Pattern Recognition Letters
journal homepage: www.elsevier.com/locate/patrec

Adaptive fusion of biometric and biographic information for identity de-duplication

Prem Sewak Sudhish[a*], Anil K. Jain[b], and Kai Cao[b]

[a] *Department of Physics and Computer Science, Dayalbagh Educational Institute, Dayalbagh, Agra, UP 282005, India*
[b] *Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA*

ABSTRACT

Use of biometrics for person identification has increased tremendously over the past decade, e.g., in large scale national identification programs, for law enforcement and border control applications, and social welfare initiatives. For such large scale applications with a diverse target population, unimodal biometric systems, which use a single biometric trait (e.g., fingerprints), are inadequate due to their limited capacity. Multimodal biometric systems, which fuse multiple biometric traits (e.g., fingerprints and face), are required for large-scale identification applications, e.g., de-duplication where the goal is to ensure that the same person does not have two different official credentials (e.g., national ID card) based on different credentials. While multimodal biometric systems offer several advantages (e.g., improvement in recognition accuracy, decrease in failure to enroll rate), they require large enrollment and de-duplication times. This paper proposes an adaptive sequential framework to automatically determine which subset of biometric traits and biographic information is adequate for de-duplication of a given query. An analysis of this strategy is presented on a virtual multi-biometric database of 27,000 subjects (fingerprints from NIST SD14 dataset and face images from the PCSO dataset) along with biographic information sampled from the US census data. Experimental results, using three-fold cross-validation, show that without any loss in de-duplication accuracy, on average, for 63.18% (of a total of 27,000) of the queries, only fingerprint capture is adequate, for an additional 28.69% of queries, both fingerprint and face are required, and only 8.13% of the queries needed biographic information in addition to fingerprint and face.

Keywords: Identity de-duplication; biometric traits; biographic information; adaptive fusion; sequential methods.

* Corresponding author. Tel.: +91-562-2801545; fax: +91-562-2801226; e-mail: pss@alumni.stanford.edu

1. Introduction

Biometric systems are becoming ubiquitous for automated person recognition. These systems are based on the measurement of purportedly unique anatomical (e.g. fingerprint) and behavioral traits (e.g. handwriting) of an individual [1]. The applications of biometric systems range from traditional security applications (e.g., forensics and surveillance) to more recent applications such as mobile user authentication and social welfare programs.

The growing prevalence of biometrics is evident from the large scale deployments at the national level, such as the Biometric Identity Management System of the US Department of Homeland Security[1] and the Aadhaar project of the Government of India[2]. The primary purpose of using biometrics in large scale applications is to ensure that no individual is able to assume more than one identity, e.g., for obtaining national identity cards and social benefits. The process of detecting and removing duplicate identities is commonly referred to as de-duplication. Besides, de-duplication is also required in many other applications, e.g., a user may assume duplicate or false identities in border protection or law enforcement applications, or to avoid harsher penalties for repeat offenders[3]. In a civilian scenario, multiple identities may be used by an individual for tax evasion[4]. Identity de-duplication is also needed when several biometric databases are merged together [2], e.g. databases from multiple law enforcement agencies.

Large-scale biometric systems that rely on a single instance of biometric evidence (unimodal systems) suffer from limitations such as limited capacity to distinguish between a large number of identities, non-universality (absence of a biometric trait), noisy data, and spoof attacks [3]. These limitations can be overcome to a large extent by using multimodal systems[5] that fuse information from multiple biometric traits [4].

In addition to biometric traits, biographic information (e.g. person's name) can also be used for de-duplication. A few prior studies [5, 6] have shown that fusion of biographic information with biometrics can improve the de-duplication accuracy. However, biographic information has the following limitations [7, 8]: (a) data entry errors by human operators; (b) lack of a standard format and standard transliteration; (c) similar biographic information (e.g. name) of different individuals; and (d) data integrity issues due to change in certain biographic information e.g., change of address.

De-duplication in multimodal systems is typically performed by comparing query biometric and biographic information against the records stored in the reference database. The final decision on presence of duplicate identity is based on the fusion of scores from comparison of biometric and biographic information. To maintain the integrity of the biometric system [9, 10, 11], the identification system should be highly accurate and computationally efficient. Consider the scenario of a national identification system (e.g., Aadhaar) with a target enrolment of over a billion individuals. Even a conservative estimate of false positive identification rate (FPIR) of the order of 0.0025% translates to tens of thousands of individuals being falsely identified as duplicates.

This paper proposes an adaptive sequential fusion of biometric and biographic information for efficient de-duplication. For each query, the proposed system selects only those biometric traits and biographic information deemed necessary, by the sequential selection strategy, to maintain de-duplication accuracy. Experimental results on a virtual multimodal database[6] (fingerprint images from the NIST SD 14 fingerprint database [12], mugshots from the PCSO database [13], along with biographic information spawned from the US Census [14] for 27,000 subjects) show the viability of the proposed scheme. More specifically, our system correctly performs de-duplication of 63.18% of the queries only using fingerprints. For an additional 28.69% of the queries, fusion of both fingerprint and face scores allows for correct de-duplication. Further, incorporation of biographic information is able to de-duplicate the identities of the remaining queries.

The rest of the paper is organized as follows. Section 2 briefly reviews prior work on biometric and biographic fusion. Section 3 describes the proposed algorithm and explains the rationale behind it. Section 4 presents our experimental results along with a comparison with published studies. The system limitations and directions for future research have been presented in Section 5.

2. Background

De-duplication of identities in large-scale applications consists of comparison and fusion of biometric and biographic information.

2.1. Fusion of multibiometric information

In most large-scale multi-modal biometric systems, the three traits most commonly used for de-duplication are fingerprint, iris and face. Fusion of complementary traits, such as a fingerprints and iris, is known to provide superior de-duplication performance [15]. In general, biometric traits can be fused at different levels: sensor level, feature level, match score level, rank level and decision level. Score level fusion which is the most widely used procedure is adopted here [16].

