Lessons from Collecting a Million Biometric Samples

Patrick J. Flynn Kevin W. Bowyer
University of Notre Dame
Notre Dame, IN 46556, USA
flynn@cse.nd.edu kwb@cse.nd.edu

P. Jonathon Phillips
National Institute of Standards and Technology
Gaithersburg, MD 20899, USA
jonathon@nist.gov

***Abstract*— Over the past decade, independent evaluations have become commonplace in many areas of experimental computer science, including face and gesture recognition. A key attribute of many successful independent evaluations is a curated data set. Desired things associated with these data sets include appropriateness to the experimental design, a corpus size large enough to allow statistically rigorous characterization of results, and the availability of comprehensive metadata that allow inferences to be made on various data set attributes. In this paper, we review a ten-year biometric sampling effort that enabled the creation of several key biometrics challenge problems. We summarize the design and execution of data collections, identify key challenges, and convey some lessons learned.**

I. Introduction

The creation of designed and curated data sets for grand challenges and independent evaluations has been an important driving force behind progress in biometrics over the last two decades. Data sets distributed to the research community foster the development of new algorithms and technologies, and they allow independent evaluations of the state-of-the-art. Data sets also contribute to the identification of future research directions, especially if the data is available to the research community with minimal restrictions on its use. This paper describes a decade-long data collection effort that collected approximately 1 million biometric samples. The biometrics data collected supported seminal grand challenges in still and 3D face recognition and iris recognition that were key steps in fielding face and iris recognition systems.

Much of the collected data is available by license to the research community, and some data set components have been downloaded hundreds of times. The collection team continues to receive many requests per month. These data sets can serve as an "on-ramp" for new research groups entering the field. Since data sets are associated with experimental designs, new algorithms can be evaluated in a context that is well established and considered canonical by the research community. However, research groups can also identify new biometrics problems by examining data sets and experimental results using them.

The focus of this paper is the design and execution of biometric data collection. Our goals are: (i) to provide guidance to prospective collectors; (ii) to highlight the key challenges that arise when planning and executing such an activity; and (iii) to highlight the research advances that resulted from our work.

II. Motivation

Before telling a story of sustained large-scale data collection and curated data set assembly, we provide the motivation for undertaking this task and highlight the key issue of metadata collection and management.

The notion of *replicability* is foundational to almost every field that employs experimental/empirical research. Scholarly communities tend to prize research that is disseminated in papers with a level of detail sufficient to enable replication, with confidence that the follow-on experiments are faithful copies of the original experiment. Conversely, research papers that omit key experimental details or provide vague descriptions that allow multiple interpretations are not as highly prized. We consider the organized collection of data using an explicit plan, followed by comprehensive curation efforts, followed by distribution to the teams conducting experiments, culminating in public release of results and licensed distribution of data, as a gold standard for management of experimental data.

The broad usefulness of a data set is strongly dependent on the amount, type, and quality of metadata accompanying the data itself. In the context of biometrics experiments, the most important item of metadata is the identity tag. Other items of metadata in common use for face recognition include gender, age, appearance characteristics (e.g., hair color, eye color), face pose, expression, light sources in use, and others. Metadata management and, in particular, error detection/correction are key challenges in data collection at scale, as metadata errors are unavoidable in most circumstances.

III. Related Biometric Data Collections

The Face Recognition Technology (FERET) evaluation [35], [32] was the first significant effort in face recognition to distribute a common data set along with an established standard protocol. Since then a variety of data sets, competitions, evaluations, and challenge problems have contributed to the face recognition field. Here we highlight a few.

The Carnegie Mellon University (CMU) Pose Illumination and Expression (PIE) face database [13] was collected in such a way as to support excellent empirical explorations of controlled interactions between illumination and pose. The Extended Multimodal Face Database (XM2VTS) and Banca Databases [1] were each released with associated evaluation protocols and competitions were organized around each [18], [20]. The European BioSecure project represents a

major coordinated effort to advance multi-modal biometrics, including face [24].

The Labeled Faces in the Wild (LFW) dataset consists of images downloaded from the web, along with a website that curates current performance results [17]. The YouTube[1] Faces dataset consists of videos of people collected in the spirit of LFW [38].

A number iris image data sets are available from research groups across the world. The Institute for Automation at the Chinese Academy of Sciences (CASIA) distributes the CASIA-IRIS-V4 data set[2]. The components of this data set include iris-at-a-distance images, handheld sensor images, images acquired using a novel illuminator design, and synthetic iris images.

