De-duplication

The Complexity in the Unique ID context

1. Introduction

Citizens in India depend on the Government for various services at various stages of the human lifecycle. These services include issuance of birth certificate, voter identity card, ration card, driving license, passport, PAN card etc. In addition the government also implements different welfare schemes like Targeted Public Distribution System (TPDS), National Rural Employment Guarantee System (NREGS), health insurance, old age pensions etc for the economic and social upliftment of the people. A Unique Identity (UID) assigned for every citizen would obviate the need for a person to produce multiple documentary proofs of his identity for availing any government service, or private services like opening of a bank account. The Unique Identity (UID) would remain a permanent identifier right from birth to death of the citizen.

UID would enable government to ensure that benefits under various welfare programmes reach the intended beneficiaries, prevent cornering of benefits by a few people and minimize frauds. UIDs are also expected to be of help in law and order enforcement, effective implementation of the public distribution system, defining social welfare entitlements, financial inclusion and improving overall efficiency of the government administration.

2. Enrollment

It is expected that the Government will enroll the citizens by capturing the Biographic and Biometric details and issue a Unique ID number. During the enrollment process, it has to be ensured that the same citizen does not get enrolled more than once. This can be done by comparing the biometrics of the citizen with all other citizens already enrolled and denying enrollment in case a match is found. Enrollment of citizens can be done in two ways.

(i) The enrollment can be done online by adopting a centralized architecture, in which all the enrollment stations in the country are connected to the central server and the biometrics of the citizen being enrolled are matched / compared with the biometrics of all the citizens already enrolled. In case a match is found, the system will not allow enrollment to be done.

(ii) The other way in which enrollment can be done is by adopting an offline enrollment method by synchronizing the data with the central server

periodically as and when internet connectivity is available or through regular backup of data by means of DVDs / Hard disks. The biometrics of the citizens captured through the offline method are then matched / compared with the biometrics of all other enrolled citizens at the central server to identify multiple enrollments of the same citizen.

What is important in both the cases is the **speed of matching** and the **accuracy of the matching** results. The speed of matching has to be very high as the number of citizens to be enrolled runs into millions. The **accuracy is equally important** as false matches will result in erroneous enrollments, delays and potential failure of the project itself. It is to be noted that, during the enrollment, the raw images of the biometrics are captured and algorithms are used for converting the images into templates which are used for comparison/ matching. The speed of matching and accuracy of matching depend on the biometric captured, the algorithm used and the matching engine deployed.

3. Biometrics

There are several Biometrics such as Fingerprints, Iris, Facial recognition, Hand Geometry, Signature, Voice patterns etc. which are being used by Governments all over the world for an extensive array of highly secure identification and personal verification solutions. Each of them has certain advantages and disadvantages which must be considered in developing biometric systems. Selecting the right biometric is critical to the success of any Identity Management project such as the Unique ID project. Key metrics that need to be evaluated for choosing a biometric include the stability of the biometric over the lifetime of a human being, Failure to Enroll (FTE) rate, False Accept Rate (FAR), and the False Reject Rate (FRR). It is important to understand the advantages and disadvantages of each biometric and the advantages of going in for Multimodal Biometric solutions.

4. Multi-Modal Biometrics

Multimodal biometrics refers to the use of a combination of two or more biometric modalities in a single identification system. Biometric systems based solely on one-modal biometrics are often not able to meet the desired performance requirements for large user population applications, due to problems such as failure to enroll, noisy data, spoof attacks, environmental conditions and unacceptable error rates. Each of the biometrics has its' relative merits and applications where they can be used. A few examples are given below.

4.1. Face

The face can be the first form of identifying a person without the need for any external device; however, a facial recognition camera may not be able to distinguish between identical twins. Facial recognition works on the system identifying 9 geometric points on a human face and international studies have confirmed very high "False Acceptance Rates" (1 in 100).

4.2. Fingerprints

Fingerprints are ideal for verification (1:1 matching) though there are Automatic Fingerprint Identification Systems (AFIS) which do identification (1:N or N:N) also. However, even in the case of identification, the "False Acceptance Rates" are about 1:100,000.

- Fingerprint de-duplication is cost effective only for small population and the cost of de-duplication goes up significantly due to manual intervention required while doing de-duplication for huge population. This is due to large number of false matches thrown up during the fingerprint de-duplication process requiring Human intervention / Back office operations to work on the probable matches and thereby adding up to the costs;
- Fingerprints are susceptible to noisy or bad data, such as inability of a scanner to read dirty fingerprints clearly. People above 60 years and young children below 12 years may have difficulty enrolling in a fingerprinting system, due to their faded prints or underdeveloped fingerprint ridges. It is estimated that approximately 5 percent of any population has unreadable fingerprints, either due to scars or aging or illegible prints. In the Indian environment, experience has shown that the failure to enroll is as high as 15% due to the prevalence of a huge population dependent on manual labor.

