TECHNICAL REPORT

ISO/IEC TR 29198

First edition
2013-12-15

Information technology — Biometrics — Characterization and measurement of difficulty for fingerprint databases for technology evaluation

Technologies de l'information — Biométrie — Caractérisation et mesure de difficulté pour bases de données d'empreintes digitales pour évaluation de technologie

Reference number
ISO/IEC TR 29198:2013(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

In exceptional circumstances, when the joint technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example), it may decide to publish a Technical Report. A Technical Report is entirely informative in nature and shall be subject to review every five years in the same manner as an International Standard.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

ISO/IEC TR 29198 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 37, *Biometrics*.

Introduction

Recently, there have been worldwide increasing activities in testing and evaluating the performance of fingerprint recognition systems or algorithms. Testing activities occur in public sector, private sector, and academic entities, typically using datasets exclusive to a given entity. This complicates comparison of test results from different entities. Methodologies for assessing the level of difficulty of test datasets should improve the comparability of performance evaluation results over different fingerprint datasets.

ISO/IEC 19795-1:2006, 5.5.3[11] states:

"In a technology evaluation, testing of all algorithms is carried out on a standardized corpus, ideally collected by a "universal" sensor (i.e. a sensor that collects samples equally suitable for all algorithms tested). Nonetheless, performance against this corpus will depend on both the environment and the population in which it is collected."

Comparison of evaluation results based on testing against different corpora may be misleading. Further, policies for inclusion or removal of low-quality data in a corpus may vary from organization to organization, such that the same algorithm tested against the same corpus may generate different results. There are also certain difficulties when trying to compare multiple evaluation results derived from different corpora. Currently there is no established methodology for characterizing the level of difficulty of datasets used in performance evaluation. The ability to characterize a dataset's level of difficulty should support predictions of operational accuracy when processing data known to be of equivalent difficulty.

The purpose of this Technical Report is to provide guidance on predicting how "challenging" or "stressing" a fingerprint dataset is for recognition, based on factors such as relative sample quality, relative rotation, deformation, and overlap between impressions. The provided guidance can be used for characterizing and measuring the relative difficulty levels of fingerprint datasets used in technology evaluation.

Following the guidance in this Technical Report, users and system evaluators in different organizations will be able to compare and place into context the performance evaluation results of the other organizations according to the level of difficulty of its dataset.

This Technical Report proposes dataset generation methods based on analysis of comparison results or scores from multiple fingerprint recognition algorithms. These dataset generation methods support creation of datasets with specific levels of difficulty and creation of datasets for use in interoperability evaluations.

ISO/IEC TR 29794-4[16] defines methods for expressing the quality score of a *single* fingerprint image. Such quality scores are typically predictive of matching accuracy. This Technical Report, by contrast, is concerned with differences in rotation, deformation, and common area between reference and probe samples.

NOTE Other modalities can be considered in the future as more information becomes available about standardized quality measurements that are suitable for predicting the performance of other biometric systems.

Information technology — Biometrics — Characterization and measurement of difficulty for fingerprint databases for technology evaluation

1 Scope

This Technical Report provides guidance on estimating how “challenging“ or “stressing“ is an evaluation dataset for fingerprint recognition, based on relative sample quality, relative rotation, deformation, and overlap between impressions. In addition, this Technical Report establishes a method for construction of datasets of different levels of difficulty. This Technical Report defines the relative level of difficulty of a fingerprint dataset used in technology evaluation of fingerprint recognition algorithms. Level of difficulty is based on differences between reference and probe samples in the aformentioned factors. This Technical Report addresses such issues as:

— characterizing level of difficulty attributable to differences between samples acquired from the same finger,

— developing statistical methodologies for representing the level of difficulty of a fingerprint dataset by aggregating influencing factors,

— comparing the level of difficulty of different fingerprint datasets,

— defining procedures for testing and reporting the level of difficulty of fingerprint datasets collected for technology evaluation,

— analysing mated pair data characteristics based on comparison scores,

— describing the archived data selection methodology for building a dataset for evaluation.

This Technical Report provides guidelines for comparing the relative level of difficulty of fingerprint datasets.

Outside the scope of this Technical Report are:

— defining the quality of individual fingerprint images,

— defining the methodologies or explicit measures for evaluating or predicting the performance of fingerprint recognition algorithms.

2 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

2.1
raw biometric sample
information obtained from a biometric sensor, either directly or after further processing

2.2
biometric reference
<template, model> one or more stored biometric samples, biometric templates or biometric models attributed to a biometric data subject and used as the object of comparison

EXAMPLE Face image stored digitally on a passport; Fingerprint minutiae template on a National ID card; Gaussian Mixture Model for speaker recognition, in a dataset.

Note 1 to entry: A biometric reference may be created with implicit or explicit use of auxiliary data, such as Universal Background Models.

Note 2 to entry: The subject/object labelling in a comparison might be arbitrary. In some comparisons a biometric reference might be used as the subject of the comparison with other biometric references or incoming samples used as the objects of the comparisons. For example, in a duplicate enrolment check a biometric reference will be used as the subject for comparison against all other biometric references in the dataset.

2.3
biometric probe
biometric data input to an algorithm for comparison to a biometric reference(s)

2.4
technology evaluation
offline evaluation of one or more algorithms for the same biometric modality using a pre-existing or specially collected corpus of samples

2.5
failure-to-enrol rate
FTE
proportion of the population for whom the system fails to complete the enrolment process

Note 1 to entry: The observed failure-to-enrol rate is measured on test crew enrolments. The predicted/expected failure-to-enrol rate will apply to the entire target population.

2.6
failure-to-acquire rate
FTA
proportion of verification or identification attempts for which the system fails to capture or locate an image or signal of sufficient quality

Note 1 to entry: The observed failure-to-acquire rate is distinct from the predicted/expected failure-to-acquire rate (the former may be used to estimate the latter).

