Can Combining Demographics and Biometrics Improve De-duplication Performance?

Himanshu S. Bhatt, Richa Singh and Mayank Vatsa
IIIT-Delhi, INDIA
{himanshub,rsingh,mayank}@iiitd.ac.in

Abstract

With the prevalent utilization of citizen databases, an individual has to prove his/her identity for accessing several services such as banking, health care, and social welfare benefits. These databases are now increasingly using demographic and biometric information to uniquely identify the individuals. It protects the core identity of citizens and facilitates them to receive the entitled benefits and rights. It is therefore important that every citizen should enroll only once in the database and be assigned only one unique identifier. De-duplication process prevents an individual from enrolling multiple times in the database. It is essential to understand the importance of constituent information (demographic and biometric) in the de-duplication process. Using a large database, this research attempts to fill the gap in existing literature by analyzing the performance of demographic and biometric information for de-duplication. The study presents the results when demographic and biometric information are individually processed and complementary information from the two modalities are combined at match score level for de-duplication under different operating scenarios.

1. Introduction

Identity science is pertinent in our daily life and individuals have to prove their identity for availing services such as welfare programs, financial inclusion programs, and border security. Further, services such as issuance of birth certificate, driving license, and passport require the individuals to uniquely establish their identity. Several large scale national ID projects including India's UIDAI (Aadhar project) [1] offer 'establishing identity and matching as a service' to banks, local government agencies and institutions that need to verify the identity of individuals. This ensures that everyone receives the entitled social welfare benefits and rights. With advances in biometrics, several countries are increasingly using both demographic and biometric information for uniquely identifying individuals. Especially, large scale ID programs associate demographic information with biometric data to assign a unique identification to every citizen [1]. Biometric information generally includes face, fingerprint, and iris, whereas, demographic information includes name, address, date of birth, place of birth, relationship with the family head, and gender [5]. Such national ID projects protect the core identity of the citizens and allow them to securely assert their identity for several services. A unique identity, in the form of a number, thus alleviates the need to produce multiple document proofs to establish one's identity.

To ensure smooth and fair dispersion of benefits under different welfare programs, prevent cornering of benefits by only a few individuals, and minimize frauds, it is essential that every citizen should get only one unique identity. Therefore, de-duplication is a critical process in such large scale projects. De-duplication involves preventing records from being stored multiple times in a database or eliminating existing multiple copies from the database. In a national ID project, information of unique identities (individuals) is stored in the database during enrollment and a unique identity is assigned to a user. However, before assigning a unique identity, the information of a new enrolling user is compared with all the existing identities to check for possible duplicate identities in the database. De-duplication has been extensively studied by researchers in database and information management systems [7, 9]. With increasing number of large scale identity programs using demographic and biometric information, the significance of de-duplication process has been realized in biometrics community as well.

Tyagi *et al.* [16] proposed a likelihood ratio based match score fusion approach to fuse biographical and biometric information for improved identification performance. However, we believe that it is important to analyze and understand the usability and applicability of a de-duplication process using both demographic and biometric information in large scale applications. This research addresses the question *"whether fusing demographic and biometric informa-*

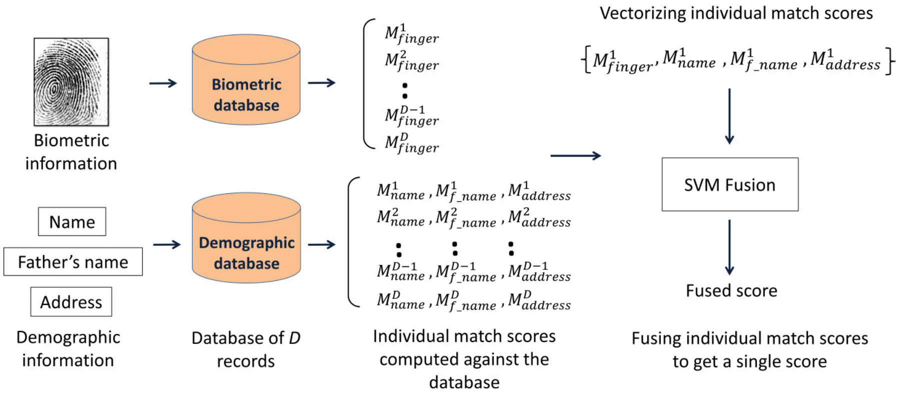
Vectorizing individual match scores
$\{M^1_{finger}, M^1_{name}, M^1_{f\_name}, M^1_{address}\}$
Biometric information
Biometric database
M^1_{finger}
M^2_{finger}
M^{D-1}_{finger}
M^D_{finger}
SVM Fusion
Fused score
Name
Father's name
Address
Demographic database
$M^1_{name}, M^1_{f\_name}, M^1_{address}$
$M^2_{name}, M^2_{f\_name}, M^2_{address}$
$M^{D-1}_{name}, M^{D-1}_{f\_name}, M^{D-1}_{address}$
$M^D_{name}, M^D_{f\_name}, M^D_{address}$
Demographic information
Database of D records
Individual match scores computed against the database
Fusing individual match scores to get a single score

