USEFULNESS OF CURRENT BIOMETRIC DATASETS

Biometrics and Forensics Data Symposium

Elham Tabassi

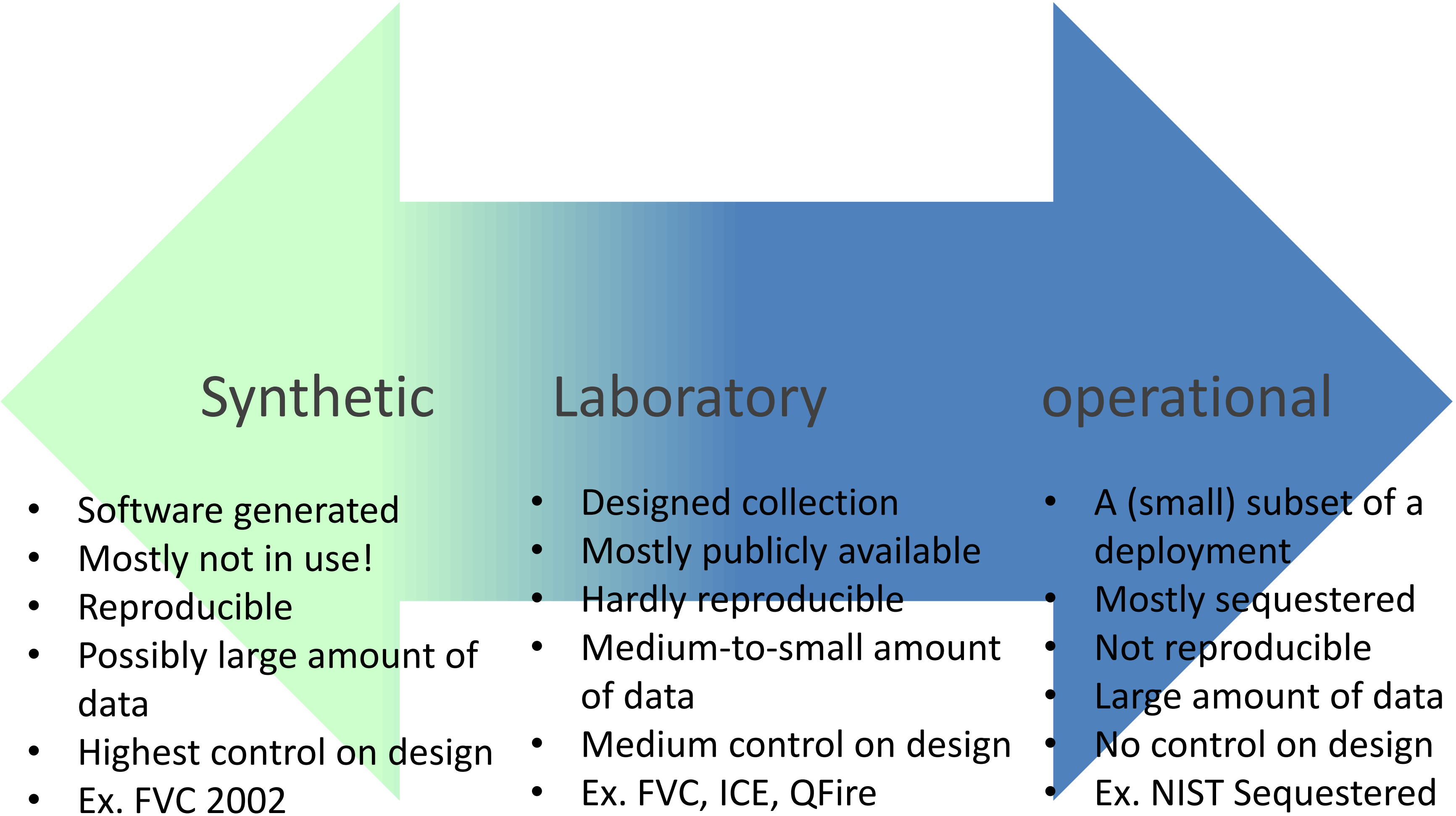
NIST / ITL / Image group

January 26, 2015

OUTLINE

- Current landscape
  - Issues + gap
- Synthetic
- Laboratory collection
- Operational
- Wish list

Current Landscape :: Data origins



Synthetic

- Software generated
- Mostly not in use!
- Reproducible
- Possibly large amount of data
- Highest control on design
- Ex. FVC 2002