The University of Beira Interior Iris (UBIRIS) database consists of two distinctive datasets containing noisy images of the iris captured in the visible wavelengths [36]. Smart Sensors, an United Kingdom corporation, distributes a set of near infrared (NIR)-illuminated iris images collected with a high quality sensor[3].

IV. Overview of Data Collection

For convenience, we will discuss our data collection activities as a set of three *epochs*, each of which aligns roughly to a group of U.S. Government sponsored evaluation activities. The first epoch is from 2002 through 2006; the data collected in this period supported the Face Recognition Grand Challenge (FRGC), Face Recognition Vendor Test (FRVT) 2006, and Iris Challenge Evaluations 2005 and 2006 [29],[34],[25]. The second epoch ran from 2007 through 2010, and supported the Multiple Biometrics Grand Challenge (MBGC) [28] and the Multiple Biometric Evaluation (MBE) 2010 [14]. The third epoch ran from 2010 through 2012, and data collected in this period supported the Intelligence Advanced Research Projects Activity's (IARPA) Biometrics Exploitation Science and Technology (BEST) Program; data from the epoch supported the the Point and Shoot Face Recognition Challenge (PaSC) [4], [7].

The decade long effort collected data from nine individual modalities (excluding a few modalities from boutique collections). The overall collection protocol was organized by academic year or semester, during which it changed minimally if at all. During a semester or academic year, the same core biometric modalities were collected. The stability of the core modalities collected resulted in multi-modal biometric collections, including identity-linked multimodal collections.

Subjects were generally allowed to report for data collection once in each week of data collection operations. Each week of acquisition was considered a session of collection, and the biometric samples of a subject collected in his or her weekly appearance is referred to as a subject session. Upon reporting to the collection site, a subject would proceed through a set of stations. At each station, a set of biometric samples were collected. For example, a session could consist of four stations, the first collecting three-dimensional (3D) scans of the face, the second collecting iris images, the third collecting still images of the face in a studio environment, and the fourth station collecting face and body imagery (video and stills) outdoors. The number of stations and the set of biometric samples collected at each station varied over the decade but was fixed within a semester of collection. A set of biometric samples from a person is multi-modal if all the samples were collected in the same session. A collection is multi-modal if all the subject sessions are multi-modal and each subject session has samples of the same modes.

With the exception of one small collection, all still images and videos in the visible spectrum were taken with consumer cameras. The 3D face images consists of both range and texture images and the iris images were collected in the NIR. The nine modalities collected are

- still images of the face,
- still images of the face and body,
- videos of the face and body,
- 3D scans of the face,
- long wave infrared (thermal) imagery of the face,
- iris images collected with a iris sensor designed for cooperative subjects,
- iris images as people walked through a portal or following a walk, stop, and walk protocol (as referred to as iris at a distance),
- still images with a profile view of the face and ear, and
- 3D scans that contained the ear and a profile image of the face.

The three epochs resulted in three large multi-modal collections. The modes in each of the multi-modal collections are

- still face, 3D face, and iris (FRGC/FRVT/ICE),
- still face, video face and body, iris, and iris at a distance (MBGC), and
- still face, still face and body, video face and body, iris, and iris at a distance (BEST).

Acronyms in parentheses identify the U.S. Government efforts that each multi-modal collection supported.

There is one special collection that does not neatly fit into the modality nomenclature, which is the Twins data set. The Twins data set was acquired at the Twins Days Festivals in Twinsburg, Ohio in 2009 and 2010. The twins data set contains both face and iris images.

Face imagery was collected under a diverse set of conditions. These conditions reflect the wide range of potential applications for automated face recognition. We modeled this range of applications by five collection scenarios. In all conditions, both frontal and non-frontal faces were collected. The scenarios are

- still images taken in a studio environment with a digital single lens reflex (DSLR) camera,

[1]The identification of any commercial product or trade name does not imply endorsement or recommendation by NIST.

