Global ID: Biometrics for Billions of Identities

Anil K. Jain, *Life Fellow, IEEE*, Sharath Pankanti, *Fellow, IEEE*, Karthik Nandakumar, Salil Prabhakar, *Fellow, IEEE*, Sunpreet S. Arora, Anoop M. Namboodiri, and Arun Ross

Abstract—The world's population, which is currently (2017) estimated to be 7.5 Billion, is very likely to surpass 10 Billion by the turn of the century. While there are several challenges when dealing with a population of this magnitude, the ability to positively establish or refute an individual's identity is likely to be one of the fundamental expectations of a global society. In this article, we systematically discuss the issues impacting the design, implementation and deployment of a large-scale biometric identification system that can effectively manage and distinguish over 10 Billion identities. In this regard, we identify four technological issues that have to be satisfactorily resolved to design such a system: system scalability, identification accuracy, response time, template security and privacy. We discuss how the lessons learned from ongoing large-scale biometric systems such as UAE's Border Crossing System and India's National ID Card Program (Aadhaar) can be leveraged and incorporated into a Global ID system that handles 10B identities. Further, we study existing large-scale pattern recognition and machine learning systems, and determine how the challenges resident in such systems can be effectively addressed for use in the proposed Global ID system. Finally, we assess the gaps that need to be addressed by the research and development community-at-large for designing the Global ID system. We conclude that the outstanding research, engineering and design topics are "Grand Challenges" and, without a serious understanding of the underlying complex issues, simplistic identity infrastructure solutions will be dwarfed by the enormity of the identity problems of the next generation.

Index Terms—Biometrics, Identity, Authentication, Large-scale, De-duplication, Grand challenge, Privacy.

1 Introduction

The word individual has its roots in the word *indivisible* signifying the distinctiveness of a person, which gives rise to the concept of an *identity*. The notions of an individual, individual's rights/duties, and accountability of an individual's actions constitute the cornerstones of any free society. Operationalizing these notions warrants assigning a unique identifier to every individual in a society and creating a mechanism to verify an individual's claimed identity when required. A formal identity is quintessential for a country's residents to avail of its public services (e.g., healthcare or welfare benefits) and opportunities (e.g., education). While the need for a robust identity management infrastructure is acknowledged by various geo-political-social entities and some solutions are being experimented, these existing systems are typically fragmented (i.e., narrowly focused on a constituent subset of individuals in the community), meant to serve a specific single purpose (e.g., benefit disbursement), and closed in nature (i.e., access to the identity functionality cannot be easily extended to the other stakeholders) [17]. Furthermore, the issue of geographic and demographic exclusion of population groups has been identified as one of the key limitations of existing identity management solutions [3]. To make such a solution available and accessible globally, its core has to be necessarily digital and based on who the person is as opposed to something that they possess or know [10].

With increasing mobility, connectivity, and wider exposure, the world is shrinking into one small heterogeneous global community. Currently, there exists no solution that establishes an absolute frame of identity reference for all individuals in the world and is instantaneously accessible to all legitimate stakeholders everywhere. We would like to think of a system that will allow anyone to be authenticated anywhere in the world, at any time. Without such a universal system, it is obvious that excluded individuals or stakeholders would have to depend on exception handling procedures. This in turn subverts the fundamental value proposition of the identity solution, which is to guarantee fairness and accountability to all its constituents. Further, with increasing incidences of fraud, security, and inequities, the lack of a universal frame of reference for personal identity could turn out to be downright dangerous and catastrophic. Realizing the gravity of this situation, the United Nations Sustainable Development Goals (SDG) includes providing "legal identity for all, including birth registration" by 2030 [6] as one of the foremost requirements. The problem of identity management concerns every sector of government and industry across the world (see Table 1 for a list of major application domains). Therefore, the solution requires a wide-ranging movement involving all the stakeholders including the government, private industry, international agencies & non-governmental organizations (NGOs) as well as the broader civil society.

In order to establish a global identity system, one needs to ensure that every person in the world is able to enroll into the system and obtain an identifier, irrespective of their social, economic and cultural background. This implies that one need to base the system on factors that are universal to individuals irrespective of the above factors. Biometrics, which refers to the automated recognition of individuals based on their biological and behavioral characteristics [10], provides such a basis on which one can build an identity solution. There are several biometric modalities with varying degrees of discriminability, stability and ease of capture. Figure 2 shows few of the most commonly used biometric characteristics that are currently in use.

At this point, we note a global digital identity verification system with a single identity per person also leads to significant concerns on privacy and security of the enrolled users. The primary concerns of such a system are: 1) its

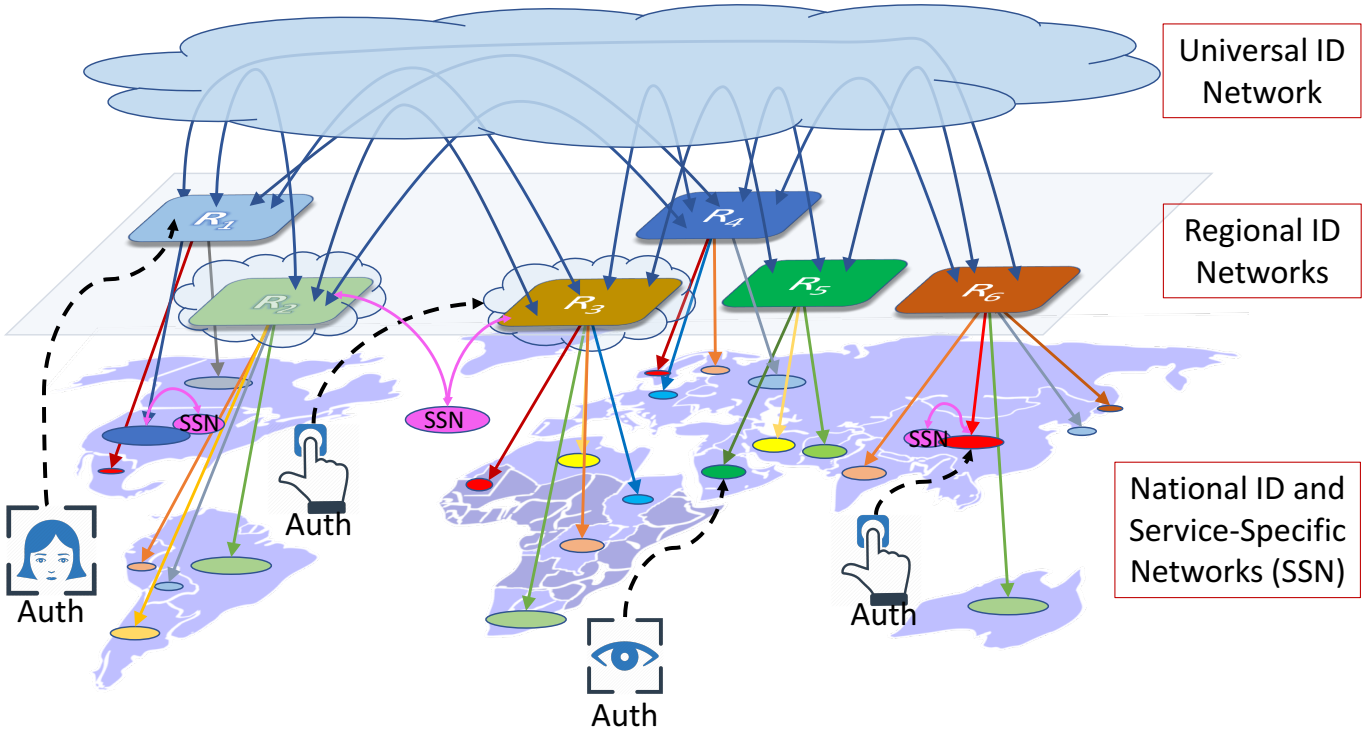

Fig. 1. A high-level overview of the Global ID system. It will be an interconnection of national, regional and service-specific networks (e.g., ATMs, railway networks) that agree upon the rules and protocols of information storage and exchange. The biometric information (core identity) would reside only within a user-trusted repository (say within the national network). The authentication process is distributed, where a request may arise from any part of the network and would be fulfilled by the user-trusted network or through a derived identity that the user has provided to a service provider (see Figure 9) at any other part of the network.

