Appeared in Proc. of 8th IAPR International Conference on Biometrics, (Phuket, Thailand), May 2015

Guidelines for Best Practices in Biometrics Research

Anil Jain
Michigan State University

Brendan Klare
Noblis

Arun Ross
Michigan State University

Abstract

Biometric recognition has undoubtedly made great strides over the past 50 years. To ensure that current academic research in biometrics has a positive impact on future technological developments, this paper documents some guidelines encouraging researchers to focus on high-impact problems, develop solutions that are practically viable, report results using sound experimental and evaluation protocols, and justify claims based on verifiable facts. The intent is to ensure that methods and results published in the literature have been properly evaluated and are practically feasible for automated or semi-automated human recognition. It is believed that following these guidelines will avoid inflated claims and support published research on a legitimate foundation that can be embraced by practitioners and peers in biometrics and related scientific disciplines (e.g., forensic science).

1. Introduction

We provide a collection of guidelines for the proper design and evaluation of biometric algorithms and systems by researchers. It is our desire that the guidelines presented here will lead to a more effective and high quality research agenda marked by a judicious choice of problems. We hope these guidelines will (i) further the pace of innovation in biometric recognition and (ii) increase the likelihood that results obtained in a laboratory setting will generalize to operational scenarios. In turn, this will lead to biometric solutions that can be rapidly transitioned into practical applications that improve both system efficacy and security on the one hand, and user convenience and privacy on the other.

This article is inspired by George Nagy's 1983 paper titled "Candide's Practical Principles of Experimental Pattern Recognition" [1], where the author used satire to emphasize some of the common mistakes and inappropriate assumptions made by researchers in experimental pattern recognition. In the same vein, but without the satire, we wish to point out certain practices in the biometrics community (as reflected in published papers) that can undermine the effectiveness of the research while conveying a false sense of progress. Just as Voltaire's magnum opus "Candide, ou l'Optimisme" (1759) highlighted some of the flaws of Leibnizian optimism prevalent at that time and espoused a practical philosophy, it is necessary for the biometrics community (including the authors of this article) to revisit its research agenda, approaches, methodologies, and empirical analysis in order to maximize the broad impact of biometrics research. In the words of Candide, "we must cultivate our garden".

2. Biometric Recognition

For the purpose of this paper, we adopt the following definition of biometrics: "*Automated recognition of individuals based on their behavioral and biological* characteristics" *[ISO/IEC JTC1 2382-37:2012].* There are two important aspects to this definition: (i) recognition of *individuals*, and (ii) the use of *automated* methods. These tasks, in turn, require the coupling of a biometric recognition system to a particular application, where individuals in a population of interest need to be recognized. Further, the intended application may, itself, impose certain restrictions and requirements on a biometric system in the form of sensing modalities, computing resources, recognition accuracy, expected throughput, mode of operation (1:1 comparison vs. 1:N comparison), cost-benefit analysis, usability, level of security and privacy, and so on. For example, there is a distinct difference between using fingerprints to apprehend suspects in law enforcement applications and using them for unlocking a mobile phone.

The choice of research topic has a significant impact on the progress in biometrics. On the one hand, when researchers focus obsessively on improving, say, the matching accuracy on a specific dataset, then the generalization ability of the ensuing solutions may be suspect; further, the chances of innovating new paradigms will be low. On the other hand, if the focus is only on "creating" new problems under the guise of innovation (say, exploring a new trait X or fusing traits Y and Z), then the practical utility of the ensuing research may be limited. Hence, we postulate that:

The choice of biometrics research problems pursued by the biometrics research community is as important as the innovative solutions proposed for a particular recognition problem. Offering a compelling justification

for pursuing a particular line of research is necessary. Comprehensive research programs that focus on both foundational and practical aspects are recommended.

Assuming a suitable choice of research problem, the following guidelines are presented as a means to ensure that research is relevant and ensuing claims are firmly grounded in facts rather than speculations. It must be noted that the biometrics field has significantly matured over the past few decades and, therefore, these guidelines are more applicable to current research and should not be retroactively used to evaluate research that was conducted during the early days of biometrics.

3. Guidelines

3.1. Fundamental tenets of biometrics

Biometric recognition is based on *two central tenets*: *distinctiveness* (individuality) and *persistence* (permanence) of biometric traits. Ideally, a biometric trait should be able to perfectly distinguish all individuals in the population of interest and from those not included in the population (i.e., the open-set problem discussed later). Further, this capability to distinguish individuals should not diminish over time. Surprisingly, our knowledge about uniqueness and persistence for even the three most commonly used biometric traits (fingerprint, face, and iris) is incomplete. Claims such as "Trait X is highly distinctive" or "Trait Y does not change over time" cannot be made in the literature, unless such claims have been reliably evaluated.

