NISTIR 8361

Best Practices in the Collection and Use of Biometric and Forensic Datasets

R. Austin Hicklin
George Kiebuzinski
Melissa Taylor

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8361

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 8361

Best Practices in the Collection and Use of Biometric and Forensic Datasets

R. Austin Hicklin
George Kiebuzinski
Noblis

Melissa Taylor
National Institute of Standards and Technology
Special Programs Office

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8361

March 2021

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology

National Institute of Standards and Technology Interagency or Internal Report 8361
Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8361, 38 pages (March 2021)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8361

Abstract

This document discusses issues that arise in the collection, dissemination, and use of biometric and forensic science datasets for research purposes and provides recommendations on how to best address the issues raised. A variety of factors should be considered when collecting data so that the resulting dataset is fit for purpose, or when deciding whether an existing dataset is appropriate for a given purpose. This document does not address privacy and human subjects aspects of data collection. These topics will be addressed in forthcoming publications.

Contents

1 Introduction

The collection and use of datasets is necessary for biometric and forensic science research, the development of new technology, the evaluation and validation of operational systems, and informed policy decisions. The complexity and expense of collecting biometric and forensic datasets make it necessary that the process be well understood and executed.

Biometric and forensic datasets have a lifecycle that passes through three stages: ***collection***, ***dissemination***, and ***use***. This document discusses some of the issues that can arise through this lifecycle, with examples of when the collection or use of such datasets have gone very wrong — and recommends best practices that should help avoid such pitfalls.

This document also includes two appendices, providing specific examples of methods of collecting datasets.

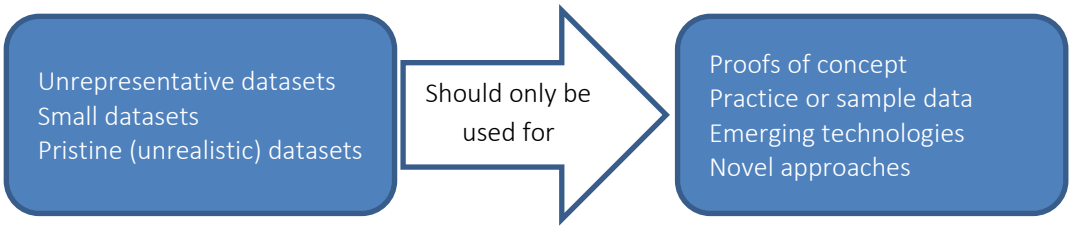
1.1 Appropriateness for use

Many of the problems that can be encountered in the collection and use of biometric and forensic datasets have to do with whether a given dataset is appropriate for a given use ("fit for purpose"). A dataset could be considered appropriate for a given use if it can address the needs of that use in terms of representativeness (of subject population and data attributes, as discussed in Section 2.3), in availability and accuracy of metadata (e.g. if the documentation and ancillary information are adequate to understand the data, and are sufficient for analysis), and size (number of samples).

Although every dataset is (presumably) appropriate for some use, not all datasets are appropriate for every use. Appropriateness needs to be considered at every stage in the lifecycle:

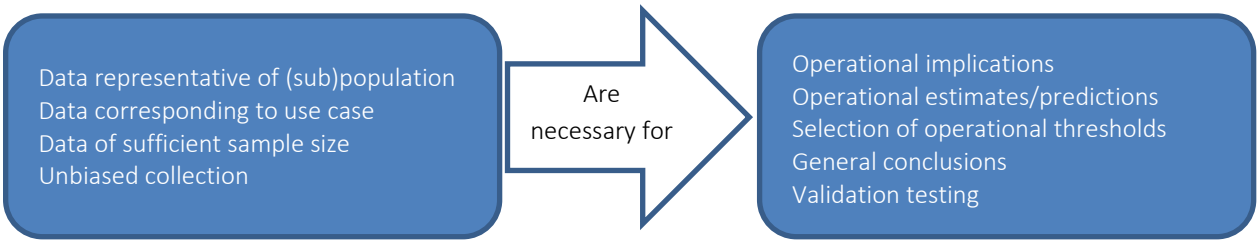
- ***Collection*** **—** Potential uses should be considered when planning for data collection as well as during collection itself.
- ***Dissemination*** **—** Documentation needs to be clear so that future users will be able to assess appropriateness —especially for uses that may not have been foreseen at the time of collection.
- ***Use*** **—** When using existing datasets, it is necessary to determine whether they are fit for purpose, or how to make the best use of available data.

Some uses have few implications regarding the representativeness or generalizability of datasets, and therefore almost any datasets may be appropriate:



Note, however, that unrealistic datasets should be used with caution even in the early stages of research and development to avoid making invalid assumptions that could bias the direction of the effort.

However, only some datasets are appropriate for making general conclusions, operational decisions, or validation:



1.2 Categorizing datasets by intended use

Datasets can be used for a wide variety of purposes, which can be grouped in these categories:

- ***General-use datasets*** — Datasets purported to be representative of a large population (or of all people), with generic or universal characteristics, for a variety of (unspecified) purposes. The main difficulty here is that attempting to be simultaneously representative of multiple demographic traits, multiple data attributes, and multiple use cases is very challenging: see Section 2.3.
- ***Specific-use datasets*** — Datasets collected to be representative of a specific population and/or representative of a specific use case. Specific-use datasets are invaluable as long as their use corresponds to the intended purpose. However, if they are reused for a different purpose, they are at risk of being misused by overgeneralizing the results.
- ***Datasets selected to be challenging*** — Datasets composed of samples that are chosen to be difficult. These can be used to stress-test human examiners or algorithms or to provide a basis for differentiation among participants in a test where more representative data would lead to near-perfect results.* These datasets are of particular risk of being misused by overgeneralizing the results: while specific-use datasets are intended to be representative of a defined population, datasets selected to be challenging can only be representative if the selection process is modelled after a specific use case. For example, if a fingerprint dataset includes "close non-mates" by selecting the first candidate returned by an Automated Fingerprint Identification System (AFIS) from every search, it can be considered both a specific-use dataset (representative of that use case) and a dataset selected to be challenging. If instead that fingerprint dataset only includes the subset of candidates returned by the AFIS that are deemed particularly difficult, representativeness is more problematic: such a dataset may be appropriately used for differentiation, but rates or other results should not be considered as representative of a broader population.
- ***Datasets of convenience*** — Datasets that contain whatever samples are available, without a planned collection. These datasets are particularly susceptible to collection bias (Section 2.4), as well as overgeneralization of results.

* *Duane Blackburn explained this as "the evaluation itself must not be too easy, nor too difficult. If the evaluation is too easy, all the systems will perform well and will group together at one end of the capabilities spectrum. If the evaluation is too difficult, none of the systems will perform well and will group together at the other end of the spectrum. In either case, the evaluation will fail to produce results that will enable you to accurately distinguish one system from another." [1]*

Best Practice #1: *Dataset documentation should indicate the original intended use(s) of the dataset at the time of collection. Intended uses should be explicitly defined, such as proof of concept, operational performance, or performance differentiators with respect to a specific operational requirement.*

Best Practice #2: *Specific-use datasets, datasets selected to be challenging, and datasets of convenience should never be assumed to be broadly representative.*

1.3 Reproducibility and generalization of results

The scientific community has recognized reproducibility to be a challenge: a variety of scientific disciplines have expressed concern regarding research that has not been reproduced (e.g., [2,3,4]). The reproducibility of research is interwoven with the data used. There are different aspects associated with the reproducibility (or replicability)* of research:

- Can the results be corroborated using the ***same*** data and the ***same*** processes or systems? (evaluation of intra-process variability, also known as repeatability)
- Can the results be corroborated using the ***same*** data and ***different*** processes or systems? (evaluation of inter-process variability)
- Can the results be corroborated using ***different*** data? (evaluation of whether the results are generalizable)

Using public datasets (or making datasets public after a study is complete) allows for the first and second (same data) types of corroboration. Public datasets do not directly address the third type of corroboration (different data), but they do assist in understanding the initial study in detail; for example, if a researcher publishes results using a public dataset, subsequent studies can use that dataset to baseline their results so that they know their procedures are in line with those of the initial study, in addition to performing analyses using new data. However, for most evaluations and research, generalizable results are the ultimate purpose: we want conclusions that we are confident apply to the specified population, and therefore corroboration using different data is very important.