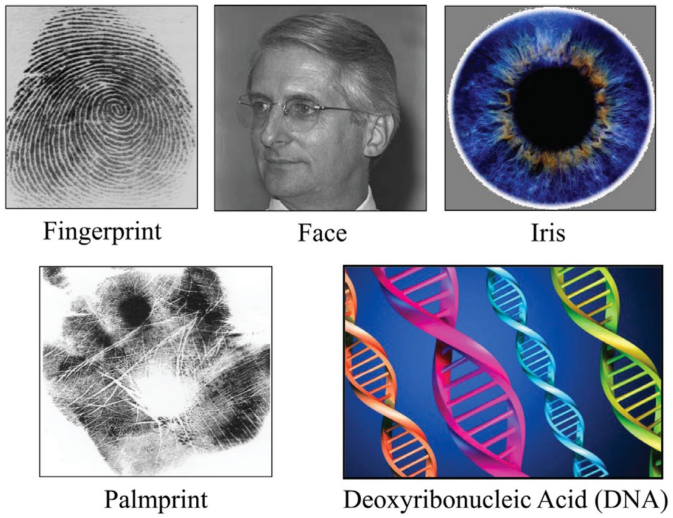

Fig. 2. Some of the popular biometric traits that have been used in large-scale identification systems. They vary in their discriminability, ease and cost of capture, stability of the trait over time and universality. Most large-scale biometric systems in existence use a combination of fingerprints, iris and face to create unique identities.

widespread use leads to the ability to track a person across multiple domains, 2) break of trust by an authentication service provider can lead to denial of services, 3) users being forced to reveal their true identity, where it is not required to do so, and 4) a security breach of the database will result in irreversible loss of data as biometric traits of users cannot be altered. We will look at potential directions to address these concerns in our proposed system.

One of the biggest challenges in establishing a global identity management system is the explosive increase in world population in the recent past. It is estimated that the world population will reach over 10 Billion before it stabilizes (see Figure 3). Note that even after world population stabilizes, the number of identities in the system will keep on increasing as it is difficult to reliably remove identities.

At the outset, we believe that the push for a global identity will have significant opposition due to the differences in level of trust between sovereign nations. However, the primary motivation of this work is to determine if it is indeed possible to define an identifier[1] or a set of identifiers, which can be used to build a person identification system that provides universal coverage and allows interoperability across a large variety of applications. In addition to a unique identifier for a person, a universal framework might also allow a person to create have several derived identifiers for privacy reasons (see Figure 9).

Several biometric ID systems were established in the last decade to achieve the goals of authentication at a national

1. In this paper, the terms *identifier* and *identity* will be used interchangeably. However, it must be noted that *identity* is a complex term that holistically encompasses the essence of an individual, while *identifier* denotes an external descriptor associated with the individual. In the case of non-biometric cues, the *identifier* is sometimes referred to as a *credential*.

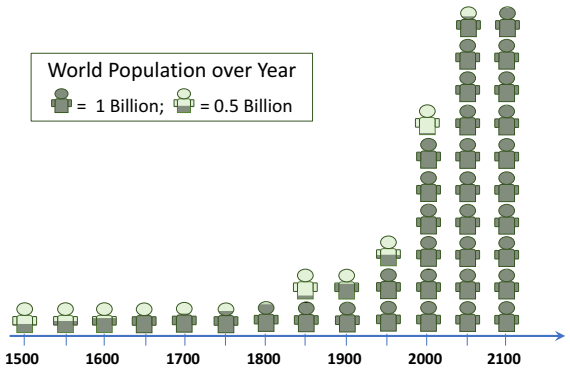

Fig. 3. World population reached 1 Billion in 1804, 6 Billion by 2000 and is expected to go over 10 Billion within this century before stabilizing or decreasing. Even at this stage, the database size of an identity system will keep increasing due to new enrolments. Each icon represents a billion people and partially shaded ones indicate fractions of 1 billion.

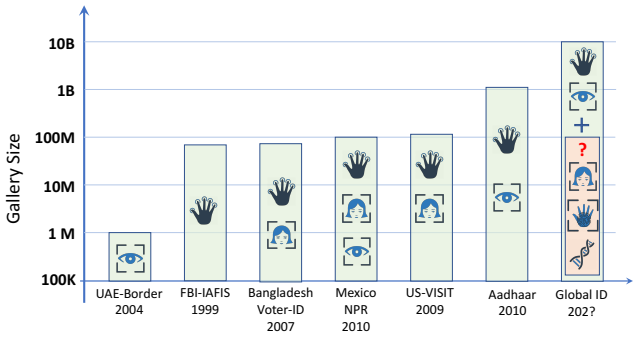

Fig. 4. Last decade has seen significant increases in the use of biometrics for large-scale identification. Several *National ID* projects have close to 100 million identities with the *Aadhaar* database crossing the 1 billion mark. A Global ID would require another order of magnitude increase in the number of identities.

level (see Figure 4). The first large-scale system that used automatic biometric matching was the FBI-AFIS system, which was followed by the IRIS-based UAE border security system and the Bangladesh voter ID system [9]. However, even the largest of them that exists currently is an order of magnitude smaller than what is required and the we need to understand the challenges that would arise to move towards a Global ID. Towards this end, the first key question to be answered is whether 10 Billion people can be uniquely distinguished, if required. Therefore, we will focus on the technological issues in scaling person identification systems to giga-scale. The other critical issue is the interoperability of a person's identifier across applications spanning geo-political-social barriers. Hence, we will also address the technological issues in facilitating ubiquity of person identification solutions.

Such a universal identity solution is different from any of the existing systems in three fundamental aspects:

1) **Scale**: The system will be an order of magnitude larger than any other system that exists in the world.
2) **Distributed nature**: The global system will necessarily be distributed across nations with each nation retaining the rights to the biometric data of their population.

TABLE 1
Application domains where identity management plays a vital role, along with an indication of scope (national vs global) and origin.

Application	Enterprise	National	Global
Border Control		✓	✓
Consumer Devices	✓		
Financial Transactions	✓		✓
Healthcare	✓	✓	
Law Enforcement		✓	✓
Forensic	✓	✓	✓
Cybersecurity	✓	✓	✓
Social Welfare		✓	

3) **Diversity**: In addition to any increase in variability of biometric data due to a global population, the system will have diversity from the modalities and sensors used by individual subsystems and the diversity of use cases that such a universal system will enable. Moreover, the individual subsystems will also be subject to the rules and regulations of the respective nations.

These differences give rise a set of challenges that needs to be addressed to allow the creation of such a system. Figure 5 summarizes these challenges. While all the factors listed above affects every challenge to some extent, we relate them here to their primary causes. We discuss each of these challenges in detail in the following sections.

To understand the challenges better, it is useful to visualize the nature of such a universal identification system (see Figure 1). Due to legal and data ownership constraints, the system has to be an aggregation of individual national ID systems and regional systems, where certain nations agree upon higher level of information exchange. These networks may use dedicated channels or interconnect through the internet.

We will discuss the design in greater detail in Section 5. The rest of the paper is organized as follows. After overviewing the high-level design goals of giga-scale person identification solutions, we will identify the major challenges in designing such systems. Next, we review the wider pattern recognition literature as well as existing large-scale biometric systems (Aadhaar case study and UAE Border Control Program [2]) to understand how the lessons learned

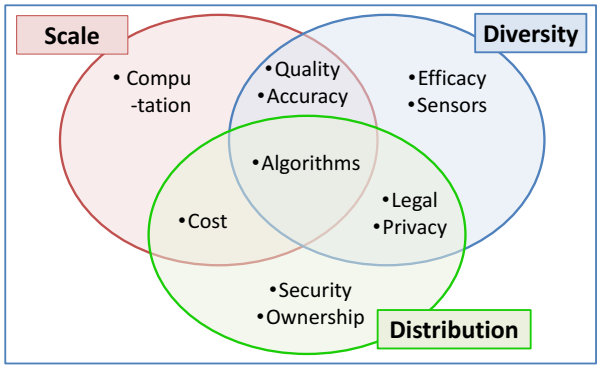

Fig. 5. Challenges of a global identity solution grouped by the primary causes of each. Each of these challenges could be compounded by multiple causes.

can be applied to solving this problem. Subsequently, we describe our proposed identity management solution for achieving universal ubiquitous person identification, in which giga-scale biometric identification is a critical component. Finally, we propose a general framework for giga-scale biometric identification systems and identify the research gaps (grand challenges) that need to be addressed.