2.7
false non-match rate
FNMR
proportion of genuine attempt samples falsely declared not to match the biometric reference of the same characteristic from the same subject supplying the sample

2.8
false match rate
FMR
proportion of zero-effort impostor attempt samples falsely declared to match the compared non-self template

Note 1 to entry: The measured/observed false match rate is distinct from the predicted/expected false match rate (the former may be used to estimate the latter).

2.9
false reject rate
FRR
proportion of verification transactions with truthful claims of identity that are incorrectly denied

2.10
false accept rate
FAR
proportion of verification transactions with wrongful claims of identity that are incorrectly confirmed

2.11
receiver operating characteristic curve
ROC curve
plot of the rate of false positives (i.e. impostor attempts accepted) on the x-axis against the corresponding rate of true positives (i.e. genuine attempts accepted) on the y-axis plotted parametrically as a function of the decision threshold

2.12
detection error trade-off curve
DET curve
modified ROC curve which plots error rates on both axes (false positives on the x-axis and false negatives on the y-axis)

2.13
performance
capability in terms of error rates and throughput rates

2.14
quality
degree to which a biometric sample fulfils specified requirements for a targeted application

Note 1 to entry: Specified quality requirements may address aspects of quality such as focus, resolution, etc. Implicit quality requirements address the likelihood of achieving a correct matching result.

2.15
quality score
quantitative expression of quality

2.16
matchability
degree to which two mated fingerprint samples can be successfully compared through multiple comparison algorithms

2.17
mated pair
set of two samples of the same biometric characteristics captured from the same source, where one is used for the reference and the other used for the test

2.18
level of difficulty
measure of a biometric dataset which represents how ‘challenging’ or ‘stressing’ the fingerprint dataset is for recognition relative to other datasets

Note 1 to entry: Fingerprint dataset “A” is more difficult than dataset “B” with respect to chosen fingerprint comparison algorithms if the performance of these algorithms is significantly lower for dataset “A” than dataset “B”. For how to assess the performance of given comparison algorithms, see ISO/IEC 19795-2.[12]

Note 2 to entry: For estimating the level of difficulty of a fingerprint dataset before testing the performance of fingerprint comparison algorithms against this and other datasets, this Technical Report defines measures that predict level of difficulty.

Note 3 to entry: This Technical Report addresses the level of difficulty for fingerprint corpora only.

2.19
singular point
either core point or delta point in fingerprint

2.20
alignment point
either a singular point or a certain minutia point which is used to align a mated pair of fingerprints

Note 1 to entry: Since each alignment point has position and orientation, the alignment process based on a pair of corresponding alignment points from a mated pair will compensate the rotation and the translation between the two fingerprints.

3 Symbols and abbreviated terms

AP	alignment point
CA	common area
DF	relative deformation
LOD	level of difficulty
RSQ	relative sample quality
SP	singular point

4 Differential factors in fingerprint samples

4.1 General

As described in ISO/IEC TR 19795-3,[13] the following properties of a fingerprint dataset have influence on the performance of fingerprint recognition:

- Sensor type (e.g. total internal reflection, capacitance, thermal, swipe, touchless, ultrasonic, etc)
- Impression type (e.g. flat, rolled, segmented slap, scanned ink-print, etc)
- Image resolution
- Environmental conditions (e.g. temperature, humidity, etc)
- Demographics (e.g. age, gender, occupation, etc)
- Finger position (e.g. thumb, index, etc)
- Template ageing
- Biological condition (e.g. skin moisture)
- Subject motivation, habituation etc.

When the dataset is homogeneous in the aspects of sensor type and impression type, the rest of the properties can be represented and quantified by a fingerprint sample quality score.

As defined in ISO/IEC 29794-1,[15] the quality of a biometric sample is the degree to which a biometric sample fulfils specified requirements for a targeted application and the quality score is a quantitative expression of the quality. However, the quality score is associated with each individual biometric sample. As such it does not incorporate differences between reference and probe samples.

As pointed out by Hicklin and Reedy,[1] the ability to match fingerprints is dependent on three characteristics: (i) number of fingers (in the case of ten-print identification), (ii) correspondence between reference and probe images, and (iii) quality of both reference and probe images. Correspondence between the two fingerprints is a function of the degree of overlap and distortion between the reference and the probe, as well as inherent friction ridge content. Image quality metrics can be used to quantify the quality of the reference and probe images separately.

For example, as shown in Figures 1 and 2, even when both finger images are of good quality, the comparison score will be low if their common area is small (Figure 1) or the relative deformation is severe (Figure 2). Furthermore, comparison of two low-quality samples may produce a higher score than comparison of a high-quality and a low-quality sample.

Figure 1 — A low similarity score will result when comparing impressions with small common area

Figure 2 — A low similarity score will result when comparing impressions with severe deformation

Considering these cases, the quality defined in ISO/IEC TR 29794-4[16] is not fully sufficient to assess the LOD of a fingerprint dataset in a technology test. In addition, the relative quality needs to be defined in order to consider the influence of other differences between mated pairs of fingerprints.

The relative level of difficulty may be applicable to selecting data for a performance evaluation. In cases where limited resources are available to conduct an interoperability performance test, it may be desirable to focus on challenging datasets because meaningful results may be generated through relatively fewer comparisons. Further, it can be used to evaluate the suitability of datasets for such evaluation. An experimenter may focus on a small amount of matchable sample pair data to make an initial assessment of the suitability of a given dataset for this purpose.

4.2 Common area

4.2.1 Introduction

The common area between mated fingerprint sample pairs can vary due to human factors. In general, a larger common area results in a higher comparison score. Figure 3 depicts the overlapping area of a pair of mated fingerprints. Possible measures for the common area are:

a) the ratio of the common area to the total area covered by the mated pair (the preferred method, discussed below), or

b) the area overlap of the convex hulls of the minutiae on each impression.