[2]http://www.cbsr.ia.ac.cn/china/Iris\%20Databases\%20CH.asp

[3]http://www.smartsensors.co.uk/products/iris-database/

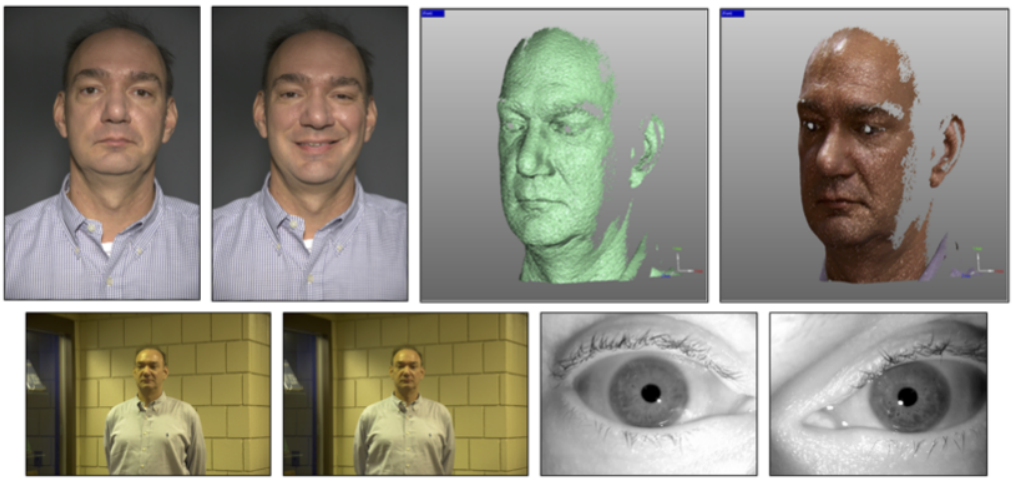

Fig. 1. An example of the types of images used in the FRCG and FRVT 2006 and the ICE 2005 and 2006. The two left frontal images in the top row were taken under controlled illumination with neutral and smiling expressions. The two left images in the bottom row were taken under uncontrolled illumination with neutral and smiling expressions. The two right images in the top row show the shape channel and texture channel pasted on the shape channel for a 3D facial image. The two right images in the bottom row show right and left iris images. All samples are from the multi-biometric dataset. Courtesy of Phillips et al. [34].

- still images taken under ambient lighting in hallways, atriums, and outdoors with a DSLR camera,
- still images taken under ambient lighting in hallways, atriums, and outdoors with handheld digital point and shoot cameras (*e.g.*, cell phones),
- videos taken under ambient lighting in hallways, atriums, and outdoors with a tripod mounted video camera, and
- videos taken under ambient lighting in hallways, atriums, and outdoors with handheld digital point and shoot cameras.

Fig. 1 shows samples used in the FRGC, FRVT 2006, and ICE 2005 and 2006. Fig. 3 and 4 show samples from the MBGC. Fig. 5, 6, 7, and 8 show samples collected in the BEST epoch.

To answer specific questions or investigate special topics, we conducted smaller 'boutique' collections. One strength of the maintaining the large collection infrastructure was that the marginal cost of collecting the boutique data sets was minimal.

Over the decade, we collected and enrolled in our data management system 986,246 samples, which does not include samples from a limited number of boutique collections. Therefore, the actual total number of samples collected was approximately 1 million. Figure 2 breaks out number of biometric samples collected and enrolled in our data management system by academic year. From the start of the effort in 2002 through June 2005, we collected 284,401 samples. At the time this was the largest laboratory data collection activity by over an order of magnitude. From November 2010 through May 2012, 423,587 samples were collected. This data was collected three days a week for 20 weeks. On average 20,648 samples per week, 6,882 per day, were collected. This increase in the number of

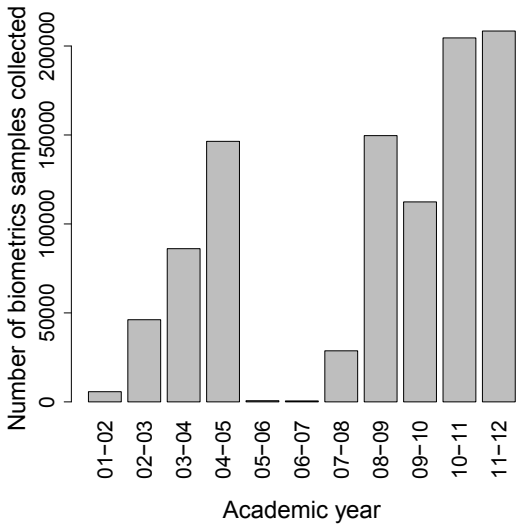

Fig. 2. Number of biometric samples collected broken out by academic year.

biometric samples that could be collected and processed reflected improvements in our collection infrastructure. The infrastructure included management of hardware, software, and laboratory personnel.

