Modest proposals for improving biometric recognition papers

James R. Matey, George W. Quinn, Patrick Grother, Elham Tabassi & Craig Watson
NIST, Gaithersburg MD
POC: james.matey@NIST.gov

James L. Wayman
San Jose State University, San Jose, CA
jlwayman@aol.com

Abstract

We present practical recommendations for improving the clarity, transparency, and usefulness of many biometric papers. Several of the recommendations can be enabled by preparing a publicly available library of state of the art Receiver Operating Characteristics (ROCs). We propose such a library and invite suggestions on its details.

1. Introduction

This paper makes specific, practical recommendations aimed toward improving the clarity, transparency, and usefulness of biometric papers. We propose good practices on how to make biometric research reproducible and how to facilitate objective comparison of results from independent studies – a key component of reproducibility. To support these practices, we propose construction of a publicly available library of state-of-the-art Receiver Operating Characteristics.

While in proof, we discovered a recent paper by Jain et al., "Guidelines for Best Practices in Biometrics Research" [14]; we recommend reading that paper in conjunction with this paper. Jain et al. focus on planning research; we focus on presentation of results; good research papers require both.

The topic of reproducible research has received considerable attention in recent years. In particular, Vanderwalle et. al. [35] open their paper with the statement:

> Have you ever tried to reproduce the results presented in a research paper? For many of our current publications, this would unfortunately be a challenging task.

Vandewalle et. al were correct in 2009 and it is unfortunate that they are still correct – we have not made that much progress.

The principle of reproducibility is a cornerstone of the scientific method and the capability to make meaningful comparisons across time and space is key to reproducibility. Data analysis, and more generally, scientific claims, should be published with their data and methods of derivation so that others may verify the findings and build upon them. If a result cannot be reproduced, it is at best incomplete and at worst "not even wrong" (a phrase often attributed to Nobel Laureate Wolfgang Pauli) in the sense that it cannot be falsified by experiment.

Attention has already been brought to the problem of irreproducibility in a range of fields outside of biometrics. There are reports that only 10% to 20% of papers in biotechnology are reproducible [5]. *Nature* keeps a running tally of high-profile failures in the reliability and reproducibility of published research [22] . In 2012, Nobel Laureate Daniel Kahneman expressed concern in a public letter that the field of "priming", an important human factors idea in biometrics, may be poorly founded after the results of several high-profile experiments could not be replicated [6].

Lack of reproducibility does not in itself imply that the research is fraudulent or the methodology is flawed. The problem is often that the publications do not enable easy replication of results that would enable the self-correcting aspects of the scientific method to operate efficiently. As a consequence, bad science can hide amongst the good without any way to differentiate the two.

Although in principle the scientific process is self-correcting, certain practical considerations can make that difficult at times. For example:

- Journals and funding agencies tend to place greater value on original research than on replication of existing results.
- Experimental methods and analyses are becoming increasingly complex, making them more difficult to replicate.
- Important experimental factors may not be under the

full control of the researchers - original or others.

- Restrictions on the original data or experimental tools may hamper or disallow distribution (e.g. privacy concerns, intellectual property issues).
- A failed replication attempt leaves us in a quandary: were the prior results in error, or did the replication fail in some way.

Although a positive correlation between citation count and ease of reproducibility has been noted [35], the long term advantages of reproducible research are to the community. Authors, particularly newer ones, should have a vested interest in adopting our recommendations since robust and transparent research is more likely to be impactful and lead to career advancement. Furthermore, the peer-review process should place the necessary emphasis on ensuring all data and tools used in a publication are freely available whenever possible.

An additional goal of the our proposals regarding Receiver Operating Characteristic (ROC) plots[1] is to help authors avoid attempted publication of poor variants of a common form of paper that can be summarized as "Yet Another Recognition Algorithm" (YARA). A key failure of many such papers is the comparison of results against "straw man" examples using small datasets. In iris a common straw-man is the Libor Masek algorithm constructed by an undergraduate student more than a decade ago [19]. We do not wish to denigrate Masek's accomplishment. It was a significant addition to the biometric literature. Likewise, the Model T was a significant addition to automobile technology. However, if we were writing a paper on an improvement to automobile technology, we would not use the Model T (let alone a homemade copy) as the standard for comparison. We would use a modern, state of the practice automobile. Likewise, if we were writing a paper on iris recognition, we would use a modern, state of the practice algorithm as the standard for comparison.