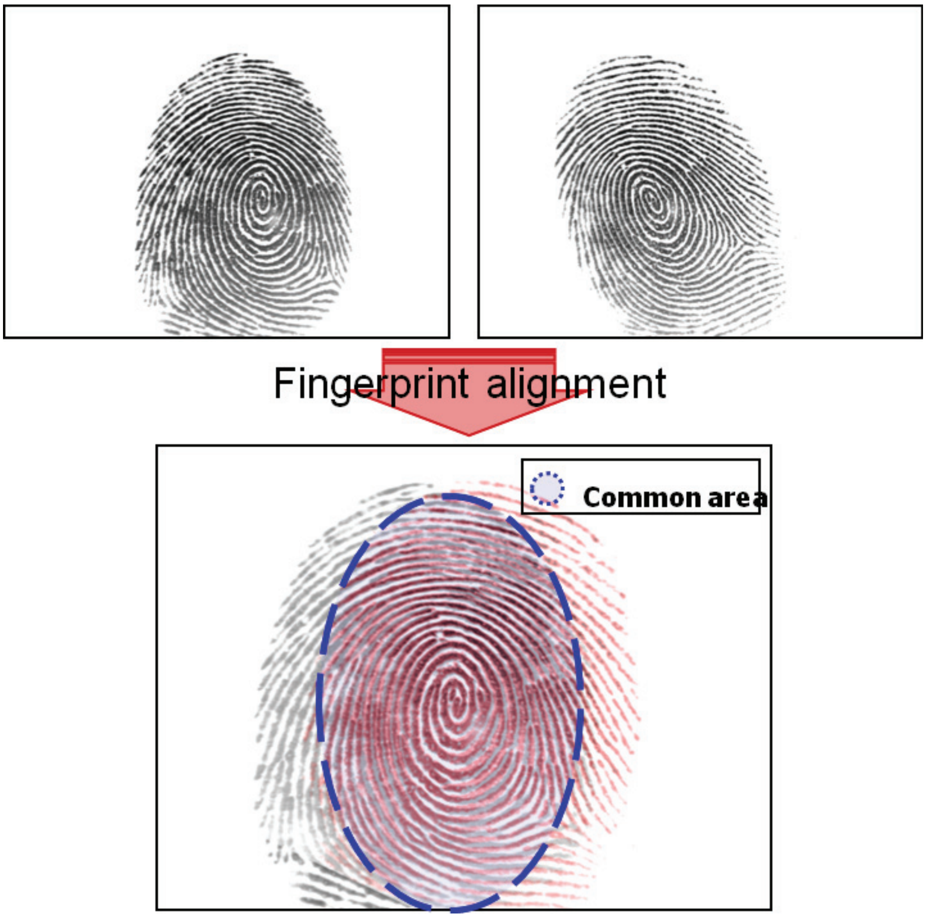


Figure 3 — One possible definition of common area based on foreground areas of mated impressions

Regardless of the comparison algorithm, the minutiae-based or the image-based, the common area is one of the major factors which influence the matching performance in fingerprint recognition. In general, the greater the common area of a mated pair, the higher the similarity score. Figure 4 shows one mated pair with a similarity score using a commercial fingerprint comparison algorithm.

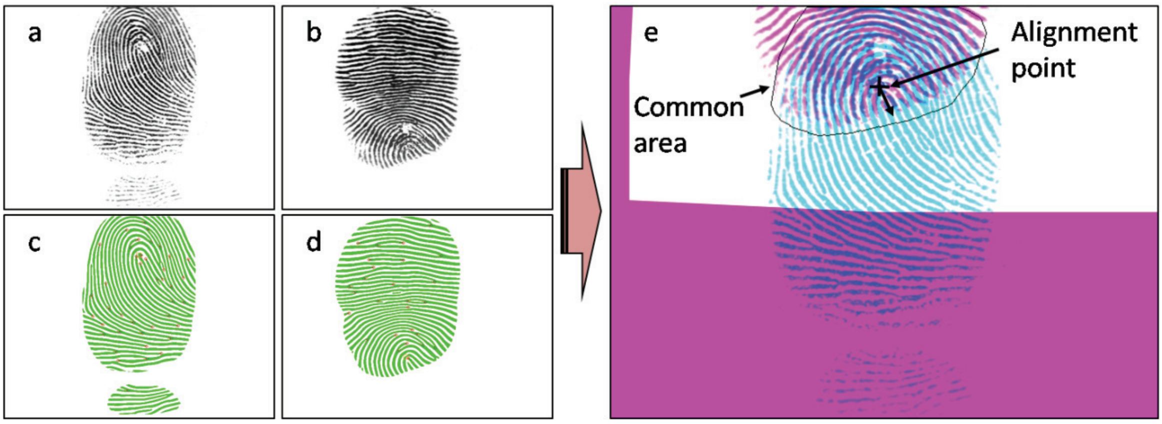


Figure 4 — Example of a mated pair with low common area whose similarity score is 0; (a) and (b) are a mated pair, (c) and (d) are processed images of (a) and (b) after minutiae extraction, and (e) is the result of aligning (a) and (b) using a core point as an alignment point

4.2.2 Definition of common area

In this document, the measure of the common area for a mated sample pair is defined as the ratio of the common area to the total area covered by the mated pair:

$$CA = \frac{P_{reference} \cap P_{probe}}{P_{reference} \cup P_{probe}}$$

where P denotes the fingerprint foreground extracted by segmentation. This metric is normalized to [0, 1], where 0 indicates that no corresponding AP pair is found.