Part of the reproducibility challenge has to do with overstating the conclusions of studies. The generalizability of a study is directly tied to how representative the dataset is of the target population and data attributes. If a study is based on a small, biased, or otherwise unrepresentative dataset, making broad assertions of the importance and general applicability of the results is unwarranted.† In some disciplines, the overall population is not well understood, or is highly heterogeneous; in such instances, researchers need to be careful not to overstate conclusions. Confidence intervals are often misused in this way: confidence intervals describe the relation between a dataset and the broader population of which it is representative; however, they say nothing about the relation between an unrepresentative dataset and the broader population.

Biometric and forensic disciplines vary in how well overall data distributions are defined. For example, exemplar fingerprints and controlled-collection face images that follow the ANSI/NIST-

* *The terms "reproducibility" and "replicability" are used by some to differentiate between these two aspects (corroboration using same data vs corroboration using different data) — unfortunately, without agreeing on which term goes with which definition. See [5]*

† *For some scenarios of particular concern, there is validity in existence proofs, which assess whether a given scenario is possible; if found to be possible, subsequent study would have to assess its prevalence.*

ITL standard [6] constrain a variety of factors that could affect sample quality, and therefore are much more homogeneous than latent prints and uncontrolled face images that do not constrain such factors.

1.4 What can go wrong: examples of problems

It is wise to be a cautious collector — and skeptical consumer — of biometric and forensic datasets. Poorly collected or misused datasets can result in ineffective technology, wasted effort, inappropriate research conclusions or policy decisions, or unrealistic expectations regarding policy or technology. The types of problems discussed— and their recommendations —generally break down into these categories:

- ***Collection***: Problems include ad hoc or uncontrolled data collection, biased procedures, inadequate quality control, and removing or filtering data. These can be addressed through planning, quality assurance procedures, and documentation.
- ***Dissemination***: Inadequate documentation means that the consumer cannot fully understand the data and limitations. This can be addressed by verifying that datasets and any revisions are accompanied by complete documentation. Standardized ways of characterizing datasets (such as taxonomy and metadata) may be of benefit.
- ***Use***: Not all datasets are appropriate for a given use, and therefore problems result from naïvely (or deliberately) treating available datasets as representative or overstating conclusions. These can be addressed by making certain that datasets are not misused or misrepresented, and that conclusions are not overstated.

Here are some examples of how collection or use of datasets can go very wrong:[*]

Study Design	Issues
In an AFIS (Automated Fingerprint Identification System) study, months of analysis were conducted before finding that about 15% of the images had been removed because a fingerprint examiner thought "they would be difficult to match." The information was in a readme file that was not bundled with the dataset.	Biased data collection Biased filtering of data Inadequate documentation
AFIS evaluations that used datasets where the "ground truth" associations between fingerprints had been made by another AFIS. This approach omits all data that the initial AFIS could not match (such "AFIS bias" means that the results are not testing the accuracy of the AFIS being tested, but whether they duplicate the results of the initial AFIS). Similarly, forensic tests conducted in which a single human examiner provided the "correct" answers used to assess other examiners.	Biased data collection Biased data attribution
A dataset collected by sampling operational workflow on a single specific day, without realizing that there was significant day-to-day variation in the types of data and users of the system, which meant that the sample was not broadly representative.	Biased data collection Assuming representative data

[*] *We are deliberately not specifying which projects or studies these examples come from: our purpose is not to criticize past work, but to provide examples of problems to avoid in the future.*

Iris data was collected from a very limited population (i.e., engineering grad students at a specific university), but gender/age/ethnicity information was not retained; subsequent evaluations drew conclusions as if the data were representative of people in general.	Assuming representative data Overstated conclusions Inadequate documentation
Evaluations in which a single sample is selected for each subject but subsequently evaluated as if there were no within-class variation (i.e., the evaluation was conducted as if there is no variation in samples from a subject). Variations of this include picking an ideal or characteristic sample for each subject, or identification system evaluations that assume there would be no within-subject variations in matcher score.	Assuming representative data Lack of intra-subject variation Overstated conclusions
Fingerprint collection under unusual environmental conditions (dust, heat, and humidity) that resulted in poor-quality data that was then pooled with data from other sites.	Unrepresentative collection problems
Small, unrepresentative datasets used as the basis for operational technology thresholds, policy decisions, or overstated conclusions in reports and journal publications.	Assuming representative data Overstated conclusions
Image artifacts caused by inappropriate use of compression. Examples: taking compressed fingerprint images, then rotating, cropping, and recompressing; overcompressing fingerprint or face images by using inappropriate WSQ* or JPEG settings.	Unrepresentative collection problems
Face recognition algorithms that showed an "other-race effect": some algorithms from east Asia were more effective at differentiating Asian faces than algorithms from the U.S., and vice versa, due to the racial proportions used in algorithm training.	Assuming representative data
Extensive time in a study was spent on statistical measures of confidence for results from an available dataset that was not representative of any population, even though subsequent dataset collections would be outside those bounds.	Assuming representative data Misuse of statistics
Face data collected from a large pool of images by using a face detection algorithm that filtered out off-angle images, images without two eyes visible, and images with poor lighting.	Biased data collection
Datasets collected without Institutional Review Board (IRB) approval, therefore never releasable.	Inadequate planning Legal issues
The understanding of fingerprints used in systems engineering for an AFIS was gained through experience with criminal data, which led to inaccurate assumptions when designing systems for use with civil fingerprints (due to different age, gender, and data quality distributions).	Assuming representative data

* *Wavelet scalar quantization fingerprint compression algorithm*

"Controlled" datasets collected without quality control procedures resulted in a large proportion of source attribution errors.	Inadequate planning Inadequate quality assurance (QA) procedures
Data collection that required humans to type or write long identifiers, or identifiers that have easily confused characters (such as "0" (zero) and "O" (capital O), or "1" (one) and "l" (lower-case L) and "I" (capital I)), resulting in administrative errors in source attribution (i.e., erroneous "ground truth" information).	Inadequate planning Inadequate QA

2 Collection

Datasets can differ in the procedures used in collection, and in the populations and attributes of which they are (or are attempting to be) representative. This section discusses the issues related to dataset collection and makes recommendations.

Collection of personally identifiable data from human subjects is considered Human Subjects Research and requires Institutional Review Board (IRB) approval. The forthcoming "Beginners Guide to Biometric and Forensic Human Subjects Research Protections" [*] will summarize requirements and best practices in this area.

2.1 Types of collection: controlled, operational, manipulated, and synthetic data

The samples included in datasets can be collected in different ways:

- ***Controlled collection*** **—** Datasets can be composed of samples collected under controlled conditions. These laboratory-collected datasets have the advantage of definitive (ground-truth) source attribution (Section 2.2), and allow for control and measurement of a variety of factors and collection of metadata that might be impossible if using operational data. However, such data may be unrepresentative, often lacking the heterogeneity of operational data. Collection of large datasets under controlled conditions can be expensive.
- ***Operational data collection*** **—** Datasets can be collected from existing operational databases, or from sampling operational workflows. Such datasets can have the advantage of being representative and very large — but have the drawback that determining source attribution with a high degree of certainty for such data is difficult without biasing the resulting data (Section 2.2). In addition, the data being collected may be subject to variations caused by differences in operational procedures, use of new or different capture devices, or changes in operational data capture policies. Note that data from one system/agency/process should not be assumed to be representative of data from other systems/agencies/processes.[†]
- ***Environmental collection*** **—** For some uses, when source attribution is not necessary, datasets can be composed of samples of naturally handled items (e.g., taken from trash bins, recycling bins, etc.)

[*] *In preparation at time of publication.*

[†] *For example, NIST's FpVTE, SDK, and PFT studies reported significantly different results for operational datasets collected from different agencies. [7-11]*

Best Practice #3: *Controlled collection should be used when source attribution with a high degree of certainty is critical, when control of collection procedures is critical, or when measurements or metadata are needed that are impractical when collecting operational data.*

Best Practice #4: *Operational data collection should be used when representativeness or quantity of data are critical. Source attribution of operational data must be carefully addressed to avoid selection bias.*

When datasets of adequate size are not available, artificial samples can be created either by manipulating existing data or by generating synthetic data:

- ***Manipulated data*** **—** Available biometric data can be manipulated to make up for inadequate or unavailable data. For example, some evaluations will degrade exemplar fingerprint images in an attempt to resemble latent prints, by adding noise, distortion, or cropping.
- ***Synthetic data*** **—** Some software programs can generate synthetic data, such as SFinGe [12], which creates artificial fingerprint images. (This differs from manipulated data in that synthetic data originates from an automated source; manipulated data modifies actual biometric data.)