2. Proposals for Paper Improvement

Our recommendations to improve the understandability, relevance and usefulness of biometrics papers are enumerated below. In advocating for these, we acknowledge that there will be niche cases where it is proper not to implement particular practices. In such cases, we recommend that authors explain their rationale for doing so in their paper.

1. **Any paper discussing the effectiveness of a modality, algorithm, or process of data acquisition should include a receiver operating characteristic plot (ROC) [1] or detection error trade-off (DET) plot [18]**. The paper should include
    - plots representing the various modalities, sensors, recognition algorithms and training data under consideration;
    - plots comparing the work with the best state of the practice, including commercial efforts.

The classic ROC plot shows true accept rate (TAR) - essentially the probability of true detections (hits) - on the vertical axis against false accept rate (FAR) - the probability of false detections (false alarms) on the horizontal axis. This is done parametrically for any real threshold t, an accept being declared when the comparison of two samples yielded a similarity score greater than or equal to t.

The ROC plot originated in radar work during WWII and much of the early literature on the topic resides in unpublished reports, e.g. [29], rather than academic journals. The ideas filtered out into the broader community after the war[17]. The psychology community adapted the concept in the 1960's [31, 32]. Other uses followed: medical diagnostics [20] , electronic signal detectors [34]. The speaker verification community's adoption of the concept in the 1980's [21] may be the first biometric use.

The biometric community uses several ROC variants obtained by plotting error rate or (1- error rate) for false rejections and by using linear, log and other transformations of the axes. In particular, the speaker recognition community found it enlightening to plot errors against errors, (false reject rate, FRR = 1 - TAR vs. FAR) developing the Detection Error Tradeoff (DET) characteristic. Moreover they plotted on a normal deviate scale to afford straight line plots for Normal genuine and impostor scores. [18]. In many modalities, algorithms do not produce normally distributed scores and researchers may find it expedient to use logarithmic scales[16] instead. For all of these, the important point is that ROC/DET plots encapsulate the relationship between hits and false alarms as a function of this threshold in a clear and concise way. As such it is of primary importance to owners of biometric systems in that it allows them to set policy on security (FAR) vs. convenience (TAR).

The motivation for using ROC/DET plots goes deeper: Edward R. Tufte, a statistician and professor emeritus of political science, statistics, and computer science at Yale University known for his contributions to information design and data visualization, states in Chapter 4 of his book *Envisioning Information* [33] that "At the heart of quantitative reasoning is a single question: *Compared to what?*". The ROC and it variants are likely the most effective and salient way of comparing biometric recognition algorithms, data collection methods and systems.

We note that reporting an equal error rate (EER) is seldom a good substitute for a full ROC plot since the EER is rarely a desirable operating point.

We also note existence of free software to expedite preparation of ROC/DET curves[3, 26].

2. **An ROC should plot salient FAR values in an easily legible form (e.g. log)**.

   The scaling issue is not just cosmetic. Whatever transformation is used (log, linear, normal), and whatever quantity (FRR or 1-FRR) appears on the ordinate axis, most applications of biometrics operate at thresholds set to attain low false accept rates. Therefore research papers should support accuracy comparisons in the regime $0.0001 \leq \text{FAR} \leq 0.01$. Many current studies instead plot $0 \leq \text{FAR} \leq 1$ on a linear axis such that salient low FAR values are almost invisible on the very left edge. This is a serious fault since it can lead researchers to optimize FRR at inappropriately high FAR values. Algorithm improvements at high FAR may not yield improvements at low FAR, as within-class and between-class separability can be traded off algorithmically.