However, the advantages of fingerprints are given below

- Fingerprint is cost effective at the time of verification. (Since at verification or the point of service, fingerprint devices of low cost can be used.
- Fingerprint can be used for forensic purposes.

4.3. Iris

- Iris recognition is the most accurate of the top three biometrics: fingerprints, facial recognition, and iris recognition. Iris recognition has a false accept rate of 1 in 1.2 million for one eye (1 in 1.44 trillion for two eyes) regardless of database size. As a result of the accuracy of Iris recognition, Iris returns a single result back. Fingerprint and face technologies generally return a

candidate list and then a manual process is required for resolving the candidate list. For this reason, Iris is the ideal biometric for applications which require real time identification. Processes such as fraud screen (to check for duplicates) enrollment for large populations can be easily handled by Iris recognition where they are very difficult for fingerprints.

- Iris recognition algorithms can search upto 20 million records in less than one second using a normal Quadcore – 2 Processor blade server. In a parallel process, using COTS hardware, Iris can perform at 1 billion matches per second. The ability to search a population database in real time and return a single match result is unique to Iris recognition technology. Due to manual candidate list resolution with face and fingerprint technologies, Iris is the only biometric which delivers operational results in real time which can be acted upon.
- Iris is an internal organ because of which there is no problem of environmental conditions affecting the Iris unlike fingerprints which may not be prominent in people who do labor work or work in harsh environments (e.g factories, farms, etc).
- The Iris of a person is stable throughout a person's life (From the age of one year till death); the physical characteristics of the Iris do not change with age, diseases or environmental conditions. Hence one time enrollment is enough for a person during his lifetime.
- One of the most important advantages of using Iris as a Biometric is the lower effort, lesser infrastructure (servers, database licenses, datacenter infrastructure etc) required for de-duplication, whereas finger print de-duplication requires more than 50 times infrastructure and more human effort. Another related cost that is normally overlooked is the infrastructure maintenance cost for running such as huge datacenter like, manpower, power consumption, annual maintenance costs for hardware and software etc.
- A comparative study of the performance of multiple biometrics done by the centre for Mathematics and Scientific Computing, National Physical Laboratory NPL of UK is given in the Table 1.

Table 1

Biometric	FAR (False Acceptance Rate)	FRR (False Rejection Rate)	FER (Failure to Enroll rate)	Scalability	Stability
Iris	1:1.2 million	0.1 – 0.2%	0.5%	1: all search	Very stable
Fingerprint	1: 100000	2.0 – 3.0%	1.0 – 2.0%	1: 1 match	Changes
Facial recognition	1:100	~10%	0.0%	1:1 match	Changes
Hand Geometry	1:10000	10 – 20%	0.0%	1:1 match	Changes

For complete Report Please Refer - Biometric Product Testing Final Report (19 March 2001, Center for Mathematics and Scientific Computing, National Physical Laboratory, UK).

The most compelling reason to adopt Multi-Modal Biometric is to introduce certainty in the recognition process, real time identification, lower effort for de-duplication and reduce the possibility of inconvenience caused by malfunctioning of a single Biometric.

Advantages of using Multi-Modal Biometrics

- The enrollment cost for Multi-Modal Biometric enrollment will be about 5 – 10% marginally higher compared to single/dual biometric enrollment. However, the total cost of solution in case of multimodal enrollment is significantly reduced due to reduced cost during de-duplication which outweighs the marginal additional cost incurred during enrollment.
- Single biometric enrollment results in Failure to Enroll (FTE) if those biometric characteristics are absent in a citizen or if they are not qualified for enrollment (due to scars, aging or illegible/ worn out / cut / unrecognizable in case of fingerprints). In Indian conditions, where more than 60% of the population is involved in manual labour, experience has shown very high FTE rates for fingerprints.

5. De-duplication

De-duplication is the processing of the biometric data of citizens to remove instances of multiple enrollments by the same citizen. During de-duplication, matching the biometrics of a citizen is done against the biometrics of other citizens to ensure that the same person is not enrolled more than once. This will ensure that each person

will have a unique identity. De-duplication will be a necessary component in the "Unique ID" project. De-duplication is discussed in the context of the two different enrollment scenarios which are given below.