4.2.3 Localizing common area for a mated pair

Given a mated pair of segmented fingerprints for matching, in order to localize the common area of a mated pair, it is necessary to locate a corresponding alignment point pair. For non-arch type fingerprints, the alignment point (AP) pair can be found from corresponding pixel-level singular points.[2] For fingerprints with no singular points including the arch type, the AP pair can be obtained from corresponding minutiae points. Figures 5, 6, and 7 show examples of computing the common area for whorl type, arch type, and loop type with missing singular point pair, respectively. Note that in each case there can be multiple AP pairs.

Each AP has position and orientation. By aligning the position and the orientation, the rotation and the translation differences between the reference sample and the probe sample can be corrected.

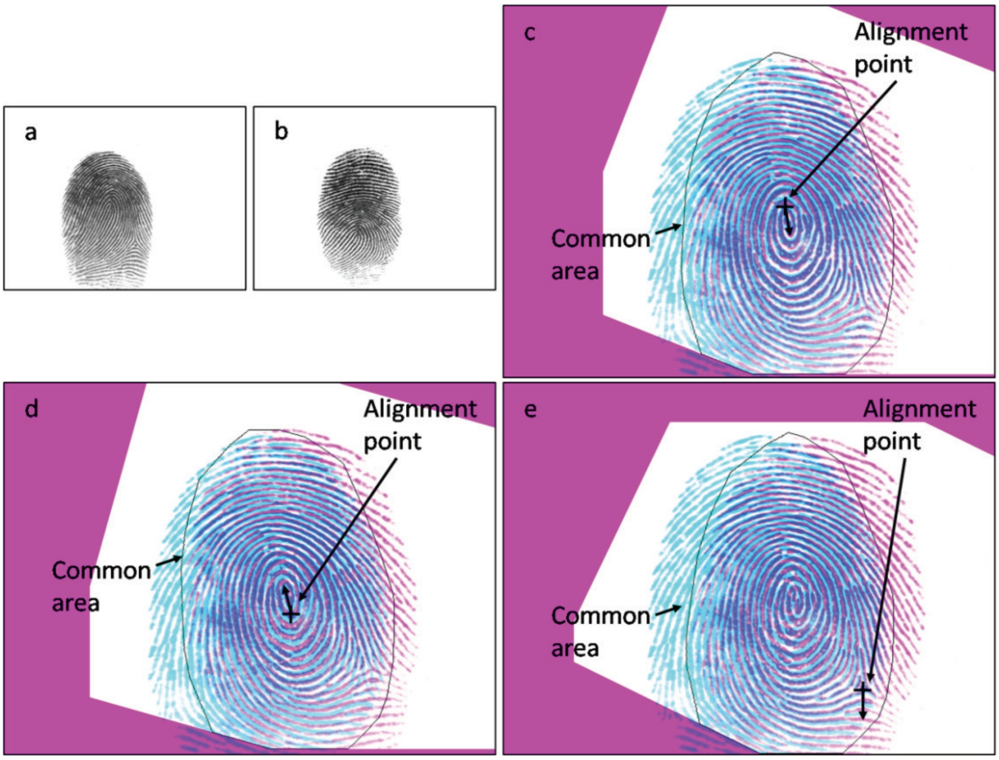


Figure 5 — Possible localization of common area for whorl type; (a) and (b) are a mated pair, (c), (d) and (e) show the resulted common areas based on different AP pairs

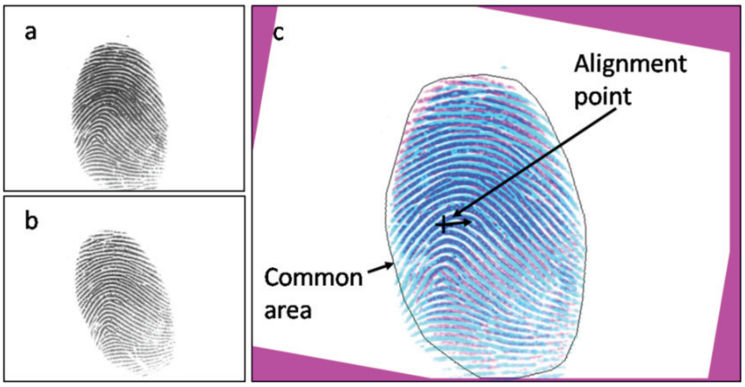


Figure 6 — Possible localization of common area for arch type aligned by a corresponding minutia pair; (a) and (b) are a mated pair, and (c) shows the resulted common area

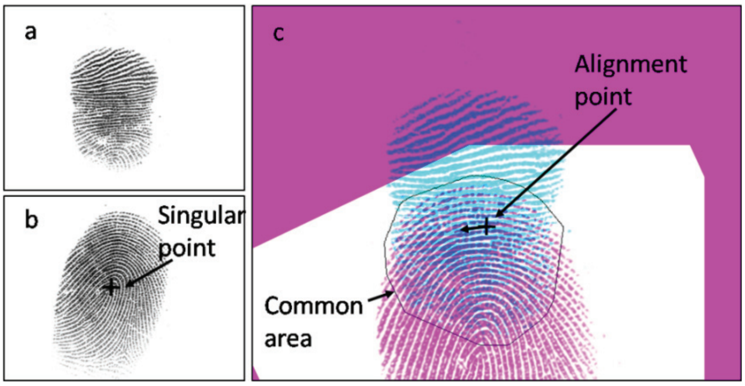


Figure 7 — Possible localization of common area for loop type with missing corresponding singular point pair; (a) and (b) are a mated pair, and (c) shows the resulted common area

4.2.4 Computation of common area for a mated pair

The computation of AP is the key step of common area measurement. The AP needs to be detected in pixel-level precision. Since most of non-arch type fingerprints contain at least one SP, pixel-level SPs are the first choice for the candidate AP. SP detection[2] are conducted in pixel-level to guarantee the accuracy of alignment of mated fingerprint pairs. Meanwhile, for arch type fingerprints and the fingerprints which miss finding corresponding SPs, corresponding minutiae are used instead. Figure 8 shows the flowchart of the computation of common area.

