More about Biometric Databases

Brad Wing – NIST
Austin Hicklin - Noblis

26 January 2015

Basic Issues

- Can existing datasets be effectively used for new research?
- When collecting data for one purpose, is it cost-effective (or even possible) to design it for multiple purposes?
- Can researchers ‘trust’ the data and metadata from existing sources?
- Will the sponsoring organizations allow the use of data from outside sources? IRB, copyrights, MOUs…

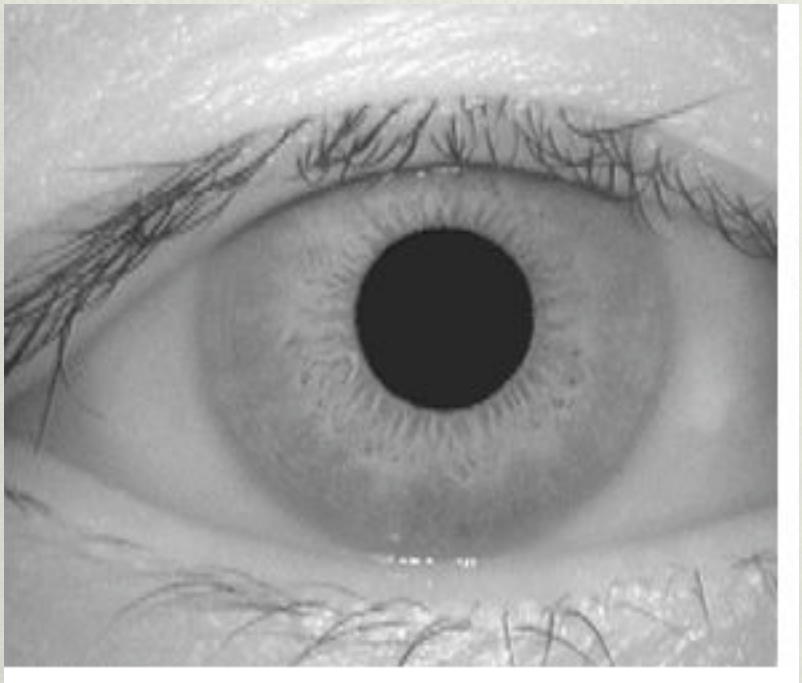
Basic Issues

- Is the size and ‘depth’ of databases sufficient?
- Has ground truth been established?
- Have parameters of data collection been sufficiently categorized?
- How to the ‘holes’ in data needs get met?
- Can multiple existing data sources be effectively used instead of collecting new data for each study or grant?

Is there data available?

- Need to ask:
  - What data characteristics are needed to test / prove / disprove a hypothesis?
  - Will the study be sufficiently affected if one or more of those characteristics is not present ?
- Database Catalogue as a resource
  - To be covered as a talk in this conference
  - Lists major characteristics of the known databases

Example: Is altered data suitable for research ?



- CASIA-IrisV1 dataset

From the website description: "In order to protect our IPR in the design of our iris camera (especially the NIR illumination scheme), the pupil regions of all iris images in CASIA-IrisV1 were automatically detected and replaced with a circular region of constant intensity to mask out the specular reflections from the NIR illuminators. Such editing clearly makes iris boundary detection much easier but has minimal or no effects on other components of an iris recognition system, such as feature extraction and classifier design."

What about ‘outliers’?

Datasets may be ‘pure’ examples of ‘normal’ conditions or contain unusual cases. For iris, this may be optical diseases, or induced conditions (permanent or temporary) such as tattooed sclera or patterned / colored contact lenses. Images from BBC http://www.bbc.com/news/magazine-30750361

Controlled degraded data

Examples from

Q-FIRE (Quality in

Face and Iris Research

-Clarkson University

(a) (b) (c) (d)

Fig. 4. Example images from motion blur videos; 7 feet slow (a), 7 feet fast (b), 15 feet slow (c), and 15 feet fast (d).

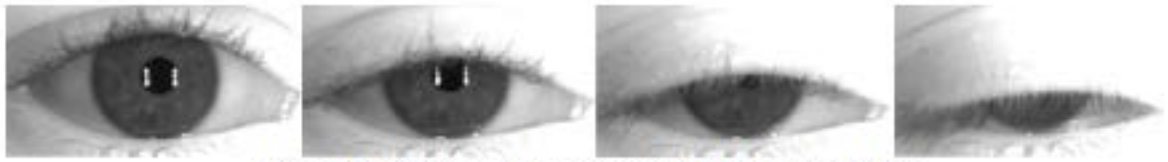


Fig. 10. Example images of blinking video.

See http://www.clarkson.edu/biosal/pdf/quality.pdf

Training to data

- Databases may become so familiar to researchers that they tune their systems to ensure ‘success’