The data set consists of biometric samples from 3,145 subjects. The demographics of the subjects are females 49% and males 51%; the self declared race is Caucasian 77%, Asian 14%, and other or unknown 9%. The majority of the subjects were undergraduate students.

One of the overarching design goals of the data collection effort was measuring the impact of time lapse between

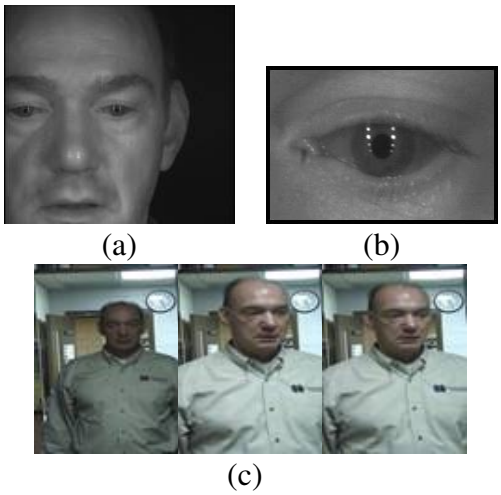

Fig. 3. Example of imagery collected as a subject walks through the portal. The image in (a) is a full 2,000 by 2,000 pixel NIR frame acquired by the IoM and (b) is the left ocular region from the frame in (a). There are approximately 120 pixels across the iris in (a) and (b). The images in (c) are three frames from an high definition video sequence of a subject walking through the portal. Courtesy of Phillips et al. [28].

Fig. 4. Example frames from video sequences. The image in (a) is from a video sequence of a subject walking towards the cameras in an atrium and (b) is a subject performing a non-frontal activity outdoors.

acquisitions. Because of the continual change in the student population, there are constraints on the size of longitudinal studies readily supported [2], [10], [11]. The longest time lapse between a subject's first and last acquisition was 3666 days (10 years). The mean time lapse was 266 days and the median was 87 days.

V. Purpose for Data Collection Activities

The data collection effort was not a monolithic activity directed to a single long-term goal. The motivation and goals of the collection activities changed over time. The changes were driven by multiple factors.

- As research in biometrics continues, performance improves, performance goals increase, and the amount of data needed to estimate performance changes.
- As research in biometrics continues, knowledge of the conditions where techniques perform poorly becomes more detailed, sophisticated, and nuanced. As a result, data collection plans can begin to target challenging conditions revealed by new experiments.
- Few research groups had any experience in large-scale biometric data collection in 2001, especially involving a large number of human subjects. As collection activity continued, a base of experience and mature supporting infrastructure was developed and led to a number of efficiencies in later years.
- Improvements in sensing, storage, distribution, and computation technology as well as process improvements allowed data collection scale to increase over time without a corresponding scale-up in personnel and cost.

The start-up or scale-up of a data collection activity followed a pattern that was executed many times over the years. The initial collection effort mounted in support of

Fig. 5. Examples of images in the PaSC taken during four sessions. Note that locations were varied between sessions, while sensor, distance to camera and pose were varied within sessions. Courtesy of Beveridge et al. [4].

Fig. 6. Cropped face images extracted from still images in the PaSC. These images demonstrate some of the complications that arise in point-and-shoot images, lighting, motion blur and poor focus. Courtesy of Beveridge et al. [4].

Fig. 7. Four snapshots from one video showing a subject carrying out an action, in this case blowing bubbles. Courtesy of Beveridge et al. [4].

Fig. 8. Sampled portions of video frames from PaSC videos indicating some of the situations that make recognition challenging. Courtesy of Beveridge et al. [7].

the FRGC and FRVT 2006 specified the collection of high-resolution facial images from 200 subjects once per week for an academic year, with repeat visits by the same subjects encouraged. The corresponding research study enabled by this collection was investigation of short-term time lapse effects on face recognition. Subsequent sponsor interest in 3D face scans and iris images triggered a sensor/vendor selection process, a protocol modification, and a pilot collection focused on sensor usability and post-collection data management workflow, to provide a base of experience for subsequent large-scale collections with the new sensors. By defining a process for collection operations changes, the amount and types of data collected, and types of sensors used, and the sophistication and capability of our data management system matured fairly smoothly over the years.

VI. Organizing the Data Collection

Our team collected, processed, ground-truthed, and prepared for distribution an average of 100,000 biometric samples per year. Since the daily average collection load increased over the years of collection in response to programmatic commitments, we evolved the key resources needed to operate this collection activity without being overwhelmed by complexity and scale at its inception. These key resources include the following items.