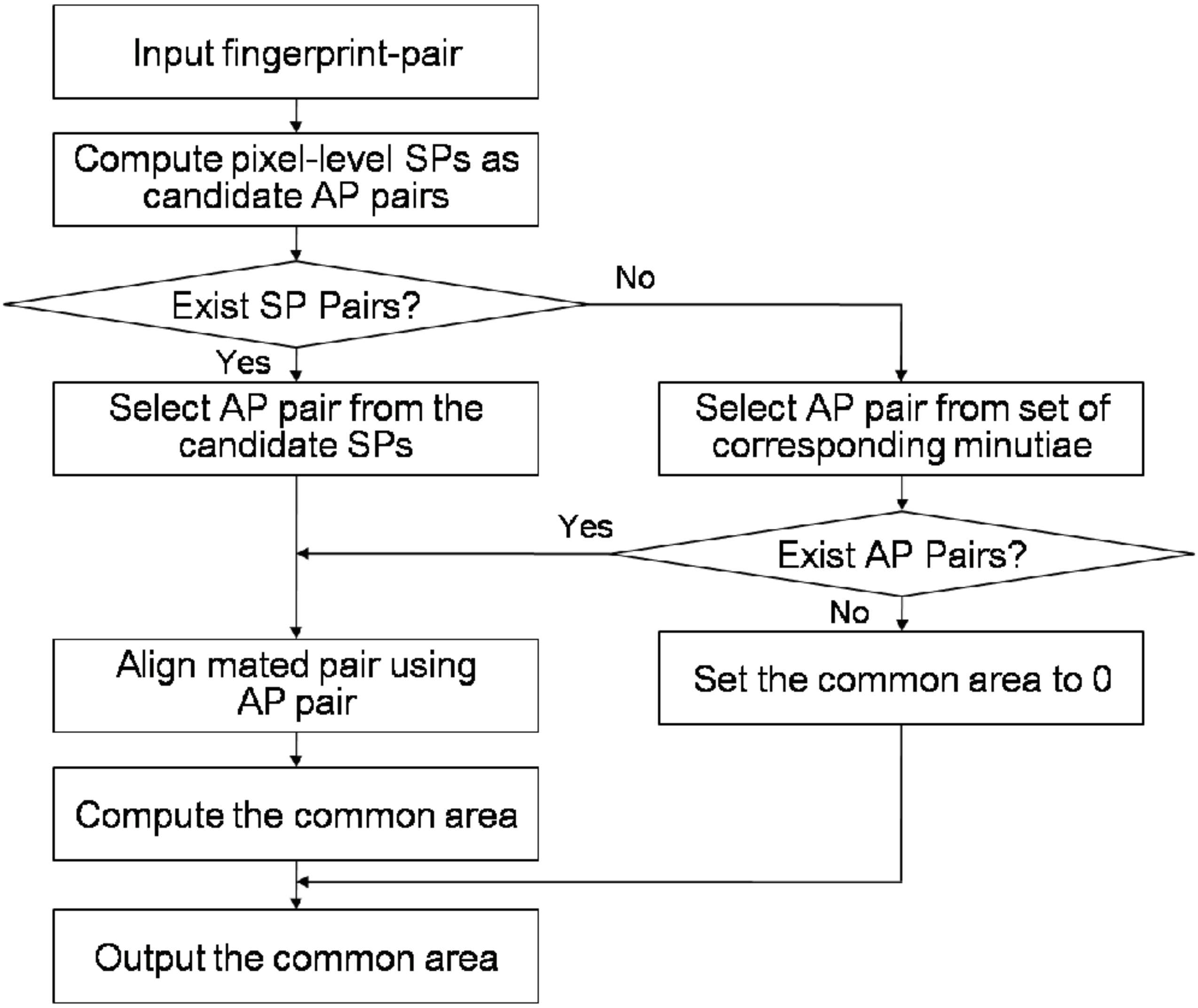

Input fingerprint-pair
Compute pixel-level SPs as candidate AP pairs
Exist SP Pairs?
No
Yes
Select AP pair from the candidate SPs
Select AP pair from set of corresponding minutiae
Yes
Exist AP Pairs?
No
Align mated pair using AP pair
Set the common area to 0
Compute the common area
Output the common area

Figure 8 — Flowchart of the computation of common area

When multiple AP candidate pairs are found in a mated pair as shown in Figure 5, the one with the maximum common area is selected as the final AP.

NOTE 1 For the fingerprints with bad sample quality, finding correct APs may fail in spite of the existence of corresponding SPs or minutiae pairs, in which case setting the common area to zero is natural.

NOTE 2 Since arch type fingerprints have no singular points, the AP pairs can be obtained from a set of corresponding minutiae pairs using any comparison algorithm. When there are multiple corresponding AP pairs, the one with the maximum common area is selected as the final AP.

4.2.5 Relationship between common area and similarity score

It is very natural to claim that the common area and the similarity score have a proportional relation. However, the similarity score is influenced by other factors such as deformation and sample quality. Figure 9 shows the scatter plots of the common area versus the similarity score for mated pairs over FVC 2000 datasets.[6] It seems true that mated pairs with high similarity scores have a large common area while mated pairs with low similarity scores do not necessarily have a small common area. Furthermore, mated pairs with a small common area tend to produce low similarity scores while mated pairs with a large common area do not necessarily produce high similarity scores.

