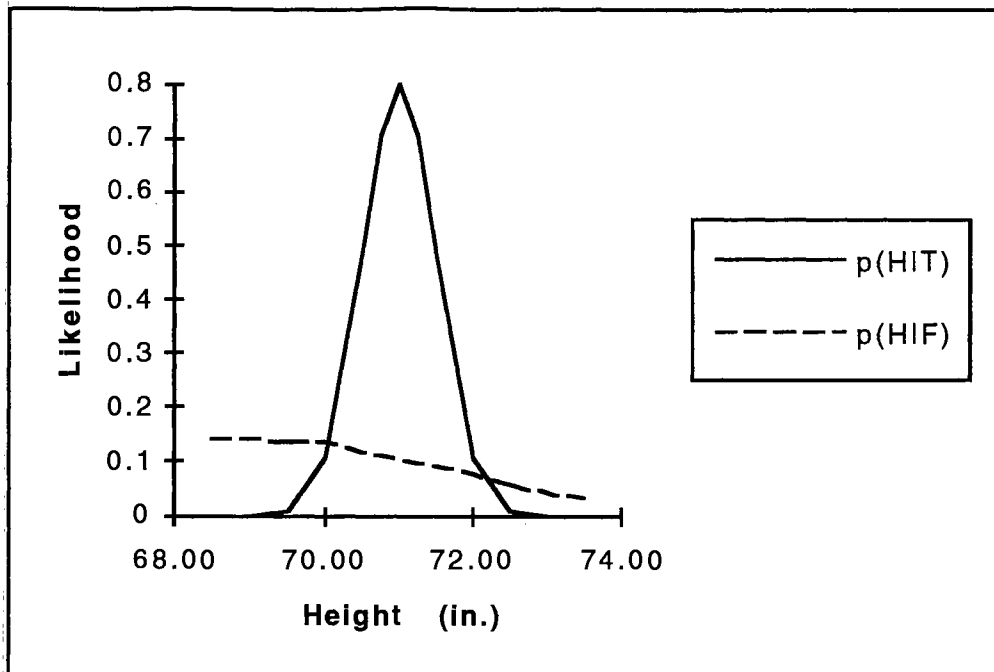


Figure 5.1: Likelihood Functions for  $H_X = 71$ .

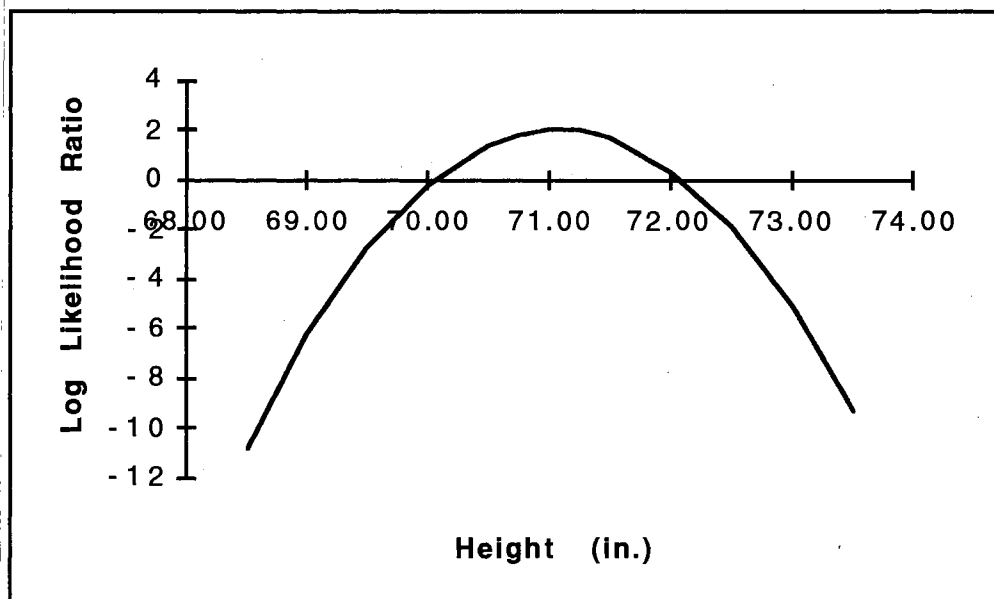


Figure 5.2: Log likelihood Ratio Function for  $H_X = 71$ .

Determination of the LLR for height is relatively simple because: (1) height is a scalar quantity as opposed to a vector; (2) intra-person variation in

measured height is mainly due to measurement errors and is therefore well modeled, independent of the identity claim; and (3) stable (adequately trained) population statistics,  $p(H)$ , are available. Biometrics such as voice and facial images present the following problems:

- Feature vectors extracted from the measurements are high dimensional. Therefore a very large number of measurements is needed to obtain stable estimates of the required likelihood functions.
- Intra-person variance is significant. Its magnitude and principal directions depend on the individual. Multiple enrollment sessions are needed to model this variance.

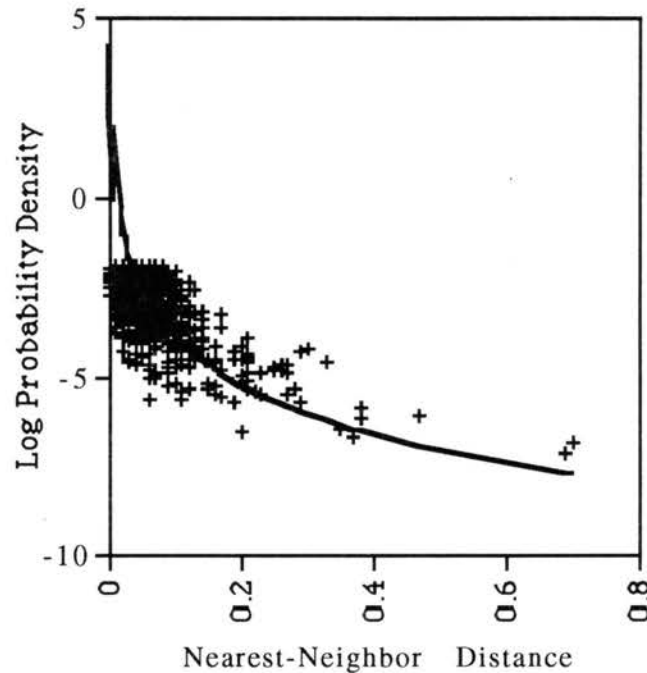
#### **5.4 Probability Density Estimation for Densely Sampled Populations**

The nature of the PDFs of speech and image feature vectors is not well understood. This lack of understanding is at least partially due to the difficulty of collecting large enough data samples to reliably estimate density in spaces of high dimensionality. In particular, evidence that speech and facial image features are well represented using known parametric distributions has not been reported. It is therefore reasonable to look to non-parametric, as opposed to parametric, statistical methods as the basis for verification decisions.

The probability density at a test point  $x$  can be estimated by considering a spherical ball of known volume,  $V$ , centered on  $x$ . The total probability mass within the ball is equal to the probability density (assuming the density is constant within the ball) times the volume of the ball, or  $p(x)*V$ . If  $N$  samples are drawn at random from  $p$ , and  $k$  fall inside the ball, then the measured relative frequency is equal to  $k/N$ . Equating these two probability estimates,

$$p(x)*V = \frac{k}{N} \quad \text{or} \quad p(x) = \frac{k}{NV} \quad (5.13)$$

at random from the same Gaussian PDF. For each test sample, Figure 5.3 plots the Euclidean distance from the test sample to the nearest of the 1000 training samples versus the true log probability density at the test sample. The solid curve is the log probability density estimate derived from Equation 5.16.



**Figure 5.3: Log Probability Density versus Nearest-Neighbor Distance for  $N(\mathbf{0}, I_2)$ .**

The maximum log density attained by  $N(\mathbf{0}, I_2)$  is equal to  $-\log(2\pi)$  or  $-1.838$ . Although actual log densities cannot exceed this value, the estimates derived from Equation 5.16 do exceed it for sufficiently small values of nearest-neighbor distance. Otherwise, Equation 5.16 provides an accurate model of the observed data.

### **5.5 Probability Density Estimation for Sparsely Sampled Populations**

We are concerned in practice with feature spaces for which  $v \geq 10$ . For  $v=10$ , one thousand samples would provide an extremely sparse covering of the

space, invalidating Equations 5.14, 5.15 and 5.16. To cover a 10-dimensional space with the same average density as the 2-dimensional example above would require roughly  $1000^{10/2} = 10^{15}$  samples. Therefore, while  $N=1000$  is a large sample size for  $v=2$ , it is a small sample size for  $v=10$ . It is generally not feasible to collect enough data to justify the use of Equation 5.16 for feature spaces of five or more dimensions.

Although the difficulty of estimating density in a high-dimensional space is increased by the relative sparseness of samples, the distance from the test point to the nearest sample (NN distance) remains the strongest data upon which to base the estimation. To develop a method of estimating probability density that is valid for high-dimensional spaces, we further examine the relationship between local density and NN distance. In previous work [4], the following relationship was conjectured:

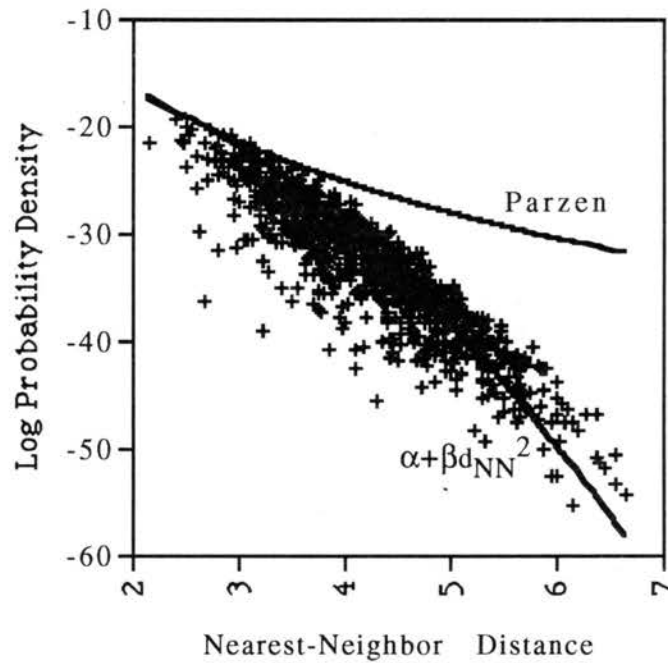
$$\ln p(\mathbf{x}) \approx \alpha + \beta (d_{NN})^2. \quad (5.17)$$

where  $\alpha$  and  $\beta$  are constants and  $(d_{NN})^2$  is the squared Euclidean distance from test point  $\mathbf{x}$  to the nearest sample. In terms of the  $\mathbf{x}$  and  $\mathbf{y}_i$  sample locations,

$$\ln p(\mathbf{x} | Y) \approx \alpha + \beta \min_{\mathbf{y}_j \in Y} |\mathbf{x} - \mathbf{y}_j|^2. \quad (5.18)$$

Equation 5.18 relates local log density to squared NN distance (as opposed to log NN distance) through an affine transformation. A Monte Carlo simulation was conducted in a manner similar to that described above in connection with Figure 5.3. In this case, 1000 samples were selected at random from the 13-dimensional Gaussian PDF  $N(\mathbf{0}, I_{13})$ . As before, nearest-neighbor distance is plotted versus true log density at each test sample location. The log density estimators derived from Equations 5.16 and 5.18 are plotted as solid curves.

Figure 5.4 shows that Equation 5.18 provides a much better fit to the data than does Equation 5.16.



**Figure 5.4: Log Probability Density versus Nearest-Neighbor Distance for  $N(0, I_{13})$ .**

Now consider the problem of evaluating the log likelihood of a set  $X$  of independent observation vectors  $x_j$  being generated by the same PDF underlying the set  $Y$ .

$$\ln p(X | Y) = \ln \prod_{x_j \in X} p(x_j | Y) \quad (5.19)$$

$$= \sum_{x_j \in X} \ln p(x_j | Y) \quad (5.20)$$

The conventional (large sample size) nearest-neighbor estimate, based on Equation 5.16, is:

$$\ln p(X | Y) \approx \sum_{x_j \in X} -\ln(NV_v) - v \ln \min_{y_i \in Y} |x_j - y_i|. \quad (5.21)$$

The small sample size estimate, based on Equation 5.18, is:

$$\ln p(X|Y) \approx \sum_{x_j \in X} \alpha + \beta \min_{y_i \in Y} |x_j - y_i|^2. \quad (5.22)$$

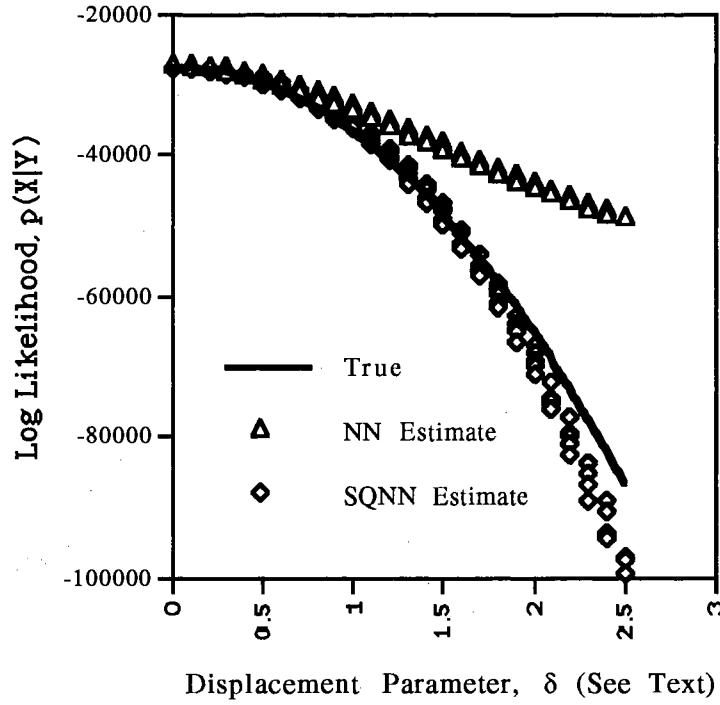
Again, the accuracy of these equations can be checked using simulated data with known PDFs. Suppose  $X$  and  $Y$  are generated from Gaussian PDFs,  $N(\mathbf{m}_X, \mathbf{C}_X)$  and  $N(\mathbf{m}_Y, \mathbf{C}_Y)$ , respectively. The true log likelihood,  $\ln p(X|\mathbf{m}_Y, \mathbf{C}_Y)$ , is given by:

$$\ln p(X|\mathbf{m}_Y, \mathbf{C}_Y) = \frac{n}{2} \ln |2\pi \mathbf{C}_Y| - \frac{n}{2} \sum_{x_j \in X} (\mathbf{x}_j - \mathbf{m}_Y)^T \mathbf{C}_Y^{-1} (\mathbf{x}_j - \mathbf{m}_Y) \quad (5.23)$$

$$= \frac{n}{2} [\ln |2\pi \mathbf{C}_Y| + \text{tr}(\mathbf{C}_Y^{-1} \mathbf{C}_X) + (\mathbf{m}_X - \mathbf{m}_Y)^T \mathbf{C}_Y^{-1} (\mathbf{m}_X - \mathbf{m}_Y)] \quad (5.24)$$

where  $| \cdot |$  signifies the determinant. A set of experiments was conducted to compare the true and estimated log likelihood values. For each data point, 1000 samples of dimension  $v=13$  were generated at random for  $Y$  using  $\mathbf{m}_Y = \mathbf{0}$  and  $\mathbf{C}_Y = \mathbf{I}_{13}$ , and another 1000 samples of dimension  $v=13$  were generated at random for  $X$  using  $\mathbf{m}_X = (\delta, \delta)^T$  and  $\mathbf{C}_X = \mathbf{I}_{13}$ . Values of the displacement parameter  $\delta$  were chosen in the range from 0 to 2.5 in increments of 0.02. A scatter plot of the true log likelihood from Equation 5.24 versus estimates from Equations 5.21 and 5.22 is shown in Figure 5.5.





**Figure 5.5: Comparison of Log Likelihood Estimators for  $N(0, I_{13})$ .**

The values of  $\alpha$  and  $\beta$  used in Equation 5.22 for this simulation were derived as follows. The true log density at each sample point  $y_i$  was evaluated using the known parameters  $m_Y$  and  $C_Y$ ,

$$\lambda_i = -\frac{v}{2} \ln 2\pi - \frac{1}{2} \ln |C_Y| - \frac{1}{2} (y_i - m_Y)^T C_Y^{-1} (y_i - m_Y). \quad (5.25)$$

The squared Euclidean distance  $d_i^2$  from each sample  $y_i$  to its nearest neighbor in  $Y$  (excluding itself) was determined as

$$d_i^2 = \min_{\substack{y_j \in Y \\ y_j \neq y_i}} |y_j - y_i|^2. \quad (5.26)$$

The constants  $\alpha$  and  $\beta$  were then chosen to minimize the squared error in the equation  $\lambda_i = \alpha + \beta d_i^2$ . This was accomplished by

$$\alpha = \frac{\sum d_i^4 \sum \lambda_i - \sum d_i^2 \sum d_i^2 \lambda_i}{N \sum d_i^4 - (\sum d_i^2)^2} \quad \text{and} \quad \beta = \frac{N \sum d_i^2 \lambda_i - \sum d_i^2 \sum \lambda_i}{N \sum d_i^4 - (\sum d_i^2)^2}. \quad (5.27)$$

The values of  $\alpha$  and  $\beta$  used in Figure 5.5 were  $\alpha = -12.49$  and  $\beta = -1.06$ . These values were determined as indicated above, using samples from the Y set only.

Figure 5.5 shows that the estimates derived from Equation 5.21 are reasonably accurate for  $d < 1$ , but overestimate the log likelihood outside this range. As the PDFs move farther apart, test points and their nearest neighbors often have significantly different densities, invalidating Equation 5.21.

Equation 5.22, however, remains accurate over a larger range of displacements of the PDFs.

The primary evidence cited in [4] in support Equation 5.17 was that a voice recognition algorithm based on Equation 5.22 gave dramatically higher performance than one based on Equation 5.21. Two additional arguments are the following. First, the density of the feature vectors under consideration is assumed to have a finite upper bound, so that the negative log density has a lower bound. However, nothing precludes arbitrarily small NN distances from occurring, particularly when the test point is in the vicinity of the distribution mode. This effect can be seen in Figure 5.3. The logarithm of  $d_{NN}$  therefore has no lower bound, whereas  $(d_{NN})^2$  does. Second, when the test point is distant from the distribution mode, the nearest sample is likely to be much nearer to the mode, so that  $(d_{NN})^2$  will be roughly equal to the squared distance from the test point to the mode. For quasi-Gaussian distributions, the negative log density will then rise in proportion to  $(d_{NN})^2$ .

Investigation of the validity of Equation 5.17 using experimental data would be very difficult because of the intractably large number of samples that would be required to accurately estimate the underlying density functions. It is possible, however, to determine the theoretical relationship between the expected value of log probability density and NN distance for known (parametric) density functions. The function relating these quantities

is developed for normalized Gaussian PDFs in Appendix A. This function approximates a logarithm in the limit of large sample size and low dimensionality, and a parabola in the limit of small sample size and high dimensionality. The parabolic approximation appears to be valid in cases of practical interest such as multi-media PIV.

## 5.6 Dimensionality Estimation

Whether a given population is considered densely or sparsely sampled depends on the number of samples observed and on the dimensionality of the subspace which they occupy. In the case of voice data, the nominal dimensionality of the feature space is 16. However, it is well known that nearly all the variance of voice feature vectors occurs within a subspace of dimension less than 16. A conventional method of estimating the dimensionality of the subspace is to count the number of significant eigenvalues of the feature vector covariance matrix. The dimensionality of speech data determined in this manner is reported to be between 8 and 12.