Note

problems

on left slap

image -- from

NIST Special

Database 29

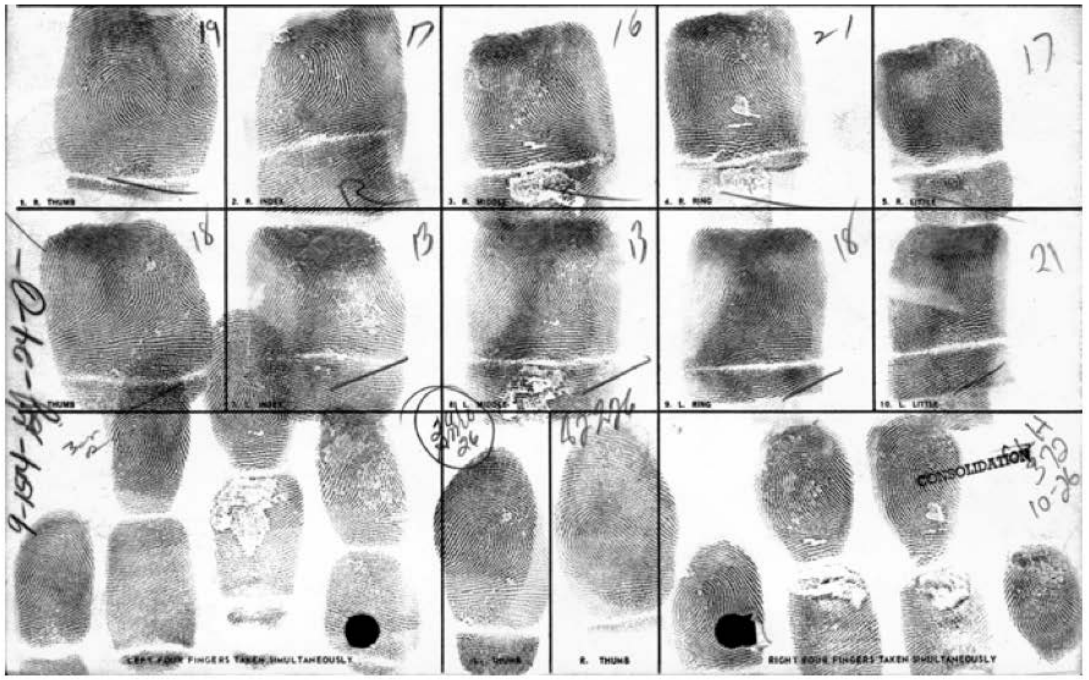


Figure 1 Image of fingerprint card a042.an2 from the database.

NIST software testing: Use of sequestered data

Testing Timeline

1. Collect Data
2. Supply K% of Data to to Capable Organizations for Development
3. Retain 100-K% as Sequestered Data
4. Development phase
5. Post vendor-reported Results
6. Host Workshop to present Analysis and Define Next Steps
7. Loop to Step 1 as needed

8. Acquire (software) implementations for formal evaluation
9. Execute SDK-based Independent Evaluation using Sequestered Data
10. Report

Dataset size

- What is the minimum size needed to answer the question?
- What is the amount of data that can be realistically / cost-effectively acquired?
- Will it be possible to extrapolate the analysis to larger dataset sizes? Should that be done?
- What are the assumed characteristics of the dataset as it grows?

- Chart from: SUMMARY OF NIST STANDARDS FOR BIOMETRIC ACCURACY, TAMPER RESISTANCE, AND INTEROPERABILITY November 13, 2002 http://biometrics.nist.gov/cs_links/pact/NISTAPP_Nov02.pdf

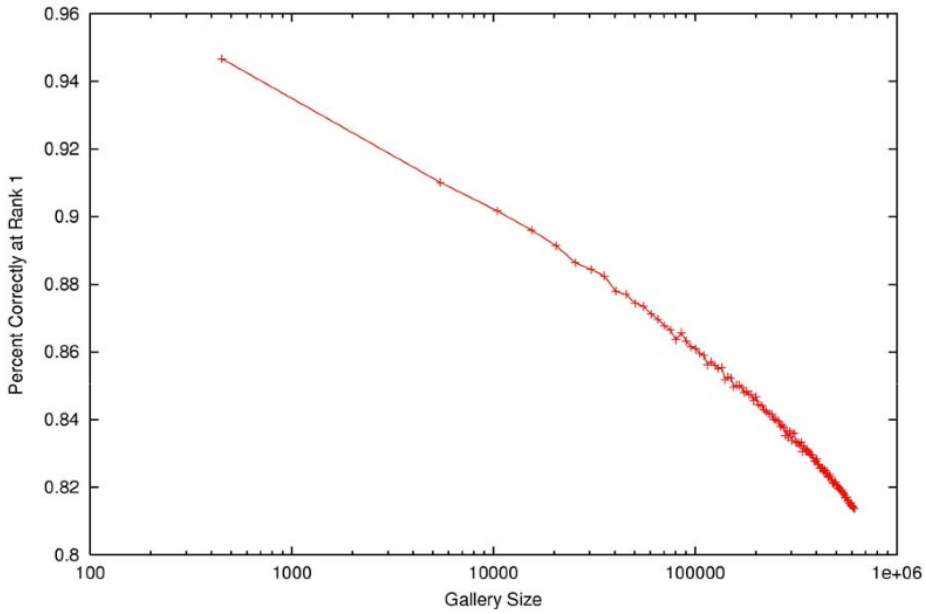
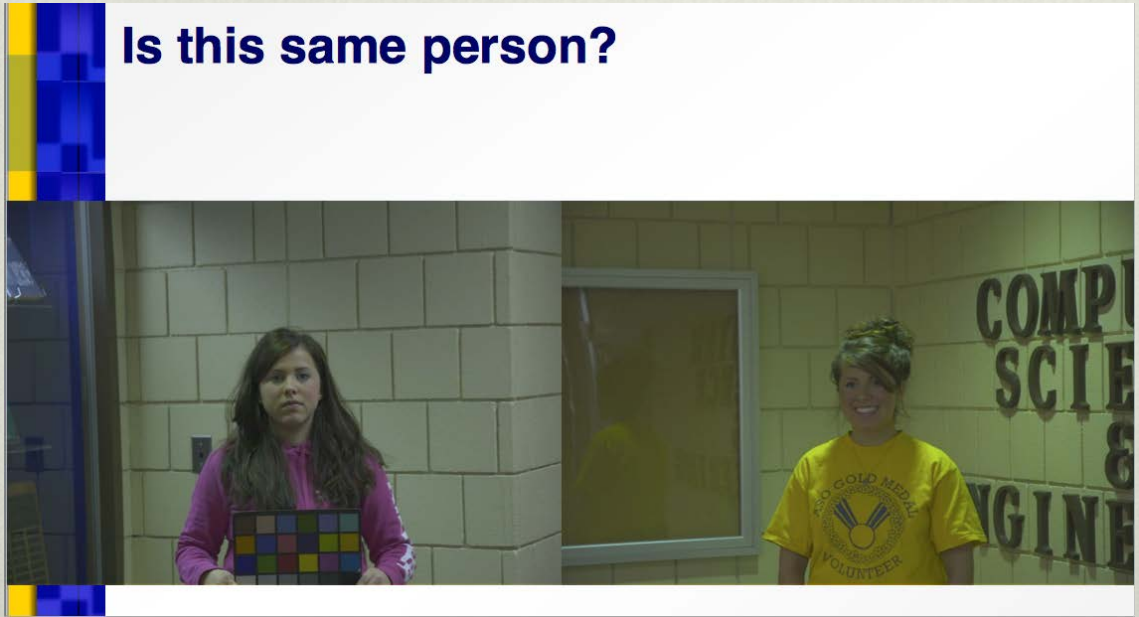