2 High-level Design Goals

While universality and ubiquity are two fundamental design goals of a global person identification system, utmost care should be taken to ensure that identifiers and personal identity information do not fall into the hands of unauthorized individuals and organizations. Hence, the following aspects must also be considered when defining policy, legal and regulatory frameworks: data storage and integrity, confidentiality and authenticity of data, integration and access control over networks for data availability, and prevention of misuse/theft of data. Thus, universality, ubiquity, security, and privacy are bound to be four primary design aspects of an identity solution. Below, we describe each of these aspects and related design issues.

Universality: Assigning an identity for every individual necessarily entails distinguishing an individual from every other individual, often termed as identification or de-duplication problem.

In a universal person identification system, it is desirable that the establishment of absolute identity reference be based on solely on intrinsic features of the individual and relying on any features extrinsic to individual will defeat the purpose of the system. For example, many identity solutions today rely on a different authority (breeder documents) such as birth certificates, to establish the absolute reference and resulting system can only be as good as the integrity of the breeder documents [7]. Biometrics, the science of identifying an individual based on her unique physical, physiological, or behavioral traits, is the key for a universal identification system, where unique identification of each individual from amongst a very large population group is required. Only biometrics can achieve large-scale identity de-duplication[2] and eliminate the generation of fraudulent identity. To make a biometric recognition system universally applicable, a large repertoire of identifier choices (e.g., fingerprint, iris, face, etc.) should be offered to the users. Ideally, the selected identifiers should also be permanent so that they can be used from birth to death and beyond. Thus, demographic or biographic details of individuals cannot be considered as universal identifiers, because these surrogate identifiers are easily susceptible to change over the lifetime of an individual.

One of the primary design issues with large scale biometric identification systems is that of capacity. In this context, capacity refers to the number of individuals who can be reliably de-duplicated based on the information extractable from their biometric identifiers. In other words, sufficient number of discriminable bits of information should be extractable from every individual, which in turn dictates the types of biometric identifiers needs to be captured by the system. At giga-scale, it is critical to know practically what identification accuracy the system has and what is its identification throughput? When the biometric identification system is not sufficient to pin-point the identity, what other techniques can help bridge the gap?

Ubiquity: Enabling anytime and anywhere identification or verification of identity is primarily an infrastructure problem. Ensuring that all computation, communication, data resources are available, responsive, and usable, is a massive task. Gigascale operations warrant decentralized, highly redundant infrastructure and core operations that are based on interoperable and open standards. In case of highly distributed biometrics and sensing systems, the existing standards may need to be extended beyond the common biometric data formats to include common standards for sensor hardware, recognition software, and acceptable signal quality [13]. Additionally, the system is required to provide identity services to increasingly complex, heterogeneous, isolated, displaced individuals, ascertaining availability of trustworthy infrastructure becomes more challenging. For example, countries have thus far focused only on building national identity systems. Such systems are rendered ineffective when the individual travels across national borders. One of the fundamental questions is whether we can have a trusted and verifiable international identity document. Moreover, when extreme individual situations are encountered (e.g., sufficient identifiers of acceptable quality cannot be provided), adequate exception handling procedures are required to be incorporated into the system operation workflow.

Data & System Security: Given the diversity of services, stakeholders and scale of system, providing everywhere identity services will have broader set of security challenges than those encountered by the existing mainstream large scale systems such as credit cards [16]. Security design issues involve incorporating appropriate layers of defense involving encryption, certifications, and access control to ensure the integrity of the system while securing the access to only authorized users thwarting diverse adversarial threats by hackers, insiders, and malware. Finally, revocation of issued identifiers and related authorizations involve some of the most challenging security scenarios. In fact, protection against insider threats, presentation attacks, and security of stored identifiers are the major stumbling blocks in the adoption of large-scale identity management systems.

Privacy: Identity solutions should enable an individual to remain anonymous, if required and when permitted by law. A closely related topic of public acceptance depends upon the perception of control. One of the key design questions is: "Who owns and controls the data and system access?". Ideally, the individual must not only have the ability to control access to and usage of identity-related data, but also be able to manage/avoid cross-linkages across various applications. Fair and transparent policies, informed consent mechanisms, secure logging and audit of all interactions with the identity management system, and strict enforcement of legal liabilities among all stakeholders are some of the critical pieces in solving the privacy puzzle. Other Issues: In addition to the four primary design pillars described above, there are many other system, legal, regulatory & financial

2. On the contrary, an individual's identity can be typically verified by a biometric (e.g., face) identifier, non-biometric identifier (e.g., ATM card), or a combination thereof.

issues that an identity solution needs to deal with.

- How should the legacy identity solutions be handled?
- Should the proposed system start from scratch or build upon existing systems?
- Public vs private sector participation
- Who will pay for such an identity solution? How will the cost be shared among various entities?
- What happens if such a system runs contrary to existing international & national laws?
- Which entities should we trust? Who takes responsibility if something goes wrong?

Because of the universal nature of the system and its significant impact on our society, these topics are going to be very significantly more complex. We acknowledge that these issues need to be addressed appropriately for a successful identity solution and keep our focus on addressing the four technological issues as they relate to building a biometrics-based universal ubiquitous person identification system.

3 Technological Issues in Scaling Biometric Identification

Biometric characteristics are known to be unique to a person and does not vary significantly over their lifetimes. Hence these characteristics can form the basis of establishing an identity that is permanently tied to a specific individual.

A biometric universal ID system will need to provide two core functionalities. While enrolling a person or creating a new identity, it needs to compare the biometric traits of the enrollee against that of every other person in the system. This process is referred to as deduplication as it ensures the same person is not enrolled twice and given two different identities. During authentication or verification, a person provides their identity code obtained during enrollment along with a biometric and the system should be able to verify whether the biometric belongs to the person with the claimed identity. Note that the enrollment process is very compute intensive and will also require a larger set of biometric traits to ensure uniqueness. The authentication process is a one-to-one matching step, which is fast and may be done with a single fingerprint or face.

The technological challenges that need to be addressed arise from the scale, diversity and the distributed nature of a global ID solution. However, some of these could also force a solution that is more robust and responsive with enhanced privacy. We note that a final solution should also take into account the political realities and legal precedences and not designed purely from a technological perspective. We now look at the primary challenges in further detail.

3.1 The Accuracy Barrier

Deduplication of identities is an open-set biometric identification problem. The accuracy of a biometric identification system can be measured in terms of two error rates: false positive identification rate (FPIR) and false negative identification rate (FNIR). A false positive identification error occurs when the biometric samples of a person who has hitherto not been enrolled in the system is incorrectly matched with biometric data from an enrolled identity. Since such an error may lead to wrongful denial of identity/service to a legitimate individual, biometric samples from potential false positives are typically subject to re-assessment by human operators. Furthermore, other available identifiers (e.g., name, demographic details, legacy identity documents, etc.) are closely scrutinized in such cases to minimize false positive identification errors. On the other hand, a false negative identification error occurs when the biometric samples of an individual who is already enrolled in the system fails to match with his/her enrolled biometric data. False negative errors typically lead to the creation of duplicate identities for the same individual, which violates the fundamental purpose of a biometric deduplication system.

Both the identification error rates, FPIR and FNIR, are inherently related to the false match rate (FMR) and false non-match rate (FNMR) of the underlying biometric matcher. If we assume that (i) identification is performed by comparing the query biometric sample with the biometric data of each enrolled identity, (ii) match/non-match decisions are made individually for each biometric comparison, and (iii) errors resulting from each biometric comparison are independent, the relationship between FPIR, FNIR, FMR, and FNMR can be expressed as follows:

$$FNIR \approx FNMR$$
$$FNIR \approx 1 - (1 - FMR)^N$$

where N is the number of individuals already enrolled in the system. Note that the above equations are simple approximations, which are valid only under the specified assumptions. In practice, the errors of different biometric comparisons are seldom independent because the same query biometric sample is compared against all the enrolled data. Furthermore, the identification process often involves multiple filtering/indexing stages to avoid the need to directly compare the biometric query against all the enrolled data. Finally, the set of match scores generated based on the query biometric sample is rank-ordered, before a threshold is applied to determine a positive or negative identification. Despite these shortcomings, equations (1) and (2) are useful as they provide reasonable approximations to understand the challenges in designing a large-scale biometric identification system.

Suppose that we wish to design a 10G scale biometric deduplication system with a FNIR of 0.1% and FPIR of 0.1%. Assuming a worst case of 1 billion people attempting to acquire a duplicate identity, a 10G system can be expected to have around 1 million duplicate identities ($10^{-3} \times 10^9$). In other words, 99.99% of the identities in the system would indeed be unique. This is a worst case estimate and the actual number of duplicates could be significantly lower if there is a penalty associated with a person trying to acquire multiple identities.