2.2. Matching of biographic information

Typical biographic information used for de-duplication consists of a person's name, his father's name, and address. While the use of biographic information in conjunction with biometric information, has been shown to improve de-duplication accuracy [5, 6], biographic information alone can lead to possible de-duplication errors. This is because different individuals may share the same biographic information, e.g., name and address. The choice of biographic similarity score depends on the biographic data type. For nominal data (e.g., gender, race), the similarity is binary ("same" or "not same"). On the other hand, for textual data approximate string matching distance (e.g., Levenshtein distance [17]), is often used. Other metrics specific to data types are also sometimes used, e.g., geospatial distance to compare addresses.

2.3. Score normalization

The similarity scores for individual biometric traits and biographic information are generated by different algorithms and, therefore, may have different upper and lower bounds (e.g., [0,100], [0,1], etc.). The standard practice, in such a case, is to normalize the scores to a common range prior to fusion. The choice of score normalization scheme (e.g., min-max or z-score normalization), in general, depends on the underlying score

[1] www.dhs.gov/obim
[2] uidai.gov.in
[3] www.law.stanford.edu/organizations/programs-and-centers/stanford-three-strikes-project/three-strikes-basics
[4] businesstoday.intoday.in/story/tax-evaders-hold-multiple-pan-cards-cag-report/1/14157.html
[5] www.nist.gov/itl/idms/nextgen_biometrics.cfm

[6] Virtual multimodal databases contain records which are created by consistently pairing a user from one unimodal database (e.g., face) with a user from another database (e.g., fingerprint). The creation of virtual users is based on the assumption that different biometric traits of the same person are independent [15].

distributions. Given the normalized scores, different fusion strategies can be used, e.g., density based, classifier based, quality based, or based on a dynamic score selection strategy [4].

2.4. Prior Work

Tyagi et al. fused biographic information (name and address) with biometric similarity scores (fingerprint and face) in NIST BSSR1 dataset [5, 18]. They showed that the recognition accuracy improves from 94.73% when no biographic information is used to 98.93% after fusion. Bhatt et al. [6] fused person's name, father's name, and address with fingerprint. The recognition accuracy when only fingerprint is used is 76.6%, which improves to 86.5% when biographic information is fused with fingerprints. There are two limitations of [5, 6]: (i) the state-of-the-art comparison algorithms were not used, and (ii) the computational efficiency was not considered. Due to the complementary nature of biometric and biographical information, a few commercial systems that fuse biometric traits and biographic data [19] are available as well. Table 1 compares published studies with the proposed study.

Table 1. Comparison of studies on fusion of biometric and biographic information

Study	Target application	Biometric trait and database	Biographic information and database	Matching algorithm and fusion strategy	Accuracy	Comments
Bolme et al. [20]	Person (celebrity) identification	1,331 face images of 118 celebrities	Textual Information (~400 words) from celebrity websites	**Biometric**: EBGM for face **Biographic**: Cosine of angle between word frequencies **Fusion**: Weighted sum	**Biometric**: 22% **Biographic**: 22% **Fusion**: 35%	All scores fused for every query. Small database.
Tyagi et al. [5]	De-duplication	Two fingerprint match (left and right index) scores of 3K subset (1.5K each for training and testing) from NIST BSSR1 [18]	Names and addresses from an electoral record dataset	**Biometric**: Precomputed fingerprint scores in BSSR1 **Biographic**: Matching algorithm not specified **Fusion**: Log-likelihood ratio	**Biometric**: 94.73% **Biographic**: 84.40% **Fusion**: 98.93%	All scores fused for each query Small database.
Bhatt et al. [6]	De-duplication	Fingerprints of 5,734 subjects (2K for training and 3,734 for testing) from various datasets. Gallery augmented with additional 10K fingerprints.	Name, father's or husband's name, address	**Biometric**: NIST NBIS [21] **Biographic**: Levenshtein distance for string matching **Fusion**: SVM	**Biometric**: 76.6% **Biographic**: 69.4% **Fusion**: 86.5%	All scores fused for each query. Reported accuracy not sufficient for de-duplication.
Proposed study	De-duplication	Fingerprints of 27K subjects from NIST SD 14 [12] augmented with face images of 27K subjects from PCSO [13].	Gender, name and father's name. Names derived from US Census data [14]. Gender is extracted from the PCSO face dataset.	**Biometric**: State of the art COTS matchers for fingerprint and face. **Biographic**: Combination of Levenshtein [17], Damerau-Levenshtein [22] and editor distances [23]. **Fusion**: Proposed adaptive sequential fusion algorithm.	**Biometric**: 99.64% **Biographic**: 97.47% **Fusion**: 100.0%	Fingerprint alone is adequate for 63.18% of the 27K queries; face required for only 36.82% of the queries; biographic information required only for 8.13% of the queries.

*Accuracy is the percentage of subjects for whom the true mate is retrieved at rank 1; COTS matcher stands for Commercial Off-the-Shelf matcher.

3. Proposed de-duplication framework

A drawback of the fusion strategies proposed in the literature, as well as those commonly used in practice, is that they are static in the sense that, once the traits are fixed, all the corresponding scores are computed and fused for every query. To address this limitation, we present a sequential selection strategy that determines, in real time, which subset of biometric and biographic information is adequate for a given query.