Figure 3. Probability of Detection at Rank 1 as a Function of 619,000 plain Index Fingers

Establishing ground truth

- Was data known at the time of capture (such as who spoke, what the words spoken were to be …) ?
- Was data established by experts (such as when established 'matches' between a latent friction ridge print and an exemplar) ?



http://www.nist.gov/oles/upload/7-Phillips_P-Jonathon-Challenges-in-Forensic-Face-Recognition.pdf

Changes in features

- Possible? (Tattoo removal or alteration)
- Should these be included in datasets?
- Impossible? (Tooth with root canal becoming a virgin tooth)
- Should these be ‘seeded’ in datasets?

http://www.newlookhouston.com/Tattoo_Removal_Photos.html

Data pedigree

- Has data been (unintentionally) altered? Such as through lossy compression of an image:

- From "**Comparison of the WSQ and JPEG 2000 Image Compression Algorithms On 500 ppi Fingerprint Imagery**" http://www.nist.gov/customcf/get_pdf.cfm?pub_id=910658

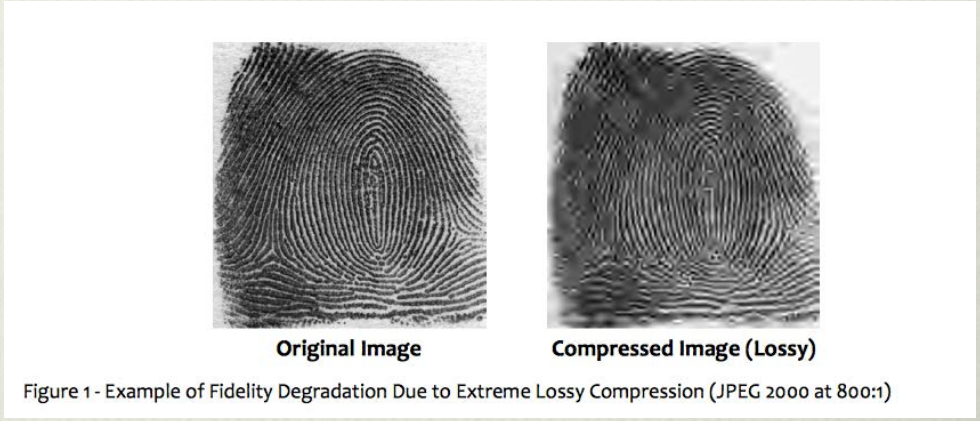
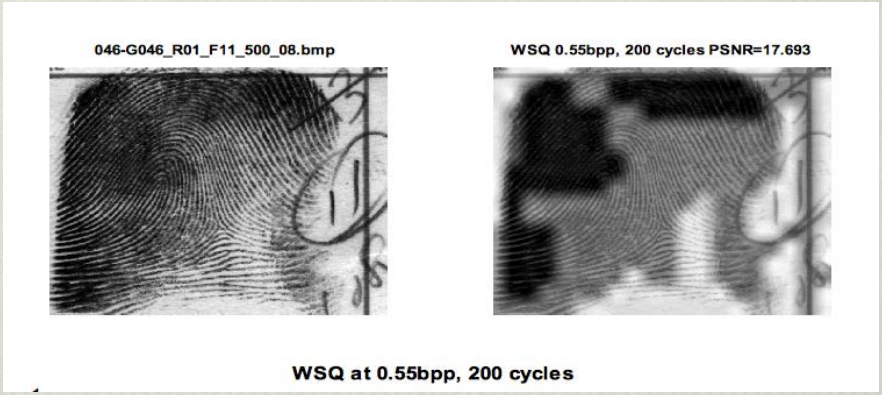
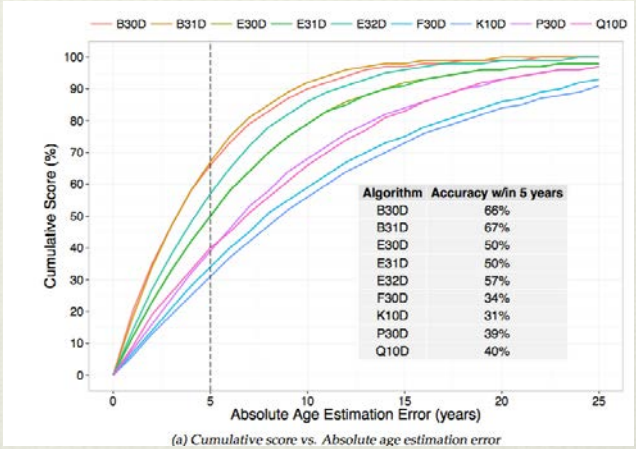


Figure 1 - Example of Fidelity Degradation Due to Extreme Lossy Compression (JPEG 2000 at 800:1)



Aging of data

- Should samples be taken sequentially, in the same or different sessions?
- Should data (such as photos) be used that were taken in 'uncontrolled' circumstances earlier?
- Is the object of the study the impact of aging or is aging a variable to be considered in analysis of the data?
- Can data from earlier in one's life be compared against current data for the modality under study?