The above method overestimates true dimensionality when the observed samples lie on curved, as opposed to linear, surfaces. Consider, for example, a two dimensional space in which all data samples lie on the unit circle. The covariance matrix has two equal eigenvalues, indicating a dimensionality of two. However, the location of any data sample can be specified exactly using only one parameter value. The true, or *intrinsic*, dimensionality is equal to one in this case because data are distributed throughout the space with only one degree of freedom.

Intrinsic dimensionality is important in nonparametric density estimation because it governs the relation of test points to their nearest neighbors. Pettis, et. al [5] developed an intrinsic dimensionality estimator

based on near-neighbor distances. Consider a set of samples,  $\{x_i\}$ ,  $0 \leq i < n$ . Let  $r_{k,i}$  be the distance from sample  $x_i$  to its  $k$ th nearest neighbor, and let  $\bar{r}_k$  be the average distance to the  $k$ th nearest neighbor,  $\bar{r}_k = \frac{1}{n} \sum_{i=1}^n r_{k,i}$ . Pettis shows

that

$$E\{\bar{r}_k\} = \frac{k^{1/d} C_n}{G_{kd}} \quad (5.28)$$

where  $G_{kd} = \frac{k^{1/d} \Gamma(k)}{\Gamma(k + \frac{1}{d})}$  and  $C_n = \frac{1}{n} \sum_{i=1}^n [np(x_i) V_d]^{-1/d}$ . Taking logs in Equation

5.28, and substituting  $\bar{r}_k$  in place of its expected value,  $E\{\bar{r}_k\}$ , gives

$$\ln G_{kd} + \ln \bar{r}_k = \frac{1}{d} \ln k + \ln C_n. \quad (5.29)$$

The term  $\log G_{kd}$  is shown to be close to 0 for all  $k$  and  $d$ , and the term  $\log C_n$  is independent of  $k$ . Therefore a plot of  $\ln \bar{r}_k$  as a function of  $\ln k$  has slope equal to  $1/d$ . Pettis et. al estimate  $1/d$  by performing a linear regression of  $\ln \bar{r}_k$  versus  $\ln k$  for  $1 \leq k \leq K$ .

### 5.6.1. An Extension of the Method of Pettis, et. al

Suppose that two sets of samples are available:  $X = \{x_i\}$ ,  $0 \leq i < n$ , and  $Y = \{y_j\}$ ,  $0 \leq j < m$ . Redefine  $r_{k,i}$  as the distance from test sample  $x_i$  to its  $k$ th nearest neighbor among the set  $Y$ . Substituting  $r_{k,i}$  (rather than  $\bar{r}_k$ ) for  $E\{\bar{r}_k\}$  in Equation 5.28, and taking logs of both sides,

$$\ln G_{kd} + \ln r_{k,i} = \frac{1}{d} \ln k + \ln c_i \quad (5.30)$$

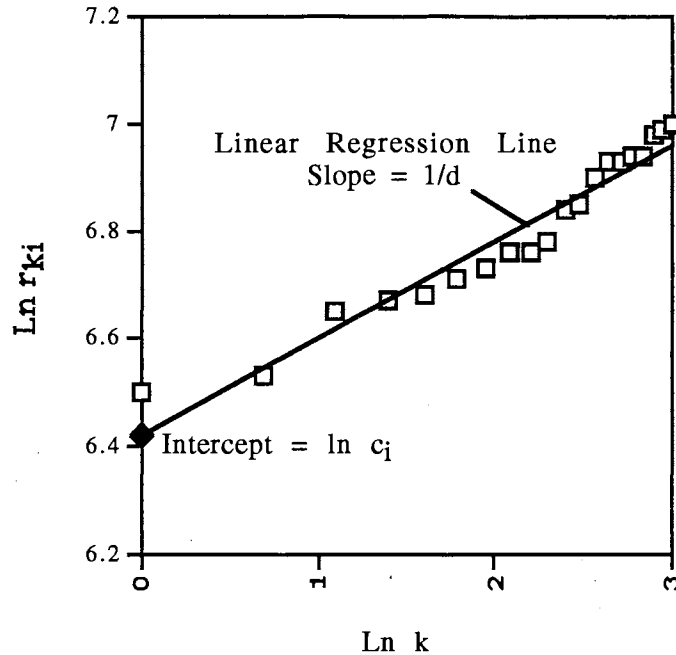
where

$$c_i = [np(x_i | Y) V_d]^{-1/d}. \quad (5.31)$$

The value of  $\ln c_i$  can be obtained for each  $i$ ,  $0 \leq i < n$ , from the approximate solution to Equation 5.30 through linear regression. As before, a plot of  $\ln r_{k,i}$  as a function of  $\ln k$  has slope equal to  $1/d$ . The additive term in the regression (which is not used by Pettis, et. al) is equal to  $\ln c_i$ . Taking logs of Equation 5.31 and solving for  $\ln p(\mathbf{x}_i | Y)$ ,

$$\ln p(\mathbf{x}_i | Y) = -\ln(nV_d) - d \ln c_i. \quad (5.32)$$

Equation 5.32 is identical to Equation 5.15, but with  $c_i$  substituted for  $r$ , which in the notation of this section would be  $r_{1,i}$ . This result is reasonable, since  $c_i$  can be interpreted as an estimator of  $r_{1,i}$ . Figure 5.6 shows a plot of  $\ln r_{k,i}$  as a function of  $\ln k$  for  $1 \leq k \leq 20$  with the linear regression line superimposed. These values of  $\ln r_{k,i}$  are for a randomly chosen frame of voice data. The reciprocal of the slope of the line equals 5.6, the approximate local dimensionality at the test point,  $\mathbf{x}_i$ . Note that the zero value on the abscissa occurs at  $\ln k = 0$ , or  $k = 1$ . The additive term in the regression,  $c_i$ , may therefore be interpreted as a smoothed or interpolated estimate of  $r_{1,i}$ . We therefore refer to density estimation by means of Equation 5.32 as the *interpolated nearest neighbor* (INN) method.



**Figure 5.6: Plot Showing Approximate Linear Relationship of  $\ln r_{ki}$  Versus  $\ln k$ , With Slope Equal To Reciprocal of Local Dimensionality.**

Equation 5.32 has two advantages over Equation 5.15. First, the linear interpolation involved in the derivation of  $c_i$  makes  $c_i$  a less "noisy" indicator of local density than  $r_{ki}$ . Second, an estimate of the local dimensionality,  $d$ , at each sample point  $x_i$  is available as a bi-product of the computation of  $c_i$ . This estimate can be used in Equation 5.15, as opposed to the assumed constant value of dimensionality that would normally be used in Equation 5.15.

### 5.7 Likelihood Ratio Estimation

Using Equation 5.17 as the estimator of local probability density, one may evaluate the LLR of the claimant versus other individuals. Combining Equations 5.12 and 5.18 gives the decision rule,

$$\sum_{x_j \in X} \alpha_C + \beta_C \min_{y_i \in Y_C} |x_j - y_i|^2 - \max_{\substack{S_k \in D \\ S_k \neq S_C}} \left\{ \sum_{x_j \in X} \alpha_k + \beta_k \min_{y_i \in Y_k} |x_j - y_i|^2 \right\} > T, \quad (5.33)$$

where  $T$  is the acceptance threshold,  $S_C$  is the claimant,  $Y_k$  is the set of enrollment frames belonging to individual  $S_k$ , and  $D$  is the set of individuals, or "cohort", for whom enrollment data is available. The vectors  $x_j$  and  $y_i$  contain pixel values within the  $27 \times 32$  face boxes. The subscripts attached to  $\alpha$  and  $\beta$  indicate that these constants are associated with the model PDFs. It is assumed that they do not vary as a function of the individual. If this is the case, the values of  $\alpha$  and  $\beta$  do not affect verification performance as measured by the ROC curve. In the remainder of this report, we use the values  $\alpha = 0$  and  $\beta = 1$ .

### 5.8 Multiple Models Per Individual

The decision rule represented by Equation 5.33 is based on the premise that one PDF characterizes each individual. Suppose that subject  $S_i$  possesses  $N$  distinct states,  $\omega_n$   $0 \leq n < N$ , and that the likelihood function associated with the combination of subject  $S_i$  and state  $\omega_n$  is  $p(X | S_i, \omega_n)$ . Distinct voice states, for example, might be assumed for "morning voice" and "afternoon voice". Facial appearance states might be associated with the presence or absence of glasses or hats. Further suppose that on any particular occasion,  $\omega_j$  is selected at random with probability  $p(\omega_n)$ . Then the likelihood function for subject  $S_i$  is:

$$p(X | S_i) = \sum_{n=1}^N p(X | S_i, \omega_n) p(\omega_n). \quad (5.34)$$

If  $p(\omega_j)$  is a uniform density, and if the sum is dominated by one term, then Equation 5.34 may be approximated by:

$$p(X | S_i) = \frac{1}{N} \max_n \{ p(X | S_i, \omega_n) \}. \quad (5.35)$$

Combining Equation 5.9 and Equation 5.35 gives the LLR test,

$$\max_n \{ \ln p(X | C, \omega_n) \} - \max_{\substack{S_i \in D \\ S_i \neq C}} \{ \max_n \{ \ln p(X | S_i, \omega_n) \} \} > T. \quad (5.36)$$

Now combining Equations 5.22 and 5.36,

$$\begin{aligned} & \max_{0 < n < N} \left\{ \sum_{x_j \in X} \alpha_C + \beta_C \min_{y_i \in Y_{Cn}} |x_j - y_i|^2 \right\} \\ & - \max_{\substack{S_k \in D \\ S_k \neq SC}} \left\{ \max_{0 < n < N} \left\{ \sum_{x_j \in X} \alpha_k + \beta_k \min_{y_i \in Y_{kn}} |x_j - y_i|^2 \right\} \right\} > T, \end{aligned} \quad (5.37)$$

where  $Y_{kn}$  is the set of feature vectors representing the  $n^{\text{th}}$  model of Subject  $S_k$ . In the experiments performed,  $Y_{kn}$  was the set of frames observed in the  $n^{\text{th}}$  enrollment session of Subject  $S_k$ .

## 5.9 Summary

The decision to accept or reject a claimed identity is reached according to Bayes decision rule by determining whether the likelihood of the data given the claimed identity is greater or less than the likelihood given the alternative (the prior probabilities being equal). The former likelihood is estimated using the claimant's model. The later likelihood is estimated using models for a set of individuals other than the claimant, called a cohort.

The individual model provides an estimate of local probability density at any sample point. The conventional nearest-neighbor density estimate, based on asymptotic large-sample arguments, was shown to be appropriate for low-dimensional feature spaces. Accurate likelihood estimation was demonstrated



(using Equation 5.21) for 1000-sample populations from known 2-dimensional Gaussian PDFs.

For problems involving small numbers of samples with high dimensionality, the asymptotic arguments are not valid. Evidence was shown that in some cases, the negative log of local probability density is more closely related to the squared nearest-neighbor distance (SQNN) than to its logarithm. An expression is developed in Appendix A for the conditional expectation of local density given nearest-neighbor distance for normalized Gaussian PDFs. This function approximates a logarithm in the limit of large sample size and low dimensionality, and a parabola in the limit of small sample size and high dimensionality.

The relevant measure of dimensionality is the intrinsic, or local dimensionality, which is the minimum number of independent parameters needed to specify the location of a point in the space occupied by the data. A method of estimating intrinsic dimensionality was examined and modified to provide estimates of both dimensionality and probability density at any sample point. The density estimator resembles the conventional nearest-neighbor estimator, but uses an interpolated nearest-neighbor distance (INN), which takes into account the distances to the  $K$  nearest neighbors.

Bayes' decision rule was formulated in terms of summations over time of the estimated local log probability densities at the test sample points for both the SQNN and INN estimators. These formulas were developed without reference to the type of data being observed. To proceed further with development of PIV algorithms, closer attention must be given to the specifics of voice and image sequence data. Using the experimental data described in the next chapter, "front end" processes will be developed by which raw input

data are converted to feature vectors. These data will then be used to test and compare the performance of the SQNN and INN "back end" processes.

## CHAPTER 6

### EXPERIMENTAL DATA

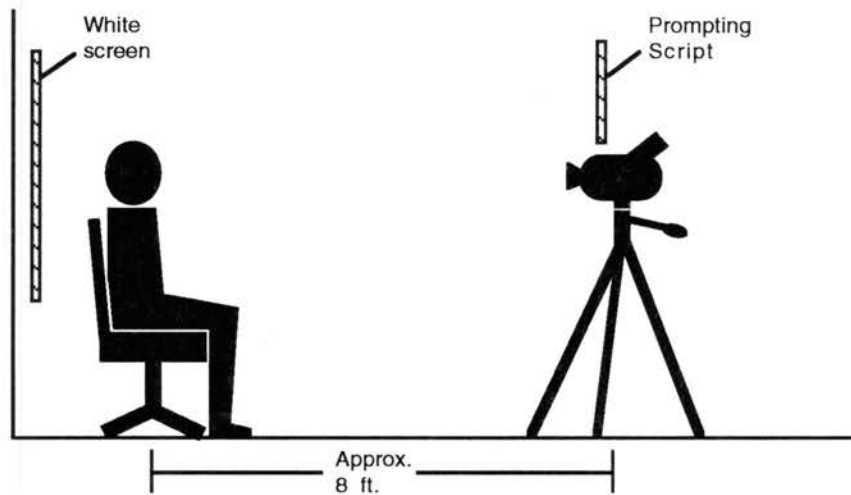
#### 6.1 Introduction

A small experimental database was collected for the purpose of developing and testing multi-media PIV algorithms. Motion video and audio recordings were made of subjects reading from scripts on multiple occasions. The experimental setup and the scripts were comparable to what might be used in a practical application of multi-media PIV.

#### 6.2 Equipment Setup

Recordings took place in a laboratory room at Oklahoma State University. The room had ceiling-mounted fluorescent lighting, and slight noticeable reverberation due to absence of carpeting or sound-absorbant furniture. No special measures were taken to control the lighting or sound characteristics. An illustration of the experimental setup is shown in Figure 6.1. The subject was seated in a chair in front of a white projection screen. An 8mm camcorder was set up on a tripod about 8 feet in front of the subject. The vertical position of the camcorder was adjusted separately for each subject to

allow for differences in the subject's height.



**Figure 6.1: Illustration of Experimental Setup.**

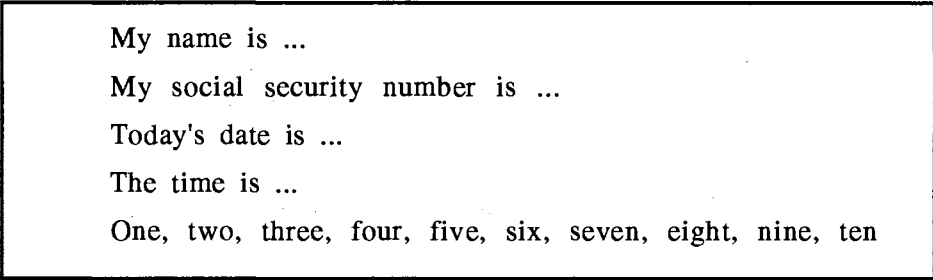
The camcorder was a Cannon model E250A with 6.7 - 80.4 mm autofocus macro zoom lens. The zoom lens was adjusted separately for each subject. An electret lapel microphone (Radio Shack model 33-3003), clipped to the subject's shirt or coat, was used instead of the camcorder's built-in microphone. The audio recording level was regulated automatically by an automatic gain control with a time constant of about 10 seconds.

### 6.3 Subjects

Twelve subjects participated voluntarily in the experiment. They consisted of OSU students (graduate and undergraduate), faculty, and staff. The subject population was diversified with respect to sex, age, and country of origin. The purpose of the experiment was explained to each subject.

### 6.4 Prompting

The subject was asked to speak several phrases in a natural voice. The phrases were prompted using a paper script that the experimenter held just above the camcorder. The script is shown in Figure 6.2.



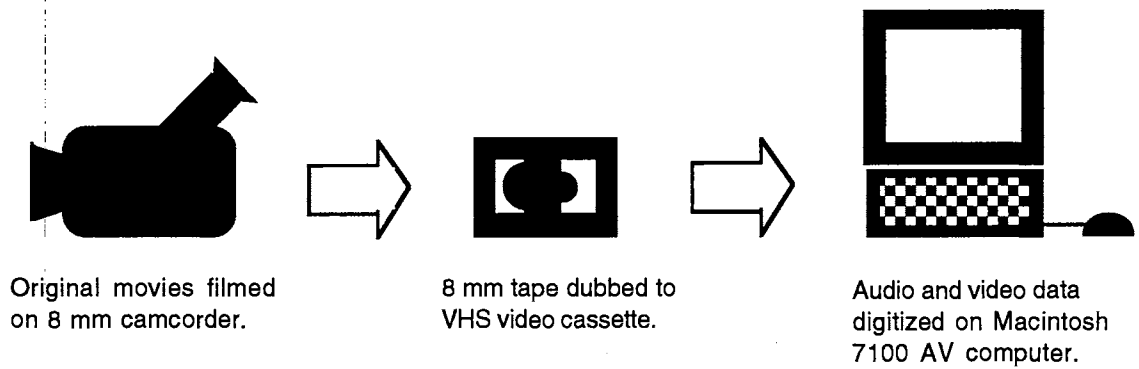
My name is ...  
My social security number is ...  
Today's date is ...  
The time is ...  
One, two, three, four, five, six, seven, eight, nine, ten

**Figure 6.2: Script Used for Prompting Subjects.**

Subjects were told that the exact choice of words to complete the phrases was not important. For example, the date could be spoken as "November second", "the second of November", or in any other normal way.

### 6.5 Initial Data Processing

After completing all recordings, the 8 mm video tape was dubbed to a VHS format tape. Each session was then digitized to a separate movie file using a Macintosh 7100 AV computer. Both the audio and video channels were digitized. These steps are shown in Figure 6.3.



**Figure 6.3: Initial Data Processing.**

FusionRecorder™ version 1.1 was used to perform the digitization. Resulting digitized movies were stored in Apple QuickTime™ format. For most sessions, the original video recording included the subject sitting down, clipping on the microphone, reading the prompts, and getting up to leave. Only the portion of each session in which the subject was reading the prompts was digitized. Therefore each digitized session started immediately with "My name is...", and ended with the digit sequence. The digitized portion of the session was not edited in any way. Relevant video and audio digitization parameters are shown in Figure 6.4.

<p><u>Video:</u></p> <p>10 frames / second</p> <p>160 (w) x 120 (h) pixels / frame</p> <p>8 bits / pixel - grayscale</p> <p>Cinepak compression - best quality</p> <p><u>Audio:</u></p> <p>22 kHz sampling</p> <p>16 bits per sample, linear quantization</p> <p>no compression</p>
---

**Figure 6.4: Video and Audio Digitization Parameters.**

Video compression was used to enable the video image sequences to be stored on the available disk drive. Each image frame represents  $160 \times 120 = 19,200$  pixels. Without compression, the video portion of a typical 20-second movie would require 3.8 megabytes, or

$$(1 \text{ byte/pixel}) \times (19200 \text{ pixels/frame}) \times (10 \text{ frames/sec}) \times (20 \text{ seconds}).$$

The audio portion requires 880 kilobytes, or

$$(2 \text{ bytes/sample}) \times (22000 \text{ samples/second}) \times (20 \text{ seconds}).$$

The total requirement is therefore about 4.7 megabytes per movie.