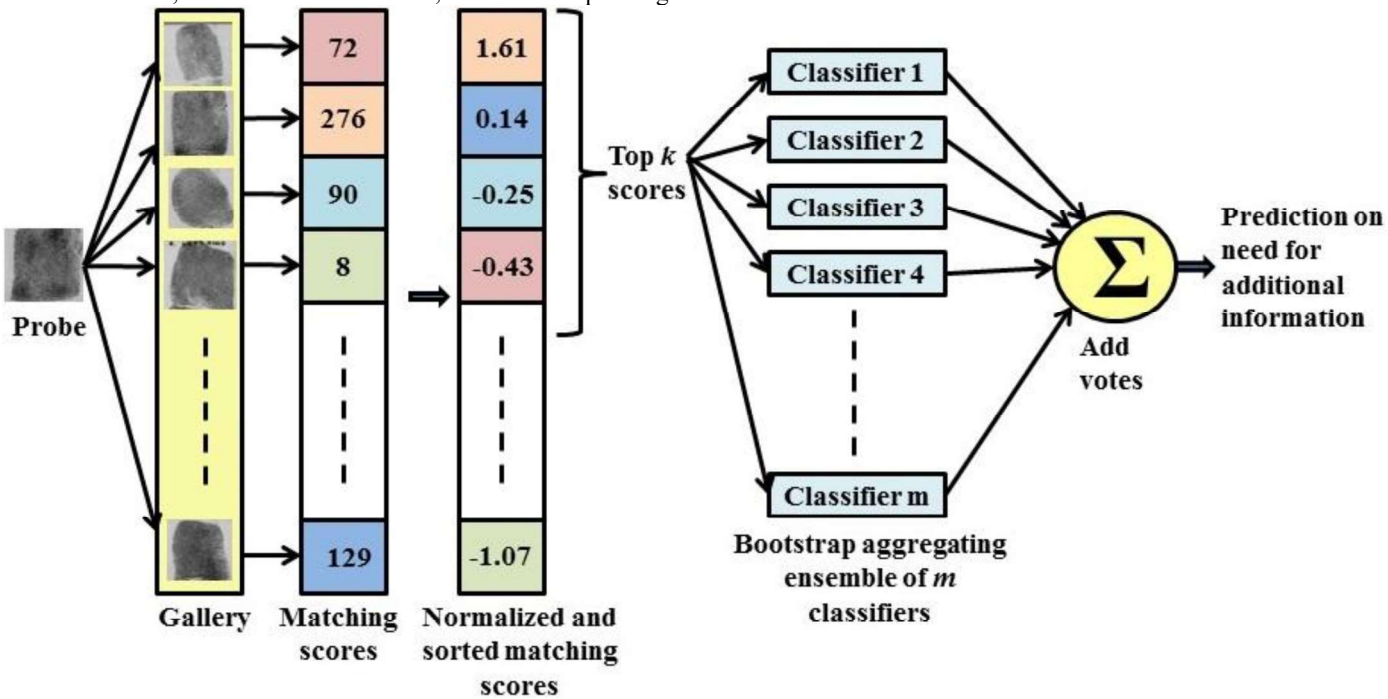


Fig. 1. Proposed de-duplication framework. Additional biometric or biographic identifiers are selected at each stage by an ensemble of logistic classifiers; fingerprint match scores are used here for illustration.

The proposed adaptive fusion algorithm is based on the principle of sequential fusion [15]. The order in which the biometric traits and biographic information is presented to the system is determined according to their discriminability[7]. Based on the comparative discriminability assessment of traits using the training data, traits are selected in the following order: (i) fingerprint, (ii) face and (iii) biographic information (see footnote 7).

Once a user's biometric or biographic information is presented to the system, an ensemble of veto-wielding [24] logistic regression classifiers [25] predict whether the corresponding rank-1 score represents a genuine match. The number of classifiers in the ensemble is chosen based on the trade-off between desired accuracy and computational effort. A schematic diagram of this prediction model is shown in Fig. 1.

Decision threshold for the ensemble is based on the probability of false accept. It is set to an extremely low value (e.g., 10^{-6}) which is learned from a training set. If all the classifiers in the ensemble agree that the rank-1 score represents a genuine match, no additional biometric trait or biographic information is needed to determine the identity of the user in the query.

The dual safeguard of using a small probability of false accept as the threshold for prediction, coupled with the authority of each classifier to exercise a veto, diminishes the chances of premature termination of adaptive fusion.

3.1. Matching of biometric and biographic Information

We utilize state-of-the-art commercial-off-the-shelf (COTS) matchers to obtain comparison scores for both fingerprint and face images. Due to licensing restrictions, we are not able to disclose the names of the vendors, but both of these matchers rank in the top three in recent NIST evaluations for fingerprint and face[8].

Levenshtein [17], Damerau-Levenshtein [22], and Editor distances [23] are used as distance metrics for biographic information. Levenshtein distance between two strings is the minimum number of single-character insertion, deletion or substitution operations required to transform one string to the other. Damerau-Levenshtein distance also allows for transposition between two adjacent characters. Editor distance is similar to the Levenshtein distance except that substitutions are treated as two separate operations – insert and delete. The edit distance is converted to a similarity value by first normalizing it in the [0, 1] range by dividing it by the maximum possible edit distance between two strings of the same lengths as the given pair of strings. The corresponding similarity between two strings is simply (*1 - normalized edit distance*). The final similarity measure for the biographic information is the mean of similarities derived from Levenshtein, Damerau-Levenshtein, and Editor distances.

3.2. Proposed adaptive fusion algorithm

Our proposed adaptive fusion algorithm is motivated by Arora et al. [26] who proposed a strategy to determine whether the true mate of a latent fingerprint query matches at rank-1 out of the top-*k* retrieved images. This was based on the "upper outlier" in the similarity score distribution, under the assumption that scores follow an exponential distribution. Intuitively, the presence of a single upper outlier is a strong indication of a true mate at rank-1 (correct decision) because of the abysmally low probability of two events occurring simultaneously viz., a false match with a very high match score and a true mate with an extremely low score. Arora et al. adopted this strategy to determine whether additional feature markup is needed for the latent query [26].

An optimal parametric distribution that fits the biometric and biographic match scores may not be available, so a soft computing approach is proposed. The proposed model consists of an ensemble of *m* veto-wielding [24] logistic regression classifiers [25]. A logistic regression classifier is based on the logistic (sigmoid) function used for the two-class classification problem. The logistic function of a variable *z* is given by eq. (1).

$$f(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

The value of the function *f(z)* lies in (0, 1), and for the purpose of classification here, is seen as the probability of the alternate hypothesis (H_1: additional biometric or biographic information is required); the probability of null hypothesis (H_0: the information presented is sufficient to reach a final decision) is ($1 - f(z)$).

