Collecting Large Biometric Datasets
A Case Study in Applying Software Best Practices

Delores M. Etter, Southern Methodist University
Jennifer Webb, Southern Methodist University
John Howard, Southern Methodist University

Abstract. The lessons and best practices that have become required operating procedure in software development groups can often be applied outside the immediate field of software engineering. This article details a groundbreaking new, multi-year, large-scale biometric dataset that is designed to improve the accuracy and robustness of iris recognition algorithms. We identify several challenges associated with this collection effort and demonstrate how the application of software best practices was able to overcome these obstacles. We believe this list of recommendations represents the current best practices for large scale, long-term biometric collections.

Why Biometrics?

With more than seven billion people now inhabiting our planet, determining an individual's identity has never been more important or more challenging. Biometric algorithms are a form of computer-aided identification that extract and compare various inherent or learned human features. They offer the ability to decipher who someone is, not by what they have, such as an ID card or what they know, such as a password, but by their fundamental intrinsic and behavioral characteristics. Not only are these harder to steal or fake but they also can offer a much lower chance of erroneous identification. For the DoD in particular, which is engaged in international conflicts that can challenge traditional friend-or-foe identification methods, these capabilities are truly transformative.

Iris Biometrics

Iris recognition is a recent technological development that has only become widely utilized in the last decade. First described by Cambridge researchers in the early 1990s, this particular biometric quantizes the intrinsic texture of the human iris in order to automatically determine if two occular images are from the same physical eye [1]. Because individuals with dark or brown irises reflect very little light in the visible spectrum, iris biometric samples are normally collected by sensors that are sensitive to light in the near infrared (NIR) range, which spans from 700 to 900 nm.

Iris recognition algorithms have shown the ability to achieve incredibly low error rates. False match rate (FMR) is the number of times that two different individuals are incorrectly declared to be the same person. False non-match rate (FNMR) is the percentage of trials where a single person appears to not match their own biometric sample, usually requiring the individual to re-submit their test sample. High-quality commercial iris systems can maintain a FMR of one in one million matches while sustaining an FNMR of one in every one thousand attempts [2].

These extremely accurate metrics make iris biometrics one of the few that are appropriate for fully automated population-scale identification programs. Table 1 details some of the large national programs initiated in the last decade. In 2007, the United States military also began utilizing mobile iris biometric technologies. These aptly named devices, known as the Handheld Interagency Identity Detection Equipment (HIIDE) and Secure Electronic Enrollment Kit (SEEK) were deployed to battlefields in both Iraq and Afghanistan to assist with base access, detainee management, local population screening, and special operations missions. By 2009, the Biometrics Identity Management Agency, which executes biometrics initiatives for the DoD, had collected more than 7.5 million iris images in the field [3].

Country	Program Name	Inception	Program Purpose	Estimated Number of Images
India	UID	2009	National ID	1.2 Billion
Indonesia	e-KTP	2012	National ID	170 Million
Mexico	MNID	2010	National ID	100 Million
Middle East (Multiple Countries)	ETS	2004	Immigration Control	50 Million

Table 1 - Population Scale Iris Biometric Programs

Best Practices for Software Development

While software development languages and tools change constantly there are some fundamental principles that have become widely recognized as best practices. At its core, software development encompasses every aspect of product creation. Consequently, best practices in software development can often be seamlessly applied to other technical areas where the goal is the creation of a finished product. This article will demonstrate how four of these concepts, automation, configuration management, documentation and quality control were utilized to address some of the complex problems associated with biometric database construction.

1. A Next Generation Multispectral Iris Biometric Dataset

Motivations

The ability to achieve a FMR of one in every one million matches is truly an impressive statistic. However, the portion of the human population that is enrolled in an iris database is increasing rapidly. Biometric processes must continue to mature so that they can meet this growing demand. This requires development in two key areas:

1. Accuracy – Iris recognition algorithms must continue to demonstrate the ability to reduce false match and non-match error rates in order to support fully automated matching in populations of several million individuals.

2. Robustness – Iris recognition algorithms must continue to sustain performance across increasingly diverse population sets and in increasingly uncontrolled collection conditions.

Recent research has suggested that iris texture changes when illuminated with different wavelengths of light [4], meaning it is possible that several different unique biometric signals can be captured from a single eye (see Figure 1). This discovery has the potential to drive the error rates associated with iris recognition even lower. For example, consider the rare case of two different individuals having matching iris texture in an image captured near 700 nm. By illuminating the two irises with light at some other frequency, it may be feasible to algorithmically determine that the two samples are different, thus avoiding a false-match error.