3. **Include threshold links on comparative DETs/ROCs.** If a DET/ROC plot includes separate characteristics for one recognition algorithm applied to two or more subsets, for example from male, female populations, or from good, fair, poor imaging conditions, the plot should include links between points of constant threshold. The example in Figure 1 shows excursions in both FAR and FRR for a face recognition applied to high quality mugshots and low resolution webcam images. DETs should be considered to shift vertically *and* horizontally. This recommendation is made because operational systems are usually configured with a fixed threshold but are nevertheless frequently applied to semantically different data sets.

4. **ROC/DET data should be made available in digital form**. This enables re-use of the data and re-plotting of the data in any of the ROC plot variants as needed. The data should be available at sufficient resolution to capture the details of the curve, not just at coarsely sampled FAR intervals; it should be available as tuples $(t, \text{FAR}(t), \text{FRR}(t))$ in a text file, made available either as supplementary data for the paper for journals which offer that capability, or on a website or by email request to the author.

5. **Comparative studies of algorithms, or sensors, should be based on 1:1 (one-to-one) comparison of pairs of biometric samples** as is done in the Labeled Faces in the Wild (LFW) [11] and the IJB-A [15] benchmarks and in the long running NIST Speaker Recognition Evaluation[28]. This recommendation differs from many papers which compare probes against linear galleries. Our contention is that core recognition accuracy can best be stated by executing pure 1:1 comparisons, and not 1:N (1-to-N) search algorithms. Our basis here is:

   - 1:N algorithms can employ additional, separate approaches - gallery normalization, indexing - that confound assessments of the core feature-based recognition power.
   - Failure analysis[7] is easier when recognition outcomes depend only on two samples, not N+1.
   - The existence of an enrolled database, the gallery, cannot be assumed in many real world applications, for example via legal prohibition[8] , or by circumstance (e.g. passport authentication).

   Of course, papers that specifically address identification technology should be exempt from this recommendation.

6. **Implement 1:N recognition only after 1:1 accuracy has been shown to be promising.** We recommend that most biometric technology should be developed and assessed in a 1:1 matching mode. Only in cases where specific 1:N technology is the goal of the research, should this step be elided. 1:1 accuracy should be estimated over K mated pairs, and L non-mated pairs rather than a gallery. The gallery concept is only needed in 1:N. Generally, $L \gg K$ in order to support measurement of FAR at low, relevant values with statistical significance.

   1:N should be allowed of course if the study's specific intent is to explore 1:N search - e.g. an indexing algorithm.

7. **CMC should not be the primary metric**. The corollary of the previous recommendation is that CMC is deprecated in this context. Many papers state accuracy as a cumulative match characteristic (CMC), which gives the proportion of searches for which the corresponding mate is found in the top R ranked gallery identities. We recommend against such an approach. Our motivation for this recommendation is:

   - CMCs decrease with increasing gallery size N (for example, as a Power Law[9]), and this renders algorithms incomparable unless N happens to be identical.

- CMC is purely a rank-based metric that discards potentially interesting comparison score information.
- As the CMC is estimated over closed-set searches where each search sample has a mated enrollment, it frequently does not represent operational reality - not everyone has a prior criminal file, for example. Additionally, closed-set tasks can lead researchers to unwittingly code linear assignment strategies. Researchers should use an open-set design where some proportion of searches do not have a mate, as described above.

8. **Disclose software availability**: Any paper describing or utilizing a method implemented in software should indicate its availability (e.g. as open source, commercial product, etc.) and how it might be acquired. If it is not available, the paper should provide an explanation of why it is not.

9. **Disclose data availability**: Any paper that makes use of a dataset, including development, training and test data, should indicate if the dataset is generally available and how it might be acquired. If it is not available, the paper should provide an explanation of why it is not. All papers should include an ROC plot based on one or more publicly available datasets to enable comparisons between different algorithms by different researchers.

   In support of the recommendations regarding availability of code and data, we turn to primary rationales for publication – (1) to enable others to reproduce the work and (2) to enable comparisons of results from different sources. Again, as pointed out by Vandewelle the availability of code and data are correlated to the citation count for articles containing such. As an anecdotal example, the Libor Masek thesis [19] is one of the most highly cited papers in iris recognition – in large part because it made available the source code developed in the thesis.