The variable *z* is a weighted sum of *k* input features used for classification, as given by eq. (2).

$$z = \sum_{i=0}^{k} \theta_i s_i, \tag{2}$$

where s_i is the i^{th} feature and θ_i is the weight assigned to s_i. The bias term s_0 is set to unity.

The proposed algorithm uses the top *k* scores (with the highest scores for the subject being de-duplicated against enrolled subjects) as inputs s_i to the logistic function. The value of *k* is chosen to avoid overfitting. The output *f(z)* is interpreted as the probability that the rank-1 score does not represent a genuine match. Additional biometric or biographic information is deemed to be necessary if this probability is above a pre-determined threshold η. Each of the *m* classifiers in the ensemble is trained on different training subsets obtained via bootstrapping [27]. Each individual classifier in the ensemble wields a veto such that the process of further matching and fusion of available biometric and biographic information is terminated only when all classifiers in the ensemble unanimously [24] predict that additional information is not required. A high level algorithm of the proposed adaptive fusion method is given in Table 2.

Table 2. High level description of proposed adaptive fusion algorithm

Input: Training set $D = \left\{\left(\langle \boldsymbol{x}_j^{(i)}{}_{raw}\rangle_{j=1}^{n}, y^{(i)}\right)\right\}$

Output: Whether or not $y^{(i)}$ is at rank-1

Normalization: z-score normalize: $\boldsymbol{x}_j^{(i)} := \frac{x_j^{(i)}{}_{raw} - mean\left(x_{j_{raw}}\right)}{standardDeviation\left(x_{j_{raw}}\right)}$

Training:

for $j \leftarrow 1$ *to* n

$\boldsymbol{fusedScore}_j^{(i)} \leftarrow \frac{1}{j}\sum_1^j \boldsymbol{x}_j^{(i)}$

$\left(\boldsymbol{fusedScore}_j^{(i)}\right)_k \leftarrow$ top k scores from $\boldsymbol{fusedScore}_j^{(i)}$

Train ensemble of m logistic regression classifiers $h_j\left(\left(\boldsymbol{fusedScore}_j\right)_k\right)$ to predict $\begin{cases} 1, y^{(i)}\ is\ not\ at\ rank-1 \\ 0, otherwise \end{cases}$

end

Implementation:

Set logistic regression prediction threshold η to an arbitrarily low value (e.g., 10^{-6})

Initialize: $j \leftarrow 0$

do:

$j \leftarrow j + 1$

$\boldsymbol{fusedScore}_j \leftarrow \frac{1}{j}\sum_1^j \boldsymbol{x}_j$

$\left(\boldsymbol{fusedScore}_j\right)_k \leftarrow$ top k scores from $\boldsymbol{fusedScore}_j$

while $\sum_m h_j\left(\left(\boldsymbol{fusedScore}_j\right)_k\right) > 0$ and $j < n$

Result: Duplicate, if exists, is at rank-1.

[7] The discriminability here refers to the accuracy of prediction whether additional information is required to be considered, and is not necessarily the same as identification accuracy.

[8] www.nist.gov/itl/iad/ig/biometric_evaluations.cfm

In Table 2, $\boldsymbol{x}_{j\,raw}^{(i)}$ is the vector consisting of raw (not normalized) match scores for training example (probe) i against the gallery for biometric or biographic information j and $y^{(i)}$ denotes the true identity of the subject. The total number of identifiers per subject is denoted by n. Since different matching algorithms generate scores in different ranges and with different distributions [15], score normalization is essential in score level fusion. To avoid upper outlier scores from compressing a majority of the biometric score distribution to a small range, z-score normalization (as opposed to min-max), as illustrated in Table 2, was used. Fusion of scores for all scenarios discussed in this and the subsequent sections have been performed on z-score normalized scores by employing the sum fusion rule [4].

A separate ensemble is trained at each stage of the algorithm using bootstrap aggregation with the same number of training examples for each classifier as the size of the training set. Suppose the available identifiers include a fingerprint, face and subject name. If the training data indicates that fingerprint has the best discriminability, followed by face, and finally subject name, then the ensemble for the first stage is trained using the fingerprint training scores, followed by score fusion of fingerprint and face.

4. Experimental evaluation and analysis

Below we describe the experiments to show the accuracy and efficiency of the proposed algorithm on a large scale benchmark dataset under various unimodal and multimodal scenarios.

4.1. Dataset

The biometric database used for experiments consists of two fingerprint images for each of 27,000 subjects from the NIST Special Database 14 [12] and two mugshot face images for each of 27,000 subjects sampled from the PCSO dataset [13]. The virtual bi-modal database was created by randomly coupling a face in the PCSO database with a finger in the NIST SD14 database. The first impression of each finger is used to form the gallery while the second impression is used as probe. For face, the image acquired at a younger age was used as the gallery and the one at an older age as probe. The biometric databases used here have been anonymized. Since no large scale benchmark datasets for biographic information is available, the biographic information was assigned to each subject first using the gender information in the face database and then randomly drawing the first name, last name and father's name by mimicking statistics from the US Census [14]. An example of a virtual subject is shown in Fig. 2.

(a) (b)

KAROLE ABLEMAN ALONSO ABLEMAN

(c) (d)

Fig. 2. Examples of biometric and biographic information collected during enrollment. (a) Fingerprint [12]; (b) face [13]; (c) subject's name [14]; and (d) her father's name [14].

In practice, the name may not always be identical because of possible human data entry errors (see Figs. 9, 10 for example). To simulate these errors, a crowdsourcing experiment on Amazon Mechanical Turk[9] was conducted by us with about a hundred workers. Each worker was required to enter the textual data by looking at the data presented to them as an image (so that they do not simply copy-paste the text). A simple statistical model that embodies the characteristics of textual variations and human errors was created. The probabilistic model consists of insertion, deletion and replacement of random characters, swapping of adjacent characters in different parts of the name, and replacement of parts of name with only the initials.

4.2. Matching of unimodal information

Two state of the art commercial matchers (COTS-A for fingerprint and COTS-B for face) were used to compute the biometric similarity scores; the average similarity value of Levenshtein, Damerau-Levenshtein, and Editor Distances (see section 3.1) was used for matching of biographic information.