Figure 1. Block diagram of the proposed learning based fusion technique for de-duplication.

tion enhances the de-duplication performance?". Different case studies are analyzed for de-duplication using (1) only demographic, (2) only biometric, and (3) both the modalities together. This study also presents the results when either demographic or biometric data is forged by a user for re-enrollment. The analysis is construed using the proposed learning based Support Vector Machine (SVM) [4] fusion algorithm which combines information from two modalities at the match score level. Section 2 elaborates the learning based fusion algorithm and Section 3 provides details about the database, experimental protocol, and analysis.

2. Learning Based Fusion for De-duplication

The major challenge for de-duplication in a large scale application is the huge number of records in the database. It is observed that the performance of a biometric based system becomes unreliable with increasing database size as it tends to accumulate false accepts. In literature, it has been shown that combining complementary information from multiple evidences leads to improved performance [14]. Therefore, to utilize this observation, demographic information is combined with biometric information to make the de-duplication process robust and scalable. Though, existing rule based fusion techniques [6] require manual attuning of de-duplication rules (such as deciding threshold for two elements to be considered a match), the proposed algorithm obviates this need. In the proposed algorithm, as shown in Figure 1, match scores from different demographic and biometric fields are combined using SVM. Different stages of the proposed algorithm are elaborated in the following subsections.

2.1. Demographic Information Processing

Demographic information is utilized as one of the modalities for detecting duplicates in a large database. Demographic data includes fields such as name, father's/husband's name, address, age, and gender [5]. Due to the lack of naming and address standards (specially in developing countries), demographic information obtained from different sources may have large variations that increase the possibilities of skipping a duplicate in the database. These variations are caused due to typographical errors, missing or unknown data, and different representations/interpretations of the same information. Table 1 illustrates such variations in the demographic data (name and address) obtained from two different sources. To convert the data obtained from different sources in a consistent and uniform format, it is first segmented into different subfields; for example, name is segmented into first, middle and last name, address is segmented into house number, street, locality, area, and pin code. Once the information from different sources is segmented and standardized, two records are matched using Levenshtein edit distance [10] between the corresponding fields of the demographic data. The Levenshtein edit distance between two strings is the minimum number of insertions, deletions, and substitutions of characters that will transform one string into the other. Mathematically, Levenshtein edit distance between two strings a and b, $l_{a,b}(|a|,|b|)$ is computed as shown in Eq. 1,

$$l_{a,b}(i,j) = \begin{cases} max(i,j), & if \quad min(i,j) = 0 \\ \\ min \begin{cases} l_{a,b}(i-1,j)+1 & , \quad else \\ l_{a,b}(i,j-1)+1 \\ l_{a,b}(i-1,j-1)+[a_i \neq b_j] \end{cases} \end{cases} \tag{1}$$

where i and j are the length of strings a and b respectively.

2.2. Biometric Information Processing

Fingerprint recognition is one of the oldest and well-known biometrics used in several applications because of its uniqueness and consistency over time [11]. Fingerprint as a biometric is more mature as compared to other biometric modalities because of the ease in acquisition, established usage, and collection by law enforcement agencies for var-

Table 1. Illustrating the variations in name and address in demographic information obtained from two different sources. *The table presents a close illustration of actual variations in demographic data from the source1 and source2.