10. **Report the dimensionality and size of the underlying feature representation**: Many approaches to recognition involve relatively low-dimensional features extracted from parent samples, In other cases, models are built to represent the input signal. In any case, the size of the data associated with a sample or an individual should be reported with both its size (e.g. in kilobytes), its data-type (e.g. four byte floats), and dimensionality (if appropriate). Such a description is not necessarily unique; a short iris2pi template might be described as a pair (code/mask)of 256x8 bit arrays or a pair (code/mask) of 256x1 byte vectors. For the case of proprietary algorithms a detailed description may not be available. Size is important to system designers faced with storage and transmission of data across interfaces, buses, or networks.

11. **Report on the computational efficiency**: While precise timing of an implementation is subject to many systematic effects, papers should include coarse estimates of duration of the feature extraction and comparison functions, and should identify the hardware and software used. Where relevant, training time should also be provided. These metrics are important in any attempt to reproduce the work.

3. Proposal for a Generally Available Set of ROC plots

An author attempting to implement our recommendation to compare their results against baseline and state of the practice results might reasonably raise the question of where they can obtain ROC data for such a comparison – they might not have access to the baseline or state of practice algorithms. We propose an answer based on the availability of data in the NIST Face Recognition Vendor Test (FRVT) [24], Proprietary Fingerprint Template Evaluation (PFT) [27] and IRis EXchange (IREX) [25] programs.

We suggest that all groups (e.g. NIST, International Biometric Group) involved in large scale testing of biometric algorithms make ROC data from such tests generally available to the community in machine readable form. At this time, we recommend inclusion of face, fingerprint and iris modalities. For each modality, we recommend publication of ROC plot data in machine readable form for current, state of practice algorithms, as well as one or more open source algorithms. For preparation of this data, we recommend the use of both large, realistic, operational, sequestered data sets such as those used for the FRVT, PFT-II and IREX programs, as well as several of the smaller, publicly available datasets such as:

- Face:
  1. NIST Special Database 32[4]
  2. NIST / IJB-A[15]
- Finger: NIST Special Datasets 14 and 29 [23]
- Iris: CASIA-Iris-Thousand [2], IrisBase [12]

The authors are accepting suggestions on details from the community at large, including recommendations on public datasets and open source algorithms. Suggestions can be directed to the first author with a subject of the form "ROC plot Suggestions: <your description>".

4. Some Technical Complications

Both the DET and ROC carry the implicit assumption that the comparisons made at every threshold are independent, identically distributed (i.i.d.). Of course, in the real world this is never the case. Models of large-scale performance of biometric systems under this assumption have been proposed [36]. The approximate accuracy of the i.i.d. model for large-scale systems constructed to perform searches of a large database of N samples as N sequential binary comparisons was established in Section 5.4 of Wilson [37]. More advanced models have been suggested which assume that the comparison scores have distributions that are not identical, but vary based on the two samples being compared [30].

Significant complications arose with the adoption of an international standard for reporting biometric performance [13]. This standard allows the redefining of false negatives and positives around the concept of a candidate list. The false negative identification rate (FNIR) was defined as the proportion of transactions by users enrolled in the system in which the user's correct identifier is not among those returned on the candidate list. The false positive identification rate (FPIR) is defined accordingly. The criteria required for an identifier to be returned on the candidate list will be system dependent and may require a comparison score above a threshold and a ranking of the score against all comparison scores to be below some cut-off value (listing identifiers with the top P comparison scores). Consequently, when an ROC or DET is plotted using these definitions of false positive and false negative, the plot becomes a function of the threshold, the size of the database, N, and the particular value of P chosen. Consequently, the plot does not necessarily converge as the size of the database and number of comparisons increases, and to date, we lack good models relating FPIR and FNIR to N, P and threshold. Some reports [10] plot error rates in simple threshold, while others [37] plot error rates using the candidate list criteria, as a function of database size, ranking cutoff and threshold. The meaning of the DET or ROC under the candidate list conceptualization must be interpreted accordingly.