Guideline 1: Without analyzing the distinctiveness and persistence of a biometric trait for the population of interest, at sufficient scale, it is unreasonable to make strong claims about the recognition accuracy and utility of the trait in large-scale applications. Any such claims must be tempered by clearly stating the caveats and scope of the reported results.

3.2. Application domain

Given the vast range of applications where biometric systems have been deployed (mobile phones, international border crossing, national ID programs), almost all aspects of a biometric system (such as choice of biometric trait, accuracy, recognition time, template size, operating environment) are conditioned on application requirements. Therefore, engineering oriented research (as opposed to fundamental science research) has to keep the end application in perspective.

Guideline 2: While an end application is not necessary for high quality research, in engineering oriented research, it is essential to keep an application in perspective in order to evaluate the utility of the research. The experimental protocols and evaluation metrics adopted must reflect the application envisioned by the researchers.

3.3. Choice of biometric trait

In principle, any anatomical, behavioral, or physiological characteristic of an individual can be used as a biometric trait. However, the choice of a trait may depend on the degree to which the following properties are satisfied with respect to the requirements of the application [2]: (i) distinctiveness, (ii) permanence, (iii) universality, (iv) collectability, (v) system performance, (vi) ergonomics, (vii) vulnerability to attacks, and (viii) usability. In practice, two-factor or three-factor authentication involving a combination of the following, viz., biometric trait(s), password, and token, may be needed.

Guideline 3: Research that explores a new biometric trait must evaluate its potential in terms of the properties enumerated above for the target application.

3.4. Comparing biometric systems

Operational systems [1] need to meet many application-specific requirements and constraints in addition to recognition accuracy. These include template size, speed or throughput (number of individuals recognized per unit of time), enrollment time, and level of system security and user privacy provided. Thus, it is necessary to consider these additional factors, besides recognition accuracy, when analyzing a new approach or algorithm.

Guideline 4: A marginal improvement in the recognition accuracy, especially on a small database, should not be heralded as a significant achievement, unless it is accompanied by an improvement in other system performance measures.

Guideline 5: When comparing the recognition accuracies of two competing algorithms, it is necessary to report the statistical significance of the performance difference.

3.5. Baseline

An inappropriate baseline for recognition accuracy

[1] The phrase "operational system" is used here to denote the utilization of a biometric system in a "real-world" application.

provides a false sense of progress. A proper baseline should be representative of the state-of-the-art performance for the problem being addressed. This comparison with the baseline can be undertaken by: (a) using publicly available code that represents the state-of-the-art matcher; (b) using a state-of-the-art commercial matcher that is known to perform well based on third party evaluations; or (c) executing the proposed method on a dataset for which the recognition accuracy of the best performing matcher is known.

Guideline 6: It is imperative that a new algorithm or system be compared against a baseline that represents the state-of-the-art for the particular recognition problem being solved.

3.6. Evaluating biometric system components

Often, it is necessary to focus on improving a single component of a biometric system, such as data acquisition, segmentation, alignment, feature extraction, or matching. In addition to ensuring a state-of-the-art baseline for the component in question, the input(s) and output(s) of the proposed and baseline components should be the same. For example, if the goal is to demonstrate the efficacy of a new minutiae matching method, then the minutiae matching module has to be isolated from the end-to-end system, and the input to the new minutiae matcher and the existing state-of-the-art matcher should be exactly the same. Further, the outputs of both the matchers should be subjected to the same scoring function.

Guideline 7: Claims regarding the superiority of a novel component within a specific biometric system can only be made if that component is isolated from the overall system and has been evaluated with respect to a proper baseline.

3.7. Choice of accuracy metric

The two most common metrics used to report biometric system matching accuracy are the *Cumulative Match Characteristic (CMC) curve* and *Receiver Operating Characteristic (ROC) curve*.[2] But, the CMC curve is only applicable for *closed-set identification* and not *open-set identification* (where the true mate of the probe may not be present in the gallery). Consequently, the CMC curve may not accurately characterize the performance of a biometric system [7]. Open-set identification performance is typically reported in terms of False Positive Identification Rate (FPIR) and False Negative Identification Rate (FNIR) [6].

When reporting the ROC curve, the intended application will dictate the *operating range* (threshold on the match score) where competing systems should be evaluated. There are not many applications where a False Match Rate (FMR) above 1.0% is acceptable, so the ROC curve should be appropriately scaled. Similarly, *equal error rate (EER)* of a system may not always provide useful information, as it is independent of the application-specific FMR. A confidence band around the ROC curve should also be reported to understand the robustness of the solution.