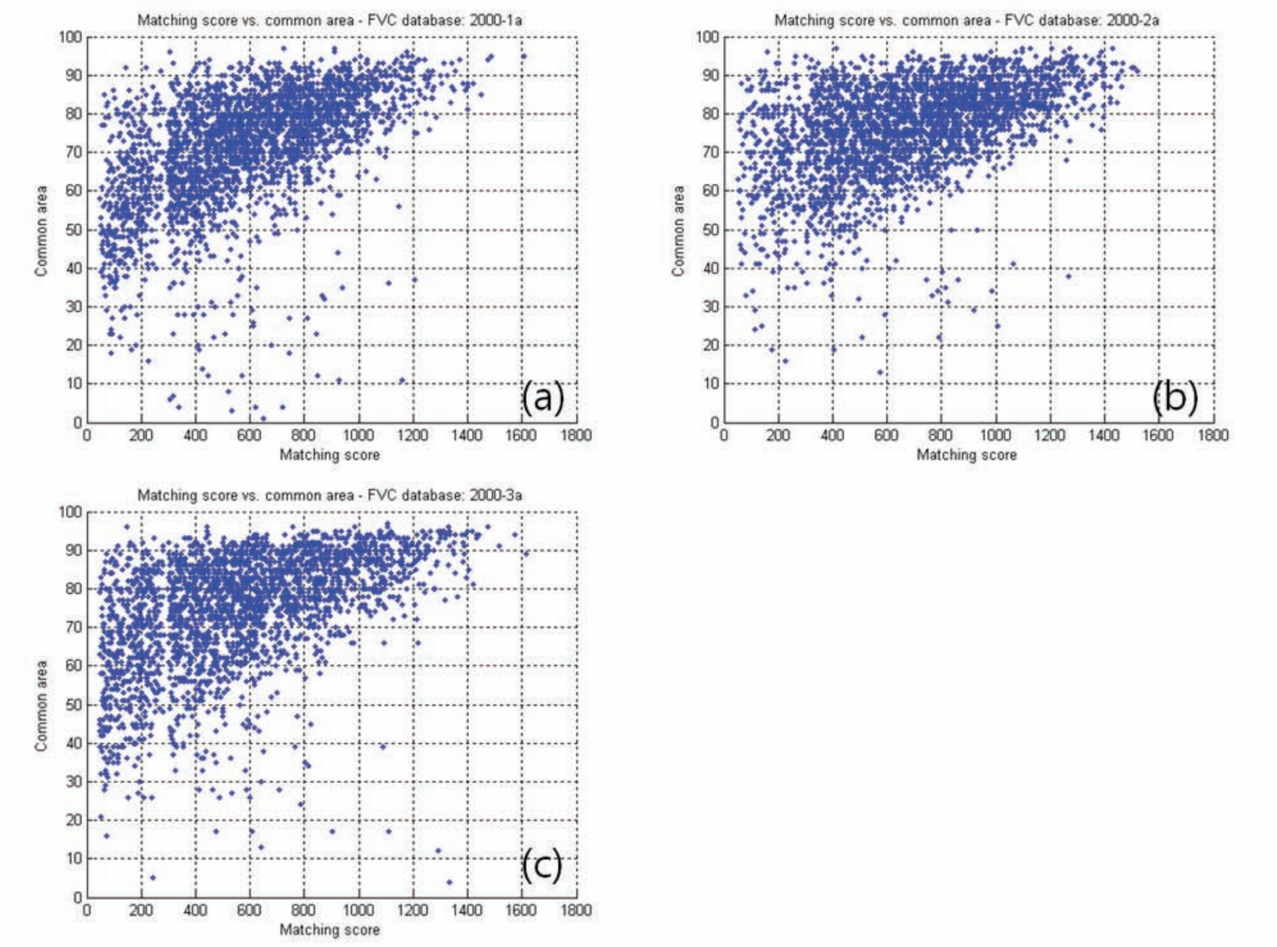


Figure 9 — Scatter plots of common area versus similarity score[9] over FVC 2000 DBs; (a) 2000-1a, (b) 2000-2a, (c) 2000-3a

4.3 Relative deformation

4.3.1 Introduction

Pressure of the finger during capture causes deformation of the fingerprint because fingers and skin are nonrigid, which prevents a perfect match even for a mated pair with 100 % common area. The existence of deformations makes the fingerprint matching more difficult. The higher the relative deformation between a mated pair, the more difficult the fingerprint matching. Therefore, the overall relative deformation of mated pairs can reflect the level of dificulty of a fingerprint dataset indirectly.

While it is difficult to measure the degree of deformation of an individual fingerprint, it is easier to measure the degree of relative deformation between a mated pair. The relative deformation can be computed by locating corresponding points or patterns such as minutiae, singular points, ridge lines and other topological patterns, followed by measuring the position and orientation differences.

Deformation of fingerprints may be both linear and nonlinear. Examples of linear deformation are rigid deformations (translation and rotations) and shear. Examples of nonlinear transformations include spline deformation. One simple measure of linear deformation is the extent to which the area of the print changes (can be estimated using the determinant of the equivalent linear deformation matrix). There are various measures of elastic deformation such as the bending energy. Possible measures of relative deformation for a mated pair of fingerprints are:

a) average of orientation differences of corresponding points after alignment of the mated pair, or

b) measure of deformation using Thin Plate Spline method.

4.3.2 Measurement of orientation field-based deformation

Assuming the continuity in the fingerprint orientation, when there is no relative deformation between a mated pair of fingerprints after alignment, the orientations will coincide at the same position. In most cases of matching, however, there exists relative deformation between a mated pair, which can be indicated by the overall differences in orientation. The orientation field-based deformation is measured over the aligned common area, and the computation of pixel-level orientation fields for the mated pair can be achieved by the multiscale Gaussion filter.[2]

The orientation field-based deformation is defined as the average of the orientation differences over the aligned common area:

$$DF = \text{Average}(\Delta\theta_{i,j}),$$

where $\Delta\theta_{i,j} = \text{abs}(\theta_i - \theta_j)$, and θ_i and θ_j refer to the ridge orientation of the aligned positions in the mated pair.