At all times, there was a key person responsible for planning the collection, scheduling the collection events, procuring necessary sensors and supplies, delegating collection tasks, and verifying that post-collection activities were done, including the final delivery of data and packaging of data sets for distribution. A characteristic property of this role is the need for a person with strong organizational skills including management of many details and unanticipated matters. The key person in this role, and the exact scope of the role changed over the years: initially, one of the Notre Dame principal investigators (PI) was the organizer. For a few years, a talented graduate student performed some of the tactical components of the role while a PI was responsible for more strategic elements of the impending collection. During the 2008-2012 collection interval, a broadly skilled staff member was placed in charge of data collections, overseeing everything except initial collection design. This staff member was aided by a second staff member for the final two years of the project; this second person oversaw the ingestion of collected data into the Biometric Research Grid (BXGrid) system [8]. Our experience is that the choice of a good collection manager and the specification of very detailed and clear plans is the key to success of sustained large-scale data collection.

For all years of our large-scale collection activity, the collection venues were staffed primarily by students, and usually undergraduates. There were many reasons for this initial design decision and our experiences have amply justified the choice. Undergraduates readily absorbed the necessary task-specific training (camera operation, use of the BXGrid system for ground truth recording, data subsetting for specific research tasks, etc.). We were able to employ some of our most talented undergraduates for multiple years in roles with increasing amounts of responsibility, providing continuity to operations. Some of these students initiated original research projects and some of those led to publications in the refereed literature.

Although the majority of our operators were undergraduates, we did assign some collection activity to graduate students from 2002 through 2008, considering it one of the mandatory duties of membership in the PI's research group. However, as collection activity scaled up in its final years, we were confident in our ability to recruit a large number (approximately fifty) of undergraduates to assume all sensor operator roles, and graduate student labor was not needed except in exceptional circumstances.

Data management infrastructure, both software and hardware, is critical. Accurate collection of large data sets at scale is only possible if there are systems that support and facilitate the collection and curation workflow. The sophistication of such systems scales in some sense with the size and complexity of the data corpus. This was a hard-won lesson for the collection team. The initial days of collection employed a simplistic manual data management system with metadata stored in simple text files along with data organized in directories named for the collection date. An National Science Foundation (NSF) funded research project in large-scale data management systems led to the development

of BXGrid [8], a database-backed and web-enabled data ingestion and management portal with redundant file storage affording robustness to disk and server failures. BXGrid is a key part of our ongoing research, as it allows subsets of our data corpus to be selected using database queries, which facilitates construction of targeted data sets for experiments.

One key to success in the large-scale data collections was conducting pilot studies and having a process for adding new sensors. Prior to starting large data collections, it is usually necessary to have a pilot data collection. The first stage in adding a new sensor is to conduct a pilot study to understand the sensor and the data it collects. After the initial pilot study, the sensor was integrated into an ongoing large-scale collection activity.

Collecting large amounts of data purely for the sake of collecting data will likely lead to wasted effort and resources. Design of a data collection needs to be motivated by a goal, or limited number of goals. Further, the goals need to be articulated in the experiment design.

Anecdotal evidence suggests that an initial raw labeling error rate of around 1 in 3,000 can occur, and that incorporating an explicit data curation stage can reduce the labeling error rate to below 1 in 25,000. This was possible in our collection for four reasons. The first two concern the ability to detect suspected errors. The large number of samples per subject in each mode and results from multiple algorithms made it easy to detect suspected labeling errors. Third, there were researchers that examined every single sample and provided feedback when suspected errors were found. Fourth, the audit trail in the acquisition process made it possible to retrospectively confirm suspected errors.

In the U.S., collecting biometric samples for research needs review by an Institutional Review Board (IRB) for human subjects approval. There are issues beyond human subjects that include legal, ethical, copyright, and institutional risk. In evaluating these issues, it is good to remember the phrase "Just because it is legal, does not mean it is a good idea." In addition, human subjects, legal, and ethical standards vary by country. Before using data for an experiment, one needs to consider the following questions: Was the data collected with appropriate human subjects approval? Is the data allowed to be distributed? Is the use of the data consistent with the human subjects approval and consent form?