Source 1		Source 2	
Name	**Address**	**Name**	**Address**
T I DHAMECHA	P-95 VILLAGE PILLANJI SAROJINI NAGAR-110023	TEJAS INDULAL DHAMECHA	92 P PILLANJI VILLAGE SRJNI NAGAR 23
H S BHATT	56A MIG FLATS POCKET F HARI NAGAR-110064	BHATT HIMANSHU SHARAD	F 56 A GRD FLOOR M I G FLATS HARI NGR 64
SHARAD BHATT	695 LIG FLATS POCKET-B HASTSAL UTTAM NAGAR-110059	BHATT SHARAD KUMAR	695-B DDA FLATS HASTSAL, UTTAM NAGAR 59
AGGARWAL PRAFUL	GH-1/178 BLOCK GH 1 ARCHNA APTARTMENT PASCHIM VIHAR-110063	PRAFULA AGGRAWAL	178 IST FLOOR D D A FLATS G H I PASCHIM VIHAR 63
ANUJ SHANKAR SAXENA	1296 BLOCK F EAST OF KAILASH	SAXENA A S	F 1296 GRD FLOOR GALI 6 EAST OF KAILASH 65

ious applications. Fingerprint matching algorithms generally use minutiae based approach which first locates minutia points and then maps their relative placement on the fingerprint. In this research, the open source NIST Biometric Image Software (NBIS) [12] is used to match fingerprints. The software consists of a minutiae detector called MINDTCT [12] which automatically locates and records ridge ending and bifurcations in a fingerprint image. For comparing two fingerprint templates, a minutiae based fingerprint matcher, BOZORTH3, is used [12].

2.3. Learning Based Fusion

Existing approaches [6] for de-duplication use different thresholds based on the uniqueness and discriminative abilities of different cues. The strict fields such as house number and pin code are given more emphasis and a smaller edit distance is required for matching, whereas, for a lenient field a larger edit distance is allowed for matching. These techniques require manual tuning of thresholds which is not a pragmatic solution in large scale applications. In this research, a learning based fusion algorithm is proposed that utilizes match scores from individual fields in demographic and biometric data for de-duplication. Individual fields in multiple modalities have varying significance in detecting duplicates from the database based on their discriminative abilities. It is our assertion that optimally combining scores from individual fields can classify a new record as a 'duplicate' or 'non-duplicate'. SVM inherently learns the significance of these scores and provides an efficient way to combine multiple scores. Therefore, SVM based fusion of individual match scores is proposed to classify an instance as 'duplicate'/'non-duplicate'. The training and de-duplication process using SVM is explained as follows:

Training the SVM: From a training set, individual match scores for each field in the demographic and biometric data are computed, represented as M_{finger} for fingerprint, M_{name} for name, M_{f-name} for father's/husband's name, and $M_{address}$ for address. Individual match scores are then vectorized to form an input vector as shown in Eq. 2

$$\mathbf{u} = \{M_{finger}, M_{name}, M_{f-name}, M_{address}\} \quad (2)$$

where $\mathbf{u}$ is the match score vector. The vector is assigned a label $z \in \{-1, +1\}$ where $\{-1\}$ represents a 'duplicate' and $\{+1\}$ represents a 'non-duplicate'. SVM is trained to classify the input match score vector as 'duplicate' or 'non-duplicate'. SVM is trained using the approach proposed by Phillips [13] and used in identification mode where a score is computed between the query and each of the database records based on the distance from the decision hyperplane. The query is declared matched to the record in the database with the minimum score. In our experiments, Radial Basis Function (RBF) kernel with gamma parameter 6 is used.

De-duplication: The proposed de-duplication process is explained below:

1. A database of D records comprising demographic and biometric information is formed.

2. When a new enrollee, q, presents his/her demographic and biometric data for enrollment, this information is compared with the corresponding information of all the existing users in the database. The match scores represented as M^q_{finger}, M^q_{name}, M^q_{f-name}, and $M^q_{address}$ corresponding to fingerprint, name, father's name, and address respectively are computed.