(a) Cumulative score vs. Absolute age estimation error

http://www.nist.gov/customcf/get_pdf.cfm?pub_id=915238

CHANGES IN THE VOICE AS WE AGE

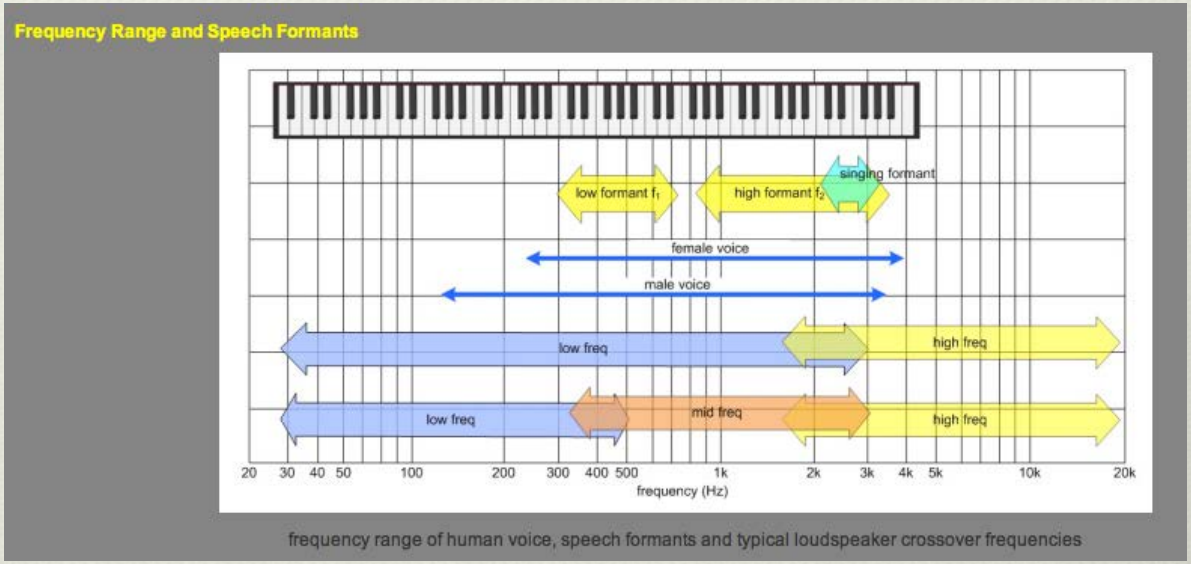
Below is a list of commonly reported voice changes as we age:

- Higher pitch voice in men
- Lower pitch voice in women
- Reduced volume and projection of the voice (or thin voice)
- Reduced vocal endurance
- Difficulty being heard in noisy situations
- Tremor or shakiness in the voice

from American Society of Otolaryngology http://www.entnet.org

Metadata about collection characteristics – voice example

Illustration from
www.Bnoack.com
(AV Info Non-Profit
Informational Website)



- Voice analysis factors:
  - Recording device type (including microphone)
  - A-D conversion ?
  - Data Compression?
  - Frequency clipping?
  - Environment
  - Redactions /
  - Discontinuities
  - Vocal content
  - Type of speech (coerced, reading, spontaneous…)
  - And more