Figures 10 through 13 demonstrate the computation of the common area and the relative deformation by aligning with different AP pairs for a mated pair. The rotation difference between the mated pair is compensated by coinciding the orientations of a corresponding AP pair. They show that both the common area ratio and the relative deformation vary depending on the AP pair. When multiple AP candidate pairs are found in a mated pair, the one with the maximum common area is selected as the final AP. Then, the relative deformation is computed using the final AP.

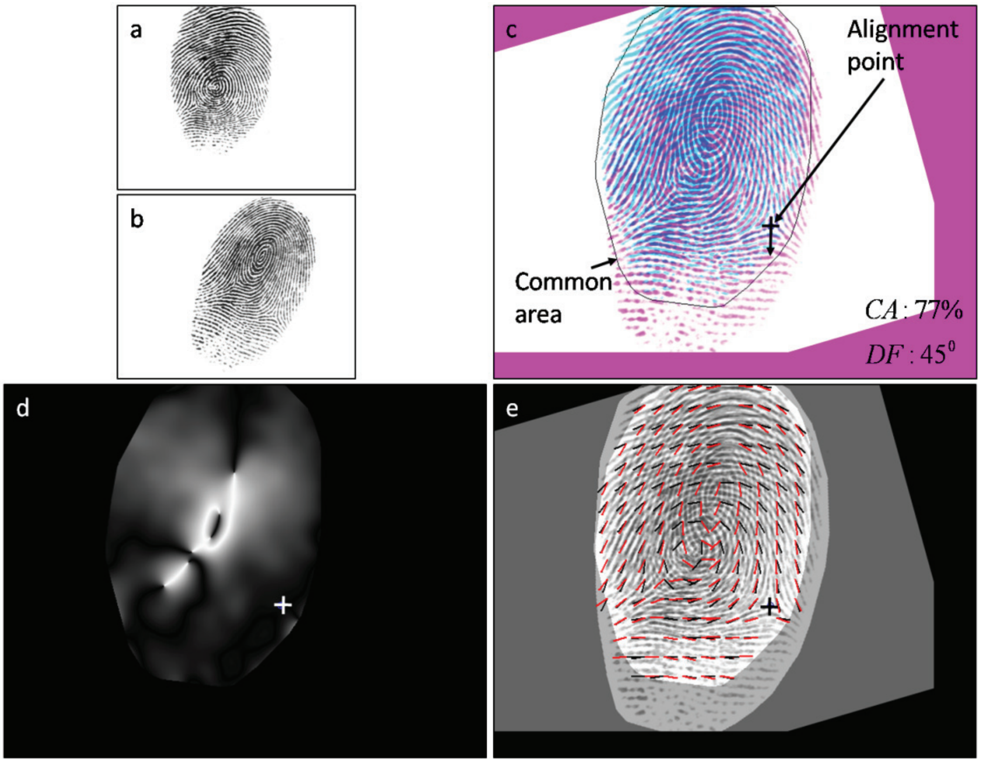


Figure 10 — Computation of common area and relative deformation aligned by the right delta point as AP: (a) and (b) are a mated pair, (c) common area, (d) pixel-level orientation difference (dark-small, light-large), (e) block-wise orientation difference

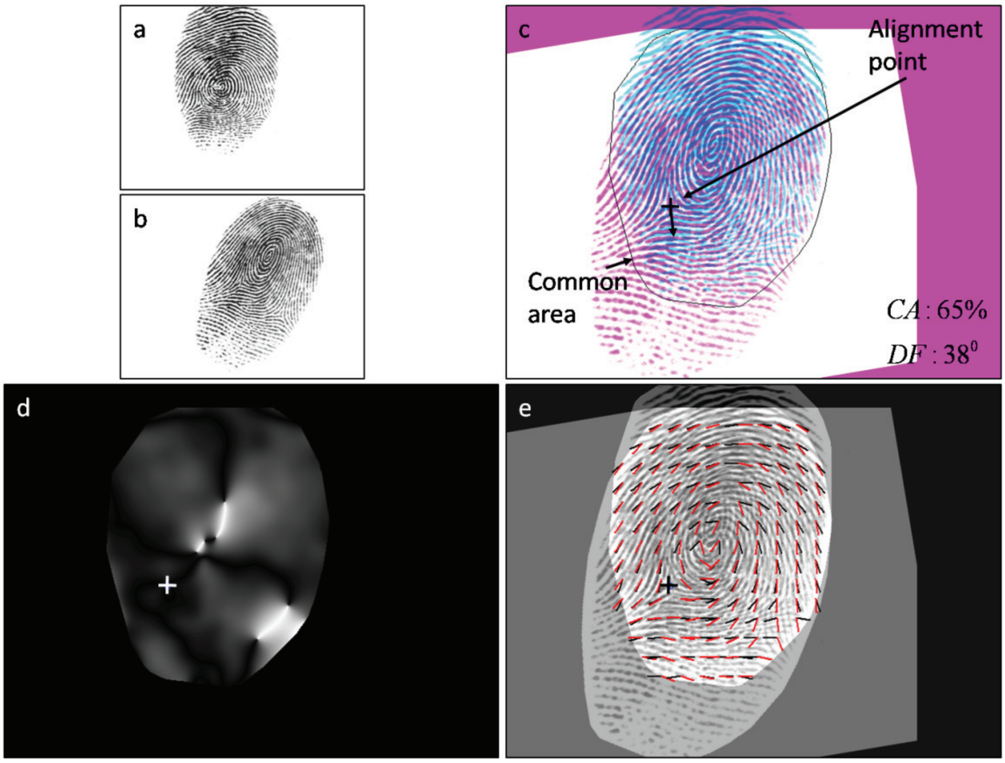


Figure 11 — Computation of common area and relative deformation aligned by the left delta point as AP: (a) and (b) are a mated pair, (c) common area, (d) pixel-level orientation difference (dark-small, light-large), (e) block-wise orientation difference