3. The individual match scores obtained for each demographic and biometric fields are vectorized as:

$$\mathbf{u_i'^q} = \{M^q_{finger}, M^q_{name}, M^q_{f-name}, M^q_{address}\} \quad (3)$$

   where $\mathbf{u_i'^q}$ is the vector of individual match scores computed between the information presented by a new enrollee q and the i^{th} record in the database. The vector $\mathbf{u_i'^q}$ is then provided as input to the SVM.

4. The trained SVM combines individual match scores and computes a single score (based on the distance from the decision hyperplane). This process is repeated for all D records in the database.

$$y^D_{i=1} = SVM(\mathbf{u_i'^q}) \quad (4)$$

where y_i is the score between the new enrollee and the i^{th} record in the database.

5. Duplicate records in the database are detected based on the final scores, $y_{i=1}^{D}$, computed by the SVM.

3. Experimental Evaluation

Several experiments are performed to measure the usability and effectiveness of demographic and biometric information for detecting duplicates in large scale databases. In our experimental evaluation, different combinations of the two modalities are used to evaluate the de-duplication performance. Section 3.1 explains the database used in this research, Section 3.2 elaborates the experimental protocol, and Section 3.3 summarizes the key results and analysis from the experimental evaluation.

3.1. Database

Publicly available fingerprint databases are combined to form a heterogeneous fingerprint database comprising 5734 classes with 2 samples per class. The heterogeneous fingerprint database comprises 3500 classes from CASIA fingerprint V5 [2], 1000 classes from MCYT [15], 1084 classes from WVU multi-modal [3], and 150 classes from FVC 2006 [8] databases. In order to simulate the complexity of a real world large scale de-duplication application, we also created an extended gallery of additional $10,000$ fingerprint classes with single sample per class obtained from a law enforcement agency. Only a single unit biometric evidence is available in many real world applications, therefore, the evaluation is performed with single unit single sample fingerprint per user.

In addition to biometric data, demographic data pertaining to 5734 individuals is collected from two different sources, termed as *source1* and *source2*. The demographic data from source1 contains name, father's/husband's name, gender, age, and address. The demographic information from source2 contains only name and address. Random distortions are introduced in father's/husband's name from source1 and assigned to the corresponding field in source2. It involves randomly replacing first and middle names with their initials, removing middle name, and introducing few typographical errors. Name, father's/husband's name, and address are used as the three demographic fields in this research. Each of the 5734 individuals is associated with two instances of the demographic information, one from source1 and another from source2 pertaining to the same individual. It presents a scenario where an individual has multiple document ID proofs that can be used to re-enroll in the database. Demographic information for the extended gallery of $10,000$ classes is also obtained from source1. The demographic data is further pre-processed using a set of rules based on the domain knowledge.

3.2. Experimental Protocol

Three different case studies are used to evaluate the efficacy of demographic and biometric information for de-duplication in large scale databases. SVM is trained using the information from 2000 individuals and the performance is evaluated on the remaining 3734 individuals. In each of the experiments, $13,734$ $(10,000 + 3734)$ individuals are enrolled in the database and 3734 users attempted to re-enroll in the database.

1. **Only demographic data is used for de-duplication:** In this case only demographic information is used to check if a user is already enrolled in the database. Many existing systems such as passport and driver's license issuing authorities (especially in developing countries) utilize only the demographic data to verify the identity of an individual and check whether there exists a duplicate in the database.

2. **Only biometric data is used for de-duplication:** In this case only fingerprint is utilized for de-duplication. It replicates the real world scenario where only biometric information is used for de-duplication or screening against a watch-list database.

3. **Both demographic and biometric data are used for de-duplication:** In this case, both demographic and biometric data are simultaneously utilized for de-duplication using the proposed learning based fusion algorithm. The paper further analyzes the scenario when a user attempts to re-enroll in the database by forging either of the two modalities as elaborated below:

   - **Demographic information is forged while biometric is genuine:** In this case, a user furnishes forged demographic information with genuine biometric information to re-enroll in the database. It represents a real world case where a user associates a stolen identity (i.e. demographic information) with his/her biometric information. Such duplicates are difficult to detect in systems that use only demographic information for de-duplication.