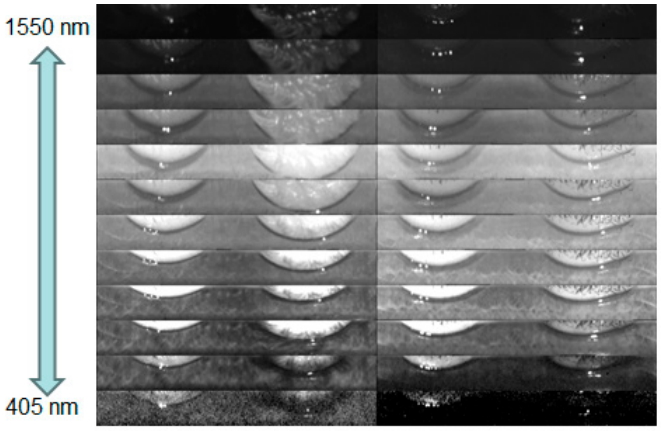


Figure 1 - Unwrapped Iris Texture Illuminated at Different Wavelengths

Approach

In order to stimulate the development of more accurate and robust iris recognition algorithms, a unique data collection was sponsored by the United States government. This collection, known as the Consolidated Multispectral Iris Dataset (CMID), has several notable characteristics that have never been explored in a single biometric collection.

1. Nontraditional Spectrum – Using a custom designed camera assembly (see Figure 2), the CMID captures six images each of the right and left eye across a spectrum that ranges from 400 to 1600 nm. The LEDs used in this experiment have been certified as eye safe by multiple radiation safety experts as well as Institutional Review Boards at both Southern Methodist University (SMU) and the government sponsor. High-resolution visible light images of the ocular region are also taken using a professional photographic camera. Lastly, an image of the left and right iris is acquired using a commercial iris collection device.

2. Duration and Repetition – The CMID collection is in its final (fourth) year with a goal of collecting each subject 16 times over that period.

3. Geographic Separation – The CMID enrolled more than 400 subjects across two geographically separated collection sites in order to increase the diversity of the collected subject pool. Roughly two-thirds of subjects are collected at the SMU research site.

4. Scale – The CMID collects more than 160 iris images per session. The final CMID dataset is expected to contain more than 1 million laboratory quality iris images.

5. Collection of Metadata – In addition to biometric samples, the CMID also captures information about the subjects enrolled in the study such as their gender, eye color, race/ethnicity, and eye health conditions.

6. Manual Segmentation – The first step in all iris recognition algorithms is to use computer vision techniques to separate iris texture from the pupil and sclera. However, these processes may fail on images captured outside the normal 700 to 900 nm spectrum. Consequently, points on the inner and outer iris boundaries are manually identified for each iris image in the CMID.

7. Manual Quality Control – Images in the CMID are also manually categorized into one or more bins based on their quality. These bins denote incidents such as blinks, image blur, and off-axis eye gaze.

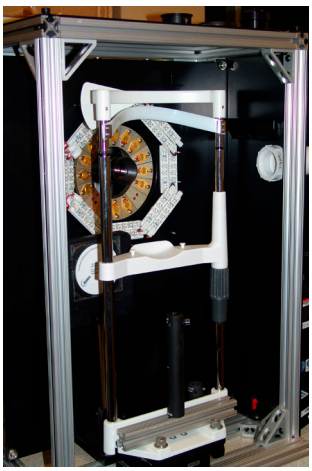


Figure 2 - Consolidated Multispectral Iris Dataset Collection Device

2. Software Best Practices For Iris Database Collection

Executing a first-of-its-kind data collection of this size and with these unique characteristics presented several novel challenges. Without exception these challenges were addressed by applying software development best practices to the biometric data collection methodology. We believe the following represents a list of the current best practices for large-scale multi-year biometric database formulation.

What Can Your Computer Do For You Today?

Automation has long been an enabling technology when developing software. For well-understood tasks, it allows engineers to reduce the possibility of human error throughout the project lifecycle. For example, nightly builds and automated regression testing ensures that this week's code modifications did not break the features added in last week's build. However automation is not synonymous with efficiency. Knowing which tasks to automate and which ones require manual engagement can make the difference between a successful project and one that is underperforming yet over budget.

In a data collection the size of the CMID, automation is a requirement, not simply a desirable feature. Software programs are responsible for nearly everything in the collection process. This includes adjusting the ocular illumination, capturing biometric samples (from all three cameras) and saving the resulting files to the correct location. In order to determine the correct image name, the software must track every variable controlled by the CMID collection (see Table 2). While a small number are entered into a graphical interface by the operator, the majority are ascertained automatically through software processes. Our goal is to prevent a human from ever having to manually save, move, or modify a biometric sample because these operations are prone to error.