Using Cinepak video compression (an Apple proprietary compression technique), a 20-second movie requires 2.1 megabytes, of which about 1.2 megabytes are allocated to the video portion. This represents a video compression ratio of 3.3 to 1. Cinepak compression was chosen from several alternative compression methods because it produced no visible degradation of the images.

## 6.6 Inventory of Sessions

A total of 48 sessions from 12 subjects were collected. The experiments described in this report used only the first four sessions of those ten speakers who had four or more sessions. In Table 6.1, subjects are identified by their initials.

Subject	Number of Sessions	Session Numbers
MA	4	1-4
GB	1	-
BB	5	5-8
BC	4	9-12
KD	4	13-16
CF	5	17-20
DH	6	21-24
YL	4	25-28
RM	4	29-32
SR	4	33-36
GW	3	-
RY	4	37-40

**Table 6.1: Inventory of Sessions by Subject.**

Data for subjects GB and GW were not used because they had less than four sessions. To identify a particular session, we will use either the session number or the notation XX-N, where XX is the subject's initials, and N is the session number, ranging from 1 to 4. For example, KD-3 is the third session of subject KD, or Session 15.

### **6.7 Subjective Observations**

The quality of all the digitized movies was sufficient to allow the experimenter (who knew the subjects) to identify each subject immediately from either the video or audio data.

As noted previously, there was slight noticeable reverberation in the room. Also, some variability in sound spectral balance was noticed from one session to another, possibly due to variations in the placement of the lapel



microphone. The ceiling lighting caused some variation in facial illumination, which was noticeable when subjects looked up or down.

An example of one image frame from MA-1 is shown in Figure 6.5.



**Figure 6.5: An Example Image Frame from MA-1.**

The following observations were made with respect to subjects' appearance and behavior:

1. BB wore glasses in BB-3, but not in the other sessions.
2. BB wore a hat in BB-2 and BB-3, but not in the other sessions.
3. BC wore a different hairstyle in BC-2 and BC-3 than in BC-1 and BC-4.
4. KD wore a hat in KD-1, but not in the other sessions.
5. SR scratched his head in SR-1.
6. SR touched his mouth and looked to the side in SR-2.
7. RY looked down at his watch in RY-2, RY-3, and RY-4.

## **6.8 Summary**

The OSU database contains at least four sessions from each of ten subjects. The data is reasonably well controlled in terms of camera and microphone position, consistency of lighting, and background. Normal variations in subjects' appearance and behavior are observed.

## CHAPTER 7

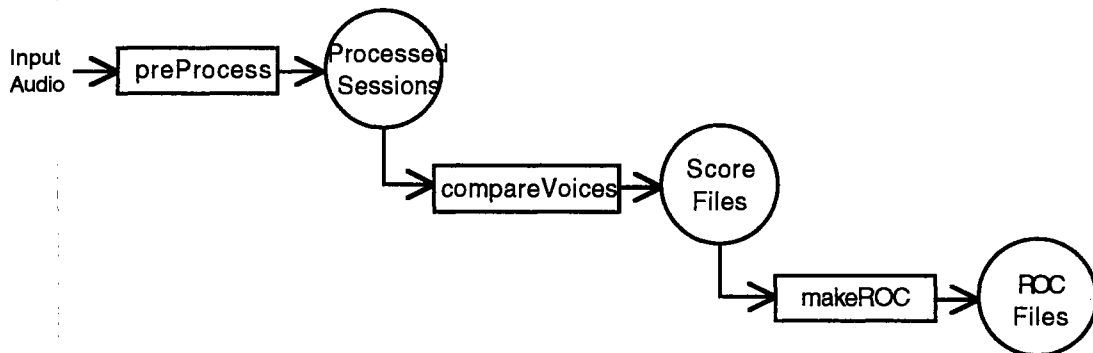
### VOICE DATA FEATURE EXTRACTION

#### 7.1 Introduction

This chapter describes the processing and analysis that was performed to convert the sampled waveform sound input data to a sequence of feature vectors to be used by the PIV "back end". The sound data from the OSU database was at first treated independently of the video data to develop a voice-only PIV algorithm. Results of testing the voice-only PIV algorithm and the multi-media PIV algorithm are presented in Chapter 9.

#### 7.2 Voice-Only PIV Algorithm

Figure 7.1 is a data flow diagram showing the sequence of processing steps involved in the voice-only PIV algorithm. Rectangles represent processes (programs), whereas circles represent data structures (disk files).



**Figure 7.1: Data Flow Diagram of Voice Data Processing.**

The voice algorithm is much simpler than the face PIV algorithm, primarily because locating the speech portion of the audio signal is less difficult than locating the face portions of the image sequence. The input signal is first passed through a pre-processor, *preProcess*, which includes several operations described in Section 7.3. Comparison of voices from different sessions is accomplished using the program *compareVoices*, described in Section 7.4, which implements the LLR measure. Finally, the score files produced by *compareVoices* are processed by *makeROC*. *MakeROC* converts the log likelihoods to log likelihood ratios as described in Section 5.7 and handles multiple models per subject as described in Section 5.8. The output of *makeROC* is a Receiver Operating Characteristic (ROC), which is used to measure verification accuracy.

### 7.3 Signal Processing

Pre-processing of the voice signal consists of four operations performed in tandem: spectral analysis, silence frame pruning, blind deconvolution, and frequency differencing. These operations are described in the following subsections.

#### 7.3.1. Spectral Analysis

The electrical signal from the microphone was sampled at a rate of 22.0 kHz. Initially, eight-bit linear quantization was used to conserve disk space. This produced audible distortion and limited voice verification accuracy. Therefore, the data was re-digitized using 16-bit linear quantization as indicated in Table 3.4. Preemphasis filtering was applied in the form of simple differencing of consecutive samples,  $y_i = x_i - x_{i-1}$ . This boosts high frequencies at the rate of 6 dB per octave, reducing the spectral dynamic

range of the speech signal. The preemphasized signal was segmented into overlapping frames, each containing 704 samples or 20.0 milliseconds. Consecutive frames were offset by 440 samples or 32.0 milliseconds. Each 704-sample frame was multiplied by a Hamming window and padded with zeros to 1024 samples. Squared spectral magnitudes were computed from a 1024-sample DFT. Dot products were then computed between these squared magnitudes and the frequency responses of each of 16 bandpass filters. Dot product  $p_{ij}$  measures the power at frame  $i$  within frequency band  $j$ .

The filters were designed to cover the range of frequencies from 350 Hz to 5000 Hz. Details of the filters are shown in Table 7.1. The specified low- and high-frequency cutoffs are the frequencies at which filter response is down 3 dB relative to the center frequency. At frequencies below 1000 Hz, the filters have a constant bandwidth of 150 Hz. Above 1000 Hz, they have a constant Q factor (ratio of center frequency to bandwidth) of 6.0. This design is consistent with studies of human perception indicating that perceived pitch of tones is proportional to frequency below 1000 Hz, and proportional to log frequency above 1000 Hz [122].

Filter Number	Low-Freq. Cutoff (Hz)	High-Freq. Cutoff (Hz)	Center Freq. (Hz)	Bandwidth (Hz)	Filter Q Factor
1	350	500	425	150	2.8
2	480	630	555	150	3.7
3	610	760	685	150	4.6
4	740	890	814	150	5.4
5	866	1024	945	157	6.0
6	1000	1183	1092	188	6.0
7	1155	1366	1261	210	6.0
8	1335	1578	1456	243	6.0
9	1542	1822	1682	280	6.0
10	1781	2105	1943	324	6.0
11	2058	2432	2245	374	6.0
12	2377	2809	2593	432	6.0
13	2745	3244	2995	499	6.0
14	3171	3748	3459	576	6.0
15	3663	4329	3996	666	6.0
16	4230	5000	4615	769	6.0

**Table 7.1: Filterbank Design Data**

The power in each filter "channel" was processed using a nonlinear compression function. Following Olano [36], the fourth root was used instead of the more traditional logarithm to avoid the extreme sensitivity of the log function at very low power levels. The resulting 16-element vector,  $(q_1, q_2, \dots, q_{16})^T$ , was  $L_2$  normalized, so that the sum of its squared elements equals a constant value (1.0) across all frames. This was accomplished as follows:

$$q_{ij} = \left[ \frac{\sqrt{p_{ij}}}{\sum_{k=1}^{16} \sqrt{p_{ik}}} \right]^{1/2} \quad (7.1)$$

The total log power within the 350 - 5000 Hz band was computed each frame as:

$$P_i = \log \sum_{j=1}^{16} p_{ij} \quad (7.2)$$

### 7.3.2. Silence Frame Pruning

Silence frame pruning was performed as follows. A histogram was computed of the log power of all frames of the input signal. A silence threshold was set equal to the 10 percentile of this histogram plus 6 dB. All input frames with log power exceeding the silence threshold were retained, while others were discarded. This algorithm assumes that the dynamic range of "silence" is 6 dB.

### 7.3.3. Blind Deconvolution

Blind deconvolution is a method of compensating for the unknown frequency response of the input channel [123]. Although the same microphone and electronics were used in all sessions, the frequency response of the microphone depends on its location and orientation relative to the subject's mouth, and whether it is obscured by clothing or other objects. Blind deconvolution was accomplished by dividing each feature,  $q_{ij}$ , by its long-term average value, and re-applying L2 normalization within each frame.

$$b_{ij} = \gamma_i \frac{q_{ij}}{\frac{1}{K} \sum_{k=1}^K q_{kj}} \quad (7.3)$$

where  $K$  is the number of frames in the session, and

$$\gamma_i = \left[ \sum_{j=1}^{16} \left( \frac{q_{ij}}{\frac{1}{K} \sum_{k=1}^K q_{kj}} \right)^2 \right]^{-1/2} \quad (7.4)$$

### **7.3.4. Frequency Differencing**

The final step of pre-processing was to compute differences between the feature values at consecutive frequencies (with "wrap around") as follows:

$$f_{ij} = \begin{cases} b_{ij} - b_{i(j-1)} & \text{if } j > 1 \\ b_{i1} - b_{i16} & \text{if } j = 1 \end{cases} \quad (7.5)$$

This step has the effect of de-correlating the elements of the feature vectors, and emphasizing spectral regions near the formants. The feature vectors used in the remaining processing were  $\mathbf{f}_i = (f_{i1}, f_{i2}, \dots, f_{i16})^T$ .

## **7.4 Voice Comparison**

The likelihood ratio score given by Equation 2.27 was applied in the voice PIV system exactly as in the face PIV system. The feature vectors  $\mathbf{f}_i$  of Equation 4.8 were used, giving a feature space dimensionality of 16.

## **7.5 Summary**

Processing steps were described by which the sampled waveform sound data is converted to a feature vector per frame. Spectral analysis is performed within 32 millisecond overlapping windows using a 16-channel FIR filterbank covering the range of frequencies from 350 Hz to 5000 Hz. The filterbank output is normalized each frame in a manner that preserves spectral shape but is independent of overall amplitude. Low-amplitude frames, corresponding to silence intervals, are eliminated. Equalization for the unknown and variable frequency response of the microphone is accomplished using blind deconvolution. Enhancement of salient spectral features is accomplished through differencing of adjacent frequency channels. Sound feature vectors are 16 dimensional and are produced at the rate of 50 per second.

## CHAPTER 8

### VIDEO DATA FEATURE EXTRACTION

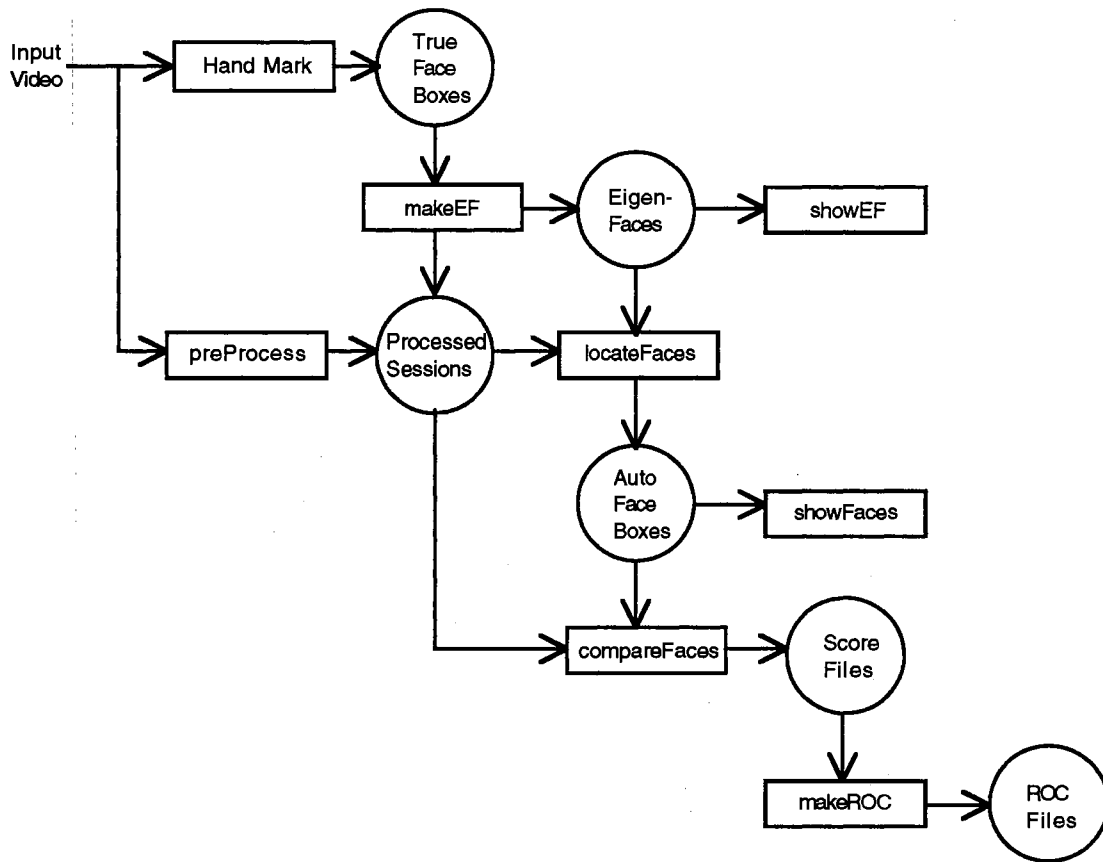
#### 8.1 Introduction

This chapter describes the processing and analysis that was performed to convert the raw video input data to a sequence of feature vectors to be used by the PIV "back end". The video data from the OSU database was at first treated independently of the audio data to develop a video-only PIV algorithm. Results of testing the video-only PIV algorithm and the multi-media PIV algorithm are presented in Chapter 9.

#### 8.2 Video-Only PIV Algorithm

Figure 8.1 is a data flow diagram showing the sequence of processing steps involved in the video-only PIV algorithm. Rectangles represent processes (programs), whereas circles represent data structures (disk files).





**Figure 8.1: Data Flow Diagram of Video Data Processing.**

Information about the subject's identity was expected to be concentrated in the part of each frame corresponding to the face. Therefore, the first intermediate goal is to locate the face within each image. This is accomplished by the program *locateFaces*, described in Section 8.5. *LocateFaces* operates on video data that has been pre-processed as described in Section 8.4. The model used to locate faces is derived from hand marking of face boundaries (or "face boxes") in a subset of the database. The hand marking process is described in Section 8.3. Comparison of faces is accomplished using the program *compareFaces*, described in Section 8.6, which implements the LLR measure. Finally, the score files produced by *compareFaces* are processed by *makeROC*. The score files produced by *compareFaces* are in the same format as those

produced by *compareVoices*. Therefore, the same program, *makeROC*, is used for both voice and face processing. The output of *makeROC* is a Receiver Operating Characteristic (ROC), used to measure verification accuracy.

### 8.3 Manual Location of Faces

The process of manual marking defines faces by example. For the purpose of this study, a face was considered to be a rectangular region bounded from the left and right by the outer corners of the eyes, from above by the top of the eyebrows, and from below by the bottom of the nose. This definition was arrived at by experimentation, and intentionally excludes the mouth, which exhibits greater within-subject variability during speech than other facial features.

A computer program, called *handMark*, was developed to manually locate each face in the entire database. *HandMark* operates as follows: A movie file for a session is opened, and the first image, similar to Figure 6.5, is displayed in a 160 x 120 pixel window. On-screen buttons are provided to zoom in or out, move up, down, right, or left, and rotate the image clockwise or counterclockwise. After each adjustment, the selected part of the image is scaled, translated, or rotated appropriately and re-displayed in the window. Using the on-screen controls, the operator adjusts the image to contain only the face. Ideally, the eyebrows are horizontal and the nose is vertical. It is sometimes not possible to locate an ideal face (even using rotation), particularly when the subject does not directly face the camera. Subjective judgement was used to determine the best alignment. The rectangle bounding the face is referred to as the "face box". After locating each face box, pushing the "Next" button causes the coordinates of the box to be saved in a data file, and the next frame to be displayed (initially within the previous frame's face

box). The data obtained from StepFrame was used as "ground truth" for training of algorithms to automatically locate faces.

## 8.4 Signal Processing

Various forms of image pre-processing were examined at each stage of the algorithm. Two general conclusions are: (1) Some form of gray-level mapping function is needed to reduce the effect of variations in illumination and/or reflectivity of the face; (2) Some form of spacial- or frequency-domain filtering is needed to enhance facial features such as the edges of the eyes, nose, and mouth. The pre-processing steps used in all experiments described below consist of histogram stretching followed by the Sobel gradient operator. These pre-processing steps are applied independently to each image frame.

### 8.4.1. Histogram Stretching

Histogram stretching is performed as follows. Suppose  $l_1$  and  $l_2$  are gray level values corresponding to specified percentiles of the histogram of the input image. Each pixel of the image is scaled in such a way that these same percentile values of the output histogram occur at  $L_1$  and  $L_2$ . If  $x$  and  $y$  are gray level values of a pixel before and after histogram stretching, they are related as follows:

$$y = \left[ \frac{L_2 - L_1}{l_2 - l_1} (x - l_1) + L_1 \right]_{0}^{255} \quad (8.1)$$

where the square brackets denote clipping within the indicated limits. In practice, the specified percentiles are 5% and 95%, and  $L_1 = 78$ ,  $L_2 = 178$ .

### **8.4.2. Gradient Filtering**

The Sobel operator is a form of spatial filtering that provides an approximation to the magnitude of the image intensity gradient at each pixel [124]. The Sobel operator uses the gray level values in a 3 x 3 neighborhood of the pixel under consideration, which is labeled as p5 in Figure 8.2.

p1	p2	p3
p4	p5	p6
p7	p8	p9

**Figure 8.2: 3 x 3 Neighborhood Used by Sobel Operator.**

The estimated gradient magnitude,  $S(p_5)$ , is given by:

$$S(p_5) = |p_7 + p_8 + p_9 - p_1 - p_2 - p_3| + |p_3 + p_6 + p_9 - p_1 - p_4 - p_7|. \quad (8.2)$$

The frame shown in Figure 6.5 is shown again in Figure 8.3, after pre-processing by histogram stretching followed by application of the Sobel operator.