Manipulated or synthetic data can be used to rapidly generate large datasets, which may be appropriate for some uses, such as the early stages of algorithm development, or if other datasets are not available. For example, when only a small representative dataset is available, artificial data may be used during a proof of concept or early stages of algorithm development so that the limited real data can be used where it is most useful. Manipulated or synthetic data also can be used to create datasets that isolate and control particular attributes [13]: for example, manipulated data can be used to evaluate what the effect of image contrast is on accuracy. However, such datasets are not recommended for general use because they cannot be assumed to be representative of real-world data. In addition, the assumptions made in the algorithms or manipulations may result in unrealistic artifacts: for example, synthetic fingerprint algorithms may create impossible pattern classes, and distortion methods used to create faux-latent prints may be unrelated to actual distortion.

Best Practice #5: *Manipulated or synthetic data should be used cautiously and only when other alternatives are not available. Manipulated or synthetic data should never be assumed to be representative and should not be used to generalize results.*

2.2 Source attribution (ground truth)

Source attribution is the process in which the source of each sample is determined. Many or most uses of biometric and forensic data require knowing which samples are mated (from the same source, such as fingerprints from the same subject and finger) or nonmated (from different sources, such as faces of different people). Although "source attribution" is often used as synonymous with "ground truth," the term "ground truth" should only be used when the source of each sample is definitively known. There are two issues to consider regarding source attribution: whether the source attribution process biases which samples are selected, and the level of confidence of source attribution. The process of determining subject attribution to a high degree of confidence may introduce a selection bias into the data.

Controlled datasets allow for ground truth source attribution, if reasonable quality assurance procedures are in place — but depending on the operational use case, it can be difficult to collect controlled data that is truly representative of operational data.

For operational datasets, source attribution can be problematic for two reasons: operational source attribution is generally conditioned on the determinations of systems or forensic examiners and therefore is less than definitive and should not be considered ground truth; and because the source attribution process can act as a filter excluding some potential samples, so that the resulting dataset is no longer representative. If source attribution for an operational dataset is made by human examiners or automated systems, the resulting dataset is limited by the capabilities of those specific examiners or matchers:

- Although human latent print examiners are highly accurate, they can make errors, may make inconclusive decisions, do not always reproduce each other's conclusions, and do not always repeat their own conclusions. Datasets with source attribution by a single examiner will reflect those issues.
- "AFIS bias" refers to using an AFIS to determine subject attribution between fingerprints. This means that the resulting dataset only includes those fingerprints that were successfully matched by that particular AFIS, making the resulting dataset useless for future evaluations of AFIS (since the dataset is biased in favor of matchable high-quality prints in general and that a specific vendor's AFIS in particular).

The potential biasing effect of source attribution can be minimized by using additional case information beyond the specific samples in question. For example, in the often-used NIST Special Database 27 [14], the fingerprints used were selected from cases in which multiple additional corroborating latents and exemplars were used in the identification, allowing for the verification of the associations between latents and exemplars in the dataset.

For datasets that will be used to evaluate systems with extremely low error rates (such as 10-print fingerprint systems), the source attribution process is particularly important because any administrative errors would prevent accurate measurement. For example, in NIST's *Studies of Biometric Fusion* [15] the underlying dataset had a 0.05% rate of data integrity problems (mislabeled samples, generally fingerprints labelled with the wrong finger position), which limited the ability to measure false rejection rates.

Best Practice #6: *The method used to determine source attribution ("ground truth") needs to be documented at the time of collection.*

Best Practice #7: *For samples not collected under controlled conditions, the source attribution method should be based on additional data rather than solely on the samples used in the dataset.*

2.3 Representativeness

The extent to which results from a dataset can be generalized depends on the representativeness of the dataset. In this context, representativeness has two aspects:

- ***Representativeness of subjects and sources*** — the extent to which the demographics of the dataset correspond to the population of interest. The relevance of demographic factors varies by modality. For example, race, sex, and age have obvious effects on facial recognition, but are irrelevant to firearms evidence. Some modalities have a more subtle interaction with

demographics: for example, for decades most work on fingerprints centered on criminals, and so generalizations about fingerprints were disproportionately influenced by male subjects (generally in their twenties through forties) and less likely to consider elderly and female subjects. Demographic factors to be considered include age, race, sex, height, weight, and occupation. Related factors that would apply for face or iris recognition may include eyeglasses or hair style. For domains with non-human sources, the classes of the sources need to be considered (e.g., manufacturer for firearm evidence).

- ***Representativeness of data attributes*** — the extent to which attributes such as type, content, and quality correspond between the dataset and the use case of interest. The relevant attributes of samples may be subcategorized:
  - Use case/process attributes include collection processes, devices, and procedures
  - Format attributes include image size, compression methods, and file format
  - Content attributes include type, features, and data quality (lossy compression could be considered both a format and content attribute)
  - Other attributes include the ratio of mated to nonmated data.

Best Practice #8: *The demographic attributes of the population included in the dataset must be considered when the collection process is designed and must be documented. Ideally, factors such as sex, age, and race should be indicated for each subject; when this is not possible due to privacy or anonymity requirements, the overall distributions of these factors should be documented for the dataset.*

Best Practice #9: *Data attributes of the samples in the dataset must be considered when the collection process is designed and must be documented.*

If collection of a dataset is for a specific use case with a clearly defined population, procedures, and data attributes, standard sampling methods can be used. *(Note it is beyond the scope of this document to attempt to address statistical sampling methods. We recommend consulting statisticians during experimental design.)*

Best Practice #10: *If a dataset is limited to or focuses on a specific population or use case, that information needs to be documented at the time of collection.*

If a dataset is being collected for general or multiple purposes, or for uses that are not well defined, the best approach is to try to define as well as possible the target population, procedures, and data attributes for a general use case — and document these as completely as possible. In such cases when it is not practical for a dataset to be truly representative, the collectors should at a minimum strive to make the dataset "realistic" in that the content and distributions of attributes could reasonably be expected in the types of use cases being considered.

Attempting to be fully representative of the combinations of multiple factors can be difficult or intractable due to the "curse of dimensionality": for example, trying to be representative of all the proportional combinations of age, sex, and race could require an extremely large dataset. It may be practical to consider factors separately by partitioning the dataset (or including metadata so the user can partition analyses): for example, for a dataset that includes fingerprints collected using a variety of livescan devices, pooling all the different devices into a single dataset implicitly assumes the relative proportions of each type, while partitioning instead allows the user to consider them separately.

2.4 Sampling bias

Data collection often perturbs the representativeness of data by disproportionately selecting some types of data (selection bias) or including only those samples that survived some process (survivorship bias). By implicitly or deliberately filtering out some of the data, the resulting dataset is skewed. The impact of sampling bias can be serious: for example, if the selection process omits most of the samples likely to fail an evaluation, the results of the evaluation will be unrealistically positive. Biased data collection can be insidious because it may not be detectable using quality metrics or other attributes of the images.

Best Practice #11: *Convenience samples and datasets of opportunity should be assumed to be non-representative (due to selection or survivorship bias).*

Best Practice #12: *The data collection process should be planned to avoid inadvertent filtering of data, disproportionate selection of some categories of data, or overweighting of some attributes.*

Best Practice #13: *If any samples that were part of the collection are deliberately removed, it is imperative to document the reasons for removal and the number of samples removed. It is preferable to include all of the data in the dataset and document known problems.*

Frequently, collection processes can bias a dataset by explicitly or accidentally filtering out the poorest-quality data. Every use has some category of unusable data: for biometric systems, unusable samples may be considered *Failure To Acquire* (FTA) or *Failure To Enroll* (FTE); for human examiners these may be considered *No Value* (NV). Although it may seem intuitive that unusable data should be omitted from a dataset, removing such data is problematic because there is significant system variation in determining which samples are considered FTA or FTE, and examiners vary in determining which samples are NV. If the worst-quality samples are removed, the resulting dataset cannot be used to evaluate the ability to deal with poor-quality data.

A key determinant of the difficulty of a dataset is the proportion of poor-quality or difficult samples included: omitting poor-quality samples biases the dataset by making it unrealistically easy, but an excessive proportion of poor-quality samples may make it unrealistically difficult. Ideally it would be desirable to know the proportion of poor-quality data in actual use, in which case the dataset can mirror that distribution. However, such distributions are often unknown. In such cases, one alternative is to allow separate analyses of the poor-quality data, by partitioning the dataset or by including metadata so that the users of the dataset can do analyses conditioned on quality.