The average similarity value of edit distances outperforms the Levenshtein distance which was used in a previous study [6]. This is primarily because in [6], variation in name spellings was assigned a similarity value based not only on the number of character differences but also on the length of the name. Intuitively, two long names with a single character distance have a higher similarity than two short names with the same character distance. The proposed algorithm for matching of biographic information outperforms the Levenshtein distance by about 4% for top-10 ranks (including rank-1 accuracy) on the data derived from the US census information [14].

The cumulative match characteristic (CMC) curves for all the unimodal identifiers used in this study (fingerprint, face, subject's name, and father's name) are shown in Fig. 3. As expected, the two biometric identifiers, face and fingerprint, outperform the two biographic identifiers, subject's name and father's name. Face and fingerprint matching performance are comparable with rank-1 accuracy of ~95%. In earlier studies (e.g., [6]) these accuracies were significantly lower. It should be noted that two state-of-the-art COTS were used for the experiments and recognition performance depends on the quality of data.

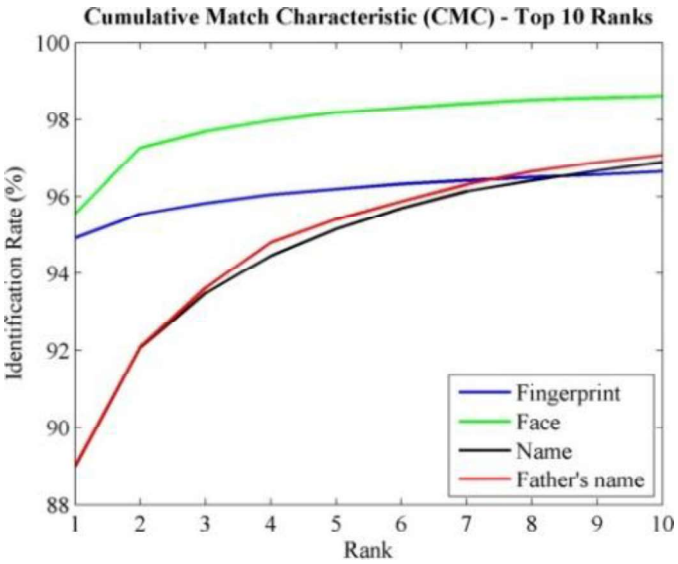
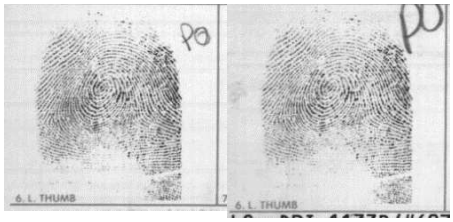


Fig. 3. CMC curves for unimodal identifiers.

An example where the rank-1 fingerprint score represents a true mate is shown in Fig. 4, while another example where fingerprint matching is not sufficient, i.e., the rank-1 fingerprint score does not represent a true mate, is shown in Fig. 5. An example where the rank-1 face score represents a true mate is shown in Fig. 6, while another example where face matching is not sufficient is shown in Fig. 7. The two face mugshot images per subject in Fig. 6 and 7 above have varying time lapses; the

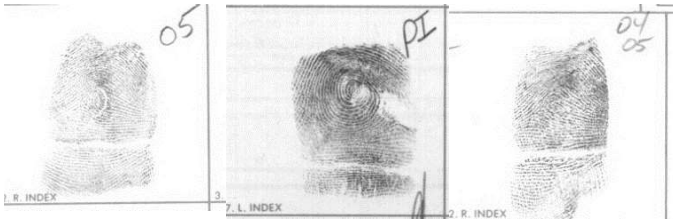
[9] www.mturk.com

age of the subject at the time of image acquisition is noted in the caption.



(a) (b)

Fig. 4. Example of successful fingerprint match where the rank-1 score represents a true mate. (a) Probe image, and (b) rank-1 retrieved gallery image.



(a) (b) (c)

Fig. 5. Example where fingerprint match alone is not sufficient. (a) Probe image; (b) rank-1 gallery image of a different subject; and (c) gallery image of true mate retrieved at rank 24,684.

(a) (b)

Fig. 6. Example of successful face match where the rank-1 score represents a true mate. (a) Probe image, age 31.0 years; and (b) rank-1 gallery image, age 28.6 years.

(a) (b) (c)

Fig. 7. Example where face match alone is not sufficient. (a) Probe image, age 47.6 years; (b) rank-1 gallery image of a different subject, age 47.4 years; and (c) gallery image of true mate, age 44.5 years, retrieved at rank 16,372.

4.3. Information fusion

A comparative evaluation, where only the biographic information or the biometric information is used, is shown in the CMC curves in Fig. 8. This figure also shows the performance gain when both the biometric traits (face and fingerprint) are fused with the biographic information (subject's name and father's name).

An example where the rank-1 retrieved biographic match is a genuine match is shown in Fig. 9, and an example where biographic matching alone does not retrieve the true mate at rank-1 is shown in Fig. 10.

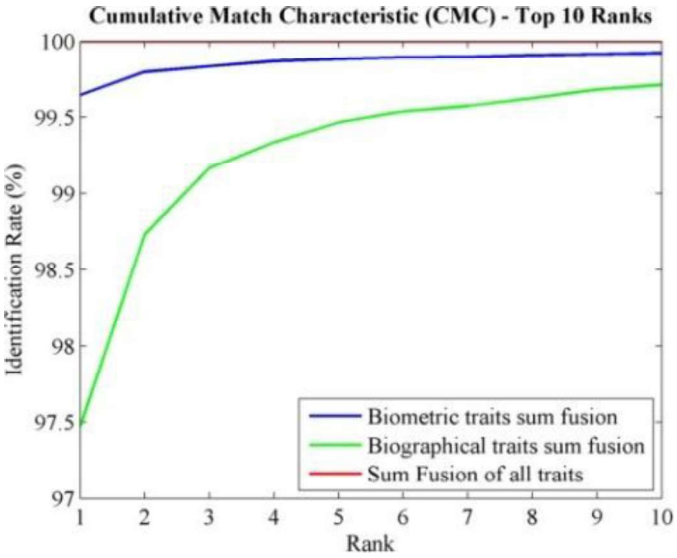


Fig. 8. CMC curves for biometric and biographic fusion. Fingerprint and face scores are fused for biometric traits, while subject's name and father's name are fused for biographical traits.