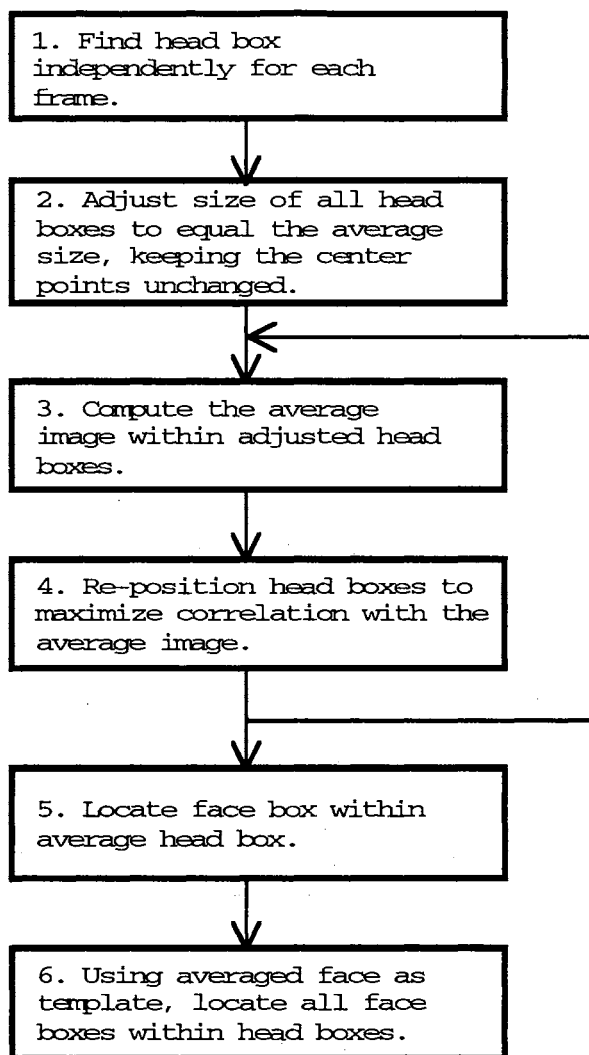


**Figure 8.3: A Pre-Processed Frame from MA-1.**

## 8.5 Automatic Location of Faces

To limit the computation involved in determining precise face location in each frame, a preliminary step was taken of locating the subject's whole head. The rectangle containing the head is referred to as the "head box". Following this step, the face box can be located by exhaustive search within the sub-image delimited by the head box.

Processing each frame in this manner, independently of the other frames, requires a large amount of computation and is error prone. Both problems are addressed by recognizing that frames within a session are in fact highly correlated. To take advantage of this, the algorithm shown in Figure 8.4 was developed.



**Figure 8.4: Estimation of Face Box Position.**

In Step 1, simple heuristic rules are employed to determine the approximate head box position independently in each frame. The top of the head box is first determined as follows: (1) the sum of pixel (grayscale) values along each row is computed; (2) from a histogram of these sums, the 10 percentile value is determined; (3) scanning down from the top of the image, the first row for which the sum of pixels exceeds the 10 percentile is deemed to be the top of the head box. Similar rules are applied to determine the left and

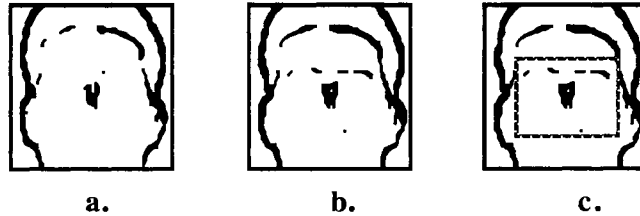
right sides of the head box. The bottom is chosen to maintain a constant ratio of height to width.

The heuristic rules above work well on average, but can behave inconsistently when given "fuzzy" edges, such as hair. In Step 2, it is assumed that each head box is centered correctly, but is subject to error in its estimated size. Step 2 sets the height and width of each head box equal to the average height and width of all head boxes, while maintaining their original center points.

Variations in sitting position or zoom lens adjustment may cause the image scale factor to change from one session to another. Within a session, however, the scale factor can be assumed to be constant. Recognizing this, and noting that as a result of Step 2, all head boxes are equal in size, Step 3 computes the average of all images within the head boxes. At each pixel position, averaging is performed across all frames in the image sequence.

The average image resulting from Step 3 may appear blurred or "out of focus" due to violations of the assumption made in Step 2 that all head boxes are centered correctly. In Step 4, the center position of each head box is adjusted to maximize the correlation of the image within that head box with the average head box image. Translation is limited to several pixels in any direction.

Steps 3 and 4 may be re-iterated until there are no further changes in the head box positions. In practice, convergence was found to occur after only one or two iterations. Figure 8.5 shows an example of an average head box image after one, two, and three iterations. Note that image sharpness and focus improve from the first to the second iteration, but remain about the same on the third. After three iterations, some residual blurriness is caused by facial movements, particularly of the mouth and eyes, within the session.



**Figure 8.5: Average Head Box Image After 1, 2, and 3 Iterations.**

After locating the head boxes, the next step is to create a "face template", with which to search for the face box within each head box. As indicated in Step 5 of Figure 8.4, the average head box image produced in Step 3 (after 3 iterations) is searched to find the face box, and the image within that face box is used as the face template. An example face box is shown in Figure 8.5c.

Step 5 is accomplished using the "eigenfaces" technique proposed by Turk and Pentland[6]. The use of eigenfaces, as opposed to simple correlation, is appropriate for this step because it is desired to locate the face consistently without prior knowledge of the identity of the subject. A virtue of eigenfaces is that it is a *subject-independent* face model.

To apply the eigenfaces method, a set of training images is analyzed to determine the mean image,  $\Psi$ , and the principal components,  $\mathbf{u}_i, i \leq 0 < L$ , of the covariance of the training images about  $\Psi$ . Any arbitrary facial image,  $\Gamma$ , not belonging to the training set can then be approximated in terms of  $\Psi$  and a relatively small number (e.g., 10) of principal components. To do this, the deviation of  $\Gamma$  from  $\Psi$  is first computed as  $\phi = \Gamma - \Psi$ . The projection,  $\phi_f$ , of  $\phi$  on the subspace spanned by the  $\mathbf{u}_i$  is given by  $\phi_f = \sum_{i=1}^L \omega_i \mathbf{u}_i$ , where  $\omega_i = \phi_f^T \mathbf{u}_i$ . The

synthesized approximation is:  $\Gamma_f = \Psi + \phi_f$ . Location of a face within a larger image can be accomplished by sliding a window over the original image, and selecting the position of the window to minimize the error function:  $\epsilon^2(x,y) = \int$



$\phi - \phi_f$ . In the following,  $\Gamma(x,y)$  denotes a subimage of  $\Gamma$  of dimensions equal to those of  $\Psi$  and  $\mathbf{u}_i$ , with upper-left corner at  $(x, y)$ . The derivation of  $\varepsilon(x,y)$  in terms of  $\Gamma(x,y)$ ,  $\Psi$ , and  $\mathbf{u}_i$  is repeated here because several errors were contained in [6]. Dependence on  $(x, y)$  is suppressed.

$$\varepsilon^2 = \|\phi - \phi_f\|^2 \quad (8.3)$$

$$= (\phi - \phi_f)^T (\phi - \phi_f) \quad (8.4)$$

$$= \phi^T \phi - \phi^T \phi_f - \phi_f^T (\phi - \phi_f) \quad (8.5)$$

$$= \phi^T \phi - \phi^T \phi_f \quad (8.6)$$

$$= \phi^T \phi - \left( \sum \omega_i \mathbf{u}_i^T \right) \left( \sum \omega_i \mathbf{u}_i \right) \quad (8.7)$$

$$= \phi^T \phi - \sum \omega_i^2 \quad (8.8)$$

Expanding the first term,

$$\phi^T \phi = (\Gamma - \Psi)^T (\Gamma - \Psi) \quad (8.9)$$

$$= \Gamma^T \Gamma - 2\Psi^T \Gamma + \Psi^T \Psi \quad (8.10)$$

Expanding the second term,

$$\sum \omega_i^2 = \sum (\phi^T \mathbf{u}_i)^2 \quad (8.11)$$

$$= \sum ((\Gamma - \Psi)^T \mathbf{u}_i)^2 \quad (8.12)$$

$$= \sum (\Gamma \mathbf{u}_i - \Psi^T \mathbf{u}_i)^2 \quad (8.13)$$

Combining the two terms, and making explicit the dependence on spacial position:

$$\varepsilon^2(x,y) = \Gamma^T(x,y)\Gamma(x,y) - 2\Psi^T \Gamma(x,y) + \Psi^T \Psi + \sum_{i=1}^L [\Gamma(x,y)\mathbf{u}_i - \Psi^T \mathbf{u}_i]^2 \quad (8.14)$$

Eigenfaces were computed from a subset of the OSU data, using the hand marked face boxes, as described in Section 4.1. The data subset consists of every twentieth frame of the first session of each of the first five subjects

(MA, BB, BC, KD, and CF). The data within each hand-marked face box was scaled to a size of 32 x 27 pixels (the average face box dimensions), and ten eigenfaces were computed. These are shown in Figure 8.6. It was found that the most significant three are adequate to locate faces reliably.



**Figure 8.6: The 12 Most Significant Eigenfaces of the OSU Data.**

The scale factor for the session, assumed in Step 1 to be constant, is unknown. To accommodate the unknown scale factor, multiple eigenface searches are performed on the average head box image, after re-scaling it using scale factors of 0.8, 0.9, 1.0, 1.1, and 1.2. Each search determines the face box position that minimizes the mean-squared error between the scaled image and its eigenface approximation. The scale factor and face box are selected that result in the global minimum error. Figure 8.5c shows an example of a face box determined in this manner.

The image in the face box determined in Step 5 is an average of the face portions of all frames in the session. In Step 6, each face box is located by using this image as a template for correlation-based matching. A procedure similar to that of Baron [86] was used. Let  $T$  be the template and  $I(x,y)$  be a rectangular region of the image with upper left corner at location  $(x,y)$ . Then define a correlation coefficient,  $C(x,y)$  as

$$C(x, y) = \frac{\langle T * I(x, y) \rangle}{\sqrt{\langle T * T \rangle \langle I(x, y) * I(x, y) \rangle}} \quad (8.15)$$

where \* represents the pixel-by-pixel product, and < > is the average operator. C(x,y) is computed for all values of x and y for which the region lies entirely within the head box. The putative location of the upper left corner of the face box is the value of (x,y) for which C(x, y) is maximum. The normalization term  $\langle I(x,y) * I(x,y) \rangle$  in the denominator was found to be important. Without this term, the maximum correlation location is biased toward dark regions of the image.

## 8.6 Face Comparison

Having located the central face region, or "face box" within each image frame, it now becomes possible to compare the sequences of facial images in two different sessions. This is accomplished using the log likelihood ratio, as computed using Equation 5.37. In applying Equation 5.37, the vectors  $x_j$  and  $y_i$  contain pixel values within the 27 x 32 face boxes. The dimensionality of the feature space is therefore 864.

## 8.7 Summary

Processing steps were described by which raw video input data is converted to a feature vector per frame. Histogram equalization and gradient filtering are included to minimize sensitivity to lighting gradients and skin reflectivity. The central portion of the subject's face is located within each video frame, allowing for unknown scale factor and head tilting, in addition to lateral movement. The face is scaled to a 27 x 32 pixel rectangular area, and the processed gray-level values with that area form the elements of feature

vector. Video feature vectors are 864 dimensional and are produced at the rate of 10 per second.

## CHAPTER 9

### ANALYSIS AND RESULTS

#### 9.1 Introduction

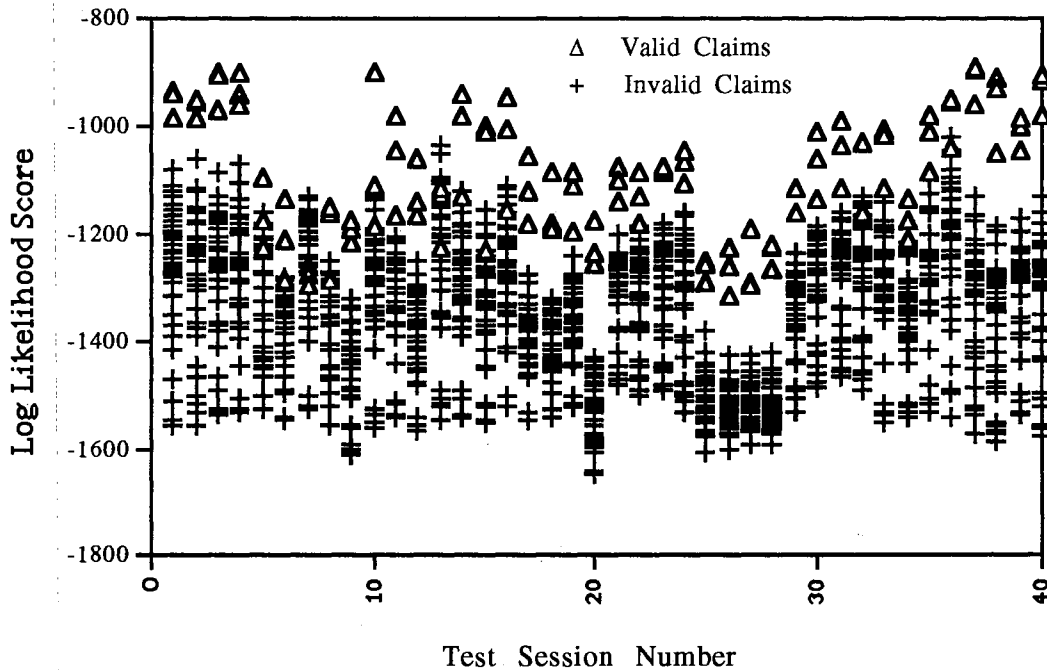
This chapter presents analysis and results derived from PIV algorithms operating on the OSU database. The PIV algorithms considered combine the voice and video feature extraction components described in Chapters 7 and 8 with the decision component described in Chapter 5.

#### 9.2 Likelihood Scoring Versus Likelihood Ratio Scoring

It is commonly believed that identity verification should only require a model of the claimant, and should not make reference to models of other individuals. The assumption underlying this theory is that the likelihood function  $p(X | C)$  is tightly localized, whereas  $p(X | \bar{C})$  is approximately a multi-dimensional uniform PDF. Therefore,  $p(X | C)$  differs from the likelihood ratio,  $p(X | C) / p(X | \bar{C})$ , only by a multiplicative constant.

With this theory in mind, consider the data shown in Figure 9.1. Each of the 40 columns was produced by computing the log likelihood of one session of the database (treated as test data) with respect to each of the other 39 sessions (treated as enrollment or model data). There are therefore  $40 * 39 = 1560$  points plotted in Figure 9.1. Each indicated log likelihood score was computed for the video data using Equation 5.22. Note that the log likelihood is not symmetric, (i.e.,  $\ln p(X | Y) \neq \ln p(Y | X)$ ). The triangles represent valid claims, in which the identity of the test subject and model subject are the same. The crosses

represent invalid claims, in which the test data is compared with model data from a different subject.



**Figure 9.1: Log Likelihood Scores for Each Test Session.**

The distribution of scores for invalid claims is clearly seen to vary from one test session to another. This is not consistent with the proposition that  $p(X | \bar{C})$  is a uniform PDF.

From the data in Figure 9.1, it can be seen that a threshold value of -1000 separates the valid and invalid cases for Subject MA (Sessions 1-4) without errors. A threshold value of about -1300 separates the valid and invalid cases for Subject YL (Sessions 25-28) without errors. However, a threshold of -1300 would allow MA to be falsely accepted in many cases, and a threshold of -1000 would reject all valid claims of YL. This illustrates the difficulty of scoring based on the likelihood function, and the advantage of the likelihood ratio.

The LLR score, computed from Equation 5.37, has the following interpretation. The LLR score for a session is equal to the best log likelihood

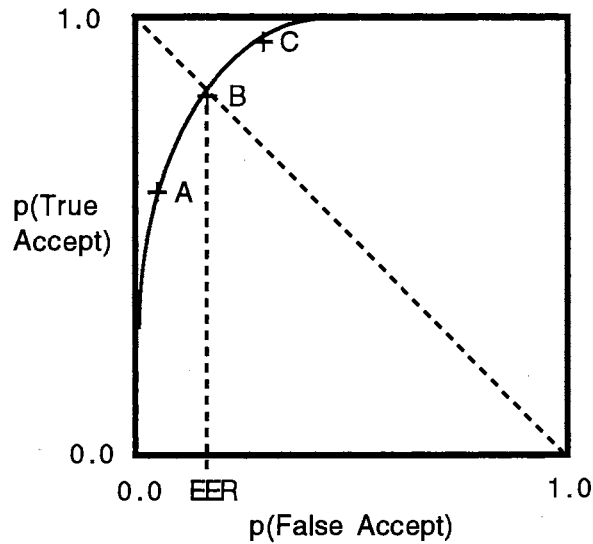
score among the claimant's models minus the best score among all other speaker's models. For a threshold of zero, this rule can be interpreted graphically in terms of Figure 9.1 as follows. A session is accepted if a triangle is the highest-scoring model, or rejected otherwise. Only two sessions (numbers 7 and 13) are rejected at a threshold of zero.

### 9.3 Intrinsic Dimensionality

The intrinsic dimensionality of the voice and video data was estimated using the method of Pettis, et. al. Estimates were made for all sessions and then averaged. A regression order of  $K = 3$  was used. For the voice data, the average intrinsic dimensionality equals 3.1. This may indicate that speech articulation of an individual exhibits three degrees of freedom, presumably associated with the tongue or other parts of the vocal tract. For the video data, the average intrinsic dimensionality equals 4.5. Similarly, this may indicate that facial expressions are dominated by the activity of 4 or 5 muscle groups.

### 9.4 ROC Performance Measurement

Accuracy of PIV systems is measured by probabilities of Type-I, or *false rejection* errors, and Type-II, or *false-acceptance* errors. Total error probability, or the sum of false-rejection and false-acceptance error probabilities, is minimized by PIV systems employing a likelihood ratio test. Since the optimum value of the threshold is not known (because an estimate of the prior probability of a valid claim is generally unavailable), it is common to measure false-rejection and false-acceptance error probabilities over a range of threshold values. A *receiver operating characteristic* (ROC) curve, as shown in Figure 9.2, provides a convenient method of displaying this information.



**Figure 9.2: Example ROC Curve**

The ROC curve plots false-acceptance probability on the horizontal axis versus true-acceptance probability (equal to one minus false-rejection probability) on the vertical axis as a function of threshold value. The endpoints are at (0, 0) and (1, 1), and the curve increases monotonically between these endpoints. The ROC curve for a perfect PIV system passes through the point (0, 1), indicating that a value of threshold exists for which all valid claims are accepted and all invalid claims are rejected. The performance of an imperfect PIV system is illustrated in Figure 9.2. The points labeled "A", "B" and "C" are three possible *operating points* corresponding to different (decreasing) threshold values. False-rejection and false-acceptance probabilities are equal at operating point B. The false-rejection or false-acceptance probability corresponding to operating point B is called the *equal-error rate* (EER). Operating points A and C represent different possible tradeoffs between the two types of errors. At operating point C, both legitimate users and imposters are more likely to be accepted than at operating points A or B.



Consider an experiment involving  $N$  subjects with  $M$  trials each. Suppose the data associated with each trial is recorded so that it can be presented repeatedly to the PIV system with different claimed identities. False rejection rates can then be estimated by presenting each of the  $N*M$  trials with the correct claimed identity and determining the fraction of rejections. False-acceptance rates can be estimated by presenting each trial with each incorrect claimed identity and determining the fraction of acceptances. A total of  $N*M*(N-1)$  simulated imposter trials is obtained in this manner. This method of estimating false-acceptance rates, termed "casual imposters" by Doddington [29], is based on the premise that imposters' behavior is independent of the identity they are claiming.

Given the LLR scores resulting from all false-rejection and false-acceptance trials, ROCs are created as follows. The LLR scores are sorted in descending order together with labels identifying whether each score resulted from a valid or invalid identity claim. Each score value is then treated in turn as a verification threshold. The reported probability of correct acceptance is the fraction of all valid trials with scores exceeding the threshold. The reported probability of false acceptance is the fraction of all invalid trials with scores exceeding the threshold. An example is shown in Table 9.1.