Moreover, if the system allows 2 million enrollments per day (Note: around 350,000 babies are born each day according to UNICEF estimates), it may generate an average of 2000 false positives, which need to evaluate manually. To achieve this desired FNIR and FPIR, the FNMR of the underlying biometric matcher should be 0.1% and the corresponding FMR should be in the order of 1 in 10 trillion (10^{-13}).

We now look into the feasibility of achieving these accuracies. One can look at the best accuracies reported for the individual modalities as a starting point.

FpVTE 2012 [15]: The most accurate fingerprint identification submissions achieved FNIR of 0.09% at FPIR of 10-3 based on ten-finger IDFlats. 30,000 search subjects were used for these results (10,000 mates and 20,000 non-mates). The number of enrolled subjects was 3 million for IDFlats.

NIST IREX III [8]: The database used in this test contains approximately 6.1 million iris images acquired from nearly 4.3 million eyes. The best algorithm had a FNIR of approximately 2.5% when a single eye was used per person (0.7% for two eyes) and the threshold was set such that there were no more than 25 false positives in every 1013 iris comparisons (i.e., FMR of 2.5×10^{-12}).

Another related issue is how to first evaluate the accuracy of such large-scale systems prior to implementation. This is critical to determine the number of biometric identifiers that should be used to achieve the desired accuracy targets. One possible solution to estimate the performance of state-of-the-art biometric systems is to synthetically generate billions of biometric samples, and do empirical approximation. For example, Maltoni et al. [11] generated a large number of single-finger samples using an improved version of SFinGe, and used the minutia cylinder code (MCC) algorithm for evaluation. They showed that verification and identification error rates at fixed thresholds (1 in 1000 or 1 in 10000) stabilize and do not increase significantly with scale (1 billion).

An alternate approach to estimate the required accuracies would be to look at large scale multimodal systems like Aadhaar that are deployed in practice. The accuracy estimate of a deployed multimodal system will incorporate various factors such as data quality, diversity and correlations between modalities. We believe that this will provide a more accurate estimate that is relevant to our problem of building a 10G system. As given in Section 4.2, the Aadhaar system has reported an FPIR of 0.01% at an FNIR of 0.1% on a gallery of 84 million users using 10 fingerprints and 2 irises. Assuming the FMR rates remain the same, this would lead to a 1.2% FPIR at an FNIR of 0.1%. In other words, one would need to manually examine around 24,000 false positives out of 2 million enrollments a day or 120 million checks in total. To be practical one needs to improve the overall FMR of the multimodal system by 1 to 2 orders of magnitude.

One needs to advance the state-of-the-art in order to achieve the required recognition performance. This could be improvements in the performance of the individual modalities (use of level-3 features of fingerprints) or the use of additional modalities.

3.2 The Response Time Barrier

Enrollment: To generate an estimate of the computational requirements to realize such a massive system, one can once again look at the time required to match the individual modalities as a reference point.

FpVTE 2012 [15]: The best median identification time for the 30,000 queries was 4.26 seconds (0.7 million ten-finger matches per second). However, the most accurate algorithm was about 10 times slower. Cappelli et al. [4] have reported a throughput of more than 35 million fingerprint (single finger) matches per second on a simple PC with four Tesla C2075 GPUs. IREX III [11]: The average search time for iris recognition was 0.5 seconds for a database size of 4 million (8 million matches per second).

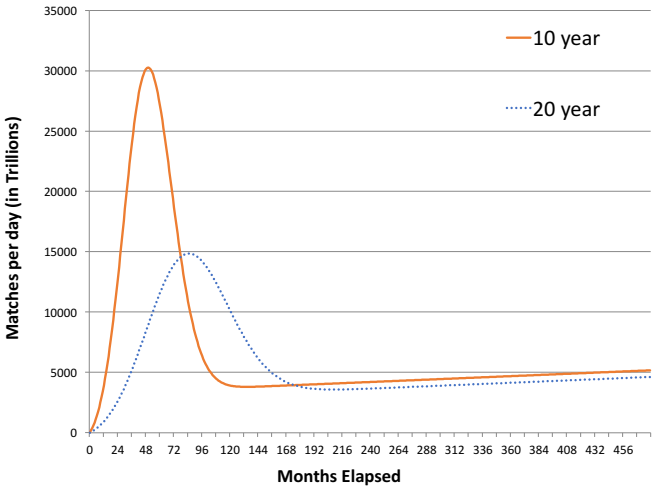

Fig. 6. Number of matches per day required to enrol 10G users in 10 and 20-year time frames, and then keep up with the birth rate. Note that the gallery size will continue to increase with enrolments.

While these times are precise, a practical system will involve a complex combination of multiple modalities and hence cannot be inferred from the individual systems. Moreover, one cannot rely on the use of specialized hardware as it would tie the hardware to a particular algorithm or vendor. Hence we look the largest functioning biometric system (Aadhaar), which uses a fusion of multiple modalities and try to extrapolate. Figure 6 provides an estimate of the number of matches per day to reach 10G, in 10 year and 20 year time frames.

An aggressive strategy of enrollment would hit a peak enrollment of around 5 million per day in 3 years, while a more realistic one would hit a peak of 2.5 million enrollments in 5 years. The computational peaks as observed in Figure 6 would be significantly different between the two cases. These requirements are around 20 to 40 times the peak computations that were required for the Aadhaar project, which was around 750 trillion matches a day. Aadhaar used around 3000 servers to meet these peak requirements (see Section 4.2). In terms of computational hardware, this would mean that a 10G database would require around 60K to 120K servers at the peak of deduplication. This is the size of a small to medium sized data center in 2017. In practice, one would have several micro data centers around the globe that would handle the regional databases as well as their mergers.

As seen from our estimates, the computational requirements are quite large and algorithmic improvements to the matching process to improve the speed without increasing the matching error rates has the potential to significantly decrease the cost of the overall system.

Authentication: Suppose that we have a universal and ubiquitous person identification system used by many or-

ganizations, what is the expected number of authentication requests that will be handled such a system? What will be the server and client side computational requirements? The other issue is the availability & reliability of communication infrastructure, especially in rural areas.

3.3 The Security Barrier

One of the most formidable challenges in designing large-scale biometric systems is database and system security. In this context, security can be broadly defined as protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction. In general, there are four major aspects of system security.

- Data confidentiality - prevent illegitimate access or disclosure of sensitive data or information.
- Integrity - guard against improper modification or destruction of the system/data and ensure non-repudiation and authenticity of information.
- Availability - guarantee timely and reliable access to and use of information.

In particular, the following issues require increased attention during system design.

- Enrollment fraud dependence on legacy documents, insider attacks (collusion/coercion), exception handling. Note that, unlike a traditional small-scale biometric system, the enrollment process of a large-scale system will be distributed across remote locations that are difficult to monitor
- Leakage of biometric information (all along the chain of custody from sensor to database)
- Spoofing and obfuscation (presentation) attacks
- Sabotage/denial-of-service/infrastructure reliability
- Device & Network security
- If the entire system is de-centralized, how to solve the trust issues among the various entities issuing identity?

3.4 The Privacy Challenge

Privacy is one of the major concerns in use of an identity verification system, where every authentication can be reliably tracked back to a specific identity. The widespread applications that a global ID system provides also multiplies these concerns and hence need to be addressed directly. An authentication or ID creation process involved two parties: the user and the identity provider or verifier. While there could be malicious third parties who try to inject themselves into the mix, we assume that this is handled as part of the security solution. The privacy issue arises primarily due to a conflict of interest between the two parties involved during the ID generation or verification process and we outline the challenges from this perspective.

- Identity Generation or Deduplication
  - *User wants to generate a core or derived ID, but the provider refuses.* Such a situation can arise if the user is unable to produce the relevant documents or if a government is actively suppressing a person or group. The technological solution should be able to address s
  - *The provider wants to do a de-duplication or identification without user's consent.* This can arise from unauthorized searches in a database based on biometric data gleaned from a user.
- Identity Verification
  - *User wants to authenticated, but the verifier refuses.* In addition to technological glitches, such a problem can arise in case of government oppression or in case of refugees who want to use the identity from their national network.
  - *The user does not want to reveal the core identity, but the verifier tries to do so.* This is one of the primary concerns of biometric identity verification and a solution should allow a user to be authenticated for a service, without revealing the core identity. This could be especially problematic in cases such as a person entering a witness protection program.
- Other Challenges
  - *Cross-database linkage.* Linking identities across multiple service providers can lead to significant loss of privacy as this can degenerate to a single universal database.
  - *Data mining on transactions linked to a single identity* can lead to loss of privacy as it generates implicit links between identities.