VII. Accomplishments

A. Grand Challenges and Evaluations

The key novel accomplishments of the FRCG, FRVT 2006 and ICE 2005 and 2006 are:

- One key goal of the FRGC was an order-of-magnitude decrease in the error rate on frontal still face images taken under controlled illumination conditions over performance reported in the FRVT 2002 [30]. The FRVT 2006 documented that this goal was achieved [34].
- The FRGC and FRVT 2006 established the first independent performance benchmarks for 3D face recognition technology.
- The ICE 2005 and 2006 were the first grand challenge and independent evaluation for iris recognition matching technology.
- The FRVT 2006 and the ICE 2006 are the first technology evaluations that compared iris recognition, high-resolution still frontal face recognition, and 3D face recognition performance.
- The FRCG and FRVT 2006 were the first competitions that systematically compared human and machine face recognition performance.
- Results from the FRVT 2006 formed the bases for the Good, Bad, and Ugly challenge problem [27].

The goal of the MBGC was to improve the performance of face and iris recognition technology from biometric samples acquired under unconstrained conditions. The MBGC is organized into three challenge problems. Each challenge problem relaxes the acquisition constraints in different directions. The Portal Challenge focused on iris recognition on the move. The goal of the Still Face Challenge was to improve accuracy from frontal and off angle still face images taken in ambient lighting indoors and outdoors. In the Video Challenge, the goal was to recognize people from video in unconstrained environments.

The data collected under the BEST program was the basis for the Point and Shoot Face Recognition Challenge (PaSC) [4]. To spur advancement in face and person recognition the PaSC focuses on still images and video taken with handheld digital point and shoot cameras. The challenge includes 9,376 still images of 293 people balanced with respect to distance to the camera, alternative sensors, frontal versus not-frontal views, and varying location. There are 2,802 videos for 265 people: a subset of the 293. The PaSC was the bases for the Handheld Video Face and Person Recognition Competition held in conjunction with the International Joint Conference on Biometrics (IJCB) 2014 [7], and the Video Person Recognition Evaluation held in conjunction with the 11th IEEE International Conference on Automatic Face and Gesture Recognition [6].

B. Scientific Knowledge and Technical Advancement

The large and diverse data collection enabled scientific investigation into fundamental properties of the biometrics. Below is a sampling of scientific discoveries that the data collections supported.

There are fundamental variations in face appearance in long-wave infrared (LWIR) over time [9]. These variations are as prominent as region A being brighter than region B in an image taken at one time, but region A being darker than region B in another image of the same face taken at another time.

The twins data collected allowed for both face and iris recognition studies. For faces, when images are taken in mobile studios in the same collection session, it is possible to distinguish twins; however, when face images are taken a year apart, it is an extremely challenging problem [23]. Twins do have similarity of iris texture, but it is not a similarity that

is seen in matching iris codes. This is one way in which iris codes are not the same as the texture of the iris [16].

Iris template aging studies that compare, for the same set of subjects, the false non-match rate for short-time-lapse matches versus the false non-match rate (FNMR) for long-time-lapse matches, find an increased FNMR with additional elapsed time [2],[10]. This basic result has since been reported by other research groups analyzing other datasets.

A change in pupil dilation between two images of the same eye results in an increase in the false non-match rate [15]. This result has since been replicated by various research groups.

Wearing contact lenses effects the performance of iris recognition algorithms [3],[39]. For clear cosmetic contact lenses there is a small increase in the FNMR. Algorithms to detect the presence of cosmetic or textured lenses can be specific to the lens brand.

In forensic comparison of iris images, humans can match iris images with substantial accuracy, though not the same accuracy on average and automated algorithms [19].

Starting with the FRGC, comparing human and algorithm performance has been systematically included into challenges and evaluations. These studies show that for frontal face images taken with DSLRs, algorithm performance is superior [33]. Human performance is superior when recognition requires fusing all identity cues present in an image or video [37], [22]. In addition, fusing human and algorithm matching score improves performance [21]. The observation that algorithms developed in by research groups in Asia are better at recognizing Asian faces and algorithms developed by groups in the West are better at recognizing Caucasian faces has practical implications [31].

Quantifying the effect of factors, covariates, and quality measures on algorithm performance is essential for understanding biometric and face recognition algorithms. A series of covariate analyses of face recognition algorithms showed that simple measures can characterize algorithm performance on a data set [5],[12]. Unfortunately, characterizations do not generalize to new algorithms and data sets. Phillips et al. [26] developed a technique for quantifying the existence of and the best case effectiveness of quality measures.

VIII. Conclusion

By collecting a designed and curated data set of 1 million samples, we have enabled the advancement of both the technology and science of biometrics. The resulting understanding provides a solid basis for decisions on when and how to field biometric systems. The resulting datasets continue to be heavily used in research on still, video and 3D face recognition, iris recognition, and gait recognition.