In the context of this paper, we consider two generic classes of matching tests: 1:1 and 1:N. In a 1:1 test a single biometric sample is compared to another single biometric sample and a match score is reported; the match threshold is not normally supplied to the match function. The process is repeated for K mated and L non-mated pairs. Using the reported scores and the ground truth for the pairs, we can construct the mated and non-mated score distributions and from those distributions construct an ROC plot using the inferred false match and false non-match rates as a function of the match threshold.

For the 1:N case we compare a probe biometric sample against a gallery of N reference samples, possibly specifying a match threshold on the call to the function implementing the 1:N match. The return values from the call depend on the algorithm implementation: a failure to match flag, a vector of all the match scores, the single best match score, and possibly some other variants.

If we can compel the 1:N matching function to return, for each probe, the scores for the top M matches in the gallery, we can set the returned number to the number in the gallery – thereby getting the match score of the probes against all N of the samples in the gallery. This generates a pseudo 1:1 result that can be combined with ground truth to generate mated and non-mated distributions. This presumes that we know enough about the implementation of the 1:N logic to be assured that each of the returned scores is the result of the same comparison operation: that the matching function is not taking advantage of its knowledge of the N comparisons to modify individual comparisons, e.g. by renormalization of scores. Hence, in some 1:N cases, given an adequate model of how the 1:N machinery is implemented, we can recover distributions corresponding to 1:1; in others, we cannot. For our proposal, this has implications for use of 1:N data and algorithms from large testing projects. This is an issue that will need to be considered in more detail as we go forward.

For the ROC plot and table presented in this paper, we limited ourselves to data from 1:1 tests, and recommend doing so whenever possible.

5. Example Graphic

Figure 1 is an example of the type of figure we are recommending. The data in figure 1 and in table 1 came from FRVT [24]. It represents a comparison of two different types of face data: mugshot and webcam with the same state of the practice recognition algorithm.

The figure illustrates several presentation best practices that should be noted:

- The data is plotted logarithmically so that the often most important region in the bottom left can be easily understood.
- Points with equal threshold values are indicated.
- The data being compared is plotted on the same graph; when plotting multiple graphs, use the same scale/aspect ratios, so that they can be easily compared.
- The legend uses both color and line type to make it possible to distinguish the lines when reproduced in black/white or if the reader is color blind (several percent of the population).

In accordance with our recommendations regarding statements of data and algorithm availability we note: (1)

The datasets on which the figure and table are based are not generally available. They are from the sequestered data used for NIST tests and for which we have no authority to redistribute. (2) The algorithm used is one provided by a vendor in the recent FRVT-2013 test reported in NIST-IR 8009 [9]. The terms of the algorithm license preclude disclosure of the vendor details.

We used this data and algorithm, in part, to illustrate an instance where the data and algorithm cannot be made available – but where those facts should never-the-less be disclosed.

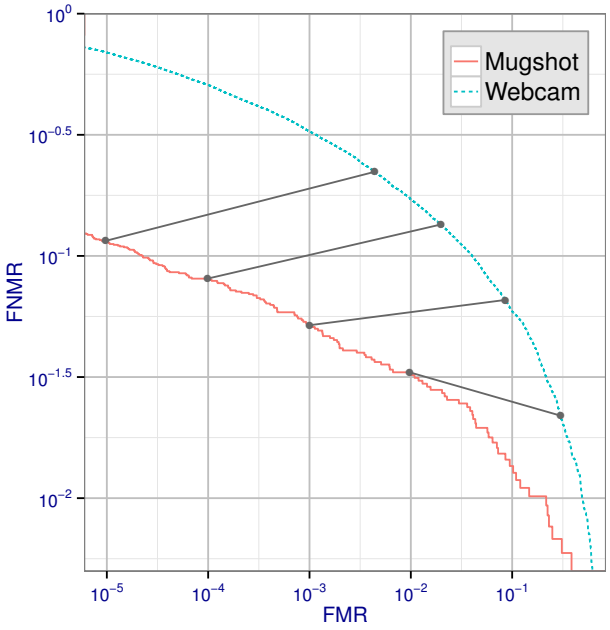

Figure 1. ROC plots comparing a state of the practice face recognition algorithm on two different data sets. The straight lines join points of equal threshold. See table 1 for a data listing.