Image Specific		Subject Specific	
·	Collection Site	·	Subject Gender
·	Source Camera (MS, commercial, photographic)	·	Subject Ethnicity
		·	Subject Eye Color
·	System Version	Session Specific	
·	Subject Identifier	·	Contacts Worn
·	Left Right or Both Eyes	·	Glasses Worn
·	Active Wavelength	·	Recent Eye Trauma
·	Pupil Control State	·	Recent Lasik Surgery
·	Capture Date	·	Recent Other Eye Surgery
·	Capture Time		

Table 2 - Consolidated Multispectral Iris Dataset Controlled Variables

One crucial aspect of this effort is the ability to automatically recall the anonymous subject identifier when individuals return for repeat collections. To accomplish this, the iris images captured by the commercial camera are run through a recognition algorithm. The result is used to determine the subject's unique identification number. While it may seem limiting to use an iris recognition system as the identification mechanism when conducting an iris data collection, this function is one of the most crucial steps in any academic biometric capture sequence. Associating the wrong number with a set of biometric images can produce a flurry of inaccurate false match and false non-match errors and call into question the validity of the entire collection.

When performing any biometric collection, system designers should rely heavily on software automation. Especially when tasks are highly repetitive and tedious, every available effort should be made to remove this burden from the human operator. Automated file operations and subject identification is guaranteed to reduce labeling errors across the lifetime of a collection project.

Control The System Configuration Or It Will Control You

Version control and configuration management have long been staples of healthy software development organizations. Software such as Subversion or Git can be used to track changes to a codebase as it matures. When bugs are discovered or misguided development paths realized, these applications allow programmers to revert back to previous stable states.

However, these concepts have rarely been applied to the collection of biometric datasets. Given the longevity of the CMID collection, the geographic separation of the two collection sites, and the deep reliance on automation during the collection process, it was highly likely that software modifications would be required as the project progressed. However, different collection software can inadvertently bias a test, making results appear to degrade or improve when in reality only the capture process has been modified. This presents a classic paradox in test methodology; if on day three of a yearlong test, a process improvement is discovered, do you implement the change at the risk of corrupting the data?

To fully document configuration control within the CMID dataset, a tracking number was integrated into the collection software. This identifier holds the date of the last system modification for a particular site that is then tagged into the name of every image collected over the four-year time span. This allows us to account for any changes in image quality or error rate that might arise from modifications to the collection system configuration.

Monitoring the configuration of the capture setup is crucial for ensuring that inevitable system changes do not bias test results. Each individual biometric sample should be tagged with the configuration tracking mechanism and related documentation provided to end users that details what these numbers mean.

The Most Important Part of the Code, Is Not Code

Documentation can often be viewed as a leading indicator of success in a software project. If the developers cannot use technical documentation to clearly communicate what a group of functions is designed to accomplish, what are the odds it will actually achieve its unuttered objectives? If a project manager cannot concisely communicate, through an end user manual, how to operate a program, can we really assume it works at all?

Meaningful documentation takes on new interpretation when conducting a long-term biometric collection. Previous iris datasets have usually produced academic papers that include voluminous specifications on what was collected but leave out the intricate details of how and why. This is possible when the collection period is relatively short and these details can be maintained in the gray matter of a select few individuals who persist with the project throughout its lifecycle. However, when seeking to maintain high-quality capture standards across thousands of individual collections, conducted by dozens of test operators, at test sites across the country, over an extended time period, the documentation will be the single-most crucial point of failure.

For our collection project, the end-user manual has been the single most modified document in our source tree. It was the first file added to our version control system and is the last file edited before a new software release. It contains detailed, click-by-click instructions on how to use the collection system. It not only tells operators how to setup the hardware and run the software, but why each step is important. It is by far the most accessed and crucial file across the entire project. It is also the hardest to find bugs in, requiring the authors and system designers to continually review the assumptions that each tester will make after reading a given step.

When conducting a long-term biometric test do not discount, save for later, or delegate to the intern, the system documentation. Starting this crucial step early and keeping this document up to date can make the difference between success and failure of the database collection.

If You Don't Care About Quality, You Can Meet Any Requirement

When conducting any long-term, highly involved process it is often easy to forget that all results, especially those arrived at with the help of human involvement, are subject to errors. Quality control is a discipline within software engineering that recognizes this inescapable fact and seeks to identify and mitigate errors in a finished software product.

In what may be a first of its kind effort, the CMID attempted to actively incorporate software quality control principles throughout the collection period. However, instead of only

applying these concepts to the finished software product, they were also applied to the deliverables of the CMID; namely the biometric images and the associated metadata.

Three specific quality control measures were taken actively throughout the four-year collection period. The first was to validate that the images being collected by the multispectral capture system would serve their end purpose, namely that they would be appropriate for conducting biometric matches. To satisfy this aim, we actively compared the NIR images collected by the multispectral camera against intra-client samples captured from the commercial iris device. The result of the majority of these operations should be a match. By tracking the rate of non-matches in this subset of images we continually validated that the camera was collecting biometric samples of an appropriate quality.