   - **Biometric is forged and demographic information is genuine:** In this case, a user attempts to re-enroll in the database by presenting genuine demographic information along with forged biometric information[1]. It represents the scenario where a user associates a stolen or fake biometric identity with correct demographic information.

[1] To know more about how fingerprints can be faked/spoofed, readers are directed to [11, 17].

- **Both biometric and demographic information are genuine:** In this case, a user attempts to re-enroll in the database by providing genuine demographic and biometric information. However, the demographic information is obtained from source2 which has variations as compared to the demographic information from source1 (already stored in the database for that user). This replicates a real world scenario where a user has multiple document IDs with variations in details such as name and address. The fingerprint (biometric) may also have variations due to rotation, pressure, moisture, and scars.

3.3. Results and Analysis

De-duplication is a critical process especially in large scale programs that involve huge number of records in the database. It ensures that every individual is enrolled only once in the database, therefore, nobody can exploit undue benefits by multiple enrollments. In this research, rank-1 de-duplication performance is reported which signifies that when a user attempts to re-enroll in the database, the correct identity (already enrolled in the database) is retrieved at the top. The key analysis and observations from the experimental evaluation are listed below:

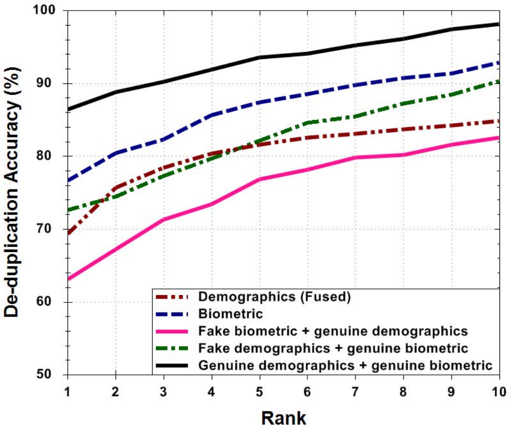

Figure 2. CMC curves showing the performance of the proposed learning based fusion algorithm for de-duplication.

Table 2. De-duplication performance using only demographic and biometric information.

Protocol	Field	Rank-1 Accuracy
Using demographic	Name	25.2%
	Father's name	28.6%
	Address	54.8%
	SVM fusion	69.4%
Using biometric	Fingerprint	76.6%

- Table 2 reports the de-duplication performance using individual fields in demographic data and when match scores from different demographic fields are combined using the proposed SVM fusion. De-duplication using demographic data achieves 69.4% rank-1 accuracy due to the lack of standards for representing name and address. Users generally possess multiple document IDs, therefore, demographic data exhibits large variations when a user provides genuine data but from different document ID proofs. Demographic data can be easily forged; hence, de-duplication based on demographic data alone may not be a good solution for large scale applications.

- Fingerprint based de-duplication yields rank-1 accuracy of 76.6% which is better than that of the demographic data. However, with large database size, the performance of fingerprint based de-duplication suffers as it tends to accumulate more false accepts with increasing database size. Moreover, the incidents of spoofing or altering fingerprints [17] are reported quite frequently that dissuade the sole use of fingerprints for de-duplication. The results in Table 2 suggest that neither demographic nor biometrics is independently sufficient to detect duplicates in large scale systems.

- Since individual modalities (demographic or biometric) are not sufficient for de-duplication in large databases, complementary information from different modalities is simultaneously utilized. The results reported in Table 3 suggest that SVM based match score fusion algorithm enhances the de-duplication performance by at least 10%. However, this improvement is observed only when genuine biometric and demographic information are furnished during enrollment.

- The results in Figure 2 and Table 3 suggest that the de-duplication performance degrades when a user forges either of the two modalities. It is observed that the performance of the proposed fusion algorithm is lower than the performance of either of the modalities (performance drops by at least 4%). This drop in performance is attributed to the fact that the modality being forged retrieves random records. Fusing information under such circumstances may allow a user to deceive the de-duplication process and re-enroll in the database.

- It is alarming to observe that by forging either of the modalities, an adversary may increase its chances of successfully re-enrolling in the database as a duplicate. It is a huge challenge in the present scenario where many large scale programs are using demographic and biometric information for de-duplication without any

Table 3. De-duplication performance using combination of demographic and biometric information.