Best Practice #14: *Samples should not be removed from the dataset just because they are poor quality. For datasets purporting to be representative, poor-quality images should be included to the extent that they occur in actual data.*

2.5 Quality control for controlled dataset collection

The primary advantage of controlled datasets is that they allow for definitive ground truth source attribution but only if reasonable quality assurance procedures are used to avoid administrative errors. Effective methods for avoiding administrative errors are generally based on planning: designing the collection process to be as foolproof as possible. For example, the collection process should limit the possibility of transcription mistakes by using preprinted labels or barcodes for

each subject. In particular, the process should be designed to avoid having humans writing or typing long numbers, or identifiers that include easily confused characters (such as "0" (zero) and "O" (capital O), or "1" (one) and "l" (lower-case L) and "I" (capital I)).

If the quality control process is not well designed, the pursuit of quality can actually make things worse. In the name of quality control, some datasets have had poor-quality or difficult data explicitly or implicitly removed, biasing the resulting data.

Best Practice #15: *The data collection process should be planned in advance to minimize the possibility of human error.*

3 Documentation, dissemination, and maintenance

The primary role of documentation and dissemination is making sure that any information relevant for potential users of a dataset is retained with the dataset.

3.1 Documentation

The documentation for a dataset should accurately describe and characterize the dataset and how it was collected. Documentation should include

- Citation
  - Dataset name and version
  - Citation requirements for use
- Description of intended use(s)
- Description of subject population
  - Population of interest (if any)
  - How subjects were selected
- Description of data collection process
  - How samples were collected
  - Whether samples were selected from a larger pool of samples
  - Whether any samples were removed (and why)
  - Whether any lossy formatting was performed (e.g., image transformations, lossy compression, cropping — see Section 3.2)
- Terms of use
  - Data use agreement
  - Conditions or limitation of use
  - Permitted uses
- Human subjects paperwork
  - Consent forms
  - Institutional Review Board (IRB) approval
- Metadata
  - Subject/source attributes (e.g., gender, ethnicity, age) — preferably for each sample, but if that is not permitted for privacy reasons, show the distributions for the dataset as a whole
  - Data attributes (e.g., collection devices or methods)
  - Environment (e.g., date of capture, indoor/outdoor, lighting condition, temperature)
  - Cross-reference tables indicating subject/source associations (i.e., which files are associated with the same people)

The dataset and documentation should be bundled to keep them together.

3.2 Data formatting

The dataset should be distributed in standard formats that are straightforward for the users to process. Some data transformations are "lossy" in that the resulting samples do not include all the information in the original samples; examples include scanning of photographs, cropping, JPEG or wavelet-based compression, or color space transformations (e.g., saving color images as grayscale). To the extent that such lossy processes are representative of a given use case, these may be acceptable, but must be documented. Standard practices should be followed to avoid processes that can lead to artifacts, such as over-compression or recompression of samples.

> **Best Practice #16:** *Any data transformation that irreversibly changes the data must be documented. The original (untransformed) data should be retained and made accessible to the users of the dataset.*

3.3 Dissemination: NIST Catalog

Publicly available datasets should be submitted for inclusion in the NIST *Biometric and Forensic Research Dataset Catalog*,[*] which will serve as a compendium of publicly available biometric and forensic datasets for finger/palmprints, iris, face, person at a distance, voice, and handwriting developed by the National Institute of Standards and Technology (NIST) in collaboration with National Institute of Justice (NIJ). The Catalog was developed to provide a compilation of publicly available databases and their associated metadata enabling researchers to search for data relevant to their research uses.

The Catalog implements a taxonomy for the classification of biometric and forensic datasets, which classifies datasets using two major dimensions, modality and data type. Modality classifies each dataset by the type(s) of biometric or forensic data (friction ridge, face, iris, person at a distance, voice, handwriting), as well as detailed subtypes for each modality. Data type classifies each dataset by the format(s) of image, video, audio, or other data files. In addition to the major dimensions, a dataset can be classified by several miscellaneous data characteristics, including dataset acquisition method, multimodal data, dataset size, source attribution method, availability of demographic information, longitudinal/multiple event data, capture method, subject cooperation, and flags for unusual attributes such as synthetic, or spoof, data, or post-mortem data.

3.4 Dissemination: Public vs sequestered datasets

Public datasets can be invaluable for research, as reference data, and for use in "public challenge" evaluations.[†] However, for competitive evaluations, public datasets are not appropriate: the data used for testing and evaluation must be collected independently and not known to the participants, so that the participants cannot take advantage of prior knowledge of the data. Public datasets can become overused: the small number of public datasets means that many researchers are using and reusing the same small number of samples, and therefore any peculiarities of those datasets have an outsized effect (NIST SD-27 latent fingerprint dataset, for example).

[*] *In the process of implementation at time of publication.*
[†] *A public challenge is a practice evaluation: an open-book proof of concept test on public data, not for substantive analysis.*

3.5 Maintenance

Datasets require long-term maintenance so that corrections and additions can be tracked. Creators of datasets should consider regular additions as new data becomes available from expanded source materials. Datasets should clearly indicate version, at a minimum by labeling the dataset and documentation with the date of creation or modification. The user should not be left with the task of differentiating among different unlabeled versions of a dataset. Updates should clearly include detailed documentation of changes made for each version. Digital object identifiers (DOIs) are strongly recommended as persistent identifiers. Checksums may be used to verify data integrity.

Best Practice #17: *Datasets should clearly indicate a version number or release date, and the version must be updated when any changes are made.*

4 Usage

The datasets we are discussing can be used for a wide variety of purposes:

- The development of automated algorithms requires data for research, development, training, and testing;
- The implementation of operational systems requires data for systems engineering, validation, and acceptance testing;
- The development and implementation of new or revised operational policies and procedures requires data for research, training, and validation testing;
- The evaluation of algorithms and systems can be comparative (with data selected to differentiate among contenders) or predictive (with data selected to estimate operational performance);
- Human examiners require data for training, individual proficiency testing, and broad-based evaluation of examiner performance.

Some uses have very general or unspecified use cases in which the subject population and attributes of the data are unknown, whereas others have detailed use cases in which the population and/or attributes of the data are precisely specified.

Given the scarcity of biometric and forensic datasets, it is unrealistic for the collectors of datasets to assume that future users will be constrained in how they use the datasets. Limited dataset availability means that people may use whatever datasets are available. The absence of documentation is likely to increase uncertainty as to the appropriateness for use.

As we discussed in Section 1.1, effective use of a dataset hinges on assessing appropriateness: determining whether existing datasets are fit for purpose. The types of data that are appropriate depend on the implications of the results:

- If the results are to be used for ***general conclusions*** about the overall discipline, the data should be as broadly representative as possible because the implication is that conclusions apply to a global population.
- If the results are to be used for making policy decisions, setting operational thresholds, or operational acceptance testing for a ***specific system***, the data should be representative of the specific population and use case being considered.

- If the results are to be used for ***evaluation*** (of systems, human experts, or validation of procedures) the type of data needed depends on whether the purpose of the evaluation is to measure or compare among alternatives.
  - An evaluation where the purpose is to ***measure*** performance requires data representative of the specific population and use case being considered: the intent is to evaluate what expected performance would be, even if there is no difference among the alternatives being evaluated.
  - Conversely, an evaluation where the purpose is to ***compare*** among alternatives requires datasets selected to be challenging: the intent is to force differentiation among the alternatives by stress testing, with results not intended to predict performance.
  - Biometric evaluations can be categorized as technology, scenario, or operational evaluations [16], of which technology evaluations are the most relevant to this discussion. Technology evaluations use pre-collected data to evaluate algorithms, as opposed to scenario and operational evaluations, which collect the data as part of the evaluation.
  - For acceptance testing in which existing systems or processes are to be replaced, the same dataset must be used on both the old and new systems in order to directly compare the performance of the old and new systems.
- For ***hypothesis testing***, the dataset needs to have metadata for each sample to allow partitioning on the specific factors being considered. Effective hypothesis testing generally requires that the specific factor can be isolated and not confounded with other factors, which is not always possible given datasets collected for other purposes. For example, a test of different types of livescan fingerprint equipment should have identical or equivalent operational procedures and environments for all such devices, so that any differences in results could be attributed solely to the type of device; a dataset collected for purposes other than evaluating livescan equipment is likely to have other associated factors, so that differences could be due to procedures or environment rather than livescan type.[*] Similarly, datasets used for testing of alternative capture procedures used with livescan equipment should be defined and documented.
- For ***research and development***, unrepresentative, small, or pristine datasets may be appropriate for some uses, such as in developing emerging technologies and novel approaches (where the purpose is to get the new technology to simply work), proofs of concept, or practice or sample data. However, as research matures, it becomes increasingly important to use realistic data so that the full range of types of data can be accommodated: use of unrealistic datasets may result in invalid assumptions or expectations.