IX. Acknowledgements

PJF and KWB received support from the following sources for 2002-2012 data collection activities: Defense Advanced Research Project Agency (DARPA), Air Force Office of Scientific Research (AFOSR), National Science Foundation (NSF), Technical Support Working Group (TSWG), Federal Bureau of Investigation (FBI), Intelligence Advanced Research Project Activity (IARPA), Army Research Laboratory (ARL), National Institute of Justice (NIJ), and the United States Government. PJP received support from the FBI and IARPA.

References

[1] E. Bailly-Bailliére and et al. The BANCA database and evaluation protocol. In *4th International Conference on Audio- and Video-based Biometric Person Authentication*, pages 625–638, 2003.

[2] S. Baker, K. W. Bowyer, P. J. Flynn, and P. J. Phillips. Template aging in iris biometrics: Evidence of increased false reject rate in ice 2006. In M. Burge and K. W. Bowyer, editors, *Handbook of Iris Recognition*. Springer-Verlag, New York, NY, USA, 2013.

[3] S. Baker, A. Hentz, K. W. Bowyer, and P. J. Flynn. Degradation of iris recognition performance due to non-cosmetic prescription contact lenses. *Computer Vision and Image Understanding*, 14:1030–1044, 2010.

[4] J. Beveridge, P. Phillips, D. Bolme, B. Draper, G. Givens, Y. M. Lui, M. Teli, H. Zhang, W. Scruggs, K. Bowyer, P. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pages 1–8, Sept 2013.

[5] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, and Y. M. Lui. Focus on quality, predicting FRVT 2006 performance. In *Proceeding of the Eighth International Conference on Automatic Face and Gesture Recognition*, 2008.

[6] J. R. Beveridge, H. Zhang, B. A. Draper, P. J. Flynn, Z. Feng, P. Huber, J. Kittler, Z. Huang, S. Li, Y. Li, M. Kan, R. Wang, S. Shan, X. Chen, H. Li, G. Hua, V. Štruc, J. Križaj C. Ding, D. Tao, and P. J. Phillips. Report on the FG 2015 video person recognition evaluation. In *Proceedings Eleventh IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

[7] J. R. Beveridge, H. Zhang, P. Flynn, Y. Lee, V. E. Liong, J. Lu, M. Angeloni, T. Pereira, H. Li, G. Hua, V. Struc, J. Krizaj, and P. J. Phillips. The IJCB 2014 PaSC video face and person recognition competition. In *Proceedings of the International Joint Conference on BIometrics*, 2014.

[8] H. Bui, M. Kelly, C. Lyon, M. Pasquier, P. J. F. D. Thomas, and D. Thain. Experience with BXGrid: a data repository and computing grid for biometrics research. *Cluster Computing*, 12(4):373–386, 2008.

[9] X. Chen, P. J. Flynn, and K. W. Bowyer. Infra-red and visible-light face recognition. *Computer Vision and Image Understanding*, 99:332–358, 2005.

[10] S. P. Fenker, E. Ortiz, and K. W. Bowyer. Template aging phenomenon in iris recognition. *IEEE Access*, 1(266-274), 2013.

[11] P. J. Flynn, K. W. Bowyer, and P. J. Phillips. Assessment of time dependency in face recognition: An initial study. In *Audio-and Video-Based Biometric Person Authentication*, pages 44–51, 2003.

[12] G. H. Givens, J. R. Beveridge, P. J. Phillips, B. A. Draper, Y. M. Lui, and D. S. Bolme. Introduction to face recognition and evaluation of algorithm performance. *Computational Statistics and Data Analysis*, 67:236–247, 2013.

[13] R. Gross, S. Baker, I. Matthews, and T. Kanade. Face recognition across pose and illumination. In S. Z. Li and A. K. Jain, editors, *Handbook of Face Recognition*, pages 193–216. Springer-Verlag, June 2004.

[14] P. J. Grother, G. W. Quinn, and P. J. Phillips. MBE 2010: Report on the evaluation of 2D still-image face recognition algorithms. NISTIR 7709, National Institute of Standards and Technology, 2010.

[15] K. Hollingsworth, K. W. Bowyer, and P. J. Flynn. Pupil dilation degrades iris biometric performance. *Computer Vision and Image Understanding*, 113(1):150–157, 2009.