Protect Subject's Privacy

Institutional Review Boards (IRBs) in US

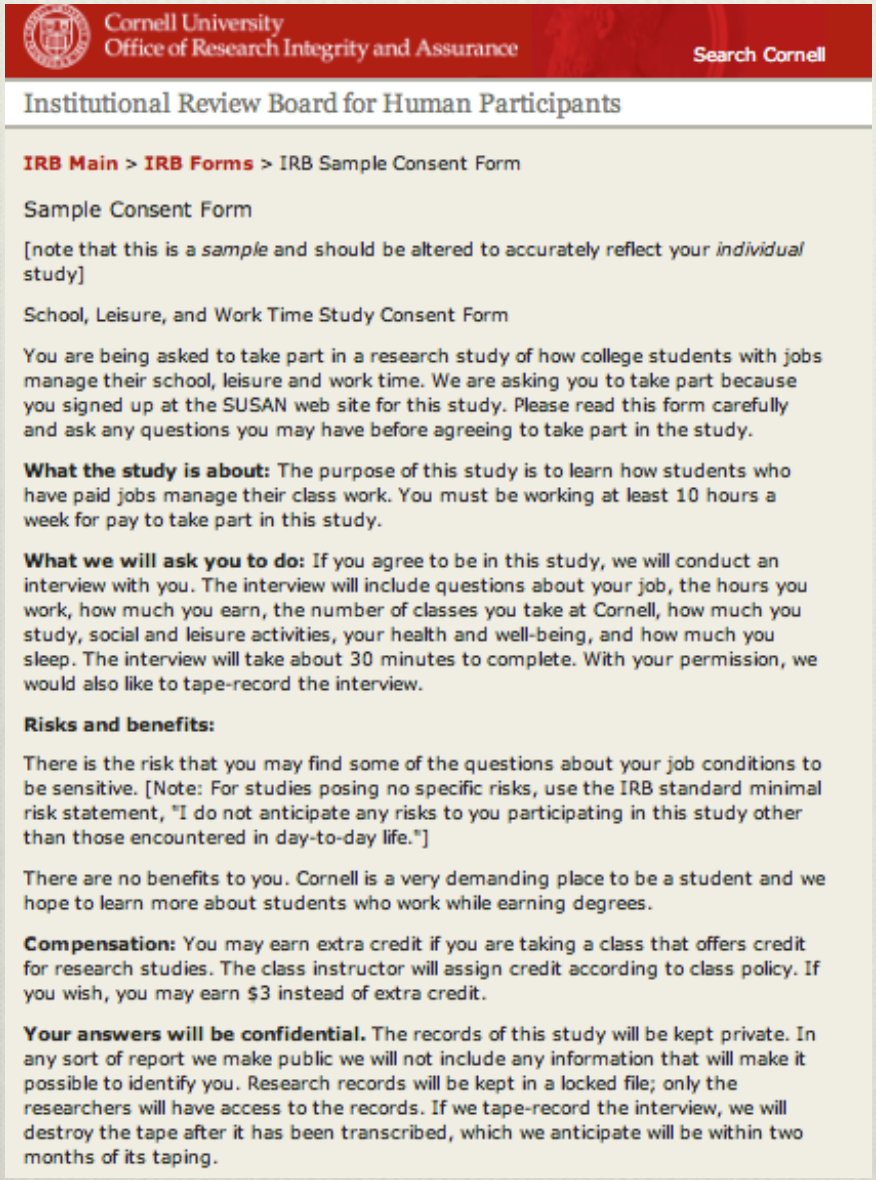
- National Research Act of 1974
- Governed by Title 45 of Federal Regulations, Part 46
- Exemptions
  - Research involving analysis of existing data and other materials if they are publicly available, or where the data can be collected such that individual subjects cannot be identified in any way

Universal Principles

- People should be informed if their personal information is being collected
- Personal information should only be used for the purpose for which it was collected
- Personal information should only be collected for a clearly identified purpose

As stated by the *Office of the Privacy Commissioner of Canada* www.priv.gc.ca

Participant Consent Sample from Cornell University



Cornell University
Office of Research Integrity and Assurance
Search Cornell

Institutional Review Board for Human Participants

IRB Main > IRB Forms > IRB Sample Consent Form

Sample Consent Form

[note that this is a *sample* and should be altered to accurately reflect your *individual* study]

School, Leisure, and Work Time Study Consent Form

You are being asked to take part in a research study of how college students with jobs manage their school, leisure and work time. We are asking you to take part because you signed up at the SUSAN web site for this study. Please read this form carefully and ask any questions you may have before agreeing to take part in the study.

What the study is about: The purpose of this study is to learn how students who have paid jobs manage their class work. You must be working at least 10 hours a week for pay to take part in this study.

What we will ask you to do: If you agree to be in this study, we will conduct an interview with you. The interview will include questions about your job, the hours you work, how much you earn, the number of classes you take at Cornell, how much you study, social and leisure activities, your health and well-being, and how much you sleep. The interview will take about 30 minutes to complete. With your permission, we would also like to tape-record the interview.

Risks and benefits:

There is the risk that you may find some of the questions about your job conditions to be sensitive. [Note: For studies posing no specific risks, use the IRB standard minimal risk statement, "I do not anticipate any risks to you participating in this study other than those encountered in day-to-day life."]

There are no benefits to you. Cornell is a very demanding place to be a student and we hope to learn more about students who work while earning degrees.

Compensation: You may earn extra credit if you are taking a class that offers credit for research studies. The class instructor will assign credit according to class policy. If you wish, you may earn $3 instead of extra credit.

Your answers will be confidential. The records of this study will be kept private. In any sort of report we make public we will not include any information that will make it possible to identify you. Research records will be kept in a locked file; only the researchers will have access to the records. If we tape-record the interview, we will destroy the tape after it has been transcribed, which we anticipate will be within two months of its taping.

Taking part is voluntary: Taking part in this study is completely voluntary. You may skip any questionsthat you do not want to answer. If you decide not to take part or to skip some of the questions, it will not affect your current or future relationship with Cornell University. If you decide to take part, you are free to withdraw at any time.

If you have questions: The researchers conducting this study are Randy Jackson and Prof. Simon Cowell. Please ask any questions you have now. If you have questions later, you may contact Randy Jackson at randy@blabmail.com or at 1-800-555-4365. You can reach Prof. Cowell at meanguy@abbey.uk or 1-800-555-4365, ext. 1000. If you have any questions or concerns regarding your rights as a subject in this study, you may contact the Institutional Review Board (IRB) at 607-255-5138 or access their website at http://www.irb.cornell.edu. You may also report your concerns or complaints anonymously through Ethicspoint (www.hotline.cornell.edu) or by calling toll free at 1-866-293-3077. Ethicspoint is an independent organization that serves as a liaison between the University and the person bringing the complaint so that anonymity can be ensured.

You will be given a copy of this form to keep for your records.

Statement of Consent: I have read the above information, and have received answers to any questions I asked. I consent to take part in the study.

Your Signature _____ Date _____

Your Name (printed) _____

In addition to agreeing to participate, I also consent to having the interview tape-recorded.

Your Signature _____ Date _____

Signature of person obtaining consent _____ Date _____

Printed name of person obtaining consent _____
Date _____

This consent form will be kept by the researcher for at least three years beyond the end of the study.