LARRY GRYANTHAN **WARREN GRYANTHAN** (a)	**LARRY GRANTHAN** **WARREN GRANTHAN** (b)
JFFREY RUSSELL **DEOUGLAS RUSSELL** (c)	**JEFFREY RUSSELL** **DOUGLAS RUSSELL** (d)

Fig. 9. Examples of successful biographic match where rank-1 score represents a genuine match. (a) Probe subject's name and father's name for biographic example 1; (b) rank-1 subject's name and father's name; (c) probe subject's name and father's name for biographic example 2; and (d) rank-1 subject's name and father's name.

R SMIT **M SMTH** (a)	**CARL SMITH** **MARK SMITH** (b)	**ROSEMARIE SMITH** **MICHAEL SMITH** (c)

Fig. 10. Example where biographic match alone is not sufficient. (a) Probe subject's name and father's name; (b) rank-1 subject's name and his father's; and (c) true mate's name and father's name retrieved at rank 211.

4.4. Analysis and performance evaluation of adaptive fusion

The 27,000 subjects in our database were randomly partitioned into three subsets (9,000 in each subset) for three-fold cross-validation. One of the three subsets was retained for testing the model, by turn, and the remaining two subsets were used as training data. We report the average performance across the three folds. In our experiments, each fold provides exactly the same performance, so the variance is zero. The ensemble in both stages (training and testing) consisted of m = 100 classifiers, with top k = 5 highest scores being supplied to each of the classifiers as input. The values of m and k have been determined empirically from the performance on training set. Note that the training and test sets here have similar characteristics.

Table 3. Identification Accuracy and Efficiency of Fusion Algorithms

Fusion Algorithm	Rank-1 Accuracy	Face matching required (% queries)	Biographic Information required (% queries)
Static fusion of all traits	100.0%	100.0%	100.0%
Adaptive fusion with single outlier detection [26]	100.0%	47.29%	18.64%
Proposed adaptive fusion algorithm	100.0%	36.82%	8.13%

The adaptive fusion with single outlier detection proposed by Arora et al. [26], which is based on standard statistical test, was

also implemented with significance level[10] $\alpha = 0.99$ for comparison. The exponential distribution was determined to be the best fit[11]. The evaluation, in terms of fusion efficiency, is presented in Table 3 and Fig. 11. Note that adaptive fusion with single outlier detection does not require any training.

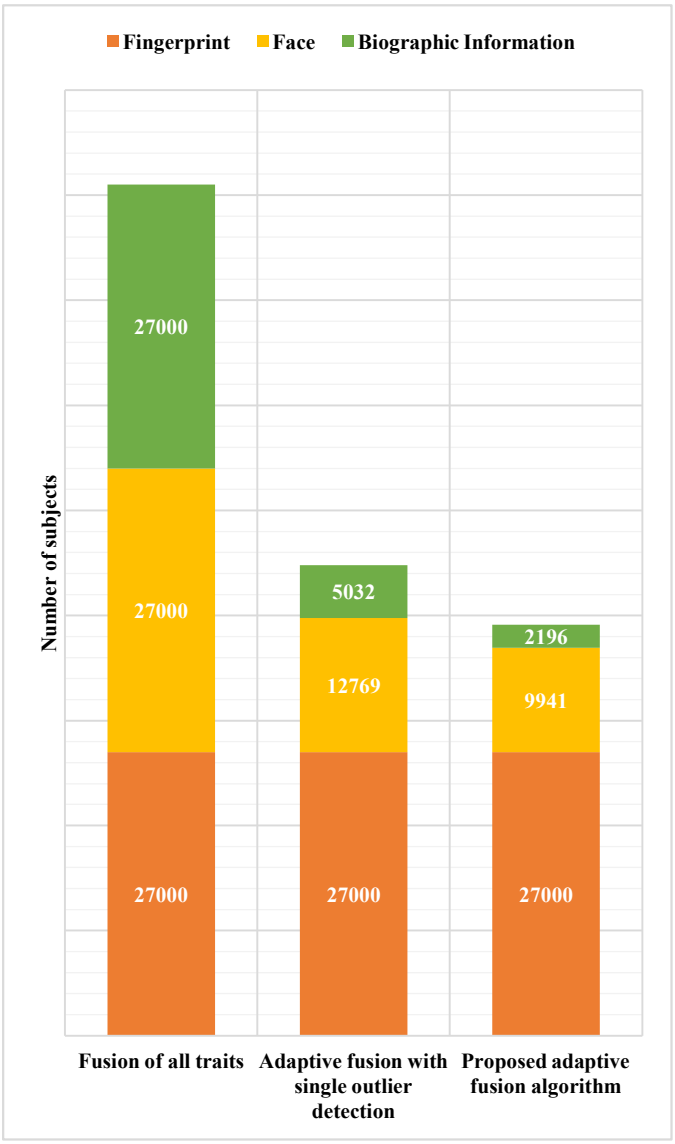


Fig. 11. Comparative evaluation of number of subjects that require matching and fusion of various biometric traits and biographic information. Total no. of subjects is 27,000. Rank-1 identification accuracy after fusion is 100% in all cases.

Example query where fingerprint alone is sufficient by itself for decision was illustrated in Fig. 4, while example query where fingerprint alone is insufficient was shown in Fig. 5. The ensemble of classifiers correctly predicts that no further information beyond fingerprint is necessary in the former case, while concludes that evidence from fingerprint is inadequate in the latter case. An example where fingerprint alone is not sufficient for rank-1 retrieval, but its fusion with face lends adequate confidence for rank-1 retrieval is shown in Fig. 12. No biographic information is needed here. Another example where fusion of fingerprint and face are not adequate for rank-1 retrieval is shown in Fig 13. Fusion of fingerprint, face and biographic information does lead to correct rank-1 retrieval.