When trying to make the best use of available data, it is imperative that the study is transparent regarding the source of the data and the extent to which the data may or may not be representative.

As discussed in Section 1.3, samples collected in uncontrolled or unstandardized conditions (i.e., most forensic and some biometric data) can be affected by a wide variety of factors. When the overall population is very heterogeneous, the representativeness of any single dataset may be debatable. For example, the attributes of latent prints vary significantly among agencies because of different crime types, so that an agency dealing with financial crimes would disproportionately have latents on paper while another agency would disproportionately have latents on bomb components; the agencies would have little overlap in the types of latents encountered, and therefore no single dataset would be representative of both. In such cases it may be wise to use

[*] *This was seen in FpVTE2003 [7], when the highest-quality livescan devices showed some of the poorest performance (and vice versa), due to differences in collection procedures.*

multiple datasets in a single study and report the results for each (as was done in NIST's FpVTE [7]), or partition results by metadata. This is better statistical practice than pooling widely variable data together: reporting an average of such data is sometimes derided as a "meaningless mean."

Appendix A Scenario-based collection of forensic samples as part of ordinary office activity

This Appendix describes a scenario approach to creating and collecting multiple modalities of forensic data in an ordinary office environment. A pilot test of the scenario should be performed so that the potential problems and technical issues may be identified prior to the full scenario data collection, updating this document with any lessons learned. Institutional Review Board (IRB) approval must be completed prior to collection of samples from human subjects, with informed consent received from each subject.

Contents

A.1 Objectives

The purpose is both

- to collect datasets of forensic samples representative of those left in ordinary office activities, and
- to assess the frequency with which usable forensic samples are left in ordinary office activities.

Forensic modalities to be collected include:

- Friction ridge (latent fingerprint and palmprint)
- Face
- Handwriting and handprinting
- Voice
- DNA
- (The scenario can easily be extended to include other modalities, such as digital signature or keyboard dynamics)

A.2 Data to be collected

Collect data samples with ground-truth source attribution for the following biometrics.

A.2.1 Exemplar samples

Exemplar samples — data collected under controlled conditions and used as a reference for identification purposes.

- Exemplar finger- and palmprints
  - Livescan
  - Inked
- Exemplar face images (passport type)
- Handwriting and handprinting samples
- Voice recordings
- DNA (swab)

A.2.2 Forensic samples

Forensic samples — data collected to simulate crime scene forensic data.

- Latent fingerprints deposited on a variety of office objects
  - Desk
  - Paper (both 8.5”x11” and index card; the latter requires using a second hand to hold in place when writing, the former allows getting full writer’s palm if the subject needs to write at the top of the paper.)
  - Keyboard
  - Mouse
  - Desk phone
  - Cell phone
  - Soft drink can
  - Desk surface
  - Pen or pencil

Note: Samples should be collected using each type of object two times for each subject. Materials should be cleaned between subjects. If the goal is to collect samples with impressions from multiple people, then objects should be prehandled by a tester. (This adds to the complexity that a specific latent impression cannot be definitively attributed to a single subject.)

- Fingerprints from cell phone fingerprint reader
- Facial images
  - Computer video camera

  - Wall-mounted video camera (surveillance)
  - Cell phone selfie
- Handwriting
  - Cursive
  - Printed
  - Signature
  - Indented impressions in pad of paper
- Voice
  - Desk phone
  - Cell phone
  - Computer microphone
  - Room surveillance microphone
  - Higher-quality microphone for exemplars at enrollment
  - Voice samples may include disguised voice as well as natural voice
- DNA
  - Saliva deposited on soft drink can
- Miscellaneous
  - Keyboard dynamics
  - Digital signature

The data would be collected using a scripted scenario (described below) for approximately 100 subjects.

A.3 Testing Approach

The test will simulate ordinary office tasks. A scripted set of actions will be followed by the test subject (role player) to provide a consistent data capture methodology.

A.3.1 Subject Preparation

Each subject will provide exemplar data for all the data modalities.

- The subject will be photographed in a controlled environment wherein the photograph is compliant with US passport data capture standards.
- The subject will be fingerprinted using an FBI-approved livescan to capture a full fingerprint and palm exemplar record and will have fingerprints and major case prints (palmprints and lower joints of fingers) collected using the traditional inking method.
- The subject will provide a buccal swab or saliva sample for DNA analysis.
- A half page of handwritten and handprinted text will be generated. The sample will include a signature and a transcribed text sample (same text for all subjects).
- The subject will have their voice recorded using both an oral interview and a reading exercise.

Each subject will be briefed as to the scenario process prior to entry into the test environment.

Each subject will be required to fill out the required privacy and informed consent forms.

A.3.2 Test Environment

The test will be conducted in an office environment that has been prepared as follows:

- Two desks will be set up (Desk A and B)
  - Each desk will have a computer (with a keyboard, mouse, and video camera), desk phone, cell phone, paper documents, a pad of lined paper, a ballpoint pen, and a pencil.
  - Both desks and objects will be cleaned and sanitized prior to the subject's entry. However, Desk B then will have all objects handled by a specific known tester (with ID retained with the data).
  - Inexpensive keyboards (now ~$4) can be disposed of after use (e.g., after superglue is used to develop fingerprints).
  - Each computer will have video cameras to capture the subject's facial image during portions of the test.
- A surveillance video camera(s) will be mounted to capture all subject physical interactions within the office setup. The surveillance camera will be used to verify that the subject has touched all the required objects and otherwise performed in accordance with the test instructions. The surveillance camera will also be used to provide facial images for facial identification.
- A surveillance microphone will be mounted in the room. All sounds will be captured by the computer and surveillance microphones; phone communications will be captured by the desk or cell phones as well.
- Lighting conditions may optionally be varied.

A.4 Scenario Test Plan

The scenario test plan includes the following sequence of steps.

A.4.1 Preparation

Steps conducted prior to each subject entering the test room.

- Verify that test office setup is correct and that all recording devices are operating properly.
- Clean the office to ensure that there are no fingerprints on desk, phones, computer keyboard, or mouse
- Testers wear gloves and masks, if DNA will be collected (note that eccrine material from sweat may permeate some porous gloves and could therefore contaminate samples)
- Place documents on desk, with preprinted headers to differentiate. Each of these will have two versions, one completely clean (never touched by anyone) and the second handled by a specific known tester
- Start surveillance cameras and microphone
- Soft drink cans:
  - 1: Provide subject with fresh, dry soft drink can that has been cleaned of fingerprints by test crew
  - 2: Provide subject with a second dry soft drink can that has been handled by a specific known member of the test crew
  - Because DNA will be collected from the soft drink cans, personal protective equipment (PPE, at a minimum gloves and masks) should be worn when handling the soft drink cans to avoid DNA contamination.

A.4.2 Subject interactions

The following steps are to be conducted for each subject. The steps are repeated for each desk (clean and "contaminated").

- Tester calls subject on phones to provide instructions.
- Computer:
  - Subject will access computer using written instructions
  - Computer video camera will record subject
  - Subject will be asked to perform tasks that result in typing and mouse use
  - Computer screen displays instructions as to what to write
- The subject will use paper pad and pen provided to make handwritten notes on the files (subject will copy the first three sentences) – instructions will be appear on computer screen.
  - Index card
    - Print and sign name with right hand
    - Print and sign name with left hand
  - 8.5"x11" paper:
    - Print and sign name with right hand
    - Print and sign name with left hand
    - With dominant hand, both print and write text from computer screen
- Tester calls subject on the phone and asks the subject to answer several scripted questions – not previously discussed with the subject (record voice sample of normal speaking voice). Results in typing, mouse, and writing samples.
- The tester will ask the subject to read the first paragraph of each document that was placed on the desk (record voice sample of reading voice).
- The tester will instruct subject to take a drink from the soda can (latent fingerprint samples and DNA sample).

A.4.3 Data collection

Steps conducted after each subject leaves the room to collect data

- Test crew will collect the documents handled by the subject, the soda can for fingerprint and DNA collection, collect latents from the phone and computer keyboard and desk.
- The test crew will use PPE when collecting any objects used for DNA samples (soft drink can).
- Annotate Data: Tester will enter comments for each test subject record (as necessary) to alert analyst of possible deviations from the test plan.
- Test crew will clean desk and keyboard to prepare test environment for next subject.

The above assumes samples are collected immediately after the subject leaves. Alternatively, a time variable may be considered: elapsed time between deposition and collection.