The title of the study should appear at the top of every page.

http://www.irb.cornell.edu/forms/sample.htm

Sample Database Use Form (License Agreement) from From Notre Dame University

http://www3.nd.edu/~cvrl/LicenseAgreements/UNDLicenseAgreementF.pdf

University of Notre Dame Biometrics Database Release Agreement (Collection F)

Introduction

The goal of our biometrics research is to develop new techniques, technology, and algorithms for the automatic recognition of humans. As part of this research, we are involved in an ongoing effort to collect a database of biometric imagery. The database is meant to aid research efforts in the general area of developing, testing and evaluating human recognition algorithms. The University of Notre Dame (UND) owns copyright of the collection of biometric images and serves as the source for distribution of the UND Biometrics Database.

Release of the Database

To advance the state-of-the-art in human recognition, portions of the database will be made available to researchers on a case-by-case basis. All requests for portions of the UND Biometrics Database must be submitted in writing to the UND Principal Investigator. To receive a copy of the imagery, the requestor must sign this document and agree to observe the restrictions listed below. In addition to other possible remedies, failure to observe these restrictions may result in access being denied for any other portion of the database. Current installments of the database may be made available to researchers via an Internet site or on CD or other media. There will be no charge for imagery made available and downloaded via the Internet.

Consent

The researcher(s) agrees to the following restrictions on the UND Biometrics Database:

1. Redistribution: Without prior approval from the UND Principal Investigator, the UND Biometrics Database, in whole or in part, will not be further distributed, published, copied, or disseminated in any way or form whatsoever, whether for profit or not. This includes further distributing, copying or disseminating to a different facility or organizational unit within the requesting university, organization, or company.
2. Modification and Commercial Use: Without prior approval from the University of Notre Dame, the UND Biometrics Database, in whole or in part, may not be modified or used for commercial purposes.
3. Requests for the Biometrics Database: All requests for the UND Biometrics Database will be forwarded to the UND Principal Investigator.
4. Publication Requirements: Those seeking to include renderings of more than 10 images from the UND Biometrics Database in reports, papers, and other documents to be published or released must first obtain approval in writing from the UND Principal Investigator. In no case should the ear images be used in a way that could cause the original subject embarrassment or mental anguish.
5. Citation: All documents and papers that report on research that uses the UND Biometrics Database must acknowledge the use of the database by including an appropriate citation that will be provided upon request by the UND Principal Investigator. (see below)
6. Publications to UND: A copy of all reports and papers that are for public or general release that use the UND Biometrics Database must be forwarded immediately upon release or publication to the UND Principal Investigator.
7. Indemnification: Researcher agrees to indemnify, defend, and hold harmless the University of Notre Dame du Lac and its Board of Trustees, officers, employees and agents, individually and collectively, from any and all losses, expenses, damages, demands and/or claims based upon any such injury or damage (real or alleged) and shall pay all damages, claims, judgments or expenses resulting from Researcher's use of the UND Biometrics Database.

NAME (in capitals) _____ SIGNATURE and DATE _____

ORGANIZATION AND ADDRESS (in capitals) _____

Ping Yan, Kevin W. Bowyer: Empirical Evaluation of Advanced Ear Biometrics. In: IEEE Computer Society Workshop on Empirical Evaluation Methods in Computer Vision (2005)

Send to UND Principal Investigator: Professor Patrick Flynn, Computer Science and Engineering, 384 Fitzpatrick Hall, University of Notre Dame, Notre Dame, IN 46556

Cross-cultural issues

- Capture procedures for some biometrics may touch sensitive cultural issues (such as removing face veils)

- Paternity and familial relationship testing through DNA – how to handle socially sensitive results and how to handle perceived differences in definitions of a family

A DNA test has confirmed Sasha Ruseva a Bulgarian Roma woman is the mother of a mystery girl known as Maria, found living with a couple in Greece

Mother-of-ten Sasha Ruseva, 35, (children, left) claims she did not sell Maria (right), who is now four years old