We address some of these challenges in a

4 Learning from Experience

Prior to taking a deep dive into the task of designing futuristic identity management systems that provide universal and ubiquitous person identification, it is important to delve into past and current approaches in designing similar or related large-scale systems in order to imbibe valuable lessons and avoid potential pitfalls. In particular, we focus on learning from the wider pattern recognition/machine learning community as well as existing large-scale biometric systems.

4.1 Lessons from Machine Learning

It is well-known that biometric recognition is fundamentally a pattern recognition/machine learning problem, which is the process of assigning class labels to given input samples based on features extracted from these samples. More specifically, biometric identification (de-duplication) is an example of open-set pattern classification. Given a query biometric sample, the goal is to assign it to one of the N known classes (enrolled identities) or assign it to the reject class (unknown identity).

While the science of pattern recognition has been evolving for nearly six decades, rapid data explosion and steady advancements in computing hardware have recently brought the issue of scalability to the fore. Scalability of a classification or machine learning problem can be measured along three dimensions, namely, (i) total number of classes/labels, (ii) number of training samples per class, and (iii) dimensionality of the feature space. A giga-scale biometric identification system pushes the boundaries of the underlying pattern recognition science to the extreme along all the three dimensions of scalability.

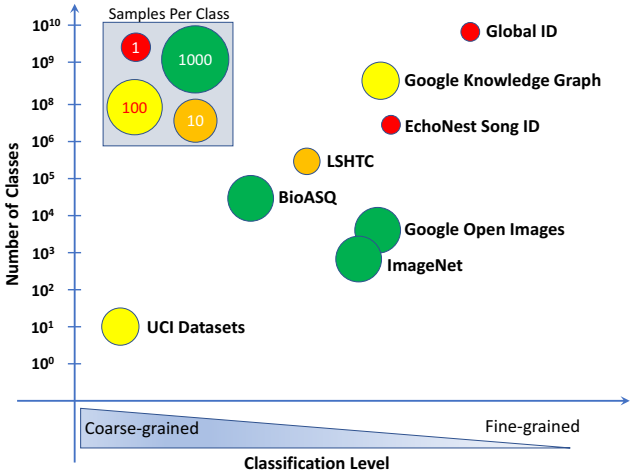

Fig. 7. Landscape of extreme classification tasks considered in the machine learning literature. The classification problem becomes challenging as one moves towards the top right, with fewer number of training samples per class.

A pattern recognition problem that deals with a large number of classes is generally known as extreme classification [1]. While most classical machine learning techniques can handle only a few classes (in the tens or hundreds), extreme classification problems involve a massive number of classes (in the thousands or millions). Real-world examples of such problems include image or video annotation (e.g., Google Open Images dataset has 9 million images with 6,000 category labels), multi-label document classification and semantic web indexing (e.g., the training data for BioAsQ [2017] challenge has 12.8 million documents with 27,773 labels at an average of 12.66 labels per document and the LSHTC challenge has 2.3 million documents with 325,056 categories), and product/ advertisement recommendation (e.g., Ads database with 9 million labels). Other potential extreme classification problems include tagging webpages with one or more of 570 million objects in the Google knowledge graph and identification of a song (from a list of 38 million songs in the Echo Nest database) based on a hummed tune. In contrast to the above example, biometric identification at the 10G scale involves classifying a query biometric sample into one of the 10 billion possible identities, which increases the challenges in designing extreme classification algorithms by several orders of magnitude (see Figure 7).

In most extreme classification tasks, the classes are organized as a hierarchical structure (either a tree or a graph) and the relationship between the class labels is known apriori [5]. Therefore, one of the foremost questions in extreme classification is how to leverage this semantic/dependence information between classes to improve the classification accuracy. Furthermore, multiple class labels are often assigned to the same sample. However, the biometric identification problem does not have any inherent hierarchical class structure (except probably fingerprint pattern classes or color of the eye or gender/ethnicity from face). It also requires a query biometric sample to be assigned to exactly one identity. Thus, it requires the design of an extreme single-label flat classifier, which is a problem that is seldom addressed in the pattern recognition literature.

Another important characteristic of biometric identification is sparsity of training data. Typically, biometric systems acquire only a single sample of the person's biometric identifier (although multiple biometric identifiers can be acquired) during enrollment. Therefore, identification of previously enrolled identities must be performed based on a single example and the system must also be capable of detecting a new identity even if it has never seen any samples from this identity before. While the former problem is commonly referred to as one-shot learning in the machine learning community, the latter is also known as zero-shot learning. The most common approach for zero-shot learning is learning some intermediate (semantic) attributes and indirectly mapping these attributes into new unseen classes [14]. Since there are no well-defined semantic attributes in the case of biometrics, the data sparsity issue combined with the extremely large number of classes presents a formidable challenge from the machine learning perspective.

Apart from the challenges related to the extreme scale of the classification task, biometric identification also poses two unique challenges that are commonly not encountered in pattern classification. The first challenge is related to the phenomenon of biometric aging, which refers to the large temporal changes in the biometric identifiers of an individual, especially in the early stages of child development. In traditional pattern recognition systems, while the object may change over time (say, a mobile phone from the 1990s will have no resemblance to today's mobile phone), it is easy to deal with this problem by adding new sub-classes. However, this is not possible in the case of biometrics. The second challenge relates to the adversarial nature of the biometric identification task. When a biometric system is applied for identity deduplication, some individuals may have a strong motivation to circumvent the biometric system by obfuscating their biometric identifiers. Hence, it is essential to design a robust pattern classifier that is capable of dealing with such threats.

Despite the extreme nature of the challenges involved, giga-scale biometric identification is a feasible pattern recognition problem due to the following reasons:

(i) Availability of some prior knowledge about the uniqueness and permanence of biometric identifiers. For example, we know from experience that detailed ridge-valley structures in a fingerprint and randomness in the iris texture are phenotypic variations, which are mostly unique to individuals and they remain relatively unchanged during a person's lifetime.

(ii) Availability of specialized sensors to capture the biometric identifiers at sufficiently high resolution and quality. Although biometric identifiers may exhibit large intra-class variations, it is possible to control quality and feature discriminability to a large extent via intuitive and ergonomic user interface design and carefully designed supervised enrollment process. This greatly reduces the complexities involved in general computer vision tasks like detecting the object in the image, large changes in scale, presence of multiple objects in the image, and ambiguity about the object of interest.

(iii) Finally, the ability to capture multiple biometric identifiers (e.g., ten fingers + two irises + face) greatly enhances the richness/discriminability of the underlying feature set.

Though the challenges encountered in giga-scale biometric identification are somewhat unique within the wider pattern recognition/ machine learning field, a number of valuable lessons can still be assimilated from this community.

(i) In the case of biometric identifiers such as face and voice, where there is little prior knowledge about the uniqueness of features, representation learning approaches from the pattern recognition community can be applied for automated feature extraction. For example, representation learning based on deep convolutional neural networks have significantly enhanced the accuracy of face recognition systems in recent years.
(ii) Advancements in the areas of dimensionality reduction, metric learning, and robust hashing/indexing can be leveraged to significantly improve the accuracy and throughput of biometric identification.
(iii) Finally, the best engineering practices employed by other large-scale pattern recognition systems in areas such as distributed storage and computing, massive parallelization, and fault tolerance can be readily adopted in the design of giga-scale biometric identification systems.

4.2 Lessons Learned from Current Biometric Systems

There are a handful of large-scale biometric recognition systems that are currently in operation around the world. We believe that a careful analysis of these existing solutions can provide valuable insights to guide the design of the proposed universal and ubiquitous person recognition systems.

4.2.1 The Aadhaar Case Study

The Unique Identification Authority of India (UIDAI) was created in 2009 with the goal of issuing Unique Identification numbers (also referred to as Aadhaar) to all residents of India. The two primary design goals of Aadhaar are: (i) elimination of duplicate and fake identities via identity de-duplication, and (ii) provide an easy and cost-effective mechanism for service providers to authenticate individuals across multiple application domains.