[10] The significance level α provides a trade-off between desired robustness and computational effort required for fusion of additional information. The nominal value of 0.99 is empirically chosen here from the training set based on 100.0% rank-1 accuracy requirement.
[11] www.mathworks.com/help/stats/model-data-using-the-distribution-fitting-tool.html

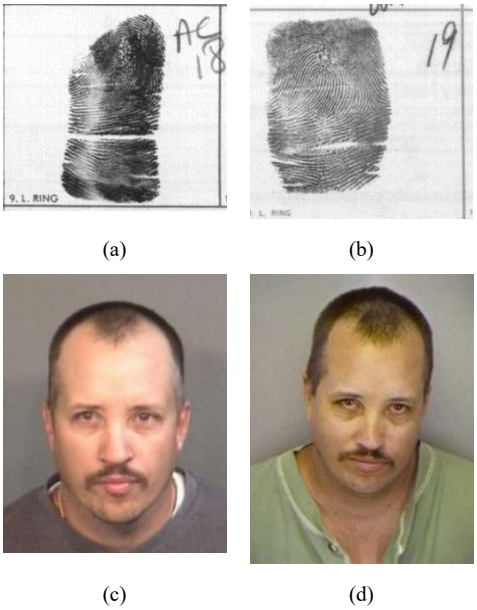


Fig. 12. Example where fingerprint alone is not sufficient for retrieval at rank-1. After score level fusion of fingerprint and face, rank-1 retrieval is successful. Biographic information is not needed for this query. (a) Probe fingerprint image; (b) gallery fingerprint image of genuine subject retrieved at rank 4; (c) probe face image, age 36.5 years; and (d) gallery face image of genuine subject, age 31.0 years, retrieved at rank 1.

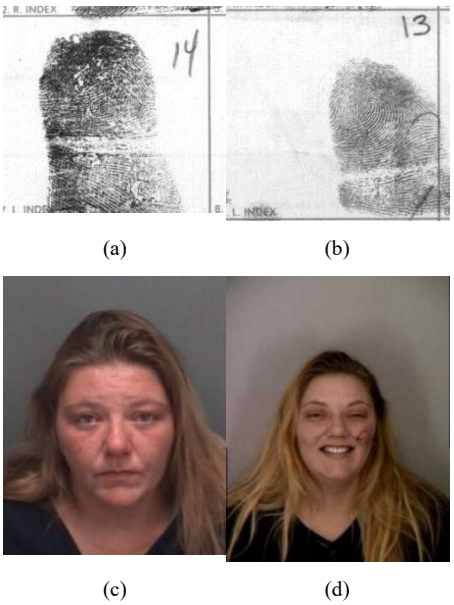


Fig. 13. Example of adaptive fusion. (a) Probe fingerprint image; (b) gallery fingerprint image of genuine subject retrieved at rank 697; (c) probe face image, age 34.5 years; and (d) gallery face image of genuine subject, age 24.3 years, retrieved at rank 17. After fusion of fingerprint, face and biographic information, true mate is retrieved at rank 1.

4.5. Predicted effort to error trade-off curve

While biometric fusion has widely been studied, to our knowledge, there has not been a systematic study of the trade-off between the effort required in computing and fusing match scores for individual traits and the benefit extended through reduction in error rates due to fusion. A new metric, called the predicted effort to error trade-off (PEET) curve, is defined here to study the relative efficiency of fusion algorithms.

The PEET curve charts the predicted effort as the percentage of subjects where fusion of additional information was predicted to be required, against the recognition error (in terms of rank-1 identification accuracy). The PEET curve comparing the efficiency of the proposed adaptive fusion algorithm to the

adaptive fusion algorithm with single outlier detection using standard statistical test [26] is shown in Fig. 14 after the first stage (fingerprint). It may be observed from the figure that the proposed adaptive fusion algorithm converges to the minimum error at a faster rate.

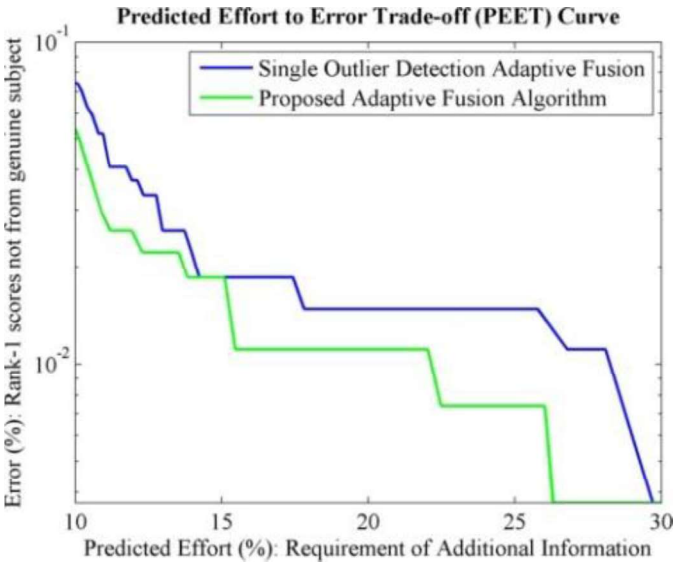


Fig. 14. PEET curve after the first stage (fingerprint) comparing the adaptive fusion with single outlier detection [26] with the proposed algorithm.

4.6. Consolidated results

Considering fingerprint, face, subject's name and father's name as the available identifiers, match scores were normalized and fused by using the various identifiers and their combinations.

Table 4. A comparison of rank-1 identification rates of individual identifiers and various combinations*.