LLR Score	Valid Claim?	Prob (False Accept)	Prob (Correct Accept)
243	1	0.000	0.000
236	1	0.000	0.025
224	1	0.000	0.050
...	...	...	...
58	1	0.000	0.875
54	1	0.000	0.900
9	1	0.000	0.925
-1	0	0.000	0.950
-7	0	0.003	0.950
-8	0	0.006	0.950
...	...	...	...
-91	0	0.442	0.950
-92	1	0.450	0.950
-93	0	0.453	0.975
...	...	...	...
-144	0	0.669	0.975
-145	1	0.675	0.975
-146	0	0.678	1.000
-147	0	0.681	1.000
...	...	...	...

**Table 9.1: Example of ROC Computation.**

The ROC data shown in Table 9.1 is perfectly sorted, with the exception of two valid trials having scores of -92 and -145. At a threshold of -1, all valid trials except these two are accepted without accepting any invalid trials. Adjusting the threshold to accept these two trials would cause 67.5% of the invalid trials to also be accepted. In this case, the ROC is said to have a long "tail", indicating the presense of one or more trials that are in some way anomalous with respect to the PIV algorithm.

**9.4.1. Integrated Error Measure**

Although the commonly-used EER measure is an important benchmark of verification error, it is independent of the tails of the ROC curve. In the example above, the EER (5%) is an optimistic characterization of performance

because of the presense of the long tail of the ROC curve. A measure of error that accounts for the tails is the integrated error measure (IEM), defined as:

$$\text{IEM} = \int_{-\infty}^{\infty} \text{PFA}(t) dt \quad (9.1)$$

where  $\text{pFA}(t)$  is the false-acceptance probability at threshold  $t$ . Graphically, IEM equals the area *above* the ROC curve as plotted in Figure 9.1.

### 9.5 Test Procedure

The likelihood ratio scoring technique poses a unique problem in testing PIV systems. In practice, the cohort would be composed of a finite set of enrolled individuals, whereas the set of potential imposters is unlimited. One might expect there to be differences in the false acceptance rate between imposters who are included in the cohort and those who are not. This was observed experimentally to be the case. Inclusion of an individual in the cohort improves the ability of the system to reject that individual when another identity is claimed. Therefore, to assure unbiased measurement of the false acceptance rate, the cohort and the set of tested imposters should be mutually exclusive. At the same time, the small size of the OSU database demands that efficient use be made of all available data.

With these considerations in mind, the following procedure was used in false-acceptance testing. For each subject presented as input to the system, the identities of the other nine subjects were treated in turn as the claim. In each trial, the cohort was composed of the eight subjects whose identities matched neither the true nor the claimed identity. This procedure is illustrated in Figure 9.3.

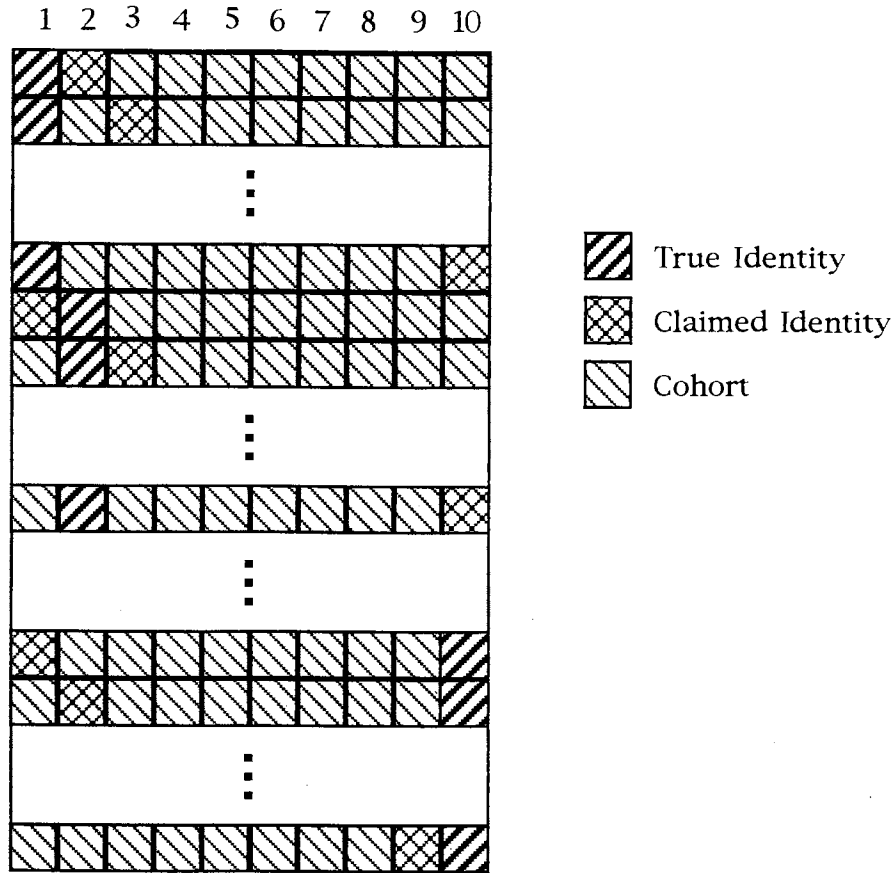


Figure 9.3: Test Procedure for False Acceptance Measurement.

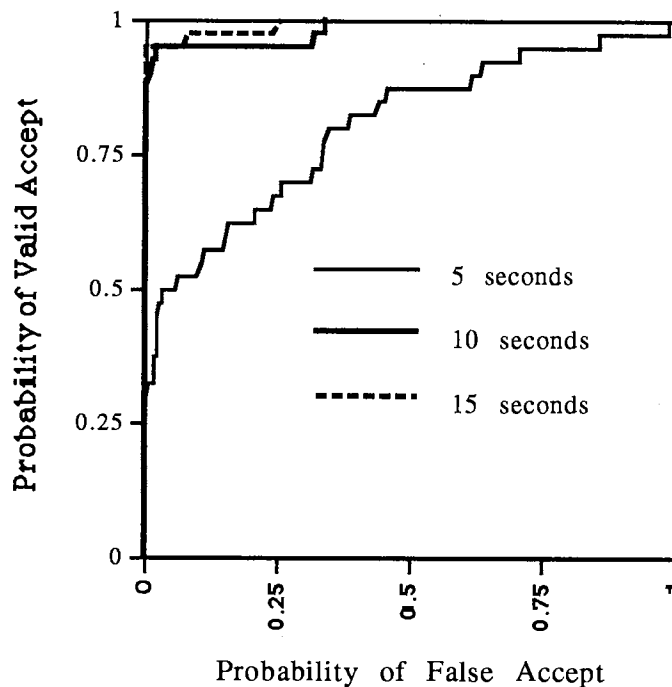
This procedure generalizes to measurement of false rejection, where the true identity and the claimed identity are one in the same. In false rejection testing, therefore, the cohort was composed of the nine subjects whose identities differed from that of the subject under test.

The test procedure simulated the use of three enrollment sessions per subject. In false-rejection testing, each of the subject's four sessions were treated in turn as the input to the system (test session), while the remaining three sessions were treated as enrollment sessions. There were therefore four false-rejection trials per subject, or a total of 40 false-rejection trials. In false-acceptance testing, the claimant's first three sessions were treated as

enrollment sessions. For each of the 40 sessions, all nine false identity claims were tested. There were therefore  $40 \times 9 = 360$  false-acceptance trials.

### 9.6 Voice-Only ROC Data

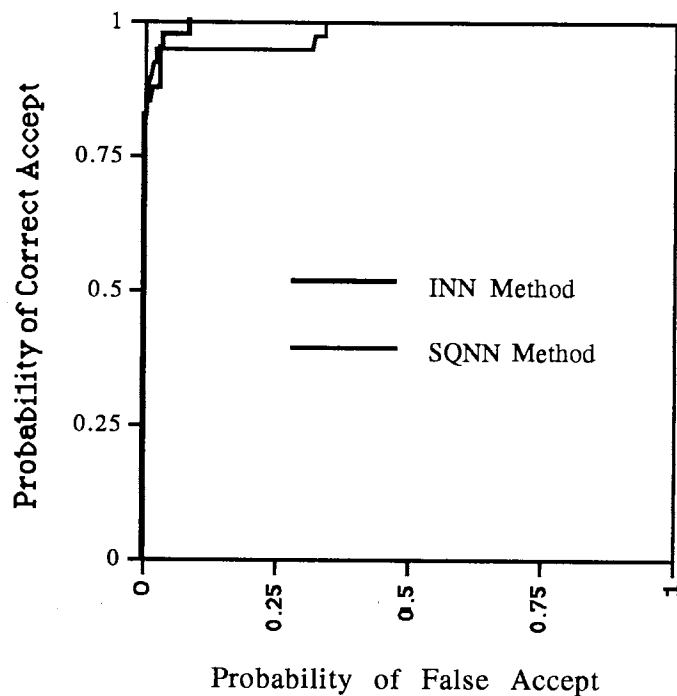
The ROC performance data for the voice-only PIV algorithm is shown in Figure 9.4. Separate ROC curves are shown for cases in which the first 5, 10, and 15 seconds of the test session are used in each trial. A large gap in performance is seen between the 5-second and 10-second cases. Increasing the test session length from 10 seconds to 15 seconds leads to a smaller performance improvement. This suggests that at about 10 seconds of speech provides an adequate sampling of the subject's voice and that diminishing new information is supplied by further observation.



**Figure 9.4: ROC Performance of SQNN Voice-Only PIV Algorithm.**

The INN method was also tested, and its ROC performance is compared with that of the SQNN method in Figure 9.5. The first ten seconds from each

session were used. The value of  $K$  used in the linear regression of Equation 5.32 was set to  $K = 20$  as a result of experimental optimization. The INN method is more accurate than the SQNN method in this test. Its IEM is 0.0056, compared with 0.0172 for the SQNN method.

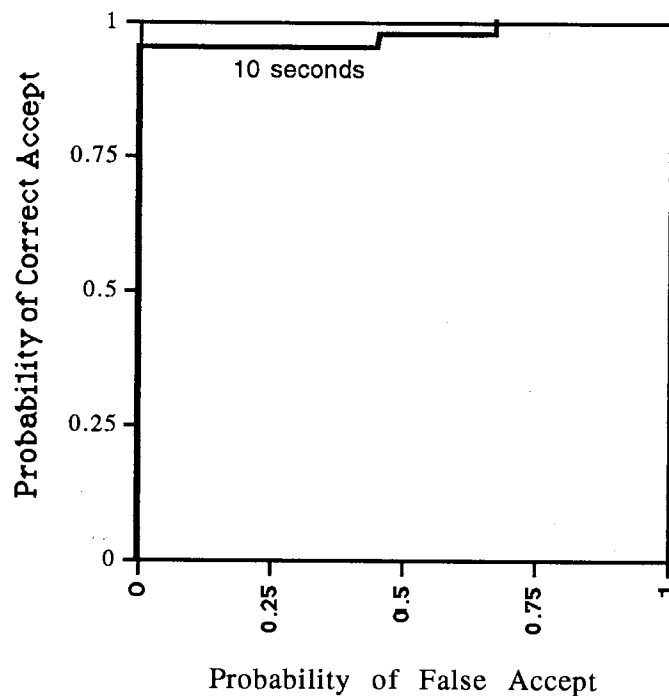


**Figure 9.5: Comparison of INN Versus SQNN Voice-Only PIV Algorithms.**

The long "tail" of the ROC for the SQNN method indicates that a small number of valid trials score very poorly (more poorly than about one third of the invalid trials). This is not the case for the INN method. The explanation may be related to the fact that the INN method makes use of distances to the 20 nearest neighbors of each test frame, whereas the SQNN method uses only the single nearest neighbor.

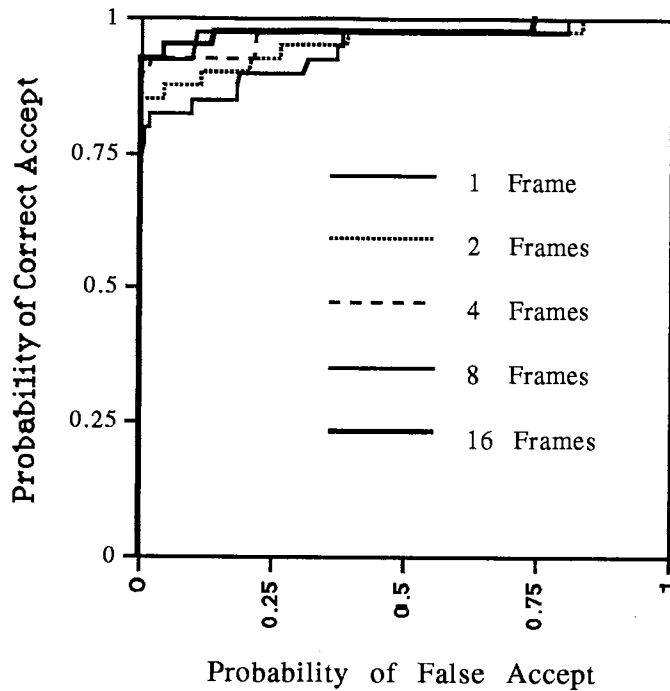
## 9.7 Face-Only ROC Data

The ROC performance data for the face-only PIV algorithm using the SQNN method is shown in Figure 9.6. Only the first ten seconds of data were used from each session. ROC performance was also measured using the first 5 seconds and the first 15 seconds of each session, but the result did not vary appreciably from that shown in Figure 9.6.



**Figure 9.6: ROC Performance of Face-Only PIV Algorithm.**

The finding that performance is relatively insensitive to the length of the session was unexpected. A likely explanation is that the data frames are highly correlated, so that performance rapidly saturates as a function of session length. Confirmation of this explanation is provided by Figure 9.7.



**Figure 9.7: ROC Performance of Face-Only PIV Algorithm.**

Figure 9.7 was obtained by comparing one frame of data from each test session with various numbers of frames from the enrollment sessions. The selected frame of test data was offset one half second from the beginning of the test session. The selected frames of enrollment data were separated by intervals of one half second, starting at an offset of one half second from the beginning of the enrollment session.

The curve labelled "1 Frame" in Figure 9.7 represents conventional comparison of faces by means of still images: A single test image is compared with a single enrollment image. Increasing the number of enrollment frames provides steady improvement in performance, up to about 16 frames. The 16-frame case corresponds to an elapsed time interval of 8 seconds. Separation of the frames by 0.5 seconds does not eliminate inter-frame correlation, but probably has some de-correlating effect. The ROC for 32 frames does not differ



appreciably from that for 16 frames, and is therefore not plotted. Use of additional test frames also does not improve accuracy. The IEM for the "16 Frame" case in Figure 9.7 equals 0.0230, compared with 0.0281 for Figure 9.6, in which 100 frames were used from each test and enrolment session. The difference is probably insignificant.

Performance of the INN method on the video data was found to be very poor. The estimated intrinsic dimensionality derived from Equation 5.32 is unreasonably high for many frames, often on the order of 100. This may be related to the very high degree of correlation among the frames. A possible explanation is that any systematic difference between sessions leads to increased distances to all nearest neighbor frames, thus increasing the apparent dimensionality. Another possible explanation of the failure of the INN method is that the enrollment data is simply too sparse to support the large-sample assumptions that underly Equations 5.15 and 5.32.

## 9.8 Fusion of Voice and Face Data

It is reasonable to assume the voice data and facial image sequence data are statistically independent in view of the fact that the face data excludes the mouth region. Assuming independence, data fusion can be accomplished without loss of accuracy by adding the log likelihood ratios for the two sources (see Section 3.4).

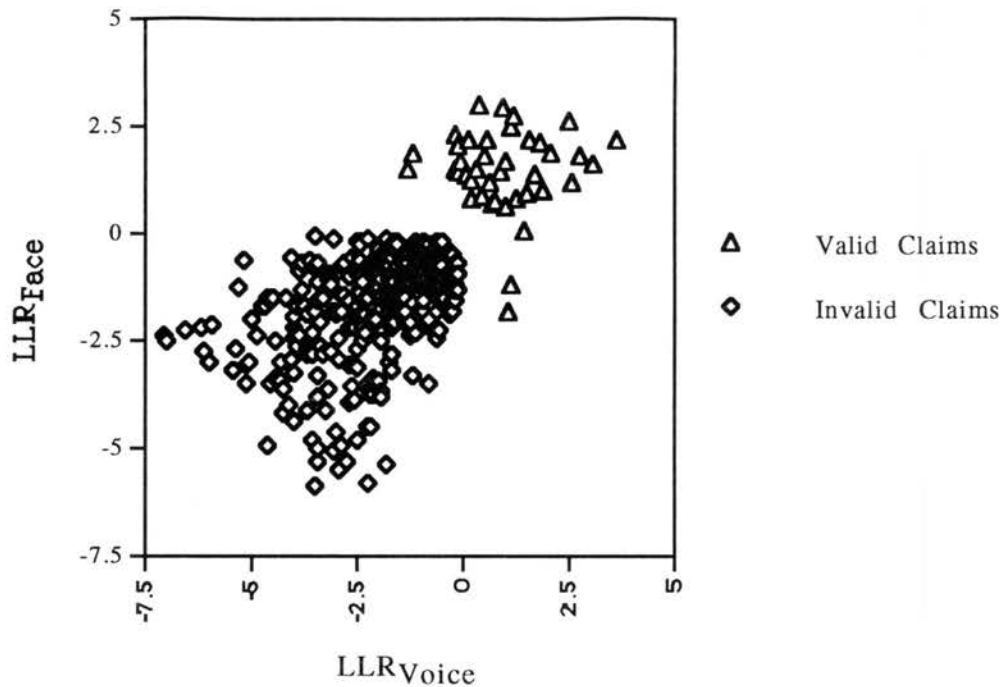
If estimated LLRs for the two sources are to be added, care must be taken to normalize the data so as to remove the effects of arbitrary scaling of the input measurements. In the case of the SQNN method,  $\alpha$  and  $\beta$  must also be estimated for the voice and image sources. This was accomplished as follows. Fixed values were assumed for the number of enrollment samples and the intrinsic dimensionality. For the voice data, the values used were:  $N = 250$  and

$d = 3$ . Values of  $\alpha$  and  $\beta$  were computed for the unit-variance Gaussian PDF with  $N = 250$  and  $d = 3$  using the method described in Section 5.5. The average nearest-neighbor distance was computed for the voice data and for the unit-variance Gaussian data. The voice data was then scaled so that its average nearest neighbor distance was equal to that of the Gaussian data. The same approach was applied to the video data, using  $N = 100$  and  $d = 4$ . The values of  $\alpha$  and  $\beta$  used in Equation 5.37 are shown in Table 9.2.

	Voice	Video
Alpha	-0.33	-0.13
Beta	-1.72	-3.18

**Table 9.2: Values of  $\alpha$  and  $\beta$  for Voice and Video Data.**

Let the LLR scores produced by the voice-only and face-only algorithms be denoted  $LLR_{\text{voice}}$  and  $LLR_{\text{face}}$ , respectively. A scatter plot of  $LLR_{\text{voice}}$  versus  $LLR_{\text{face}}$  is shown in Figure 9.8. Valid claims are represented by triangles, and invalid claims are represented by diamonds. Valid and invalid claims are clearly separable. Therefore, 100% verification accuracy can be obtained using a decision rule based on a combination of  $LLR_{\text{face}}$  and  $LLR_{\text{voice}}$ , whereas less than perfect verification is obtained using either  $LLR_{\text{face}}$  or  $LLR_{\text{voice}}$  alone.