UIDAI decided to collect biometric identifiers, including all ten fingerprints, the two irises, and the face image of each resident, at the time of enrolment. These identifiers are compared against previously enrolled identifiers of residents available in the database to perform identity de-duplication. For identity verification or authentication, a single-finger or iris capture was deemed sufficient for comparison to the enrolled biometric identifiers of the resident.

In the ensuing years, the Aadhaar program has amassed the largest known database of biometric identifiers, issuing over 1.15 billion Aadhaar numbers, as of June 2017[3]. Operating at such a large scale with primary reliance on biometric identifiers presented the many significant challenges [18]. These challenges are discussed below along with the solutions adopted to overcome them.

Identification Accuracy: A single biometric identifier (such as a fingerprint or iris) was found to be an insufficient source of information for performing identity de-duplication with high accuracy at a large scale. Multiple unique identifiers, are therefore, used in conjunction with each other (ten fingerprints and two irises) to maximize de-duplication accuracy. Interestingly, the UIDAI has explicitly designated fingerprint and iris as core biometric identifiers. Furthermore, the Aadhaar system uses three different biometric technology service providers to independently compare a new enrolment query against previously enrolled data to improve the accuracy. Demographic de-duplication is additionally used to reduce de-duplication errors.

Aadhaar also provides us with an estimate of accuracies achievable in a real-world diverse population using 10 fingers and 2 irises. As per reports published by UIDAI [Need reference], on a gallery size of 84 million users, their solution was able to achieve an FNIR of 1 in 1000 and an FPIR of 1 in 10,000.

Population coverage: Certain population segments (elderly, manual laborers and farmers) have worn out fingerprints that are difficult to capture, whereas others may have defective eyes making iris acquisition difficult. Enrolment requests with the absence of one or more biometric identifiers are logged and identity de-duplication is performed using identifiers that can be captured from a resident (from the ten fingerprints and the two irises). No resident is denied the issuance of Aadhaar number because there is a failure to capture of a particular biometric identifier. Another aspect of the coverage is the ability of a person to produce official documentary proof of any aspect of identity, which is often a barrier for poor and marginalized communities. Aadhaar requires only minimal documentation to be produced for enrollment. Note that to generate a new identity, one only needs to connect the identity to the person's biometric and external identifiers such as name are superfluous to the concept of identity itself.

Data quality control and security: Certification standards for biometric data capture devices as well as standards for data storage and interchange have been established via the Standardization Testing and Quality Certification (STQC) Directorate. The operators at enrolment centers across the country use certified devices and quality control software to enroll residents. The captured biometric data is encrypted using state-of-the-art encryption methods before transfer to the central repository for de-duplication. Provisions are also in place to ensure data privacy and integrity during transfer and storage at the central repository. Automated checks are built into the enrolment process (e.g. comparing captured biometric data from a resident to operator's biometrics) to minimize both intentional and unintentional operator errors. Duplicate entries found during an enrolment attempt are first screened to determine if they are due to the enrolment process. If not, manual adjudication by a human expert is used for making the final decision.

Privacy and Function Creep: One of the fundamental fears in large-scale identification systems is due to their ability to uniquely identify a person when they want to remain anonymous. Such a system, where biometrics are collected

3. https://portal.uidai.gov.in/uidwebportal/dashboard.do

for the purpose of identity generation, can potentially be used in criminal investigations. To avoid such function creep, the Aadhaar system is designed such that the core biometrics can only be input to the system, but never come out. Moreover, the core biometric matching system is only allowed to answer yes/no questions: Is this person already enrolled? or Does these biometric traits match a specific identity?. Such a strategy can prevent the use of an identity verification system from being misused. To avoid the possibility of guessing someone's identity number and then verifying using biometrics, the numbers themselves are made random [12]. *Interoperability*: To facilitate interoperability, the biometric technology stack is designed ground up to be generic, open and agnostic to biometric service providers. It has provisions for including new service providers as well as rules for directing queries to service providers based on their performance. This removes dependence on specific service providers, and ensures the architecture can scale seamlessly by integrating new stakeholders in the ecosystem, when needed. Furthermore, open source APIs have been made available to build solutions leveraging the biometric authentication capability. This is expected to facilitate rapid adoption of the exposed biometric identity verification services across different application domains.

Computational Requirements: Another useful metric one can derive from the Aadhaar experience is the computational requirement, especially for deduplication, for a large-scale biometric system. The UIDAI was able to achieve around 1.5 million enrollments (deduplications) per day on a gallery of 500 million users at its peak using a set of 3000 general purpose servers [Need reference]. This provides us with a basis for estimating the computational requirements for the proposed 10G system.

Next, we discuss the architecture implemented by UIDAI to deliver identity-as-a-service. UIDAI maintains a central ID repository (CIDR) to store the captured biometric (and demographic) data of residents. An entity/agency that wishes to authenticate the identity of a person must register as an Authentication User Agency (AUA). AUAs must use UIDAI certified equipment for biometric data capture (e.g., fingerprint reader) at service delivery points (e.g. a bank counter). The entire authentication pipeline (including software application, communication and security protocols, data handling procedures) must follow the standards specified by UIDAI. An approved AUA can generate authentication requests containing captured biometric (e.g. fingerprint or iris) along with Aadhaar number. The request is typically routed to a registered Authentication Service Agency (ASA), which is an enabling intermediary agency having a secure connection with the central ID repository (CIDR). Note that the AUA can also register as an ASA and directly connect to CIDR. ASAs ensure that authentication requests submitted by AUAs are compliant before routing them to CIDR. The biometric data encapsulated in an authentication request received at CIDR is compared against the corresponding enrolled biometric for the Aadhaar number specified in the authentication request. Based on the result of the comparison, a yes/no decision is returned to the ASA, that in turn relays it back to the AUA that generated the request. The identity is hence verified at service delivery point.

4.2.2 The UAE Border Crossing System Case Study

In 2001, United Arab Emirates (UAE) initiated the use of iris recognition technology for streamlining border crossing and tracking expellees [2]. At present, all visitors/travelers to UAE are required to undergo iris scanning at all entry points to the country, including land, air and sea. The iris image of the traveler (during both entry and exit?) is compared against 2.3 million irises in the centrally maintained watch-list database using a secure national network infrastructure called the IrisFarm. To speed up the comparison process, IrisFarm uses many parallel search engines called IrisEngines, with each engine capable of performing more than 500,000 iris comparisons per second.

The architecture does not require high speed links for communication. The system can perform real-time iris comparisons even with links as slow as 33.6 Kb/sec. Furthermore, it is capable of automatically synchronizing iris enrollments from multiple locations (e.g., prisons) and offers continuous search capability, even as new data is being enrolled into the watch-list database at various locations. Separate hardware is used for enrollment, recognition, and database maintenance. This facilitates load sharing and fault tolerance. Hardware can be readily added to expand capabilities depending on requirements, thereby ensuring scalability. Data is periodically backed up and a hot-standby is always maintained with real time switching ability, in case the primary database experiences any issues.

4.2.3 Insights from the two case studies

The case studies point to the fact that a biometric system at 10B scale is theoretically possible and technically feasible with technological advances that are mentioned below. One of the important aspect that emerges from the two case studies is the improvement needed in accuracy and efficiency of a matching system.

- A 10G system will need to achieve improvements of 1 to 2 orders of magnitude in overall FMR rates compared to systems that are currently deployed.
- Biometric image (or data) quality is the most significant factor that will determine the accuracy and practicality of such a large-scale biometric authentication system.
- While a dedicated hardware can provide a higher speed to cost ratio as in the case of IrisFarm, the use of generic hardware would be more practical in a distributed multimodal system.
- In addition to accuracy and speed, careful attention needs to be provided to a variety of factors to make a system practical. These include security, reliability, privacy, population coverage, scalability and standards to ensure inter-operability.

5 Proposed Solution for Global ID

Even if technological issues (accuracy, throughput, and security) can be eventually surmounted, legal and trust roadblocks are unlikely to be go away in the foreseeable future. Therefore, the most feasible system appears to be a centralized person identification system at the regional/national level and having a common set of policies, protocols, and standards to enable inter-operability and integration between these regional systems (see Figure 8).