Identifier or subset	Rank-1 identification rate
Fingerprint	94.93%
Face**	95.56%
Name	89.00%
Father's name	89.02%
Fingerprint + face	99.64%
Name + father's name	97.47%
Fingerprint + name + father's name	99.84%
Face + name + father's name	99.98%
Fingerprint + face + name + father's name	100.00%
Proposed adaptive fusion algorithm***	100.00%

***Total 27,000 queries with gallery consisting of another instance of the same 27,000 subjects; Sum fusion of z-score normalized scores.**
****Some of the face images in the PCSO** [13] **dataset are mislabeled, but have been used to replicate a realistic scenario.**
*****Aggregate results of three-fold cross validation.**

A summary of rank-1 identification rate for the various identifiers alone and their combinations is presented in Table 4.

5. Summary

The de-duplication of identities is necessary in any biometric identification system. We have proposed an algorithm that adaptively and sequentially fuses scores from biometric and biographic information for identity de-duplication. Experimental results show the proposed algorithm not only achieves high accuracy but also results in computational efficiency. In particular, our system correctly predicts that for 63.18% of the queries (in total 27,000 queries) only fingerprint is sufficient to be identified at rank-1. For an additional 28.69% of the queries, fusion of fingerprint and face scores is needed, while biographic information is needed only for the remaining 8.13% of the queries.

It would be desirable to extend this study for databases involving a larger number of subjects. Operational databases, such as the Aadhaar Project in India, typically have ten-print fingerprints, both irises, face, and biographic information. Another avenue for further study would be to incorporate biometric quality in the proposed fusion algorithm. For example, the sequence in which the identifiers are considered for a subject may also be based on the quality of the individual identifiers captured for that particular subject, instead of the globally most reliable identifier. Another avenue for research is open set identification.

References

[1] M. Fairhurst, J. Fierrez and P. Campisi, "Future trends in biometric processing [Editorial]," *IET Comput. Vis.,* vol. 5, no. 6, pp. 335-337, 2011.

[2] "India to merge ID databases," *Biometric Technol. Today,* vol. 2012, no. 5, p. 12, May 2012.

[3] A. K. Jain and A. Ross, "Multibiometric systems," *Commun. ACM,* vol. 47, no. 1, pp. 34-40, 2004.

[4] A. Ross and A. K. Jain, "Information Fusion in Biometrics," *Pattern Recogn. Lett.,* vol. 24, no. 13, pp. 2115-2125, 2003.

[5] V. Tyagi, H. Karanam, T. Faruquie, L. Subramaniam and N. Ratha, "Fusing biographical and biometric classifiers for improved person identification," in *Proc. Int. Conf. Patt. Recog.*, Tsukuba, Japan, 2012.

[6] H. Bhatt, R. Singh and M. Vatsa, "Can Combining Demographics and Biometrics Improve De-duplication Performance?," in *Proc. IEEE Conf. Comput. Vis. Patt. Recog. Workshops*, Portland, OR, USA, 2013.

[7] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar and S. Fienberg, "Adaptive Name Matching in Information Integration," *IEEE Intell. Syst.,* vol. 18, no. 5, pp. 16-23, 2003.

[8] A. K. Elmagarmid, P. G. Ipeirotis and V. S. Verykios, "Duplicate Record Detection: A Survey," *IEEE T. Knowl. Data En.,* vol. 19, no. 1, pp. 1-16, 2007.

[9] S. Agarwal, "Duplicate Aadhaar numbers within estimates: UIDAI," *Livemint,* 05 March 2013.

[10] "Aadhaar de-duplication myth busted. Any answers, Mr Nilekani?," *Moneylife,* 15 October 2013.

[11] N. R. Madhusudhan, "Many People Face Delays in Receiving Aadhaar Cards," *The New Indian Express,* 2 December 2013.

[12] NIST Craig Watson Advanced Systems Division, "NIST Special Database 14," [Online]. www.nist.gov/srd/nistsd14.cfm.

[13] "Pinellas County Sheriff's Office," [Online]. www.pcsoweb.com/.

[14] United States Census Bureau, "Documentation and methodoloy for frequently occurring names in the U.S.," 1990. [Online]. www2.census.gov/topics/genealogy/1990surnames/nam_meth.txt.

[15] A. A. Ross, K. Nandakumar and A. K. Jain, Handbook of Multibiometrics, Springer, 2006.

[16] S. Z. Li and A. K. Jain, Eds., Encyclopedia of Biometrics, vol. 1, Springer, 2009, p. 615.

[17] V. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics - Doklady,* vol. 10, no. 8, pp. 707-710, 1966.

[18] NIST Information Technology Laboratory, "Biometric Scores Set," [Online]. www.nist.gov/itl/iad/ig/biometricscores.cfm.

[19] WCC Services US Inc., "Fusion of Biometric and Biographic Data In Large-Scale Identification Projects," [Online]. www.wcc-group.com.

[20] D. Bolme, J. Beveridge and A. Howe, "Person Identification Using Text and Image Data," in *Proc. IEEE Int. Conf. Biometrics: Theory, Applications, and Systems*, Crystal City, VA, USA, 2007.

[21] NIST Information Technology Laboratory, "NIST Biometric Image Software," [Online]. www.nist.gov/itl/iad/ig/nbis.cfm.

[22] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Commun. ACM,* vol. 7, no. 3, pp. 171-176, 1964.

[23] E. Polityko, "Calculation of distance between strings," 18 November 2007. [Online]. www.mathworks.com/matlabcentral/fileexchange/17585-calculation-of-distance-between-strings.

[24] H. Eulau, "Logic of Rationality in Unanimous Decision Making," *Nomos VII: Rational Decision,* pp. 26-54, 1964.

[25] S. Walker and D. Duncan, "Estimation of the probability of an event as a function of several independent variables," *Biometrika,* vol. 54, no. 1-2, p. 167–179, 1967.

[26] S. Arora, E. Liu, K. Cao and A. Jain, "Latent Fingerprint Matching: Performance Gain via Feedback from Exemplar Prints," *IEEE T. Pattern Anal.,* vol. 36, no. 12, pp. 2452-2465, 2014.

[27] A. K. Jain, R. C. Dubes and C.-C. Chen, "Bootstrap Techniques for Error Estimation," *IEEE T. Pattern Anal.,* vol. 9, no. 5, pp. 628-633, 1987.