6. Summary

We presented several suggestions for improvements in the preparation of biometric papers that center around appropriate use of ROC/DET plots to enable comparisons of the the performance of biometric recognition systems, including algorithms and data acquisition. To facilitate such comparisons, we propose implementation of a repository of ROC/DET data that includes results from open source baseline algorithms as well as state of the art commercial algorithms using both public data sets and large, operational, sequestered datasets. We look to the community for recommendations on which baseline algorithms and public data sets should be included for each of the major modalities: face, finger, iris.

7. Acknowledgments

Our thanks to Prof. Rob Ives (USNA), Mike Garris (NIST) and Dr. Joe Campbell (MIT Lincoln Laboratory) for suggestions that improved the clarity and readability of this paper; we specifically thank Dr. Campbell for information regarding the origins of the Receiver Operating Characteristic.

8. Disclaimers

In no case does mention of any commercial product, trade name, or vendor, imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

Opinions and recommendations made in this report are those of the authors and do not necessarily represent the policy of NIST or the US Government.

Table 1. ROC data for figure 1. The data has been interpolated on a grid to fit into the space available. Contact authors for the full data set. Columns 2 and 3 are FNMR; columns 4 and 5 are thresholds.

FMR	Mugshot	Webcam	WebCam-t	Mugshot-t
0.000001	0.158	0.835	3971.0	7646.3
0.000002	0.145	0.792	3823.0	7109.2
0.000004	0.129	0.750	3669.7	6645.8
0.000008	0.118	0.708	3521.3	6220.6
0.00001	0.115	0.691	3472.5	6082.5
0.00002	0.103	0.638	3325.1	5695.5
0.00004	0.087	0.581	3180.2	5335.0
0.00008	0.081	0.524	3038.8	5001.4
0.0001	0.080	0.508	2990.0	4905.1
0.0002	0.072	0.447	2850.3	4608.5
0.0004	0.064	0.399	2713.2	4330.1
0.0008	0.056	0.345	2579.3	4067.1
0.001	0.052	0.327	2536.5	3984.8
0.002	0.043	0.278	2406.9	3736.4
0.004	0.038	0.230	2281.9	3497.6
0.008	0.034	0.185	2164.7	3266.7
0.01	0.033	0.171	2124.2	3193.7
0.02	0.028	0.134	2012.1	2971.1
0.04	0.023	0.101	1906.4	2752.2
0.08	0.016	0.069	1807.4	2535.1
0.1	0.014	0.059	1775.9	2465.2
0.2	0.010	0.035	1680.9	2245.2

References

[1] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[2] CASIA. Biometrics Ideal Test (databases). `http://biometrics.idealtest.org/index.jsp`. Accessed 2015-03-27.

[3] CRAN. Comprehensive R Archive Network. `http://cran.r-project.org/`.

[4] S. Curry, D. Founds, J. Marques, N. M. Orlans, and C. N. Watson. MEDS - Multiple Encounter Deceased Subject Face Database - NIST Special Database 32. NIST Interagency Report 7679, National Institute of Standards and Technology, 2011. http://www.nist.gov/itl/iad/ig/sd32.cfm.

[5] Economist. Problems with scientific research: How science goes wrong, 2013. http://www.economist.com/node/21588069/print. Accessed 2015-03-27.

[6] Economist. Unreliable research: Trouble at the lab, 2013. http://www.economist.com/node/21588057/print. Accessed 2015-03-27.

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

[8] G. Government. Passgesetz [Passport Act], April 1986. German law, as amended.

[9] P. Grother and M. Ngan. Face recognition vendor test (FRVT:) performance of face identification algorithms. *NIST Interagency Report*, 8009, 5 2014.

[10] P. Grother, E. Tabassi, G. Quinn, and W. Salamon. Performance of iris recognition algorithms on standard images. *NIST Interagency Report*, 7629:1–120, 2009.