Guideline 8: The matching accuracy of a biometric system should be reported using ROC and CMC curves. A CMC curve should not be reported without the accompanying ROC curve, unless the intended target application operates in the closed-set identification mode. For open-set identification, a graph plotting the FPIR against the FNIR at various thresholds must be reported.

3.8. Choice of datasets

Biometric researchers often have a choice of databases to use that differ in terms of number of subjects, number of images/subject, capturing environment, etc. When evaluating the potential of a new matcher or new biometric trait, the data should exhibit the intra-class variations that are likely to be observed in practical applications. When selecting a database, researchers have to focus on the problem being addressed and use the database that is relevant to the problem. Real operational or laboratory data [5] should be preferred for evaluation rather than synthesized data. In case it is difficult to obtain real data, the generation of synthesized data should consider the end-to-end process in the biometric system. For example, to demonstrate that a face recognition algorithm is invariant to occlusion, it is not appropriate to introduce synthesized occlusions on well-aligned face images, because in this case, the impact of occlusion on face detection and alignment cannot be considered (also see Section 3.6).

The number of subjects in the dataset is known to impact the recognition accuracy. This is borne out in the FRVT report [6] that points out: *"As more identities are enrolled into a biometric system, the possibility of a false positive increases due to lookalike faces that yield extreme values in the tail of the non-mate score distribution"*. Thus, the potential of a newly proposed biometric recognition algorithm should be evaluated on datasets with a similar scale as that of the intended

[2] Besides the ROC curve, the Detection Error Tradeoff (DET) curve can also be used to summarize verification performance.

application.

Guideline 9: Large and challenging datasets corresponding to real operational or laboratory data should be used for evaluation in order to demonstrate the benefits of the proposed algorithm.

3.9. Generalization across datasets

Recognition accuracy can vary dramatically based on the type of data used in evaluation. In this regard, some of the findings of NIST FpVTE2003 [3] are quite noteworthy:

- *"The FRR [false reject rate] for a system often varied by a factor of 2 or more between different datasets."*
- *"Projections from measurements on one type of data to operational performance on another type of data are questionable."*
- *"Accuracy on controlled data is significantly higher than accuracy on operational data."*

To ensure the robustness of biometrics systems, it is necessary to train and evaluate them on data with characteristics similar to what would be encountered in the end application [5]. For example, are the demographic of the subjects and data quality representative of the intended use case population? An ill-advised practice in data collection is to capture all samples of a same subject in a single session that limits intra-class variability; this can lead to significant overestimation of the accuracy of a biometric recognition system.

Guideline 10: Data used for evaluating biometric systems should be representative of the population and environment where the biometric system will be deployed. Operational evaluation should be done with the data at the site where the system will be fielded. Further, the data for testing should be acquired over multiple sessions spanning over a period of time.

3.10. Training, validation, and test sets

The vast majority of biometrics applications involve training the system using a transfer-learning paradigm where models must first be developed using data from a set of subjects that does not overlap with subjects in the target population; further, a separate validation set may be needed to tune the learned model parameters. There are three primary reasons for ensuring that subjects in the training, validation, and test sets do not overlap: (i) the specific target population is typically unknown at the time the system is trained, (ii) operational databases are dynamic, generally making it infeasible to update models, and (iii) operational databases often lack the ground truth required for training.

Guideline 11: To avoid positive bias in stated recognition results, subjects contained in the training, validation, and testing sets should be non-overlapping.

3.11. Experimental protocol

Details regarding experimental design, database size and characteristics (e.g., number of subjects, number of images/subject, demographic distribution of the subjects), data collection environment and sensor types, cross-validation, and comparison metrics used should be clearly stated. This is essential for others to reproduce the results [4].

Data quality is known to impact the recognition accuracy of a biometric system. For example, the 2013 Face Recognition Vendor Test (FRVT) report states: *"Improvement of image quality is the largest contributing factor to recognition accuracy"* [6].

Guideline 12: Published results must include pertinent details about the experimental protocol and characteristics of the dataset. Error bars, denoting the variance in performance, must be reported. Further, scenarios where an algorithm or method fails must be documented.

3.12. Establishing the ground truth

In the case of biometric data, labels associated with individual samples include subject identity and possibly some demographic attributes (e.g., gender, age, and race). But how reliable is this ground truth information? The FpVTE report [3] emphasizes this issue about the reliability of ground truth labels in the context of fingerprints:

"Incorrect mating information is a pervasive problem for operational systems as well as evaluations, and limits the effective system accuracy. The effective accuracy of a system is bounded by the mating error rate of the underlying data. Mating errors were found in every source used in FpVTE... For example, the number of consolidations (cases in which the same person has fingerprint sets under different names or IDs) found and removed in FpVTE was 0.49%. If these had not been found and corrected, then FAR could not have been measured below 0.5%." [3]

Guideline 13: Ground truth labels must be carefully reviewed for their correctness.