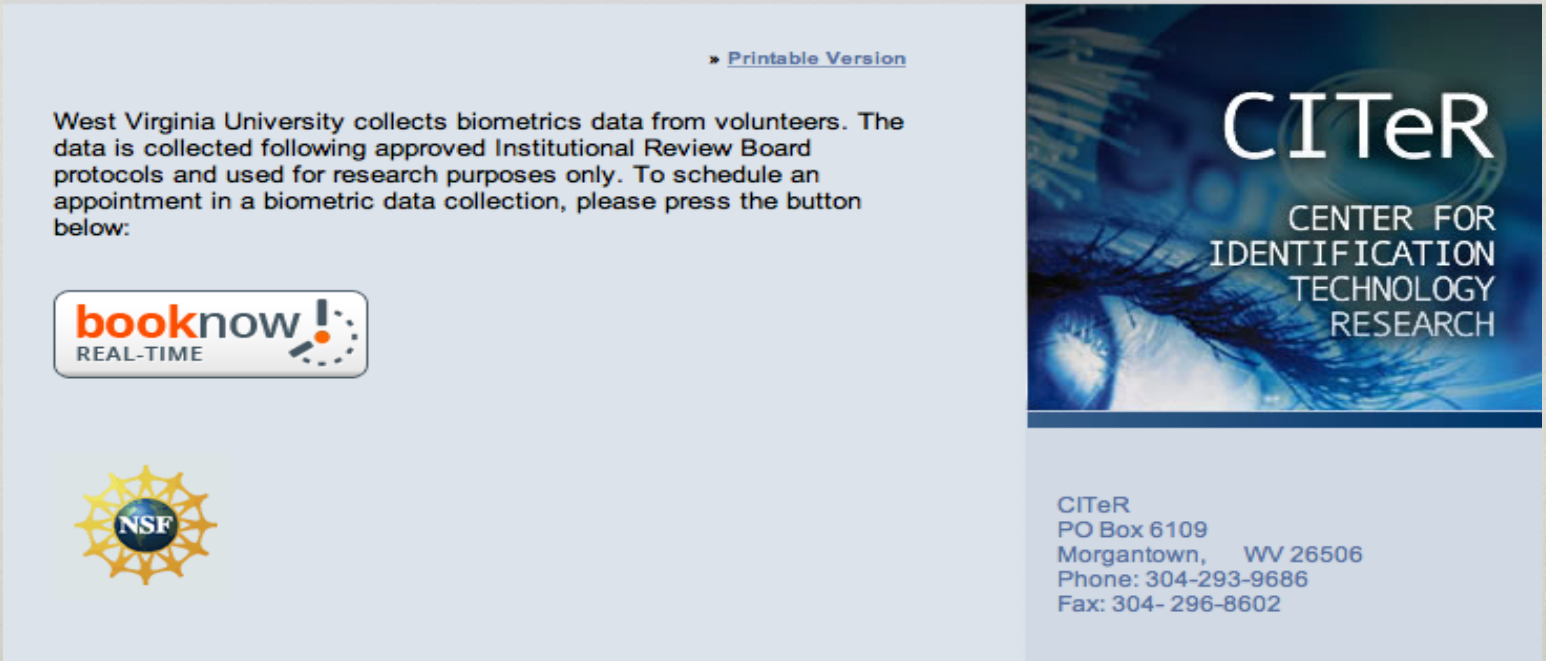
http://www.dailymail.co.uk/news/article-2476772/DNA-test-confirms-Bulgarian-woman-IS-natural-mother-blonde-haired-Maria.htm 1

Cost Considerations

- Standards for how to collect data; amount needed, etc. Such as ISO/IEC 29794 that directly impact costs (number of participants, number of separate visits, etc.)
- Mike Thieme will discuss standards later in this conference
- Collection 'pristine' data for each experiment ….
  - Paying participants and /or collection agents
- Spend time and money to verify existing datasets …
  - Signing use agreements
  - Review of data

Impacts upon Teaching

- Should data be collected as part of the teaching process? (Hear talks Stephanie Schuckers and Jeremy Dawson)
  - How to design experiments
  - How to perform statistical analysis
  - How to deal with regulatory requirements



Publishing study results

- Possible limitations based upon the terms of use of the datasets
  - Can the images (or voice samples, etc.) be made publicly available in journals and other publications?
  - How is reproducibility testing by other organizations to be done?
  - Etc.
- Should data be aggregated for presentation to avoid privacy issues? Etc.

From:
IREX III Supplement I
Failure Analysis

http://www.nist.gov/customcf/get_pdf.cfm?pub_id=910385

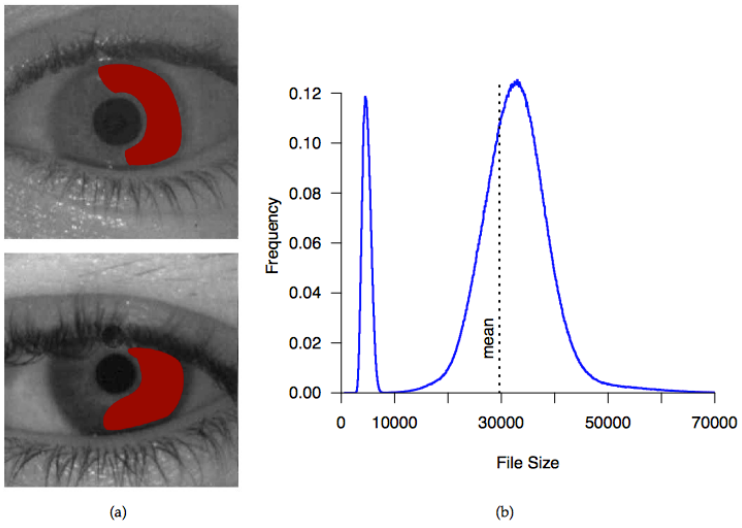
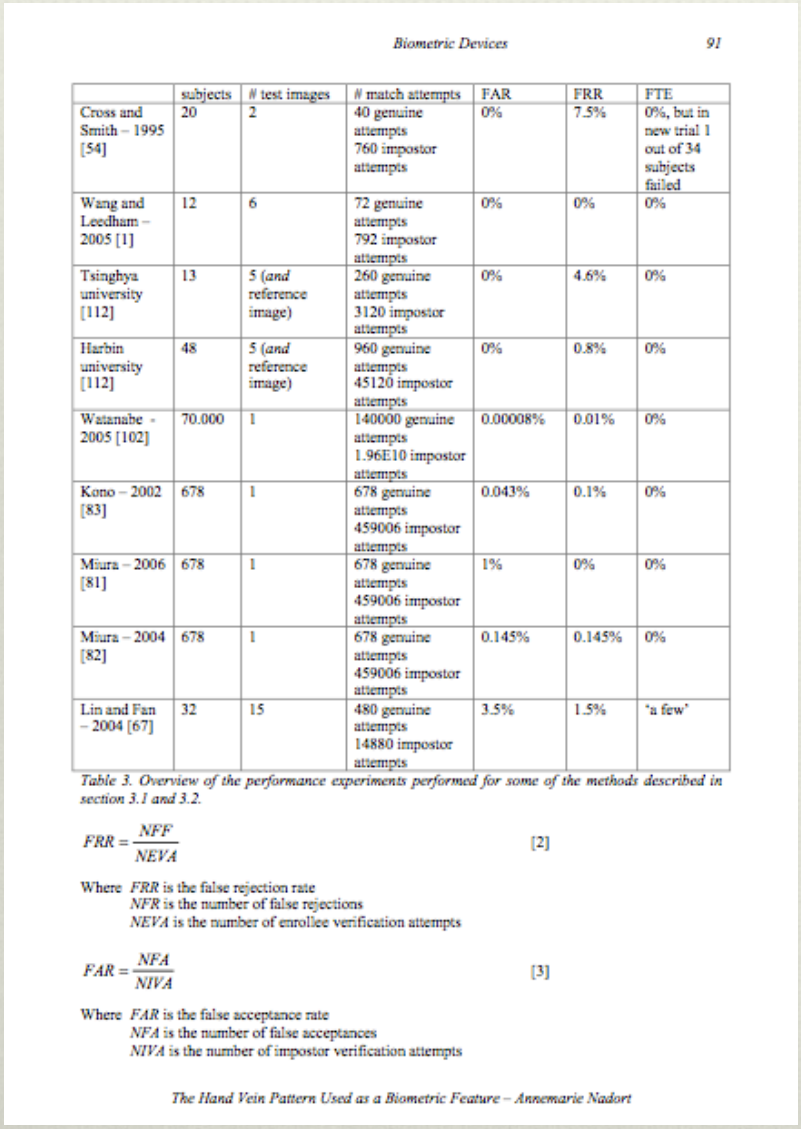