Laboratory

- Designed collection
- Mostly publicly available
- Hardly reproducible
- Medium-to-small amount of data
- Medium control on design
- Ex. FVC, ICE, QFire

operational

- A (small) subset of a deployment
- Mostly sequestered
- Not reproducible
- Large amount of data
- No control on design
- Ex. NIST Sequestered evaluation data POE

Current Landscape :: Issues + Gaps

- Non-uniform usage of publicly available data prevents reproducible research
  - Selective subset of dataset
    - Removal of some images or subjects without reporting
  - Selection of enrolled (gallery) + search (probe) sets
    - Are comparison scores independent?
  - Varying number of representations per source
- Non-intended purpose
  - e.g., reporting accuracy on a unusually low quality dataset or goat study on frequent travelers.
- Legacy vs. emerging technologies

SYNTHETIC DATA :: SOFTWARE GENERATED DATA

FROM SCRATCH OR MANIPULATING A PRISTINE IMAGE.

Advantages

- Making images with specific controlled defects, where the type and exact amount of the impairment are known.
- Ground-truth known and traceable.
- Can generate many many images
  - Repeatable
- Mostly public + no privacy issues
  - Can promote reproducible research
- Most useful for developing or evaluating algorithms for detecting specific defects (i.e., quality algorithms)

Issues

- The world is too complex to be synthesized.
  - Synthetically impaired images would not be a fair representation of the real-world low-quality images.
- Fails to capture the interaction of several simultaneous defects in an image, as is the case in real-world non-laboratory data.
- Metric for assessing the representativeness of the synthetic data to real-sensed fingerprints.

LABORATORY DATA :: DESIGNED DATA COLLECTION

VARYING CAPTURE DEVICE SETTINGS, OR ENVIRONMENT CONDITIONS, OR SUBJECTS' BEHAVIOR BY DESIGN.

Advantages

- Producing real-world (or real-sensed) images.
- Allows for designing the type and amount of impairments – to some extent.
- Can support ongoing collection if subjects can be brought back
- Mostly public
  - Can promote reproducible research
- *Can* be used as a proxy for real data

Issues

- Precise control of acquisition is challenging, so inevitably ground truth will be noisy.
- Keeping the confounding variables, i.e., subject/acquisition parameters, uniform is unattainable.
  - Over or under representation of subpopulation or image characteristic
- Care must be taken to account for data integrity and balance
  - Correct subject IDs + Equal number of representation per source
- Cost grows very quickly with size
- Human subject review + approval

Operational data :: Real-world data

COLLECTED AT OPERATIONAL DEPLOYMENTS

Advantages

- True representation
  - Capture technology, capture environments
- Real defects or impairments
  - or several simultaneous defects
- Possibly large number of data available
- Ultimate target for all research / development / evaluation

Issues

- Ground-truth of subject IDs
  - Same source different ID
  - Different source same ID
- None or very limited ground truth on source or cause of low quality
- Possible sampling issue
  - Over or under representation of subpopulation or image characteristic
- May or may not be diverse
- Often sequestered
  - Cannot promote reproducible research

WISH LIST

FOR AN ALL-PURPOSE DATA COLLECTION

General

- Representative of real-world operational data
- Large number of subjects/sources
- Multiple representations
- Reliable meta data
  - sex, date of birth, date of capture, capture technology, resolution, finger position, nationality or race, pressure, moisture, rotation, etc.
- Diverse
  - Age, sex, capture technology, race, etc.
- Ongoing, extendable
  - Longitudinal studies
  - Emerging technologies, e.g., contactless fingerprints

The devil is in the details!

- What do real-world operational data look like?
  - How to sample to get a true representative?
- How large is large?
- How to assure data integrity?
  - Reliable ground-truth IDs
  - Reliable ground-truth image characteristics
- Mark-up or annotating data
  - E.g., minutia location

We can/should do

- Accurate characterization of operational real-world data
  - To learn `clusters’ of data
  - Design data collection to target the learnt `clusters’
    - Perhaps via uniform data collection protocol

- Better understanding of required sample size
  - And the associated uncertainty in measuring the error rates

- Improve uniformity of reporting
  - Improving data integrity in laboratory collection
    - Guidance document on consolidation
  - Guidance on enrolled (gallery) and search (probe) compositions

THANK YOU.

tabassi@nist.gov

301 975 5292