3.13. Biometric fusion

Fusion of biometric traits was first utilized in the Bertillonage system [2] in the late 19th century. Additionally, fingerprint systems (such as AFIS) combine the scores from the friction ridge patterns of all ten fingers to recognize an individual. While, in principle, any two traits can be fused, it is preferred to combine traits that can be acquired in a single presentation (e.g., face and iris). Further, in the context of score-level fusion, using the sum rule with proper normalization has been observed to result in competitive performance.

Guideline 14: The improvement in recognition accuracy as a result of biometric fusion should be weighed against the associated overhead involved, such as additional sensing cost, enrollment and recognition times, computing resources, usability, etc.

3.14. Vulnerabilities of a biometric system

The two most commonly studied vulnerabilities in the context of biometrics involve presenting a spoof or altered biometric sample at the sensor, and tampering with the biometric templates stored in the system database. A number of countermeasures have been developed to detect or deflect these vulnerabilities. The efficacy (e.g., detection rates and speed) and robustness of proposed countermeasures must be systematically evaluated. The metrics for evaluation will depend upon the specific nature of the attack under consideration. Further, the feasibility and impact of an attack in the context of the overall system must be discussed [8]. For example, in the case of spoofing, the probability of success of an attack must at least be compared against a *zero-effort attack* where an impostor's biometric trait may match against an enrolled user by chance.

Guideline 15: Solutions for biometric system security must also assure that there is no significant loss in recognition accuracy.

4. Summary

Biometrics is a rapidly evolving field engaging a number of researchers from diverse academic fields and communities including pattern recognition, computer vision, signal processing, cryptography, and forensic science. Biometric recognition systems are being widely used in a number of applications ranging from international border crossings to unlocking mobile devices. Over the past couple of decades, a large number of scholarly articles have been published covering various topics in biometrics. However, there is a perceived gap between the requirements postulated by intended biometric application domain and the focus and solutions offered in many of these publications. While academic research should not be constrained by application requirements, in order to maximize its impact and usability, it is important to identify application domain(s) where the proposed research can be of potential value. To this end, we have offered some guidelines to researchers with regards to choice of problem, selection of biometric trait, and evaluation methodology. We reiterate that it is necessary and important for biometrics researchers to embrace a holistic research agenda that encompasses both the foundational science and practical engineering aspects of the technology.

Acknowledgments: *The authors are grateful to the ICB 2015 General Chairs and Program Chairs for their insightful feedback and suggestions. They would also like to thank Sunpreet Arora, Lacey Best-Rowden, Jianjiang Feng, Vincent Hsu, Scott Klum, Ajay Kumar, Zhifeng Li, Xiaoming Liu, Karthik Nandakumar, Charles Otto, Kris Ranganath, Richa Singh, Elham Tabassi, Kar-Ann Toh, Umut Uludag, Mayank Vatsa, Anne Wang, Soweon Yoon, A. Yoshida, P.C. Yuen and Qijun Zhao for providing valuable nuggets.*

References

[1] G. Nagy, "Candide's Practical Principles of Experimental Pattern Recognition," IEEE Transactions on PAMI, Vol. 5, No. 2, pp. 199 – 200, March 1983.

[2] A. K. Jain, A. Ross, K. Nandakumar, *Introduction to Biometrics*, Springer Publishers, 2011.

[3] C. Wilson et al., "Fingerprint Vendor Technology Evaluation 2003," NIST Technical Report, NISTIR 7123, June 2004

[4] J. M. Wicherts and M. Bakker, "Publish (your data) or (let the data) perish! Why not publish your data too?," Intelligence, Vol. 40, Issue 2, pp. 73 – 76, March–April 2012.

[5] A. J. Mansfield and J. L. Wayman, "Best Practices in Testing and Reporting Performance of Biometric Devices: Version 2.01", National Physical Laboratory Report, CMSC 14/02, United Kingdom, August 2002.

[6] P. Grother and M. Ngan, "Face Recognition Vendor Test (FRVT): Performance of Face Identification Algorithms," NIST Interagency Report 8009, May 2014.

[7] B. Decann and A. Ross, "Can a Poor Verification System be a Good Identification System? A Preliminary Study," Proc. of IEEE International Workshop on Information Forensics and Security (WIFS), (Tenerife, Spain), December 2012.

[8] N. K. Ratha, J. H. Connell, R. M. Bolle, "Biometrics Break-ins and Band-aids," Pattern Recognition Letters, Volume 24, Issue 13, pp. 2105-2113, September 2003.