**Figure 9.8: Scatter Plot of  $LLR_{voice}$  Versus  $LLR_{face}$ .**

If the enrollment data are considered to be representative of all subjects, and if the voice and face measurements are assumed to be mutually independent, then the LLR of the joint voice and face measurements equals the sum of the separate LLRs. Symbolically,  $LLR_{joint} = LLR_{voice} + LLR_{face}$ . In Figure 6.6, contours of constant  $LLR_{joint}$  are lines with slope equal to -1. The acceptance region associated with a decision rule based on  $LLR_{joint}$  is therefore the region above and to the right of a line with slope -1. This decision rule does not achieve perfect separation of valid from invalid sessions for any value of the threshold.

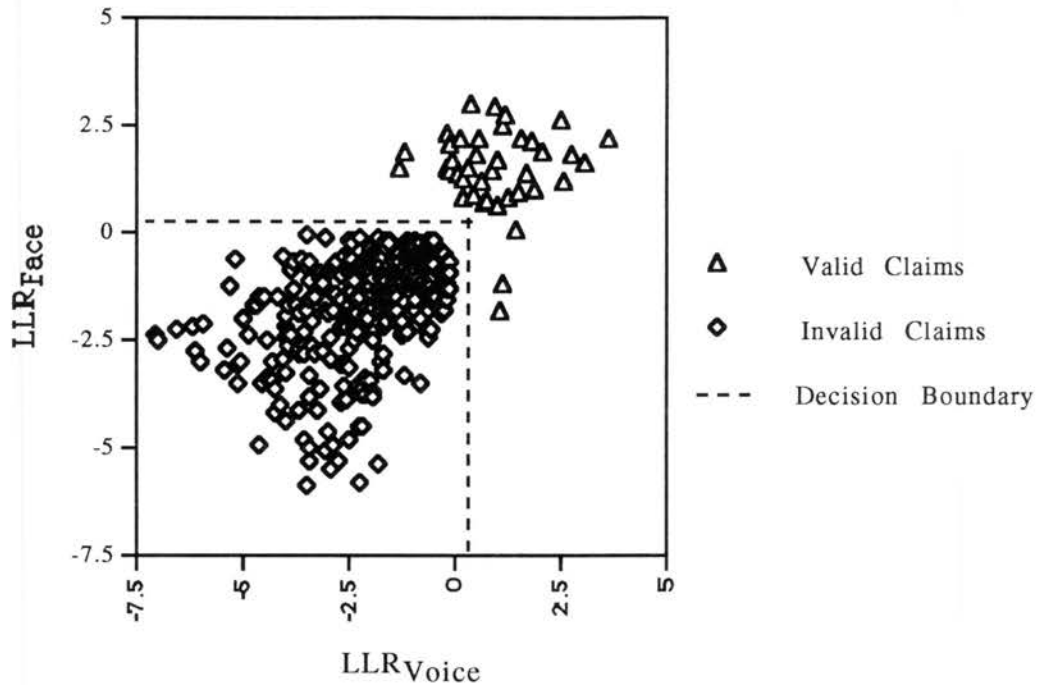
Of the valid claims, the two "outlier" sessions with respect to  $LLR_{face}$  are BB-3 and KD-1. BB-3 was the only session in which Subject BB wore glasses. Similarly, KD-1 was the only session in which Subject KD wore a hat. In both cases the test subject's appearance was affected by factors not represented in that individual's enrollment data. The two outlier sessions with respect to

$LLR_{voice}$  are YL-1 and RM-1. Listening to these sessions revealed that YL-1 has a reverberant quality not present in the other sessions. This is likely to have been caused by placement of the microphone too far from the speaker's mouth. Nothing unusual was perceived in listening to session RM-1.

The fact that three of the four outliers can be explained by failure of the enrollment data to represent conditions encountered in the test data suggests that more weight should sometimes be placed on one type of evidence than the other. Although sessions BB-3 and KD-1 are outliers with respect to  $LLR_{face}$ , they are in the normal range with respect to  $LLR_{voice}$ . Similarly, YL-1 and RM-1 are outliers with respect to  $LLR_{voice}$ , but are in the normal range with respect to  $LLR_{face}$ . In addition, the invalid claims are fairly tightly clustered in Figure 9.8, with no outliers. Based on these observations, it appears reasonable to accept the identity claim if either  $LLR_{face}$  or  $LLR_{voice}$  exceeds a threshold value. An equivalent decision rule is to accept the claim if  $LLR_{max}$  exceeds a threshold, where

$$LLR_{max} = \max(LLR_{voice}, LLR_{face}). \quad (6.1)$$

The acceptance region associated with  $LLR_{max}$  is illustrated in Figure 9.9. Using the decision rule  $LLR_{max} > 20$ , 100% verification accuracy is achieved.



**Figure 9.9: Scatter Plot Showing Decision Boundary  $LLR_{max}=20$ .**

### 9.9 Summary

In this chapter, data was shown comparing likelihood scoring with likelihood ratio scoring. The difficulty of applying likelihood scoring using a fixed decision threshold was demonstrated, as was the relative superiority of likelihood ratio scoring, based on Equation 5.37. The data presented was inconsistent with the premise that the denominator of the likelihood ratio is constant.

The intrinsic dimensionality of the voice and video data in the OSU database was estimated to be 3.1 in the case of the voice data, and 4.5 in the case of the video data. Possible physical explanations were conjectured.

Performance of the voice-only and video-only PIV algorithms were measured. A summary of the results is shown in Table 9.3. Using 10 seconds from each session and the SQNN method, the equal-error rate of both

algorithms is 5%. The INN method gives somewhat better performance for the voice data, but poor performance for the video data. Possible explanations were conjectured.

For the video-only algorithm, a single frame of test data and 16 frames of enrollment data give the same performance as 10 seconds of test data and 10 seconds of enrollment data. This is believed to be the result of a high degree of inter-frame correlation. Consistent improvement is observed, however, as the number of enrollment frames increases from 1 to 16. This demonstrates the value of image sequences, as opposed to still images, in the enrollment process.

Data Type	PDF Est. Method	Enrollment Length	Test Length	EER (%)	IEM
Voice	INN	10 seconds	10 seconds	3.3	.0056
Voice	SQNN	10 seconds	10 seconds	5.0	.0172
Face	SQNN	10 seconds	10 seconds	5.0	.0281
Face	SQNN	1 frame	1 frame	15	.0584
Face	SQNN	16 frames	1 frame	5.0	.0230
Voice+Face	SQNN	10 seconds	10 seconds	0	0

**Table 9.3: Summary of Key Results.**

Finally, a PIV algorithm using both voice and video data was tested. Data fusion is employed at the LLR level. The identity claim is accepted if either LLR exceeds the threshold value. A rationale for this strategy was given.

Using the combined algorithm, no errors occur in verification testing using the OSU database.

## CHAPTER 10

### CONCLUSION

A substantial body of literature exists in various fields relevant to multimedia PIV. Both voice and face PIV algorithms have been under development for nearly 20 years, resulting in numerous and diverse approaches. Face verification algorithms, until now, have used only still images, as opposed to image sequences. Very recently, a PIV algorithm was reported that combines voice and still facial images, achieving better accuracy with the combination than with either voice or face information alone. The current work demonstrates that further improvement can be obtained by using facial image sequences.

In this research, the sound and image data originating from a movie clip are sampled and processed to form feature vectors within periodically occurring frames over the length of the movie. The availability of sound and image data streams spanning a common interval has enabled a unified approach to be taken in processing the two information sources.

Voice and facial image data are reasonably regarded as random, as opposed to deterministic, observations of underlying characteristics the individual. Bayes' decision rule provides an optimal criterion for accepting or rejecting a user's claimed identity based on the available observations. Application of Bayes' decision rule leads to a likelihood ratio test, and reduces the problem to designing estimators for the likelihood functions  $p(X | C)$  and  $p(X | \bar{C})$ . These likelihood functions depend in turn on the local probability



density at each sample point. Estimation of local density is made difficult by the high dimensionality of the measurement spaces and by the limited availability of training data. Two approaches to this problem were investigated, and found to be useful under different conditions of practical interest.

An experimental database was collected for the purpose of developing and testing multi-media PIV algorithms. Using a camcorder, motion video and audio recordings were made of subjects reading from scripts on multiple occasions. Although this database is small (only ten subjects), it provides a reasonable demonstration of the concepts presented here.

Separate multi-media PIV algorithms for voice and facial image data were simulated and tested. Using 10-second samples of either voice data or facial image data alone, equal error rates of about 5% were observed. False-rejection errors in both cases were attributed to conditions existing in the test data that were not represented in the model data (presence or absence of eyeglasses, for example). No test data was observed to be simultaneously anomolous with respect to both voice and face. Data fusion was therefore accomplished by selecting as the final score the greater of the two log likelihood ratios based on the voice and face data. Verification accuracy of 100% was shown on the experimental database.

The approach may applicable to a wider class of hypothesis-testing or classification problems involving high-dimensional measurement sequences.

### **10.1 Summary of Accomplishments**

The development of practical multi-media PIV systems presents numerous challenges and opportunities for scientific and technological innovation. Some of these challenges and opportunities have been identified

and addressed here. The main accomplishments of this research are listed below, in decreasing order of importance.

1. **Use of Facial Image Sequences.** Facial image measurements used in previous PIV systems have been still images, as opposed to image sequences. Facial movements are a source of errors, rather than a source of information, to these systems. It was demonstrated here for the first time that the additional information provided by image sequences leads to higher verification accuracy.
2. **Unified Approach to Multi-Media Processing.** The previous work of Higgins, Bahler, and Porter [4] in the area of voice verification was further investigated, and applied to verification using facial image sequences. Error rates on the order of five percent were observed using either voice or image sequences separately. Combining the two sources, 100% accuracy was achieved on a small database. The success of this experiment demonstrates the feasibility of applying a unified approach to processing voice and facial image measurement data.
3. **Improved Likelihood Function Estimator.** The estimator of intrinsic dimensionality reported by Pettis, et. al [5] was extended to enable estimation of the likelihood of a sequence of multi-dimensional observations relative to a set of training data. This estimator was compared experimentally with the SQNN estimator of Higgins, Bahler, and Porter. Its performance is superior to that of SQNN for voice data, but inferior for image sequence data. An explanation was hypothesized.

4. **Multi-Media PIV Database.** Using a camcorder, motion video and audio recordings were made of ten subjects reading short phrases from scripts. Each subject participated in at least four sessions on different days. The data was digitized and is available in digitized form to other researchers.
  
5. **Literature Survey.** The scientific literature was surveyed in subjects relevant to PIV technology and applications. A conclusion of the survey is that development of multi-media PIV systems is a logical evolutionary step that is needed to satisfy an increasing demand for network security. Evidence was cited [1] that voice and facial appearance carry separate information about the identity of the subject. Surprisingly, no quantitative studies were found on PIV performance of humans using multi-media information.

## **10.2 Suggested Future Research**

Multi-Media Personal Identity Verification is a fascinating, multi-disciplinary subject that holds great opportunity for further advances. The goal of developing a convenient, inexpensive, and robust multi-media PIV system remains to be accomplished in the future. Reaching this goal may involve work in signal processing, computer science, probability and statistics, applied psychology, and other fields.

One of the most important and difficult challenges is to develop algorithms that maintain high accuracy in uncontrolled environments. Voice and facial images are convenient media because their measurement does not involve expensive instrumentation, precise behavior by the subject, or

physical contact with the subject. This convenience comes at the cost of concomitant measurement variability. Facial image measurements are subject to variability with respect to lighting, distance and orientation relative to the camera, the optical quality of the camera, background objects, and presence or absence of glasses, hats, beards, etc. Voice measurements are subject to variability with respect to microphone and subject positioning, microphone sound quality, background noise, room reverberation, and colds or other factors affecting the subject's voice. These sources of variability will need to be accommodated, either through improved modeling or through development of signal processing methods that are insensitive to them.

The multi-media PIV algorithm developed here was simulated, but not made to process input data in real time. It appears to be feasible to implement the algorithm in real time on current processors (comparable to the Intel Pentium) with the following modifications. First, more efficient methods should be employed for locating faces in the input images (with unknown position, rotation, and scale factor). This is the most computation intensive part of the algorithm. Burt's "coarse-to-fine" multi-resolution template matching strategy [92] could be used for this purpose. Alternately, the approach of Turk and Pentland [6] based on spacio-temporal filtering could be employed. The second modification is to use only a relatively small subset of the input image frames. The experimental results reported here suggest that 16 frames at one-half second intervals is sufficient to provide good performance. It may be desirable to sample a diversity of facial expressions (eyes open/closed, mouth open/closed, etc.). Achieving this diversity could involve a symbiotic coupling of the face-location and frame-selection processes.

To obtain a more realistic evaluation of performance, the PIV algorithm should be tested using a much larger population of subjects separate from those used in development of the algorithm. Other than the OSU database, suitable databases for development or testing of multi-media PIV algorithms do not currently exist. Creating such databases would be a major step toward enabling further PIV technology development. The environment in which the OSU data was collected was relatively benign in terms of conditions such as lighting and background noise. It would be useful to include some controlled variability of these and other "nuisance variables" in future databases.

Although the content of the speech material in the OSU database was controlled by providing prompts, the PIV algorithms do not take advantage of this knowledge through any type of linguistic modeling. It is known that knowledge of the spoken text leads to improved accuracy of speaker verification systems. There is some evidence [94, 98] that it may also be relevant to face verification by providing a means to predict mouth movements. A natural extension of the current work would be to apply text-dependent verification methods using both voice and facial image sequence measurements.

Related to this, another promising direction is development of "liveness" tests to verify that the spoken utterance matches the prompt, and that the observed mouth movement is consistent with the sound. Methods similar to the lipreading recognizer of Petajan [95] could be employed for this purpose. Liveness tests are needed to detect and reject counterfeiting attempts involving photographs and/or tape recordings. Liveness tests would logically be developed concurrently with text-dependent verification approaches as referred to in the previous paragraph.

In this research, frames of data are treated as statistically independent over the duration of a session. Each frame contributes equally to the log likelihood measures for the claimant and alternative hypotheses, which are accumulated over time. The log likelihood ratio, which is the difference between the accumulated log likelihoods, also increases in magnitude over time, reflecting an accrual of evidence. This characteristic of the log likelihood ratio scores is appropriate when the input data frames are independent because evidence is indeed being accrued at a constant rate. The assumption of independent frames is problematic, however, in the case of facial image sequences, which are obviously correlated from frame to frame. A rationale is needed for de-weighting of log likelihood ratio scores to account for inter-frame correlations. This is particularly important when combining information sources with diverse degrees of inter-frame correlation.

The important role of intrinsic dimensionality has been observed in this work. The PIV algorithms developed here assume that intrinsic dimensionality remains constant throughout the feature space. The validity of this assumption should be tested experimentally. Deviations from uniform intrinsic dimensionality would indicate the potential for performance improvement through algorithm modifications to estimate and accommodate local variations. This investigation could be conducted within either a text-dependent or text-independent framework.

In a more theoretical vein, it would be enlightening to investigate the relationship of local density to nearest-neighbor distance for a variety of PDFs with broader or narrower "tails" than the Gaussian PDF, as well as correlated Gaussians and Gaussian mixtures. The current work provides some rationale for the SQNN approximation, but does not prove that SQNN is optimal in any

sense. Development of an approximation to Equation A.17 that is valid in the limit of high dimensionality might be useful in this regard.

## BIBLIOGRAPHY

1. Brunelli, R. and D. Falavigna, *Personal identification using multiple cues*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1995. v. 17, n. 10: p. 955-966.
2. Zavala, A., *Determination of facial features used in identification*, in *Personal Appearance Identification*, A. Zavala and J. Paley, Editor^Editors. 1972, Thomas Books: Springfield, Illinois.
3. Higgins, A., L. Bahler, and J. Porter, *Speaker Verification Using Randomized Phrase Prompting*. Digital Signal Processing, 1991. v. n. .
4. Higgins, A., L. Bahler, and J. Porter. *Voice Identification Using Nearest-Neighbor Distance Measure*. in *Intl. Conf. on Acoustics, Speech and Signal Processing*. 1993. Minneapolis. p.
5. Pettis, K., et al., *An Intrinsic Dimensionality Estimator from Near-Neighbor Information*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1979. v. 1, n. 1: p. 25-37.
6. Turk, M. and A. Pentland, *Recognition in Face Space*, in *Automatic Recognition of Objects*. 1990, SPIE Milestone Series:
7. Longstaff, T. and E. Scuultz, *Beyond Preliminary Analysis of the WANK and OILZ Worms: A Case Study of Malicious Code*. Computers and Security, 1993. v. 12, n. 1: p. 61-77.
8. *The Criminal Use of False Identification*. 1976.
9. Mandell, L., *The Credit Card Industry: A History*. Twayne's Evolution of American Business Series, 1990, Boston: G. K. Hall and Co.
10. Bertillon, A., *Identification of Criminals*. Foundations of Criminal Justice, 1889, New York: AMS Press.
11. Boice, R., *Observational Skills*. Psychological Bulletin, 1983. v. 93, n. : p. 3-29.
12. Yarmey, A.D., *Verbal, visual, and voice identification of a rape suspect under different levels of illumination*. J. of Applied Psychology, 1986. v. 71, n. 3: p. 363-370.
13. Yarmey, A.D. and A.L. Yarmey, *Face and Voice Identification in Showups and Lineups*. Applied Cognitive Psychology, 1994. v. 8, n. : p. 453-464.
14. Shepard, J. and H. Ellis, *Face Recognition and Recall Using Computer-Interactive Methods with Eye Witnesses*, in *Processing Images of Faces*. 1992, Ablex: Norwood, NJ. p. 129-146.
15. Harmon, L., *The Recognition of Faces*. Scientific American, 1973. v. 229, n. : p. 71-82.



16. Ellis, H., *Face recall: A Psychological Perspective*. Human Learning, 1986. v. 5, n. : p. 189-196.
17. FBI, *The Identification Division of the FBI: A Brief Outline of the History, Services, and Operating Techniques of the World's Largest Repository of Fingerprints*. 1991. v. n. .
18. Blais, P., *Personal Communication* 1993,
19. Stock, R. *Automatic Fingerprint Reading*. in *Proc. 1972 Carnahan Conf. on Crime Countermeasures*. 1972. Lexington, KY. p.
20. Wegstein, J. *The M40 Fingerprint Matcher (NBS Tech. Note 878)*. 1975.
21. Asai, K., et al. *Fingerprint Identification System*. in *Second USA-Japan Computer Conf.* 1975. p. 30-35.
22. Wilson, T. and P. Woodard, *Automated Fingerprint Identification Systems: Technology and Policy Issues (Report NCJ-104342)*. 1987, Bureau of Justice Statistics.
23. *The FBI Fingerprint Identification Automation Program*. 1991.
24. Hoshino, Y., et al. *Automatic Reading and Matching for Single-Fingerprint Identification*. in *65th Intl. Assoc. for Identification Conf.* 1980. Ottawa, Canada. p.
25. Takeda, M., et al. *Finger Image Verification Method for Personal Verification*. in *Conf. on Computer Vision and Pattern Recognition*. 1990. p. 761-766.
26. Sherlock, B., D. Monro, and K. Millard, *Algorithm for Enhancing Fingerprint Images*. *Electronics Letters*, 1992. v. 28, n. 18: p. 1720-1721.
27. Schneider, J. and D. Wobshall. *Live Scan Fingerprint Imagery Using High Resolution C-Scan Ultrasonography*. in *Proc. 25th 1991 Intl. Carnahan Conf. on Security Technology*. 1991. Taipei, Taiwan. p. 88-95.
28. Lampton, C., *DNA Fingerprinting*. 1991, New York: Franklin Watts.
29. Doddington, G. *Speaker Verification Final Report (RADC-TR-76-179)*. 1974.
30. Merillat, P.D. *Secure Stand-Alone Positive Personnel Identity Verification System*. 1979.
31. Maxwell, R. *An identity verifier evaluation of performance (Report SAND-87-2279C)*. 1987.
32. GAO. *Electronic Funds Transfer: Use of Biometrics to Deter Fraud in the Nationwide EBT Program*. 1995.