[11] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.

[12] IrisBase. IrisBase. http://www.smartsensors.co.uk/irisweb/. Accessed 2015-03-27.

[13] ISO. IEC 19795-1: Information technology-biometric performance testing and reporting-part 1: Principles and framework. Technical report, ISO/IEC, 2006.

[14] A. Jain, B. Klare, and A. Ross. Guidelines for best practices in biometrics research. In *IEEE International Conf. Biometrics, Phuket, Thailand*, 2015.

[15] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 1931–1939, June 2015.

[16] A. J. Mansfield and J. L. Wayman. *Best practices in testing and reporting performance of biometric devices*. Centre for Mathematics and Scientific Computing, National Physical Laboratory Teddington, Middlesex, UK, 2002.

[17] J. Marcum. US Air Force Project Rand Memorandum RM-754: A statistical theory of target detection by pulsed radar. http://www.rand.org/pubs/research_memoranda/RM754.html, December 1947.

[18] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The det curve in assessment of detection task performance. Technical report, DTIC Document, 1997.

[19] L. Masek and P. Kovesi. Matlab source code for a biometric identification system based on iris patterns. Technical report, The School of Computer Science and Software Engineering, The University of Western Australia, 2003.

[20] C. E. Metz. Basic principles of ROC analysis. *Seminars in nuclear medicine*, 8(4):283–298, 1978.

[21] J. Naik, L. P. Netsch, and G. R. Doddington. Speaker verification over long distance telephone lines. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pages 524–527. IEEE, 1989.

[22] Nature. Challenges in Irreproducible Research. http://www.nature.com/news/reproducibility-1.17552. Accessed: 2015-6-25.

[23] NIST. Biometric Special Databases and Software. http://www.nist.gov/itl/iad/ig/special_dbases.cfm. Accessed 2015.03.23.

[24] NIST. Face Recognition Vendor Test. http://www.nist.gov/itl/iad/ig/frvt-home.cfm. Accessed 2015-03-27.

[25] NIST. IREX Overview. http://www.nist.gov/itl/iad/ig/irex.cfm. Accessed 2015-03-27.

[26] NIST. Multimodal Information Group Tools. http://www.nist.gov/itl/iad/mig/tools.cfm.

[27] NIST. Proprietary fingerprint template evaluation II. http://www.nist.gov/itl/iad/ig/pftii.cfm. Accessed 2015-03-28.

[28] NIST. Speaker Recognition Evaluation. http://www.itl.nist.gov/iad/mig/tests/spk/.

[29] D. O. North. An analysis of the factors which determine signal/noise discrimination in pulsed-carrier systems. Technical Report PTR-6C, RCA Laboratories. Available from Hagley Museum and Library www.hagley.org, 1943.

[30] M. E. Schuckers. Using the beta-binomial distribution to assess performance of a biometric identification device. *International Journal of Image and Graphics*, 3(03):523–529, 2003.

[31] J. A. Swets, editor. *Signal Detection and Recognition by Human Observers*. John Wiley and Sons, New York, 1964.

[32] J. A. Swets and G. D. M. *Signal Detection Theory and Psychophysics*. John Wiley and Sons, New York, 1966.

[33] E. R. Tufte. *Envisioning Information*. Graphics Press, 1990.

[34] H. Urkowitz. Energy detection of unknown deterministic signals. *Proceedings of the IEEE*, 55(4):523–531, 1967.

[35] P. Vandewalle, J. Kovacevic, and M. Vetterli. Reproducible research in signal processing. *Signal Processing Magazine, IEEE*, 26(3):37–47, 2009.

[36] J. L. Wayman. Error-rate equations for the general biometric system. *IEEE Robotics and Automation Magazine*, pages 35–48, March 1999.

[37] C. Wilson, R. Hicklin, H. Korves, B. Ulery, M. Zoepfl, M. Bone, P. Grother, R. Micheals, S. Otto, and C. Watson. *Fingerprint Vendor Technology Evaluation, FpVTE*. National Institute of Standards and Technology, NISTIR 7123 edition, June 2004.