[16] K. Hollingsworth, K. W. Bowyer, S. Lagree, S. P. Fenker, and P. J. .Flynn. Genetically identical irises have texture similarity that is not detected by iris biometrics. *Computer Vision and Image Understanding*, 115:1493–1502, 2011.

[17] G. Huang, M. Ramesh, T. Berg, and E. Learned Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report Tech. Rep. 07-49, University of Massachusetts at Amherst, October 2007.

[18] J. Matas and et al. Comparison of face verification results on the XM2VTS database. In *Proceedings of the International Conference*

on Pattern Recognition, volume 4, pages 4858 – 4853, Barcelona, Spain, September 2000.

[19] K. McGinn, S. Tarin, and K. W. Bowyer. Identity verification using iris images: Performance of human examiners. In *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS 13)*, 2013.

[20] K. Messer and et al. Face authentication test on the BANCA database. In *Proceedings of the International Conference on Pattern Recognition*, volume 4, pages 523–532, 2004.

[21] A. O'Toole, H. Abdi, F. Jiang, and P. J. Phillips. Fusing face recognition algorithms and humans. *IEEE Trans. on Systems, Man and Cybernetics Part B*, 37:1149–1155, 2007.

[22] A. J. O'Toole, P. J. Phillips, S. Weimer, D. A. Roark, J. Ayyad, R. Barwick, and J. Dunlop. Recognizing people from dynamic and stable faces and bodies: Dissecting identity with a fusion approach. *Vision Research*, 51:74–83, 2011.

[23] J. Paone, P. Flynn, P. J. Phillips, K. Bowyer, R. Vorder Bruegge, P. Grother, G. Quinn, M. Pruitt, and J. Grant. Double trouble: Differentiating identical twins by face recognition. *IEEE Transactions on Information Forensics and Security*, 9:285 – 295, 2014.

[24] D. Petrovska-Delacretaz, G. Chollet, and B. Dorizzi. *Guide to Biometric Reference Systems and Performance Evaluation*. Springer, Dordrecht, 2009.

[25] P. Phillips, K. W. Bowyer, P. J. Flynn, X. Liu, and W. T. Scruggs. The Iris Challenge Evaluation 2005. In *Second IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2008.

[26] P. J. Phillips, J. R. Beveridge, D. S. Bolme, B. A. Draper, G. H. Given, Y. M. Lui, S. Cheng, M. N. Teli, and H. Zhang. On the existence of face quality measures. In *IEEE Biometrics: Theory, Applications and Systems (BTAS)*, 2013.

[27] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. S. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer. An introduction to the good, the bad, and the ugly face recognition challenge problem. In *Proceedings Ninth IEEE International Conference on Automatic Face and Gesture Recognition*, 2011.

[28] P. J. Phillips, P. J. Flynn, J. R. Beveridge, W. T. Scruggs, A. J. O'Toole, D. Bolme, K. W. Bowyer, B. A. Draper, G. H. Givens, Y. M. Lui, H. Sahibzada, J. A. Scallan III, and S. Weimer. Overview of the Multiple Biometrics Grand Challenge. In *Proceedings Third IAPR International Conference on Biometrics*, 2009.

[29] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 947–954, 2005.

[30] P. J. Phillips, P. J. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and J. M. Bone. Face recognition vendor test 2002: Evaluation report. Technical Report NISTIR 6965, National Institute of Standards and Technology, 2003. http://www.frvt.org.

[31] P. J. Phillips, F. Jiang, A. Narvekar, and A. J. O'Toole. An other-race effect for face recognition algorithms. *ACM Trans. Applied Perception*, 8(2), 2011.

[32] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. PAMI*, 22:1090–1104, October 2000.

[33] P. J. Phillips and A. J. O'Toole. Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, 32(74-85), 2014.

[34] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale results. *IEEE Trans. PAMI*, 32(5):831–846, 2010.

[35] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing Journal*, 16(5):295–306, 1998.

[36] H. Proença and L. A. Alexandre. UBIRIS: a noisy iris image database. In *13th International Conference on Image Analysis and Processing*, pages 970–977, 2005.

[37] A. Rice, P. J. Phillips, V. Natu, X. An, and A. J. O'Toole. Unaware person recognition from the body when face identification fails. *Psychological Science*, (in press).

[38] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 529–534, 2011.

[39] D. Yadav, N. Kohli, J. S. Doyle, R. Singh, M. Vatsa, and K. W. Bowyer. Unraveling the effect of textured contact lenses on iris recognition. *IEEE Transactions on Information Forensics and Security*, 9:851–862, 2014.