33. Miller, B., *Biometric Identification*, in *IEEE Spectrum* 1994, p. 22-30.
34. Flom, L. and A. Safir. *Iris recognition system*, U.S. Patent number 4,641,349. 1987.
35. Daugman, J. *High Confidence Personal Identification by Rapid Video Analysis of Iris Texture*. in *IEEE Intl. Carnahan Conference on Security Technology*. 1992. p. 50-60.
36. Olano, C. *An Investigation of Spectral Match Statistics Using a Phonetically Marked Database*. in *Intl. Conf. on Acoustics, Speech and Signal Processing*. 1983. p.
37. Furui, S., *Cepstral Analysis Technique for Speaker Verification*. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 1981. v. 29, n. 2: p. 254-272.
38. Atal, B., *Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification*. *J. Acoustic Soc. of America*, 1989. v. 55, n. 6: p. 1304-1312.
39. Koenig, B., *Spectrographic Voice Identification: A Forensic Survey*. *J. Acoustical Soc. of America*, 1986. v. 79, n. : p. 2088-2090.
40. Bellman, R., *Dynamic Programming*. 1957, Princeton, NJ: Princeton Univ. Press.
41. Sakoe, H. and S. Chiba, *Dynamic programming algorithm for spoken word recognition*. *IEEE Trans. on ASSP*, 1978. v. 36, n. 1: p. 43-49.
42. Rosenberg, A., *Evaluation of an automatic speaker verification system over telephone lines*. *Bell System Tech. Journal*, 1976. v. 55, n. : p. 723-744.
43. Furui, S., *Comparison of speaker recognition methods using statistical features and dynamic features*. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 1981. v. 29, n. : p. 342-350.
44. Bridle, J. and M. Brown, *Connected Word Recognition Using Whole Word Templates*. *Proc. Inst. of Acoustics (U.K.)*, 1979. v. n. : p. 25-28.
45. Ney, H., *The Use of a One Stage Dynamic Programming Algorithm for Connected Word Recognition*. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 1984. v. 32, n. : p. 263-271.
46. Stockham, T., T. Cannon, and R. Ingebretsen, *Blind Deconvolution Through Digital Signal Processing*. *Proc. IEEE*, 1975. v. 63, n. 4: p. 678-692.
47. Bahl, L., *et al. Recognition of a continuously read natural corpus*. in *ICASSP*. 1979. Washington, D.C. p. 442-444.

48. Baker, J., *The DRAGON system - an overview*. IEEE Trans. on Acoustics, Speech, and Signal Processing, 1975. v. 23, n. : p. 24-29.
49. Rabiner, L., *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. Proceedings of the IEEE, 1989. v. 77, n. 2: p. 257-286.
50. Picone, J., *Continuous speech recognition using hidden Markov models*. IEEE ASSP Magazine, 1990. v. July, n. : p. 26-41.
51. Juang, B.H., *On the hidden Markov model and dynamic time warping for speech recognition - a unified view*. Bell System Tech Journal, 1984. v. 63, n. 7: p. 1213-1242.
52. Rabiner, L., J. Wilpon, and B.H. Juang, *A Model-Based Connected Digit Recognition System Using Either Hidden Markov Models or Templates*. Computer Speech and Language, 1986. v. 1, n. 2: p. 167-197.
53. Rosenberg, A., C.H. Lee, and S. Gokcen. *Connected word talker verification using whole word hidden Markov models*. in ICASSP 91. 1991. p. 381-384.
54. Matsui, T. and S. Furui. *Concatenated Phoneme Models for Text-Variable Speaker Recognition*. in Intl. Conf. on Acoustics, Speech and Signal Processing. 1993. Minneapolis, MN. p.
55. Webb, J. and E. Rissanen. *Speaker Identification Experiments Using HMMs*. in Intl. Conf. on Acoustics, Speech and Signal Processing. 1993. Minneapolis, MN. p.
56. Kao, Y.H., J. Baras, and P. Rajasekaran. *Robustness Study of Free-Text Speaker Identification and Verification*. in Intl. Conf. on Acoustics, Speech and Signal Processing. 1993. Minneapolis, MN. p.
57. Pruzansky, S., *Pattern matching procedure for automatic talker recognition*. J. Acoustic Society of America, 1963. v. 35, n. : p. 354-358.
58. Pfeifer, L. *Feature analysis for speaker identification*. 1977.
59. Paul, J., et al. *Development of analytical methods for a semi-automatic speaker identification system*. in Carnahan Conf. on Crime Countermeasures. 1975. p.
60. Rosenberg, A., C.H. Lee, and F. Soong. *Sub-Word Unit Talker Verification Using Hidden Markov Models*. in Intl. Conf. on Acoustics, Speech and Signal Processing. 1990. Albuquerque, NM. p. 269-272.
61. Gillick, L., et al. *Application of Large Vocabulary Continuous Speech Recognition to Topic and Speaker Identification Using Telephone Speech*. in ICASSP 93. 1993. Minneapolis. p. 471-474.
62. Higgins, A. and R. Wohlford. *A new method of text-independent speaker recognition*. in ICASSP 86. 1986. Tokyo, Japan. p.

63. Markel, J. and S. Davis, *Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base*. IEEE Trans. on Acoustics, Speech, and Signal Processing, 1979. v. 27, n. 1: p. 74-82.
64. Soong, F., et al. *A Vector Quantization Approach to Speaker Recognition*. in *Intl. Conf. on Acoustics, Speech and Signal Processing*. 1985. Tampa, FL. p. 387-390.
65. Schwartz, R., S. Roucos, and M. Berouti. *The Application of Probability Density Estimation To Text Independent Speaker Identification*. in *Intl. Conf. on Acoustics, Speech and Signal Processing*. 1982. Paris, France. p. 1649-1652.
66. Gish, H. and M. Schmidt, *Text-Independent speaker identification*, in *IEEE Signal Processing* 1994, p. 18-32.
67. Dempster, A., N. Laird, and D. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*. J. of the Royal Statistical Society, 1977. v. 39, n. 1: p. 1-22.
68. Reynolds, D. and R. Rose, *Robust text-independent speaker identification using Gaussian mixture speaker models*. IEEE Trans. on ASSP, 1995. v. 3, n. : p. 72-83.
69. Samal, A. and P. Iyengar, *Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey*. Pattern Recognition, 1992. v. 25, n. 1: p. 65-77.
70. Galton, F., *Numeralized Profiles for Classification and Recognition*. Nature, 1910. v. 83, n. : p. 127-130.
71. Harmon, L., et al., *Identification of Human Profiles by Computer*. Pattern Recognition, 1978. v. 10, n. : p. 301-312.
72. Wu, C.J. and J.S. Huang, *Human Profile Recognition by Computer*. Pattern Recognition, 1990. v. 23, n. 3: p. 255-259.
73. Preston, K., *Computing at the speed of light*. Electronics, 1965. v. 38, n. : p. 72-83.
74. Brunelli, R. and T. Poggio, *Face Recognition: Features versus Templates*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1993. v. 15, n. 10: p. 1042-1052.
75. Kanade, T., *Computer Recognition of Human Faces*. 1977, Birkhauser Verlag.
76. Govindaraju, V., D. Sher, and R. Srihari. *Locating Human Faces in Newspaper Photographs*. in *Proc. Conf. on Computer Vision and Pattern Recognition*. 1989. p. 549-554.

77. Ballard, D., *Generalizing the Hough transform to Detect Arbitrary Shapes*. Pattern Recognition, 1981. v. 3, n. 2: p. 111-122.
78. Yuille, A., P. Hallinan, and D. Cohen, *Feature Extraction from Faces Using Deformable Templates*. Intl. J. of Computer Vision, 1992. v. 8, n. 2: p. 99-111.
79. Kass, M., A. Witkin, and D. Terzopoulos, *Snakes: Active Contour Models*. Intl. J. of Computer Vision, 1988. v. n. : p. 321-331.
80. Huang, C.L. and C.W. Chen, *Human facial feature extraction for face interpretation and recognition*. Pattern Recognition, 1992. v. 25, n. 12: p. 1435-1443.
81. Phillips, D. and A. Smith, *Bayesian faces via hierarchical template modeling*. J. of the American Statistical Association, 1994. v. 89, n. 428: p. 1151-1163.
82. Lanitis, A., C. Taylor, and T. Cootes, *Automatic face identification system using flexible appearance models*. Image and Vision Computing, 1995. v. 13, n. 5: p. 292-401.
83. Cootes, T., *et al.*, *Use of active shape models for locating structures in medical images*. Image and Vision Computing, 1994. v. 12, n. 6: p. 355-365.
84. Kamel, M., *et al.*, *System for the recognition of human faces*. IBM Systems Journal, 1993. v. 32, n. 2: p. 307-319.
85. Sutherland, K., D. Renshaw, and P. Denyer. *Automatic face recognition*. in *Intelligent Systems Engineering*. 1992. Edinburgh. p. 29-34.
86. Baron, R., *Mechanisms of Human Facial Recognition*. Intl. J. Man-Machine Studies, 1981. v. 15, n. : p. 137-178.
87. Nakamura, O., M. Shailendra, and T. Minami, *Identification of Human Faces Based on Isodensity Maps*. Pattern Recognition, 1991. v. 24, n. 3: p. 263-272.
88. Lades, M., *et al.*, *Distortion invariant object recognition in the dynamic link architecture*. IEEE Trans. on Computers, 1993. v. 42, n. 3: p. 300-311.
89. Sirovich, L. and M. Kirby, *Low-Dimensional Procedure for the Characterization of Human Faces*. J. of the Optical Society of America, 1987. v. 4, n. 3: p. 519-524.
90. Cheng, Y.Q., K. Liu, and J.Y. Yang, *A Novel Feature Extraction Method for Image Recognition Based on Similar Discriminant Function (SFD)*. Pattern Recognition, 1993. v. 26, n. 1: p. 115-125.
91. Fukunaga, K., *Statistical Pattern Recognition*. 2nd ed. 1990, San Diego: Academic Press.

92. Burt, P., *Smart Sensing With a Pyramid Vision Machine*. Proc. IEEE, 1988. v. 76, n. 8.
93. Burt, P. and E. Adelson, *The Laplacian pyramid as a compact image code*. IEEE Trans. on Communications, 1983. v. 31, n. 4: p. 532-540.
94. Brooke, N., *Mouth Shapes and Speech*, in *Processing Images of Faces*. 1992, Ablex Publishing Corp.: Norwood, NJ.
95. Petajan, E. *Automatic Lipreading to Enhance Speech Recognition*. in *Global Telecommunications Conference*. 1984. Atlanta, GA. p. 265-272.
96. McGurk, H. and J. MacDonald, *Hearing lips and seeing voices*. Nature, 1976. v. 264, n. : p. 746-748.
97. Kabre, H. *Audiovisual speech recognition using fuzzy shape filters model*. in *Eurospeech*. 1995. Madrid. p. 307-310.
98. Chen, T. and R. Rao, *Audio-visual interaction in multimedia*. IEEE Circuits and Devices Magazine, 1995. v. November, n. : p. 21-26.
99. Brunelli, R., *et al.*, *Automatic person recognition by acoustic and geometric features*. Machine Vision and Applications, 1995. v. 8, n. : p. 317-325.
100. Rosenberg, A. and F. Soong, *Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes*. Computer Speech and Language, 1987. v. 22, n. : p. 143-157.
101. Poggio, T. and F. Girosi, *Networks for approximation and learning*. Proceedings of IEEE, 1990. v. 78, n. : p. 1481-1497.
102. LeGoff, B., T. Guiard-Marigny, and C. Benoit. *Read my lips... and my jaw! How intelligible are the components of a speaker's face?* in *Eurospeech*. 1995. Madrid. p. 291-294.
103. Ekman, P., *et al.*, *Face, voice, and body in detecting deceit*. J. of Nonverbal Behavior, 1991. v. 15, n. 2: p. 125-135.
104. Tien, J., T. Rich, and M. Cahn, *Electronic Fund Transfer Systems Fraud (Report NCJ-100461)*. 1985, Washington, DC: US Dept. of Justice.
105. Bar-Shalom, Y., *Comparison of Two-Sensor Tracking Methods Based on State Vector Fusion and Measurement Fusion*. IEEE Trans. on Aerospace and Electronic Systems, 1988. v. 24, n. 4: p. 447-457.
106. Tenney, R. and N. Sandell, *Detection With Distributed Sensors*. IEEE Trans. on Aerospace and Electronic Systems, 1981. v. 17, n. 4: p. 501-510.
107. Sadjadi, F., *Hypothesis Testing in a Distributed Environment*. IEEE Trans. on Aerospace and Electronic Systems, 1986. v. 22, n. 2: p. 134-137.

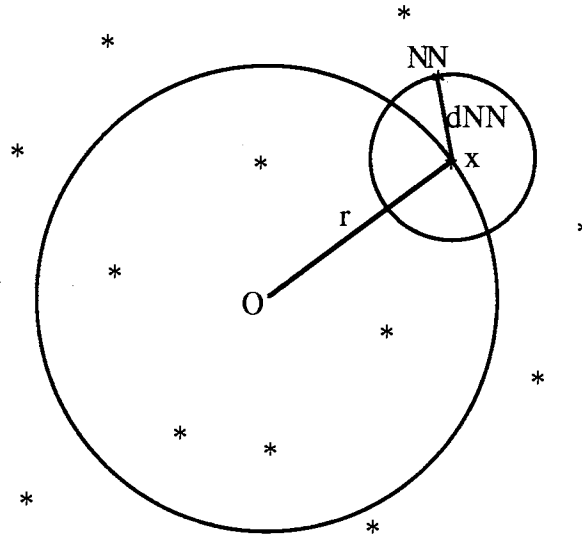
108. Chair, Z. and P. Varshney, *Optimal Data Fusion in Multiple Sensor Detection Systems*. IEEE Trans. on Aerospace and Electronic Systems, 1986. v. 22, n. 1: p. 98-101.
109. Reibman, A. and L. Nolte, *Optimal Detection and Performance of Distributed Sensor Systems*. IEEE Trans. on Aerospace and Electronic Systems, 1987. v. 23, n. 1: p. 24-30.
110. Diffie, W. and M.E. Hellman, *New Directions in Cryptography*. IEEE Trans. on Information Theory, 1976. v. 22, n. : p. 644-654.
111. Massey, J.L., *An Introduction to Contemporary Cryptology*. Proc. IEEE, 1988. v. 76, n. 5: p. 533-549.
112. Simmons, G.J., *A Survey of Information Authentication*. Proc. of the IEEE, 1988. v. 76, n. 5: p. 603-620.
113. Rivest, R.L., A. Shamir, and L. Adelman, *A Method for Obtaining Digital Signatures and Public-Key Cryptosystems*. Communications of the ACM, 1978. v. February, n. .
114. Nechvatal, J. *Computer Security: Public-Key Cryptography*. 1991.
115. Schneier, B., *The Cambridge Algorithms Workshop*. Dr. Dobbs Journal, 1994. v. 20, n. 4: p. 18-24.
116. Stallings, W., *SHA: The Secure Hash Algorithm*. Dr. Dobb's Journal, 1994. v. April, n. : p. 32-33.
117. Treasury, U.S.D.o. *Applications of Computer Card Technology*. 1991.
118. DNA, *Statement of Work for "Identification Verification Technology/Methodology"*. 1993. v. n. .
119. NSA. *Statement of Work for "YOHO Speaker Authentication"*. 1986.
120. Hoyt, J., *Personal Communication* 1992,
121. *Vital and Health Statistics*. 1980.
122. Rabiner, L. and F. Juang, *Fundamentals of Speech Recognition*. 1993, Englewood Cliffs, NJ: Prentice Hall.
123. Oppenheim, A., R. Shafer, and T. Stockham, *Nonlinear filtering of multiplied and convolved signals*. Proc. IEEE, 1968. v. 56, n. : p. 1264-1291.
124. Gonzalez, R. and R. Woods, *Digital Image Processing*. 1992, Addison-Wesley.
125. Abramowitz, M. and A. Stegun, *Handbook of Mathematical Functions*. 1970, New York: Dover.

126. Deutsch, R., *Estimation Theory*. 1965, Englewood Cliffs, N. J.: Prentice Hall.
127. Press, W., *et al.*, *Numerical Recipes in C*. 1988, Cambridge University Press.



**APPENDIX: Density versus NN Distance for Gaussian PDF**

To explore the relationship between probability density and nearest-neighbor distance, consider first the case of the standardized Gaussian probability function,  $p = N(\mathbf{0}, I_v)$ , where  $v$  equals the number of independent dimensions. Suppose we have observed  $N$  samples generated from the density function  $p$ , and we wish to estimate the local density,  $p_x = p(\mathbf{x})$ , at a test point  $\mathbf{x}$ . This is illustrated in Figure A.1. We measure the squared Euclidean distance,  $d^2$ , between  $\mathbf{x}$  and each of the  $N$  samples. The squared Euclidean distance to the closest of these samples (NN, the nearest neighbor) is denoted as  $d_{NN}^2$ . Consider a statistical ensemble of trials in which, in each trial, a test point  $\mathbf{x}$  and  $N$  samples from  $p$  are jointly selected at random. Both  $p_x$  and  $d_{NN}^2$  may then be treated as random variables:  $p_x$  due to random selection of test point  $\mathbf{x}$ , and  $d_{NN}^2$  due to random selection of  $\mathbf{x}$  and random sampling of  $p$ .



**Figure A.1: Illustration of density estimation using NN distances.**

Let the probability distribution function of  $d^2$  be  $F_{d^2}(\delta)$ . That is,  $F_{d^2}(\delta) = \text{prob}(d^2 \leq \delta)$ . The probability of a randomly chosen sample from  $p$  falling

outside a ball of radius  $\delta$  centered at  $\mathbf{x}$  equals  $1 - F_{d^2}(\delta)$ . The probability of all  $N$  independent samples in population  $T$  falling outside the ball equals  $(1 - F_{d^2}(\delta))^N$ . The probability of any of the  $N$  samples falling inside the ball equals  $1 - (1 - F_{d^2}(\delta))^N$ . Therefore, the distribution function  $F_{d_{NN}^2}(\delta)$  of  $d_{NN}^2$  can be expressed as:

$$F_{d_{NN}^2}(\delta) = 1 - (1 - F_{d^2}(\delta))^N \quad (\text{A.1})$$

The value of  $F_{d^2}(\delta)$  can be determined as the integral of  $p$  within the ball of radius  $\delta$  centered at  $\mathbf{x}$ . Since  $p$  is Gaussian,

$$F_{d^2}(\delta) = P_{\chi^2}(\delta | v, r^2), \quad (\text{A.2})$$

where  $r^2$  equals the squared Euclidean distance from the origin to  $\mathbf{x}$ , and  $P_{\chi^2}(\delta | v, r^2)$  is the non-central chi-squared distribution with  $v$  degrees of freedom and non-centrality parameter  $r^2$  [125].