Latent processing methods:

- Direct nondestructive/unprocessed capture of latents (Alternate light source, RUVIS or visible light)
- Porous materials (paper)
  - Ninhydrin or 1,2-indanedione (1,2-IND)
- Nonporous materials

- Latent prints on all nonporous items can be developed with fingerprint powder
- Latent prints on replaceable nonporous items can also be developed using superglue fuming
- Replaceable items: can, pencil, pen; potentially keyboard and mouse
- Nonreplaceable items: desk, desk phone, cell phone

If processing is not done on site, samples would need to be packed to avoid disturbing the prints: clearly label the items and package the items so they cannot move around in the container or rub against other objects in the container.

A.5 Evaluation phases

A.5.1 Pilot phase

Prior to start of actual scenario testing, a small pilot will be conducted to determine whether the data capture methodology is effective. This will include:

- Dry-run collection of all samples listed above
- Can latent prints be recovered from the target surfaces effectively?
- Is the cleaning method effective?
- Is voice capture via phone and separate device working correctly?
- Is the computer capturing video data as expected?
- Is the surveillance camera(s) capturing the area of interest?
- Is data being recorded and stored correctly?

The following issues and considerations will be addressed during the pilot stage of the scenario data collection:

- Determine how video is to be used to determine subject deviations from prescribed scenario.
- Can other DNA samples be collected and can DNA contamination be addressed?
- How should unexpected or variable conditions be addressed, such as
  - sneezing into hands?
  - variations in temperature and humidity?
  - possible impact of washing hands before data collection?
  - will subjects be asked not to touch face or hair? Or deliberately touch them to purposefully deposit sebaceous prints?
  - possible variation in matrix due to use of makeup or lotion?

A.5.2 Revise test plan based on Pilot

The test script and specific objects used should be reviewed and potentially revised based on the results of the dry run/pilot.

A.5.3 Testing phase

Execute test plan. Ensure that data collection is periodically checked for data integrity and completeness: if necessary, revise test plan or add quality assurance checks based on ongoing lessons learned. For minor changes in the test plan, document when changes were made (for example, documenting that starting on date *X*, additional metadata *Y* was collected, affecting samples numbered *Z* or later). If the test plan is revised in any way that notably affects the samples, create defined subsets of the dataset and document the distinctions.

A.6 Overall Test Plan

The overall test plan should address these issues.

A.6.1 Data collection specifications

Detail the methods by which samples are collected and processed, including equipment, materials, and skills needed.

- Video recording specifications (computer camera and surveillance camera(s))
- Voice recording specifications (microphone in test area and in exemplar creation area)
- Latent processing specifications and procedures
- Voice recording quality requirements/device selection
- Surveillance data recording (format, compaction if any)
- Phone recording methodology
- Determination of privacy impact requirements
- A review of the data to be captured by independent data analysts for data completeness and data integrity

A.6.2 Test subject requirements

- Assess test subject requirements: consider target population, demographic distribution, and desired number of subjects

A.6.3 Administration

- Time and staffing estimates necessary to conduct the scenario test
- Selection, scheduling, and administration of test venue
- Selection and administration of materials and equipment
- Selection and administration of staff:
  - Administrative staff to manage test
  - Trained crime scene investigation staff to collect samples
  - Technical staff to digitize, format, and archive samples

A.6.4 Human Subjects and Privacy Approvals

- Complete Human Subjects Protections paperwork, in coordination with the Institutional Review Board (IRB)
- Identify the appropriate privacy review authority for this test. Complete required approval request forms.
- All human subjects and privacy approvals must be in place prior to the start of testing.
- Ensure maintenance of the approval and consent records.

A.6.5 Recruit test subjects

- Recruit and select test subjects
- At the time of recruiting the subjects, test scripts and a test plan description should be given to the subjects so that they understand what is expected from them, how much time will be required, and a determination of acceptance of the test process by the subjects.

A.6.6 Data collection and storage

Data collection and storage will require the definition of the following:

- Processes and procedures for collection and processing of data
- Procedures for handling and storage of physical items
- Procedures for digitization, formatting, and storage of digital samples (images, video, audio)
- Plan for data organization and archiving
- Collection and retention of metadata (e.g., quality metrics) as part of dataset
- Plan for potential updates (e.g., maintaining points of contact so that errors found by users can be addressed)
- Plan for long-term maintenance (including storage, distribution, review, and updating)

A.6.7 Report

Reporting should address the following:

- Compile, archive, and disseminate dataset.
- Publish report including scenario plan and test plan, listing all data captured and describing quality controls applied to data and demographics of test sample. Critique test process and make recommendations for test improvement.
- List on *Biometric and Forensic Database Catalog* website. (Under development)

Appendix B Efficient collection of latent prints under controlled conditions

This Appendix describes processes to collect latent prints for evaluation that resemble casework but have "ground truth" source attribution with an extremely high degree of confidence.

Contents

B.1 Introduction

Collection of latent prints in a controlled setting may appear to be straightforward — however, it is difficult to collect latents with the range of heterogeneous attributes and quality that are seen in casework, and at the same time have exacting quality assurance procedures that guarantee that the source of each latent is known definitively. This document describes processes to collect latent prints for evaluation that resemble casework but have "ground truth" source attribution with an extremely high degree of confidence. For further guidance on the collection of latent fingerprints for research, see Sears et al., "A methodology for finger mark research" [17] and the International Fingerprint Research Group (IFRG) "Guidelines for the Assessment of Fingermark Detection Techniques" [18].

The quality, types, and attributes of latent prints vary enormously agency to agency, and by crime type. Therefore, collecting latents that seek to simultaneously represent multiple agencies and crime types is problematic.

This Appendix is focused on making recommendations for the collection of latent datasets that are intended to be useful for a variety of future applications. One overarching recommendation is to overcollect: collect a superset of the latents for any anticipated need, with a wide variety of attributes and quality, retaining this metadata with the dataset. Any future users of the dataset then

may select subsets from within the dataset based on distributions of these attributes, or conduct analyses based on these attributes. This passes to the users of the dataset the ability (and responsibility) to select a subset of prints appropriate for their specific uses.

B.2 Collecting latents with varied attributes and quality

Latents in casework can vary in a variety of attributes, which should be considered to make the collected latents resemble casework as much as possible. The following sections recommend ways of collecting latents to address each of these attributes. A key recommendation is the need to document these attributes at the time of collection: such metadata allows the user of a dataset to understand the distribution of attributes and assess the representativeness of the dataset, which would be unknown otherwise.

B.2.1 Area of skin

When we think of latent prints, we generally think of latent fingerprints from the distal (outermost) segment, relatively centered — in part because these correspond to the exemplar fingerprints that we are familiar with. However, latents should be representative of all the ways our hands (and even feet) contact objects and leave impressions. Latent datasets should include the tips, sides, and lower segments of the fingers, as well as palms.

B.2.2 Deposition type (Type of contact)

Dataset collection should also consider the type of contact between the skin and the object, which can be addressed in terms of distortion, pressure, and interaction with the object:

- ***Distortion*** — Latent prints can be deposited in a variety of ways that can result in distortion. In addition to collecting ordinary (plain) latents, consider having the subject leave samples in which the hand or finger is rolled, twisted, or slid after contact — retaining which deposition type was used as metadata.
- ***Pressure*** — The subject should consider leaving prints under light, normal, or heavy pressure (retained as metadata). Differing pressure can result in a variety of unusual impression characteristics: with very heavy pressure ridges and furrows may invert black for white; dots and incipients may vanish with light pressure and appear to be normal ridges with heavy pressure; minutiae that appear to be ridge endings may become bifurcations with heavy pressure; with excess matrix and heavy pressure, the impression may lose much ridge detail; under light pressure an actively sweating subject may leave only a series of dots (from the pores) and little or no impression of the ridges themselves.
- ***Interaction with the object*** — The plain or rolled contact made when collecting exemplar fingerprints are only a subset of the various ways in which hands make contact with objects. When planning a collection, observe how objects are held and touched in ordinary activities, and duplicate these. Consider collecting latents in ordinary interactions such as how a hand holds a mouse, knife, pen, phone, door handle, or glass; how fingerprints are left when tearing duct tape; how when writing on an index card the fingertips from the other hand are used to hold the card still; how prints are left when opening and closing a window.

B.2.3 Overlapping impressions in an image

Images collected from crime scenes frequently include multiple latents in each image. Such images are important to include in evaluations, because a key aspect of the examination process (for both

human and automated processing) is determining the impression of interest and separating it from other impressions in the same image, especially when the impressions overlap.