The second quality control step was also applied to the iris images produced by the collection system. This activity involved identifying the samples that exhibited problematic characteristics, such as blinking, off-axis gaze or motion blur. Tracking these metrics allowed us to actively coach human behaviors on a per-subject basis, which hopefully increases the usability of the dataset. We can also include the categorizations of each image to researchers, allowing them to filter in or out certain classes of imagery, depending on the focus of their analysis.

The final quality control step was designed to validate that the manually chosen points on the inner and outer iris boundaries are accurate representations of these perimeters. As briefly mentioned, every image in the CMID collection is presented to an operator who, with the help of computer software, selects a number of points on the inner and outer iris boundary (see Figure 3). This work is performed by a small team of dedicated staff but is nevertheless very tedious in nature. Consequently, we actively monitor the quality of the segmentations by allowing 1% of the total multispectral imagery to be manually segmented by two or more of the operators. The two different segmentations are compared using an area of overlap metric. By tracking this metric we can not only identify segmentation operators who may need additional training but can also use it to make intelligent estimations as to the overall accuracy of the segmentations across all types of illumination.

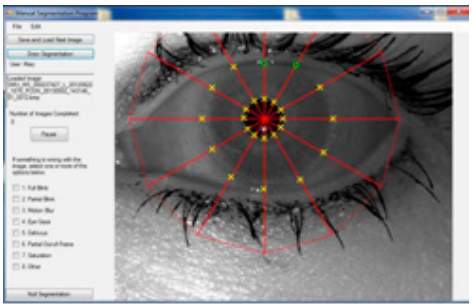


Figure 3 –Manual Segmentation Program.

Actively monitoring the quality of a long-term, large-scale biometric collection is crucial to its eventual success or failure. Simply monitoring raw numbers or gigabytes of data collected, without validating that the samples are well suited for their purpose nearly guarantees disaster. The capture system should be designed around quality control tests (not the other way around) and these tests should produce automated, well-understood metrics that can be tracked by the administrative team. This allows for an understanding of how the test is progressing from a quality standpoint, not simply from a sheer numbers point of view.

3. Conclusions

Software development has a long history of both success and failure. From either case, we learn valuable lessons about the correct way to approach problems, implement solutions and react to the unexpected. It is important to remember that these lessons can often be applied outside the field of software development to assist in other engineering and technical challenges. We have demonstrated how several of these well-established principles have helped resolve some of the complex issues that face research teams when conducting long-term, large-scale biometric collections.

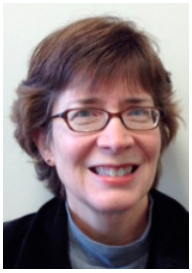
REFERENCES

1. Daugman, John. "Biometric personal identification system based on iris analysis." Patent 5,291,562. 01 March 1994.
2. Daugman, John. "Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons." Proceedings of the IEEE 94.11 (2006): 1927-1935.
3. Quinn, George et al. "IREX IV: Evaluation of Iris Identification Algorithms". NIST Interagency Report 7949 (2013).
4. Boyce, Christopher, et al. "Multispectral iris analysis: A preliminary study51." Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on. IEEE, 2006.

ABOUT THE AUTHORS

Dr. Delores Etter has been the Texas Instruments Distinguished Professor in Engineering Education and the Executive Director of the Caruth Institute for Engineering Education in the Bobby B. Lyle School of Engineering at SMU since 2008. She previously held academic positions at the U.S Naval Academy, the University of Colorado at Boulder, and the University of New Mexico. She also was the Assistant Secretary of the Navy for Research, Development and Acquisition from 2005 to 2007, and was the Deputy Under Secretary of Defense for Science and Technology from 1998 to 2001. She is also a member of the National Academy of Engineering.

E-mail: DEtter@smu.edu



Dr. Jennifer Webb is a senior researcher in Southern Methodist University's Biometrics Lab, where she has been involved with collection and processing of SMU's Multispectral Iris Image data set for the past four years. Prior to SMU, she worked at Texas Instruments with error-resilient video compression and radar systems analysis. She holds a Ph.D. in Digital Signal Processing from the University of Illinois at Urbana-Champaign and a Master's degree in Computing Science from Texas A&M University.

E-mail: WebbJ@smu.edu

John Howard is currently a Ph.D. candidate in the computer science department at Southern Methodist University. His areas of interest are biometrics, pattern recognition and big data analytics. He also works full time as a research scientist, contracting for various groups in the United States Government. He has extensive knowledge in the areas of computer vision, software development, statistical analysis, and distributed computing.

E-mail: JJHoward@smu.edu