Figure 9: On the left are examples of highly compressed images. The graph on the right shown the distribution of file sizes for the overall dataset.

Company-sponsored data



Biometric Devices 91

	subjects	# test images	# match attempts	FAR	FRR	FTE
Cross and Smith – 1995 [54]	20	2	40 genuine attempts 760 impostor attempts	0%	7.5%	0%, but in new trial 1 out of 34 subjects failed
Wang and Leedham – 2005 [1]	12	6	72 genuine attempts 792 impostor attempts	0%	0%	0%
Tsinghya university [112]	13	5 (*and* reference image)	260 genuine attempts 3120 impostor attempts	0%	4.6%	0%
Harbin university [112]	48	5 (*and* reference image)	960 genuine attempts 45120 impostor attempts	0%	0.8%	0%
Watanabe - 2005 [102]	70.000	1	140000 genuine attempts 1.96E10 impostor attempts	0.00008%	0.01%	0%
Kono – 2002 [83]	678	1	678 genuine attempts 459006 impostor attempts	0.043%	0.1%	0%
Miura – 2006 [81]	678	1	678 genuine attempts 459006 impostor attempts	1%	0%	0%
Miura – 2004 [82]	678	1	678 genuine attempts 459006 impostor attempts	0.145%	0.145%	0%
Lin and Fan – 2004 [67]	32	15	480 genuine attempts 14880 impostor attempts	3.5%	1.5%	'a few'

Table 3. Overview of the performance experiments performed for some of the methods described in section 3.1 and 3.2.

$$FRR = \frac{NFF}{NEVA} \quad [2]$$

Where *FRR* is the false rejection rate
NFR is the number of false rejections
NEVA is the number of enrollee verification attempts

$$FAR = \frac{NFA}{NIVA} \quad [3]$$

Where *FAR* is the false acceptance rate
NFA is the number of false acceptances
NIVA is the number of impostor verification attempts

The Hand Vein Pattern Used as a Biometric Feature – Annemarie Nadort

- Who does the research benefit?
- Why should people participate?
- Who 'owns' rights --- DNA lines, etc.
- Can / should company-sponsored data collection be used by researchers?

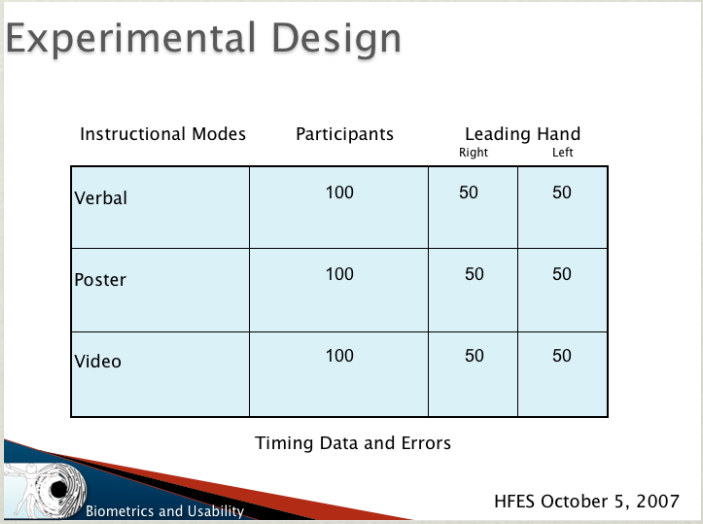
From: *The Hand Vein Pattern Used as a Biometric Feature*
Master Literature Thesis by Annemarie Nadort, vreije Universiteit Amsterdam

Other factors

- When collecting data
  - ergonomics
  - accessibility
  - universality
  - understanding of instructions
- Formatting of data
  - APIs for access, etc.

Those wishing to submit software for MINEX testing shall be required to provide NIST with an SDK (Software Development Kit) library which complies with the API (Application Programmer Interface) specified in this document. At a minimum, the SDK submitted must provide functionality to create MINEX compliant templates based on individual fingerprint images. Support for matching pairs of MINEX compliant templates is encouraged, but optional.

http://www.nist.gov/itl/iad/ig/upload/MINEX_API_041612.pdf



http://zing.ncsl.nist.gov/biousa/docs/HFES%20instructions.ppsx