	Rank-1 Accuracy (%)		
	Demographic	**Biometric**	**SVM Fusion**
Fake biometric + genuine demographic	69.4	0.0	63.1
Genuine biometric + fake demographic	0.0	76.6	72.6
Genuine biometric + genuine demographic	69.4	76.6	86.5

mechanism to address its menaces. The results suggest that fusing demographic and biometric data may not always enhance the de-duplication performance.

- When both demographic and biometric are forged, information from any of these modalities does not facilitates the de-duplication process. Therefore, with forged demographic and biometric information, an individual can potentially re-enroll in the database.

4. Conclusion and Future Work

This research presents a study on the usability and relevance of simultaneously using demographic and biometric information for de-duplication in a large database under different operating scenarios. It presents a learning based match score fusion algorithm for combining complementary information from demographic and biometric data. The proposed SVM based fusion obviates the need for manual tuning of individual rules for every field in the demographic and biometric data. The results suggest that the proposed algorithm for fusing demographic and biometric information is robust and scalable for de-duplication in large scale projects only when the user provides genuine demographic and biometric information. However, it also suggests that forging either demographic or biometric data may deceive current de-duplication process and leads to duplicates in the database. Therefore, further research efforts are required to develop mechanisms that can prevent multiple enrollments of the same individual.

In a large scale system, new individuals enrolling in the database continuously change the data distribution i.e. the 'duplicate'-'non-duplicate' match score distribution which degrades the performance of existing algorithms for de-duplication. Therefore, as a future work, we are working on developing an algorithm that can adapt the fusion rules to accommodate the variations in the data distribution with increasing database size. We are also extending the proposed algorithm to accommodate multiple biometric evidences such as multiple finger units and face images.

References

[1] Adhaar project by UIDAI: http://uidai.gov.in/. 1

[2] CASIA-FingerprintV5: http://biometrics.idealtest.org/. 4

[3] S. Crihalmeanu, A. Ross, S. Schuckers, and L. Hornak. A protocol for multibiometric data acquisition, storage and dissemination. Technical report, WVU, 2007. 4

[4] N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, USA, 2000. 2

[5] Demographic Data Standard and Verification Procedure (DDSVP) Committee Report: http://uidai.gov.in/UID_PDF/Committees/UID_DDSVP_Committee_Report_v1.0.pdf. 1, 2

[6] J. Dinerstein, S. Dinerstein, P. K. Egbert, and S. W. Clyde. Learning-based fusion for data deduplication. In *Proceedings of International Conference on Machine Learning and Applications*, pages 66–71, 2008. 2, 3

[7] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16, 2007. 1

[8] Fingerprint Verification Competition FVC 2006 : http://bias.csr.unibo.it/fvc2006/. 4

[9] K. Goiser and P. Christen. Towards automated record linkage. In *Proceedings of Australasian Conference on Data Mining and Analystics*, pages 23–31, 2006. 1

[10] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966. 2

[11] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition*. Springer, USA, 2009. 2, 4

[12] NIST Biometric Image Software NBIS : http://www.nist.gov/itl/iad/ig/nbis.cfm/. 3

[13] P. J. Phillips. Support vector machines applied to face recognition. In *Proceedings of Advances in Neural Information Processing Systems*, pages 803–809, 1999. 3

[14] A. Ross and A. Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115–2125, 2003. 2

[15] D. Simon-Zorita, J. Ortega-Garcia, J. Fierrez-Aguilar, and J. Gonzalez-Rodriguez. Image quality and position variability assessment in minutiae-based fingerprint verification. *IEE Proceedings - Vision, Image and Signal Processing*, 150(6):402–408, 2003. 4

[16] V. Tyagi, H. P. Karanam, T. A. Faruquie, L. V. Subramaniam, and N. Ratha. Fusing biographical and biometric classifiers for improved person identification. In *Proceedings of International Conference on Pattern Recognition*, pages 2351–2354, 2012. 1

[17] S. Yoon, J. Feng, and A. K. Jain. Altered fingerprints: Analysis and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):451–464, 2012. 4, 5