Case I: Enrollment using a centralized architecture

In the case of enrollment using a centralized architecture, the biometrics of the citizen have to be matched against the biometrics of all the previously enrolled citizens. The matching has to be done soon after the biometrics are captured to check whether the same citizen has been enrolled earlier. In case a match is found, the citizen will not be enrolled into the system. To accomplish this, the speed of matching has to be very high and without any false accepts. To illustrate the complexity, let us take a case where 200 million citizens have already been enrolled, and a new citizen is now waiting to be enrolled into the system at the enrollment station.

(1) When Fingerprints are used as the Biometric.

The number of matches to be performed and the time taken is shown in the table below.

Scenario	**No of matches**	**Time taken (Assuming 10 blade servers with a total matching capacity of 5 million per sec)**
No. of matches if 1 finger (say left thumb) is matched against all left thumbs of previously enrolled citizens	200 million	40 secs
No. of matches if all 10 fingers are matched against the respective fingers of all the previously enrolled citizens	2000 million	400 secs (6.67 minutes)
No. of matches if all the 10 fingers are matched against all the fingers of all the previously enrolled citizens	20000 million	4000 secs (1.11 hours)

(2) When Iris is used as the Biometric.

The number of matches to be performed and the time taken is shown in the table below.

Scenario	No of matches	Time taken (Assuming 10 blade servers with a total matching capacity of 200 million per sec)
No. of matches if 1 eye (say left eye)against all left eyes of previously enrolled citizens	200 million	1 sec
No. of matches if both eyes are matched against the respective eyes of all the previously enrolled citizens	400 million	2 secs
No. of matches if both eyes are matched against both eyes of all the previously enrolled citizens.	800 million	4 secs

Thus it can be seen that the Iris based online enrollment is many times faster compared to the fingerprint based enrollment. Moreover, the fingerprint based De-duplication throws up false matches which have to be crosschecked with the photo or other parameters before deciding the accuracy of the match.

Case II: Enrollment using a De-centralized architecture

In the case of enrollment using a De-centralized architecture, the biometrics of citizens captured during a certain period have to be matched against the unique ID enrollment database of all the previously enrolled citizens. The matching has to be done by aggregating the data from each of the decentralized enrollment stations and matching against the de-duplicated biometrics of all the previously enrolled citizens. To illustrate the complexity, let us take the case where 200 million citizens have already been enrolled, and a data of 1 million citizens has been aggregated from the enrollment stations. The data of the 1 million citizens will have to be matched against the 200 million citizens to avoid multiple enrollments.

(1) When Fingerprints are used as the Biometric.

The number of matches to be performed and the time taken is shown in the table below.

Scenario	No of matches	Time taken (Assuming 10 blade servers with a total matching capacity of 5 million per sec)
No. of matches if 1 finger (say left thumb) against all left thumbs of previously enrolled citizens	200 trillion	463 days Or 1.27 years
No. of matches if all 10 fingers are matched against the respective fingers of all the previously enrolled citizens	2000 trillion	4630 days Or 12.67 years
No. of matches if all the 10 fingers are matched against all the fingers of all the previously enrolled citizens	20000 trillion	46296 days Or 126.84 years

(2) When Iris is used as the Biometric.

The number of matches to be performed and the time taken is shown in the table below.

Scenario	No of matches	Time taken (Assuming 10 blade servers with a total matching capacity of 200 million per sec)
No of matches if 1 eye (say left eye)against all left eyes of previously enrolled citizens	200 trillion	11.57 days
No of matches if both eyes are matched against the respective eyes of all the previously enrolled citizens	400 trillion	23.15 days
No of matches both eyes are matched against both eyes of all the previously enrolled citizens	800 trillion	46.30 days

It can thus be seen that the de-centralised enrollment will lead to large de-duplication times. By increasing the number of servers, it is possible to do the de-duplication within a day. To achieve the same timelines using fingerprints, it would

take 50 times the number of servers adopted for Iris based De-duplication. In addition to the increase in timelines and hardware, the number of false matches thrown up in Fingerprint based De-duplication would require large manpower to use other comparisons such as photos to eliminate the false matches.

6. Conclusion

Choosing the right biometrics plays a very important role for ensuring the success of the Unique ID project. While Iris as a biometric ensures high matching speeds and high degree of accuracy which are very essential for large Unique ID projects, fingerprint as a biometric will be economical for verification at the Point of Service. Thus the use of Multi-Modal biometrics will enable Governments to reap the advantages of both in the most optimal manner.