Collecting controlled latents with multiple impressions while guaranteeing ground-truth source attribution is achievable (with care) if all latents are from the same source:

- Multiple impressions from the same subject and area of skin — The image includes multiple impressions, but all the impressions are from the same finger (or area of the hand) from the same person, so source attribution is not in doubt. Such "double-taps" are common in actual casework.

When source attribution of the finger/palm position is less critical, there are other alternatives:

- Multiple impressions from the same subject — The image includes multiple impressions, all of which are from the same subject. Source attribution of the subject is not in doubt, but the finger/palm position is not definitive.
- Multiple impressions from a second subject — The image includes multiple impressions from two (specified) subjects. Source attribution of the subject can be limited to two people but is not definitive; the finger/palm position is not definitive.

When latent images include multiple impressions, the metadata should indicate the sources of all impressions.

B.2.4 Types of objects (Substrate, background, shape)

- ***Substrate*** — The substrate of a latent refers to the material on which the latent was deposited. A variety of substrates should be included — and noted in metadata. ANSI/NIST-ITL [6] defines these categories of substrates:
  - Porous substrate — such as paper, cardboard, or unfinished/raw wood
  - Nonporous substrate — such as plastic, glass, metal (painted), metal (unpainted), glossy painted surface, tape (adhesive side), tape (non-adhesive side), or aluminum foil
  - Semi-porous substrate — such as rubber or latex, leather, photograph (emulsion side), photograph (paper side), glossy or semi-glossy paper or cardboard, satin or flat-finish painted surface
  - Other/unknown substrate
- ***Background*** — In addition to the substrate, the background patterns or textures can have a major effect on the utility of a latent (for example, plain white paper vs check stock). The background can interfere with latents in several ways: the visible patterns (e.g., printed text, check stock) can be difficult to separate from the ridge detail; texture (e.g., leather, crumpled paper) can make gaps and local differences in pressure in the latent; color can interfere with processing (e.g., since latent prints developed using ninhydrin are purple, prints on purple paper must be processed differently). Consider a variety of appearances of objects and backgrounds.
- ***Shape*** — In collecting latents, think beyond flat surfaces. Consider curved objects and other shapes. Note that latents on curved objects may require multiple images of the same latent.

Note the substrate can affect whether the end user will be able to assess orientation or lateral reversals (i.e., flipped left-right): for example, lateral reversals are obvious on paper with printed text, whereas orientation or lateral reversals cannot readily be determined for fingerprints on folded clear tape.

> *Lessons learned: One of the obvious ways most controlled collection latent datasets differ from casework is in the variety of objects where the latents are found. Casework latents can be on combs, cartridge cases, dolls, duct tape … any object that may be encountered at a crime scene.*
>
> *One approach that works well is go to hardware stores or dollar stores and buy many cheap random objects, books, maps, etc., and clean them for use as substrates with newly collected samples.*

B.2.5 Matrix

The matrix is the substance that forms the impression. In collecting latents, the matrix type, amount, and depletion should be considered along with the effect of subject activity. Because some people do not readily leave fingerprints, and because of depletion, do not assume that simply having people handle objects will result in large numbers of useable latents.

- ***Type*** — Overwhelmingly the type of matrix is (or is assumed to be) perspiration and/or body oils. ANSI/NIST-ITL lists various other matrices that may be considered: blood, paint, ink, oil or grease, dirt or soil, other visible contaminants, impression in pliable material (e.g., wet paint or glue), contaminant removal via touch.
- ***Amount*** — The amount of matrix can affect the quality of the latent substantially. Either extreme (very heavy or very little matrix) can result in a latent with little or no ridge detail. The amount of perspiration and oil left in fingerprints varies significantly among people: some people leave very few prints when they touch objects, but some almost always do.
- ***Depletion*** — When people handle objects, the amount of matrix left diminishes with each impression. Asking a subject to leave multiple latents in a row is known as a "depletion series" in which the amount of matrix decreases with each deposition. Depletion series are a very useful way of finessing concerns about matrix issues because the recipient then can select from a full range of heavy through light to no matrix. Start by having the subject rub his/her finger or palm across his/her forehead, then make a series of multiple impressions; generally, each impression will become less clear, and the last impressions may contain no ridge detail. Depletion series should be clearly indicated in the metadata: generally, the users of a dataset would only select one latent out of each series.

> *Lesson learned: one collection effort used preprinted paper cards, one per finger per deposition method and collected a depletion series for each. This worked very quickly: for each right index finger the subject made 5 latent plain impressions, then 5 rolled impressions, then 5 twisted impressions, then 5 slide impressions.*

- ***Subject activity*** — If subjects are actively perspiring, the amount of matrix increases, but also the appearance of the ridge detail can change, because a tiny drop of sweat is left at each pore. Pores may also change their appearance based on temperature, pressure, or distortion. Consider collecting some latents on hot days and/or after subjects exercise.

B.2.6 Processing method

The processing method has a significant effect on latents, and therefore should be varied — and recorded in the metadata. ANSI/NIST-ITL (Field 9.352) lists dozens of processing methods: although not all of these need to be addressed in each data collection, there should be much more variety than black powder on nonporous materials and ninhydrin on porous materials.

Note that some processing methods can tonally invert the image (black for white), such as cyanoacrylate (superglue) on a dark surface, or RUVIS imaging. This should either be corrected prior to dissemination or clearly documented.

B.3 Exemplar collection methods

Although collection of exemplars does not have as many variables as does latent collection, the exemplars are notably affected by the collection methods or devices used.

Fingerprint exemplars should include rolled and plain impressions. Full sets of exemplars include all palmar surfaces (including the tips, proximal and medial segments, and sides of fingers).

Exemplars should at a minimum include full sets collected using

- Inked cards (if inked exemplars and latents are collected in the same session, collect latents first so that vestiges of ink are not present in the latents)
- FBI-certified livescan devices

Collection using additional devices should be considered:

- Non-optical livescan devices
- Small-platen devices
- Non-contact devices

B.4 Other collection considerations

B.4.1 Population

Because the attributes of fingerprints can vary significantly by age, gender, occupation, etc., the dataset should, at a minimum, document the age and gender of the subjects, either as an overall distribution or per sample.

B.4.2 Multiple exemplars and latents per source

There can be notable within-subject variability among exemplars: the appearance of images from the same finger can differ in clarity and distortion, as well as by collection method. Therefore, it is highly recommended that multiple exemplar sets be collected for each subject.

For latents, the within-subject variability is even more significant. Consider collecting multiple impressions for each area of skin (not just one per finger and per palm).

B.4.3 Feature annotation and examiner assessments

Friction ridge impressions can have various types of annotations, including

- Regions of interest (ROI)
- Minutiae (in each separate image)
- Corresponding minutiae (in two images)
- Cores and deltas

- Extended feature set features (dots, incipients, etc. as defined in ANSI/NIST-ITL)
- Pattern classification

If features are annotated, ANSI/NIST-ITL format must be used.

In addition, the impressions can be accompanied by other metadata, such as examiner assessments of value (e.g., no value vs value for comparison) or quality (e.g., good, bad, or ugly).

Values from the FBI's ULW LQMetric can be used as a standard method of assessing fingerprint quality.

The method by which features are marked can vary:

- Features can be marked by automatic feature extraction algorithms, by individual examiners, or by groups of examiners ("juried" or consensus feature sets).
- Features can be marked based solely on what is visible in the image, or they can be "ideal" features based on examination of multiple images known to be from the same source (not generally operationally feasible).

The method(s) by which features are marked must be documented:

- Because there is significant inter-examiner variation in minutia markup, the markup of a single examiner should not be taken as definitive.
- Automatic feature extraction algorithms cannot be taken as definitive, and different algorithms cannot be assumed to have the same results. Depending on the end use of the dataset, use of a specific feature extraction algorithm may introduce bias into the dataset (e.g., if evaluating automated fingerprint matchers that may or may not be tuned to specific feature extraction algorithms).
- Consensus feature sets should not be taken to be representative of operational results. Although consensus feature sets reduce inter-examiner variation in minutia markup, they should not be presented as definitive.

B.5 Quality assurance procedures

Quality assurance (QA) procedures serve to minimize the possibility of error. The specific issues that need to be addressed include:

- Guaranteeing source attribution information
- Avoiding administrative errors
- Avoiding processes that bias data selection

B.5.1 Guaranteeing source attribution information

The core reason that we have controlled collection of latent prints is to guarantee — in an absolute, "ground truth" sense — that we know the subject and area of skin that left a given latent. If source attribution of each latent is not collected in a way to guarantee that it is definitive, the value of the dataset is significantly reduced.