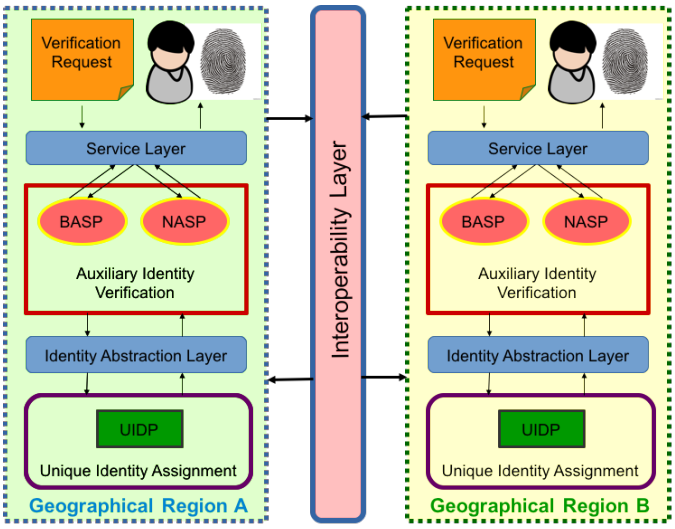

Fig. 8. A possible architecture of a global person identification system. Each system (national or regional) can independently carry out the deduplication and authentication tasks, while the interaction between systems will be based on mutually agreed upon interfaces and formats. Note that the specifics of the individual identity systems need not be identical.

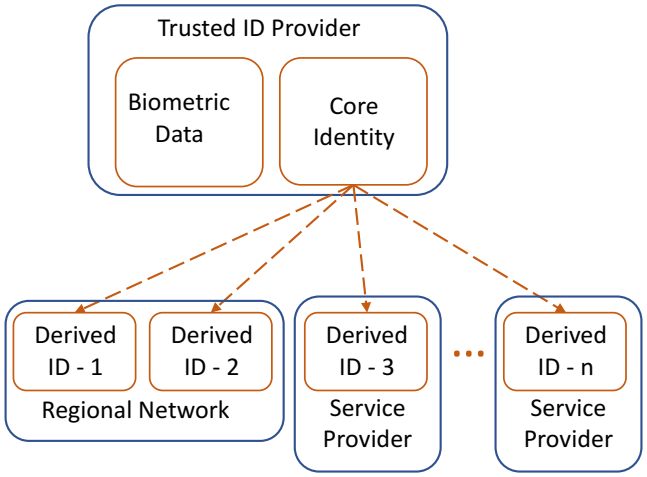

Fig. 9. A Global ID framework should provide the user to select a trusted identity provider to create a core identity using their biometric characteristics. In addition, one should also be able to generate several derived identities that are given to specific service providers for authentication. The information content available with a derived identity may be restricted to enable privacy enhanced identity verifiction.

This will facilitate ubiquity to a large extent. In addition to the central or trusted identity provider, a user might also create multiple derived identities with restricted auxiliary information, which can be used for authentication purposes (see Figure 9).

Three types of identity providers can be envisioned in the proposed system.

1) **Unique Identity Provider (UIDP)**: The primary task of UIDPs is to perform biometrics-based deduplication and issue a **core unique identity** to every individual. UIDPs could be national/state governments or large public/private-sector organizations authorized by the governments. To perform de-duplication, UIDPs have to store raw biometric samples. For universality, every individual in the world should be enrolled with one of the UIDPs. Furthermore, UIDPs around the world should be inter-operable to facilitate ubiquity or portability of identity. Consensus will be required on biometric identifiers to be used, the quality standards, data formats, feature extraction, matching, and security protocols.
2) **Biometrics-based Authentication Service Provider (BASP)**: The primary task of BASPs is to issue an **auxiliary** identity to enrolled users that can be authenticated at a latter point of time based on biometrics (or a combination of biometrics and other authentication factors). Since there is no need for this auxiliary identity to be unique (i.e., same person can have multiple auxiliary identities), de-duplication and storage of raw biometric data are not required. Only storage of secure templates will be sufficient. ***Linkage of auxiliary identity to the core unique identity should NOT be mandatory***, though the person can voluntarily choose to do so (Exceptions could be made in the case of court orders. Even in this case, linkage is possible only when both core and auxiliary identities are based on the same biometric identifiers and the features are compatible). Note that UIDPs can also double up as BASPs (e.g., India's Aadhaar system). Hence, regulations need to be in place to prevent linkage of core and auxiliary identities of an individual without explicit consent. Furthermore, linkage between multiple auxiliary identities of the same individual should not be possible without the explicit consent of the individual. This will enable individuals to preserve their privacy.
3) **Non-biometric Authentication Service Provider (NASP)**: Similar to BASPs, NASPs can also issue auxiliary identities, but the authentication will be based on non-biometric factors such as tokens and knowledge. Non-biometric authentication may be sufficient for many low-risk transactions. As in the case of BASPs, linkage to the core unique identity or other auxiliary identities can be purely voluntary.

Both BASPs and NASPs should be ***federated identity management systems***, so that external organizations can rely on their service to manage the auxiliary identities and authenticate individuals based on their auxiliary identity.

The aforementioned discussion presents the following open questions that need to be satisfactorily resolved. Should the de-duplication of identity be global or limited to only a specific UIDP? For example, suppose that John Doe is a citizen of country C1 (with a core unique identity X assigned by UIDP U1) and he now moves to a different country C2 (where identity is managed by UIDP U2). It is obvious that if John Doe wishes, he should be able to transfer/port his core unique identity X from U1 to U2. At the very least, he should be able to voluntarily link his core unique identities in U1 and U2. The key design question is: how to achieve this portability/linkage of core unique identity across UIDPs, when the UIDPs inherently do not trust each other? Either a third-party centralized entity or secure distributed protocols (e.g., blockchain?) will be required to ensure the integrity of identity linkages across

UIDPs. On the other extreme, what happens if John Doe wants to hide his identity X from UIDP U2? Should UIDP U2 still be able find out his core identity X by querying other UIDPs around the world? If yes, how can we do this reliably and securely? These questions

6 Recommendations

The following research issues have to be effectively addressed by the research community:

Research Challenge #1: Invariant biometric representation

a) For each biometric identifier, we need an invariant feature representation (ideally, a compact, fixed-length representation) with an efficient similarity metric that can achieve the required accuracy and throughput targets of a giga-scale biometric identification system. Currently, iriscodes with Hamming distance metric comes closest to this goal. But even in this case, the dimensionality is high, multiple rotations need to the tried, and we need to deal with ancillary information such as masks.

b) The representation should allow biometric identification of children (less than 5 years old) with minimal template updates during the lifetime of an individual (e.g., during transition to adulthood or when pathological changes occur to the biometric identifiers).

c) Such a representation will also make it possible to design an optimal multibiometric identification (search) algorithm, which maximizes both accuracy and throughput (preferably, also considering the massive parallelization of computations).

Research Challenge #2: Cryptographic algorithms for secure biometric storage and comparison

a) A provably-secure cryptographic scheme that enables comparison of biometric data in the encrypted domain ***without any compromise on accuracy and throughput***. Once the biometric features are encrypted, there should be no need to decrypt the plain-text features in the future.

b) A scalable, secure, distributed biometric database that can be shared by entities (e.g., nations) without any trust assumptions and which can ensure the integrity of biometric comparisons.

Research Challenge #3: Hardware for biometric capture

a) Sensors that can rapidly capture multiple biometric traits (ten fingers + iris + face ++) of a person with minimal user interaction/cooperation, without compromising on the portability, ergonomics, and cost.

b) The sensors should also be able to guard against any type of presentation attack.

References

[1] R. Agrawal, A. Gupta, Y. Prabhu, and M. Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *Proceedings of the 22nd international conference on World Wide Web*, pages 13–24. ACM, 2013.

[2] A. Al-Raisi and A. Al-Khouri. Iris recognition and the challenge of homeland and border control security in uae. *Telematics and Informatics*, 25(2):117–132, 2008.

[3] P. Biscaye, S. Coney, E. Ho, B. Hutchinson, M. Neidhardt, C. Anderson, and T. Reynolds. Review of national identity programs, epar request no. 306, university of washington, september 2015.

[4] R. Cappelli, M. Ferrara, and D. Maltoni. Large-scale fingerprint identification on gpu. *Information Sciences*, 306:1–20, 2015.

[5] M. M. Cisse. *Efficient Extreme Classification*. PhD thesis, Université Pierre et Marie Curie, 2014.

[6] Department of Economic and Social Affairs, United Nations. Transforming our world: The 2030 agenda for sustainable development. https://sustainabledevelopment.un.org/sdg16.

[7] A. Gelb and J. Clark. Identification for development: The biometrics revolution. CGD working paper 315, center for global development, january 2013. http://www.cgdev.org/content/publications/detail/1426862.