### A.1

#### Expectation of $p_{\mathbf{x}}$

#### Conditional

We wish to estimate the density  $p_{\mathbf{x}}$  at sample point  $\mathbf{x}$ , given the squared Euclidean distance to the nearest sample from population  $T$ . The minimum mean-squared error estimate equals the conditional expected value of  $p_{\mathbf{x}}$  given  $d_{NN}^2$  [126].

$$E_{p_{\mathbf{x}} | d_{NN}^2}(\delta = d_{NN}^2) = \int_0^{\infty} \rho p_{p_{\mathbf{x}} | d_{NN}^2}(\rho | \delta) d\rho \quad (\text{A.3})$$

Using Bayes' rule,

$$= \int_0^{\infty} \rho \frac{p_{d_{NN}^2 | p_{\mathbf{x}}}(\delta | \rho) p_{p_{\mathbf{x}}}(\rho)}{p_{d_{NN}^2}(\delta)} d\rho \quad (\text{A.4})$$

Expanding the denominator in terms of conditional probabilities of  $d_{NN}^2$  given  $p_x$ :

$$= \frac{\int_0^{\infty} \rho p_{d_{NN}^2 | p_x}(\delta | \rho) p_{p_x}(\rho) d\rho}{\int_0^{\infty} p_{d_{NN}^2 | p_x}(\delta | \rho) p_{p_x}(\rho) d\rho} \quad (\text{A.5})$$

Solution of Equation (A.5) requires two PDFs:  $p_{p_x}(\rho)$  and  $p_{d_{NN}^2 | p_x}(\delta | \rho)$ . Expressions for these two functions are derived in the following two sections.

## A.2 Density of $p_x$

Note that  $p_x$ , the density of the feature space at test point  $x$ , is a random variable that depends on  $x$ . The probability density of  $p_x$ , denoted as  $p_{p_x}(\rho)$ , is a probability density function of a probability density! Because  $p$  is assumed to be a normalized Gaussian function, the value of  $p_x$  depends only on the distance  $r$  of  $x$  from the origin. The density at radius  $r$  is

$$p_x(r) = \frac{e^{-r^2/2}}{(2\pi)^{v/2}} \quad (\text{A.6})$$

Figures 4.4a and 4.4b are sketches of  $p_x(r)$  and  $p_{p_x}(\rho)$ , respectively. At  $r=0$ ,  $p_x$  reaches a maximum value of  $p_{MAX} = (2\pi)^{-v/2}$ . Random selection of test point  $x$  often results in a value of  $r$  for which  $p_x$  is nearly zero. This is particularly true for large values of dimensionality,  $v$ . The function  $p_{p_x}(\rho)$  therefore reaches a maximum value near  $\rho=0$ . Since  $p_x \leq p_{MAX}$ ,  $p_{p_x}(\rho) = 0$  for  $\rho > p_{MAX}$ .

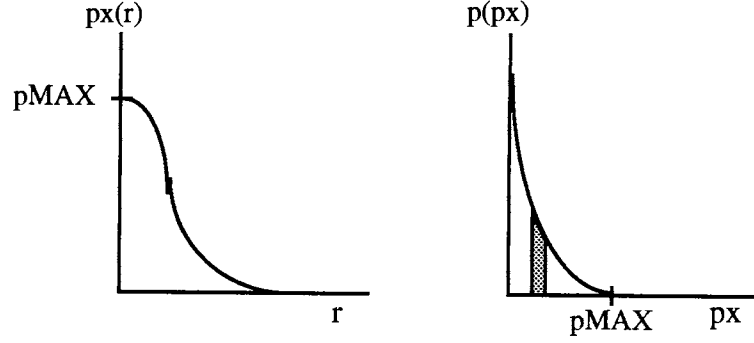


Figure A.2: (a) sample density, (b) density of  $p_x$ .

Consider a spherical shell centered at the origin with radius  $r$  and thickness  $\Delta r$ . The probability mass,  $m$ , contained in the shell is equal to its volume (surface area times thickness) times the density within the shell,  $p_x$ . The surface area,  $S_v$ , of a  $v$ -dimensional sphere of radius  $r$  is

$$S_v = \frac{2(\pi^{v/2})}{\Gamma(v/2)} r^{v-1}.$$

Therefore,

$$M = \frac{2(\pi^{v/2})}{\Gamma(v/2)} r^{v-1} \Delta r p_x. \quad (\text{A.7})$$

The shaded vertical section in Figure 4.4b corresponds to density values within the spherical shell. The density at radius  $r+\Delta r$ , for small  $\Delta r$ , is

$$\frac{e^{-(r+\Delta r)^2/2}}{(2\pi)^{v/2}} \approx \frac{e^{-(r^2+2r\Delta r)/2}}{(2\pi)^{v/2}} = p_x e^{-r\Delta r}$$

The change of density,  $\Delta p_x$ , between the inner and outer surface of the shell is

$$\Delta p_x = p_x - p_x e^{-r\Delta r} = p_x(1 - e^{-r\Delta r}) \approx p_x r \Delta r.$$

The area represented by the shaded vertical section is then

$$A = p_x(r) \Delta p_x \approx p_x(r) p_x r \Delta r. \quad (\text{A.8})$$

The area  $A$  must equal the probability mass  $M$  within the shell. Equating the terms in Equations A.7 and A.8,

$$p_{p_x}(r)p_x r \Delta r = \frac{2(\pi^{v/2})}{\Gamma(v/2)} r^{n-1} \Delta r p_x.$$

Solving for  $p_{p_x}(r)$ ,

$$p_{p_x}(r) = \frac{2(\pi^{v/2})}{\Gamma(v/2)} r^{n-2},$$

or, in terms of density values,

$$p_{p_x}(\rho) = \frac{2(\pi^{v/2})}{\Gamma(v/2)} (-2 \ln[(2\pi)^{v/2} \rho])^{(v-2)/2}. \quad (\text{A.9})$$

### A.3 Conditional Density of $d_{NN}^2$

To determine  $p_{d_{NN}^2|p_x}(\delta | \rho)$ , first note that conditioning on  $p_x$  is equivalent to conditioning on  $r^2$  (the squared distance of  $x$  from the origin):

$$p_{d_{NN}^2|p_x}(\delta | \rho) = p_{d_{NN}^2|r^2}(\delta | \lambda_x) \quad (\text{A.10})$$

$$= \frac{d}{da} [ p_{d_{NN}^2|r^2}(a | \lambda_x) ]_{a=\delta} \quad (\text{A.11})$$

$$= \frac{d}{da} [ 1 - (1 - p_{d^2|r^2}(a | \lambda_x))^N ]_{a=\delta} \quad (\text{A.12})$$

$$= \frac{d}{da} [ 1 - (1 - p_{\chi^2}(a | v, \lambda_x))^N ]_{a=\delta} \quad (\text{A.13})$$

$$= \frac{d}{da} [ 1 - (1 - \int_0^a p_{\chi^2}(\xi | v, \lambda_x) d\xi)^N ]_{a=\delta} \quad (\text{A.14})$$

$$= N [ 1 - p_{\chi^2}(\delta | v, \lambda_x) ]^{N-1} p_{\chi^2}(\delta | v, \lambda_x) \quad (\text{A.15})$$

In Equation A.10, the variable  $\lambda_x$  is related to  $\rho$  by:  $\lambda_x = -2 \ln(\rho(2\pi)^{-v/2}) = -2 \ln \rho - v \ln(2\pi)$ . Equation A.11 follows from the definition of the density function. Equation A.12 uses Equation A.1 to relate the distribution of nearest-

neighbor distances (out of a population of  $N$  samples),  $P_{d_{NN}^2|r^2}(a | \lambda_x)$ , to the distribution of *all* distances between samples,  $P_{d^2|r^2}(a | \lambda_x)$ . Equation A.13 substitutes a non-central chi-squared distribution for  $P_{d^2|r^2}(a | \lambda_x)$  using Equation A.2. Equation A.14 replaces the distribution function with an integrated density function. Finally, Equation A.15 carries out the differentiation.

Substituting Equations A.9 and A.15 into Equation A.5, and simplifying,

$$E_{p_x|d_{NN}^2}(\delta) = \frac{\int_0^{pMAX} \rho N [1 - P_{\chi^2,2}(\delta|v, \lambda_x)]^{N-1} P_{\chi^2,2}(\delta|v, \lambda_x) \ln(\rho) d\rho}{\int_0^{pMAX} N [1 - P_{\chi^2,2}(\delta|v, \lambda_x)]^{N-1} P_{\chi^2,2}(\delta|v, \lambda_x) \ln(\rho) d\rho} \quad (A.16)$$

where  $pMAX = (2\pi)^{v/2}$ .

#### A.4 Numerical Evaluation of $E_{p_x|d_{NN}^2}$

Numerical evaluation of Equation A.16 is difficult because the  $\ln(\rho)$  term, present in both the numerator and denominator, is unbounded at the lower limit of integration. We therefore introduce a change of variable,  $\psi = -\ln(\rho)$ , or  $\rho = e^{-\psi}$ . Equation A.16 then becomes

$$E_{p_x|d_{NN}^2}(\delta) = \frac{\int_{-\ln(pMAX)}^{\infty} e^{-\psi} N [1 - P_{\chi^2,2}(\delta|v, \lambda_x)]^{N-1} P_{\chi^2,2}(\delta|v, \lambda_x) \psi e^{-\psi} d\psi}{\int_{-\ln(pMAX)}^{\infty} N [1 - P_{\chi^2,2}(\delta|v, \lambda_x)]^{N-1} P_{\chi^2,2}(\delta|v, \lambda_x) \psi e^{-\psi} d\psi} \quad (A.17)$$

where  $\lambda_x = -2\psi - v \ln(2\pi)$ . Evaluation of Equation A.17 was accomplished using Romberg's integration method [127].

## A.5 Approximation of Median NN Distance

From Equation A.1, the median nearest-neighbor distance is equal to the value of  $\delta$  for which  $F_{d_{NN}^2}(\delta) = 1 - (1 - F_{d^2}(\delta))^N = \frac{1}{2}$ . Taking logs of both sides and rearranging,

$$\ln(1 - F_{d^2}(\delta)) = \frac{\ln(1/2)}{N}.$$

For large  $N$ , a good approximation is

$$F_{d^2}(\delta) = \frac{-0.693}{N}. \quad (\text{A.18})$$

In this expression,  $F_{d^2}(\delta)$  is the cumulative distribution of inter-sample distances. For the Gaussian case,  $F_{d^2}(\delta)$  is a weighted sum of non-central chi-squared distributions:

$$F_{d^2}(\delta) = \int_0^{\infty} P_{\chi^2}(\delta|v, \lambda) p_{\chi^2}(\lambda) d\lambda$$

The value of  $\delta$  which satisfies Equation A.18 can be determined using a numerical root-finding technique such as the Van Wijngaarden-Decker-Brent method [127].

## A.6 Interpretation

Figure A.3 plots the negative log density estimated from Equation A.17 as a function of  $d_{NN}$  (not  $d_{NN}^2$ ). The median value of  $d_{NN}$  (from Equation A.18) is used as a normalizing factor, so that an x-axis value of 2 indicates an NN distance twice as great as the median NN distance. The data shown are for a three-dimensional space ( $v=3$ ), with one sample and one million samples. The  $N=1$  curve appears to be a parabolic shape, consistent with the affine

connection. The  $N=1000000$  curve has a complex shape, changing its direction of curvature twice within the range plotted. At distances several times the median NN distance, the  $N=1000000$  curve has negative curvature, consistent with the use of a log function as in Equation 5.15. Note that for small  $d_{NN}$ , the two estimators approach a limiting density value that is slightly lower than the actual maximum density of a Gaussian ( $2\pi^{-v/2}$ ). It is possible (although unlikely) for a  $d_{NN}$  value near zero to be observed at a test point distant from the mode of the distribution. To account for this possibility, the estimator never reaches the theoretical maximum.

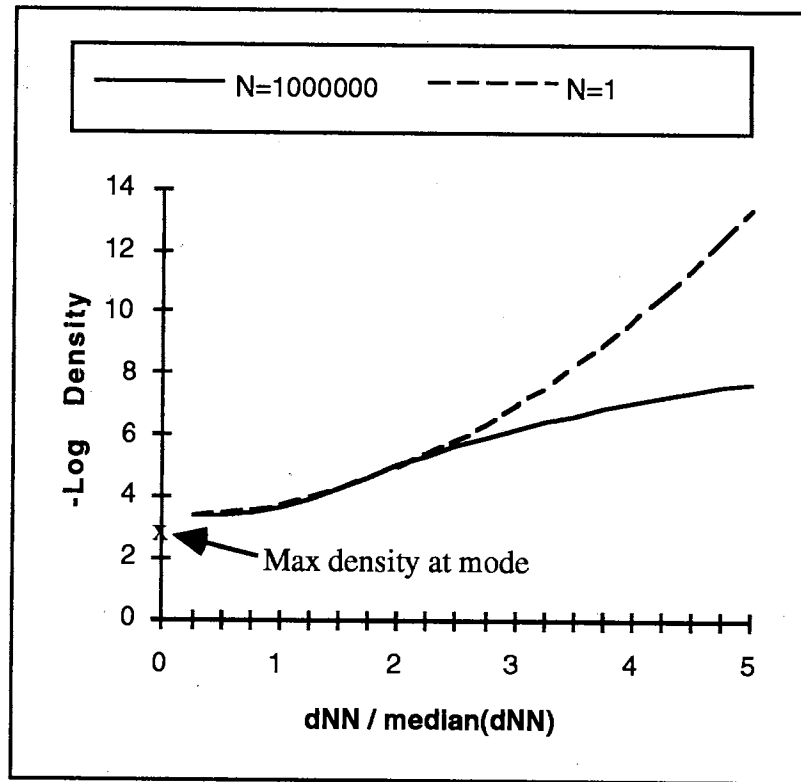


Figure A.3: Density estimators for  $v=3$ .

Figure A.4 shows a similar pair of curves for  $v=13$ . The  $N=1$  curve is again approximately parabolic and the  $N=1000000$  curve is again a complex shape. The  $N=1000000$  curve has positive curvature at distances near the



median of  $d_{NN}$ . Negative curvature at distances of several times the median of  $d_{NN}$  is not apparent in this case as it was for  $v=3$ .

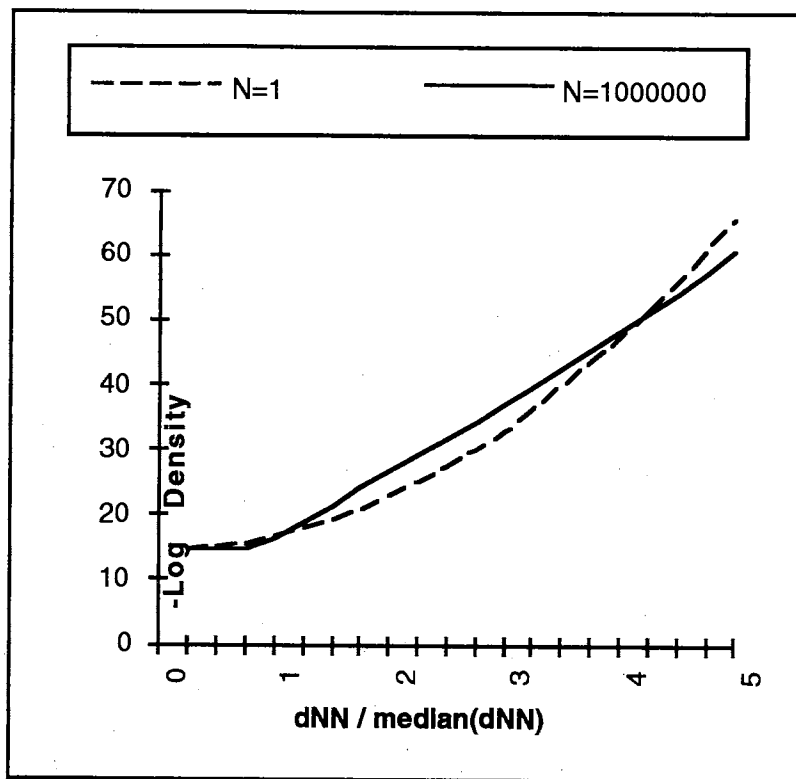


Figure A.4: Density estimators for  $v=13$ .

In most applications,  $N=1000000$  is an impractically large number of samples, although for  $v=13$ , it is still much too small to justify the asymptotic arguments leading to Equation 5.15. Figure A.5 shows the optimal estimator for  $N=1000$  and  $v=13$ . This represents a combination of dimensionality and sample size that is of practical interest. The curve is parabola like, with no changes in the direction of curvature. Note the similarity of this curve to the data and parabolic fit shown previously in Figure 5.4

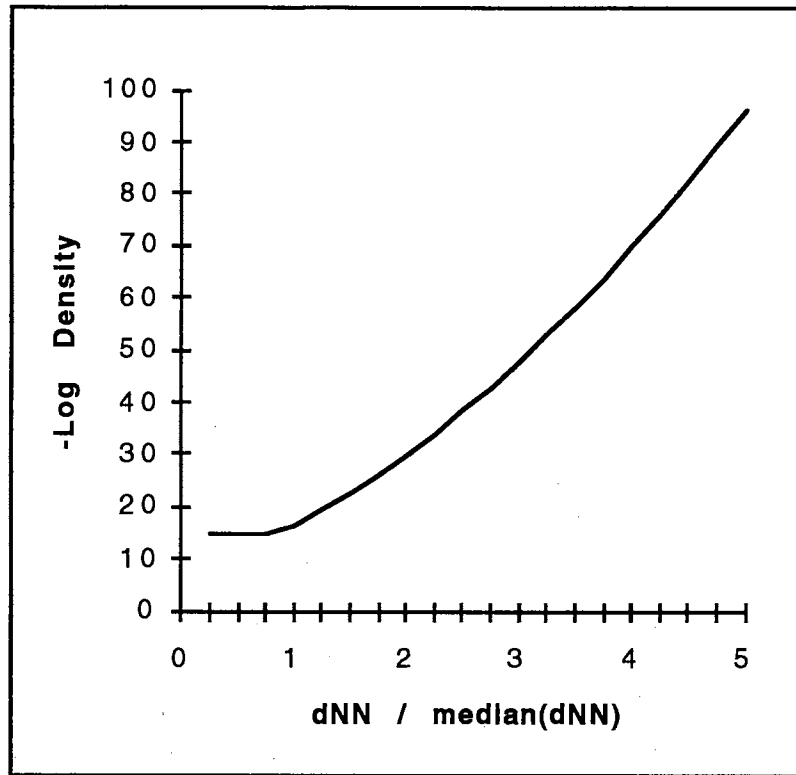


Figure A.5: Density estimator for  $v=13$ ,  $N=1000$ .

Examination of a number of curves such as those plotted above for various combinations of sample size and dimensionality suggest that the optimal estimator is logarithm-like in the limit of large sample size and low dimensionality, and parabola-like in the limit of small sample size and high dimensionality. Many practical problems approximate the latter limit.

VITA

Alan Lawrence Higgins

Candidate for the Degree of

Doctor of Philosophy

Thesis: MULTI-MEDIA PERSONAL IDENTITY VERIFICATION

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in San Diego, California, on November 2, 1952, the son of Larwrence N. Higgins and Ruth I. MacEachern.

Education: Graduated from Point Loma High School, San Diego, in 1970; received Bachelor of Arts and Master of Science degrees in Applied Mechanics and Engineering Science from the University of California at San Diego in 1974 and 1977, respectively; completed the requirements for the Doctor of Philosophy degree at Oklahoma State University in July 1996.

Experience: From 1977 to 1978, Mr. Higgins was a Development Engineer with Scripps Institute of Oceanography, where he developed instrumentation and mathematical models for predicting sand transport on beaches. In 1978 he joined Bolt, Beranek and Newman as a Scientist in the Speech Group, where he developed algorithms for narrowband speech compression. From 1981 to 1996 Mr. Higgins has been employed with ITT Industries, where he has held positions including Senior Scientist and Manager of the Speech Department. His primary technical interests are in the areas of speech and speaker recognition. Since 1985, he has been Principal Investigator on seven government R&D contracts, involving nearly 20 man-years of effort. He has authored numerous technical papers and holds six U.S. patents.