It is critical to have rigorous QA procedures in place for the collection of datasets that will be used for the testing for rare events, such as erroneous IDs by examiners, or erroneous exclusions reproduced by a majority of examiners. Without rigorous QA procedures, administrative errors could dwarf the error rates that we attempt to measure.

The worst case scenario (with respect to a dataset collected for evaluation) would be to conduct a study, find a sample associated with an erroneous ID, and then realize that there would be no way to be certain that it was not actually a non-mate caused by an administrative error.

Never use any method other than data from the original collection to determine ground-truth source attribution. If human examiners or AFIS results are used to "correct" ground truth, they are essentially acting as filters eliminating the hardest comparisons, which is biasing the dataset, not improving its quality.

The best way to guarantee source attribution is to design the collection process so that administrative errors can be avoided.

B.5.2 Avoiding administrative errors

The collection process requires procedures to minimize the possibility of administrative error. In general, this involves:

- Methods making it easy for the subject and finger/palm position to be checked by the subject or collection staff at the time of collection.
- Collection of metadata so that the subject and finger/palm position can be double-checked after the fact.
- Avoiding contaminating samples.

Recommendations to consider:

- Devise a simple, easy-to-check subject ID and sample numbering method (see Section B.5.4).
  - Consider preprinted labels.
  - Avoid having people copying IDs by hand to minimize transcription errors.
- Label every sample with subject ID, finger/palm position, sample number, and key metadata (e.g., deposition type). This should be visible for the subject and collection staff at the time of collection, and in the context images.
- Capture "context images": for every sample collected, include a larger context image that includes labels for subject ID and metadata. This allows after-the-fact review of any images in doubt. The sample images used in the dataset are then cropped out of the context images so that the labels are not visible — but the context images are retained long-term in case there is a future need for review.
  - ANSI/NIST-ITL allows storing context images (type-20 records) in association with the specific samples (e.g., storing the crop coordinates by which a type-13 latent image is associated with a type-20 context image)
  - Context images also provide for collecting multiple simultaneous impressions, which would then be cropped into several latent images per context image.
  - Context images should always include standard rulers for resolution verification.
- Staff must wear gloves in handling samples; subjects might wear gloves on the hands not being collected.

> *Lessons learned: Although the idea of flipping images left-right to add some level of anonymity sounds appealing (e.g., flip the right thumb image and call it a left thumb), it adds an additional level of administrative complexity (e.g., did I already flip this image?) and is much more problematic in dealing with inked cards (you must collect the fingers in their actual positions, but the images must be cropped to eliminate the text).*

B.5.3 Avoiding biased data selection

As discussed in the main paper, anything done that removes some samples from a dataset risks biasing the resulting data:

- Frequently datasets omit latents that an examiner considered to be no value: removing such data is problematic because examiners vary greatly in determining which samples are of no value. If the worst-quality samples are removed, the resulting dataset cannot be used to evaluate the ability to deal with poor-quality data.
- Using an AFIS to determine subject attribution means that the resulting dataset only includes those samples for which that particular AFIS succeeds and all subsequent work is conditioned on that, making the resulting dataset useless for future evaluations of AFIS — or of examiners.

Retain all images in the overall dataset, including ones that an examiner considers to be of no value, because examiners (and systems) differ extensively in this determination. Similarly, removing the easiest comparisons can bias the resulting dataset.

B.5.4 File name conventions

The file naming convention should be carefully thought through. Inconsistent or poorly designed file naming conventions will be a source of frustration for the life of the dataset. The name of the file should:

- Identify the subject and area of friction ridge skin using standard terminology (e.g., Finger 05, not “right little” or “right pinky”)
- Differentiate among different impressions for the same subject/area (e.g., latents from exemplars, and among different latents)
- Enable unambiguous association of context images and sample images (i.e., make it clear when one image was cropped from another)
- Have unique identifiers for each image (e.g., for association of metadata in accompanying tables)

Recommendations (based on seeing a variety of naming conventions that were easy vs. difficult to use in practice):

- Long and complicated numbering systems are error-prone (hard to check; greatly increase the chance of transcription errors).
- Avoid naming systems where “0” (zero) and “O” (capital O), or “1” (one), “l” (lower-case L), and “I” (capital I) can be confused.

- Use naming systems that allow the names to be easily parsed in a spreadsheet.
- Use separators and fixed widths so the filenames are easier to parse and distinguish.
- Do not use variable length filenames.
- Do not use spaces in filenames, or characters that are not usable across operating systems.
- Do not use the same names for files in different directories.

One approach that has worked well:

- Name the files so that all file names sort by dataset, then subject, then finger/palm position
  - For example, all files for dataset DS, subject 123, Finger 07 start "DS123_F07…" (or "DS123_P23..." (Palm 23 = left full palm)
- This approach is easy to work with because everything up to this point is source attribution, everything after is descriptive of collection, which makes it much easier to visually review or parse in a spreadsheet
- After the dataset/subject/position, each distinct L (latent) or E (exemplar) gets an impression number (DS123_F07-L17 is the 17th latent for subject 123 finger 07), so that all latents and exemplars for a given position sort together.

Lessons learned: Make sure that when the dataset is archived, the archive also includes the documents explaining the source of the samples, method of collection, file name conventions, and cross-references to other data.

Acknowledgements

The authors would like to thank the following for their insights and suggestions: Elham Tabassi, Maria Antonia Roberts, Robert Thompson, Susan Ballou, Heidi Eldridge, Edward Bartick, Richard Cavanagh, Gregory Fiumara, and Robert Ramotowski.

References

1 Blackburn D (2001) Evaluating Technology Properly— Three Easy Steps to Success. *Corrections Today,* 63.
2 Ioannidis JPA (2005) Why Most Published Research Findings Are False. *PLoS Med,* 2(8): e124. doi:10.1371/journal.pmed.0020124 .
3 Nature (2014) Challenges in irreproducible research. *Nature,* 7 October 2014. http://www.nature.com/news/reproducibility-1.17552.
4 Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science*, Vol. 349 no. 6251.
5 Liberman M (2015) Replicability vs. reproducibility — or is it the other way around? Language Log (http://languagelog.ldc.upenn.edu/nll/?p=21956, accessed 18 March 2020)
6 National Institute of Standards (2015) American National Standard for Information Systems: Data format for the interchange of fingerprint, facial & other biometric information. ANSI/NIST-ITL 1-2011:Update 2015.
7 Wilson C, et al. (2004) Fingerprint Vendor Technology Evaluation 2003. *NIST Interagency Report* 7123.
8 Watson C, et al. (2014) Fingerprint Vendor Technology Evaluation. *NIST Interagency Report* 8034.
9 Watson C, et al. (2005) Two Finger Matching With Vendor SDK Matchers. *NIST Interagency Report* 7249.
10 NIST. Proprietary Fingerprint Template Evaluation II (PFTII) results. (https://www.nist.gov/itl/iad/image-group/proprietary-fingerprint-template-evaluation-ii-pftii-results)
11 NIST. Proprietary Fingerprint Template Evaluation PFT III results. (https://pages.nist.gov/pft/results/pftiii/)
12 University of Bologna Biometric System Laboratory. SFinGe: Synthetic Fingerprint Generator. http://biolab.csr.unibo.it/sfinge.html (accessed 18 March 2020)
13 Orlans N, Piszcz A, Chavez R (2003) Parametrically controlled synthetic imagery experiment for face recognition testing. *Proc. of the 2003 ACM SIGMM workshop on Biometrics methods and applications*, pp. 58–64.
14 NIST Special Database 27/27A. Fingerprint minutiae from latent and matching tenprint images. No longer distributed by NIST. https://www.nist.gov/itl/iad/image-group/nist-special-database-2727a
15 Ulery B, Hicklin RA, Watson C, Fellner W, Hallinan P (2006) Studies of Biometric Fusion. *NIST Interagency Report* 7346.
16 Phillips PJ, Martin A, Wilson CL, Przybocki M (2000) An Introduction to Evaluating Biometric Systems. *IEEE Computer*. 33. 56 - 63. 10.1109/2.820040. http://www.hh.se/download/18.70cf2e49129168da0158000129674/Intro_to_Evaluate_Biometrics.pdf
17 Sears VG, Bleay S, Bandey H, Bowman VJ (2012) A methodology for finger mark research. *Science & Justice : Journal of the Forensic Science Society*. 52. 145-60. 10.1016/j.scijus.2011.10.006.
18 Almog J, Cantu A, Champod C, Kent T, Lennard C (2014) Guidelines for the assessment of fingermark detection techniques International Fingerprint Research Group (IFRG). *Journal of Forensic Identification*. 64. 174-197.