[8] P. J. Grother, G. W. Quinn, J. R. Matey, M. L. Ngan, W. J. Salamon, G. P. Fiumara, and C. I. Watson. IREX III: Performance of iris identification algorithms, NIST interagency report 7836. http://biometrics.nist.gov/cs_links/iris/irexIII/IREXIII_full.zip, Apr. 2012.

[9] M. S. Islam and A. Grönlund. The Bangladesh national biometric database: A transferable success? In *Proc. 1st International Conference on Electronic Government and the Information Systems Perspective*, EGOVIS'10, pages 189–203, Berlin, Heidelberg, 2010. ACM, Springer-Verlag.

[10] A. K. Jain, A. Ross, and K. Nandakumar. *Introduction to Biometrics*. Springer, 2011.

[11] D. Maltoni. Invited keynote talk. In *15th International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2016.

[12] Ministry of Law and Justice, Govt. of India. The aadhaar act, the gazette of india, march 2016. https://uidai.gov.in/images/the_aadhaar_act_2016.pdf.

[13] OneName Corporation. Requirements for a global identity management service. http://www.w3.org/2001/03/WSWS-popa/paper57.

[14] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, pages 1410–1418, 2009.

[15] C. I. Watson, G. P. Fiumara, E. Tabassi, S. L. Cheng, P. A. Flanagan, and W. J. Salamon. Fingerprint Vendor Technology Evaluation, NIST Internal Report 8034, dec 2014. http://nvlpubs.nist.gov/nistpubs/ir/2014/NIST.IR.8034.pdf.

[16] D. Weitzner. In search of manageable identity systems. *IEEE Internet Computing*, Oct. 2006.

[17] World Bank. Identification for development (ID4D) integration approach. http://documents.worldbank.org/curated/en/812441468191048923/pdf/98383-REVISED-WP-P115610-OUO-9-Box393205B.pdf.

[18] F. Zelazny. The evolution of india's uid program, cgd policy paper 008, center for global development, august 2012. https://www.cgdev.org/files/1426371_file_Zelazny_India_Case_Study_FINAL.pdf.

Anil K. Jain is a university distinguished professor in the Department of Computer Science and Engineering at Michigan State University. His research interests include pattern recognition and biometric authentication. He served as the editor-in-chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (1991-1994). He is the coauthor of a number of books, including Handbook of Fingerprint Recognition (2009), Handbook of Biometrics (2007), Handbook of Multibiometrics (2006), Handbook of Face Recognition (2011), BIOMETRICS: Personal Identification in Networked Society (1999), and Algorithms for Clustering Data (1988). He served as a member of the Defense Science Board and The National Academies committees on Whither Biometrics and Improvised Explosive Devices. He received the 1996 IEEE TRANSACTIONS ON NEURAL NETWORKS Outstanding Paper Award and the Pattern Recognition Society best paper awards in 1987, 1991, and 2005. He has received Fulbright, Guggenheim, Alexander von Humboldt, IEEE Computer Society Technical Achievement, IEEE Wallace McDowell, ICDM Research Contributions, and IAPR King-Sun Fu awards. He was elected a member of the National Academy of Engineering in 2016.

Sharath Pankanti is Principal Research Staff Member in Cognitive Computing Department at the Thomas J. Watson Research Center. He received Ph.D. degree in Computer Science from the Michigan State University. Sharath has led a number of safety, productivity, education, healthcare, and security focused projects involving biometrics, multi-sensor surveillance, rail-safety, driver assistance technologies that entail object/event modeling, detection and recognition from information provided by static and moving sensors/cameras. His work contributed to world's first large scale biometric civilian fingerprint identification system in Peru. He is a co-author of over 150 peer-reviewed publications in many reputed venues, including Scientific American, IEEE Computer, IEEE Spectrum, Comm. ACM, and Proc. IEEE. He is also co-inventor of more than 100 inventions. His efforts have been recognized by IBM as significant accomplishments including Master Inventor and Outstanding Accomplishment Awards. Dr. Pankanti co-edited the book, *Biometrics: Personal Identification* Kluwer, 1999 and co-authored, *A Guide to Biometrics*, Springer 2004. He is Fellow of IEEE, IAPR, and SPIE and has served as part of IEEE Distinguished Visitor and ACM Distinguished Speaker programs.

Karthik Nandakumar is a Research Staff Member at IBM Research Collaboratory - Singapore. Prior to joining IBM in 2014, he was a Scientist at Institute for Infocomm Research, A*STAR, Singapore for over six years. He received his B.E. degree (2002) from Anna University, Chennai, India, M.S. degrees in Computer Science (2005) and Statistics (2007), and Ph.D. degree in Computer Science (2008) from Michigan State University, and M.Sc. degree in Management of Technology (2012) from National University of Singapore. His research interests include computer vision, statistical pattern recognition, biometric authentication, image processing, and machine learning. He has co-authored two books titled Introduction to Biometrics (Springer, 2011) and Handbook of Multibiometrics (Springer, 2006). He has received a number of awards including the 2008 Fitch H. Beach Outstanding Graduate Research Award from the College of Engineering at Michigan State University, the Best Paper award from the Pattern Recognition journal (2005), the Best Scientific Paper Award (Biometrics Track) at ICPR 2008, and the 2010 IEEE Signal Processing Society Young Author Best Paper Award.

Salil Prabhakar Dr. Salil Prabhakar is the General Manager at Delta ID Inc. Dr. Prabhakar served as President and Chief Executive Officer at Delta ID Inc., from 2011 to and as Technical Advisor of Digitalpersona. Dr. Prabhakar received his Ph.D. degree in 2001 from the Department of Computer Science and Engineering at Michigan State University. His research interests include pattern recognition, image processing, computer vision, machine learning, biometrics, data mining and multimedia applications. He is the coauthor of more than 50 technical publications and has two patents pending. Dr. Prabhakar received his BTech degree in Computer Science and Engineering from Institute of Technology, Banaras Hindu University, Varanasi, India, in 1996. He is a Fellow of IEEE and IAPR.

Sunpreet S. Arora received the B.Tech. (Hons.) in Computer Science from the Indraprastha Institute of Information Technology, Delhi (IIIT-D) in 2012, and a Ph.D. in Computer Science and Engineering from Michigan State University in 2016. He is currently a Senior Biometrics Researcher at Visa Inc., Foster City, CA. His research interests include biometrics, pattern recognition, image processing and machine learning. He received the best paper award at BIOSIG-2016, and the best poster award at the BTAS-2012. He is a member of the IEEE.

Anoop M. Namboodiri is an Associate Professor at the International Institute of Information Technology, Hyderabad, and is associated with the Centre for Visual Information Technology. He received his PhD in Computer Science and Engineering from Michigan State University. He has coauthored more than 70 technical publications and holds two patents. He has worked with several large-scale biometric identification programs including the Aadhaar program in India. His research interests include machine learning, biometrics, document understanding, and computer vision. He is a member of the IEEE.

Arun Ross is a Professor in the Department of Computer Science and Engineering at Michigan State University (MSU) and the Director of the i-PRoBe Lab. Prior to joining MSU in 2013, he was a faculty member at West Virginia University. He also served as the Assistant Site Director of the NSF Center for Identification Technology and Research (CITeR) between 2010 and 2012. Arun received the B.E. (Hons.) degree in Computer Science from the Birla Institute of Technology and Science, Pilani, India, and the M.S. and Ph.D. degrees in Computer Science and Engineering from Michigan State University. He is the coauthor of the textbook *Introduction to Biometrics* and the monograph *Handbook of Multibiometrics*, and the co-editor of *Handbook of Biometrics*. He is a recipient of the IAPR JK Aggarwal Prize, the IAPR Young Biometrics Investigator Award (YBIA), the NSF CAREER Award, 2005 Biennial Pattern Recognition Journal Best Paper Award and the Five Year Highly Cited BTAS 2009 Paper Award. Arun served as a panelist at a counter-terrorism event that was organized by the United Nations Counter-Terrorism Committee. He was an Associate Editor of IEEE Transactions on Information Forensics and Security (2009-2013), and IEEE Transactions on Image Processing (2008-2013). He currently serves as Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology, Senior Area Editor of IEEE Transactions on Image Processing, Area Editor of the Computer Vision and Image Understanding Journal, Associate Editor of the Image and Vision Computing Journal, and Chair of the IAPR TC4 on Biometrics. He is a senior member of the IEEE.