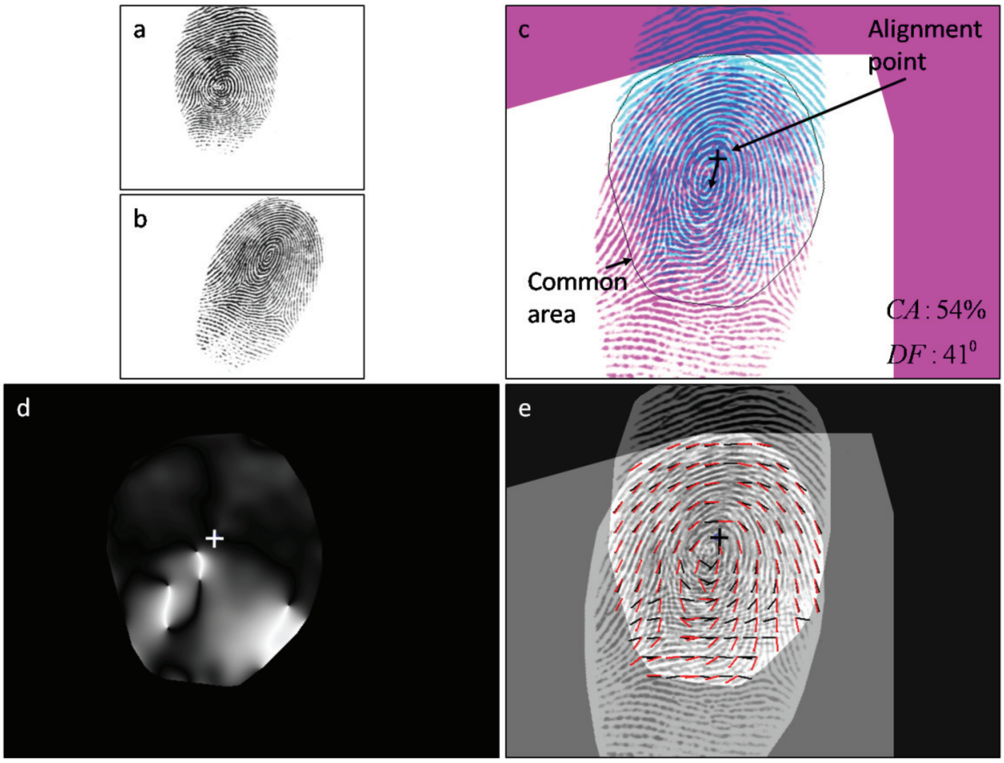


Figure 12 — Computation of common area and relative deformation aligned by the upper core point as AP: (a) and (b) are a mated pair, (c) common area, (d) pixel-level orientation difference (dark-small, light-large), (e) block-wise orientation difference

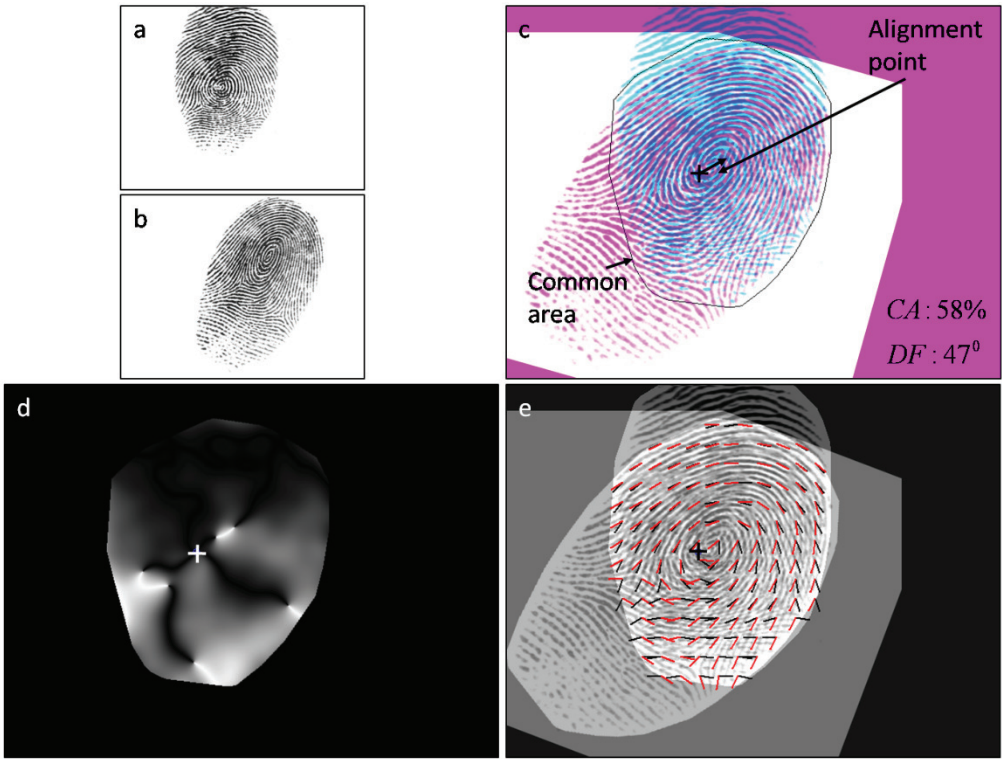


Figure 13 — Computation of common area and relative deformation aligned by the lower core point as AP: (a) and (b) are a mated pair, (c) common area, (d) pixel-level orientation difference (dark-small, light-large), (e) block-wise orientation difference

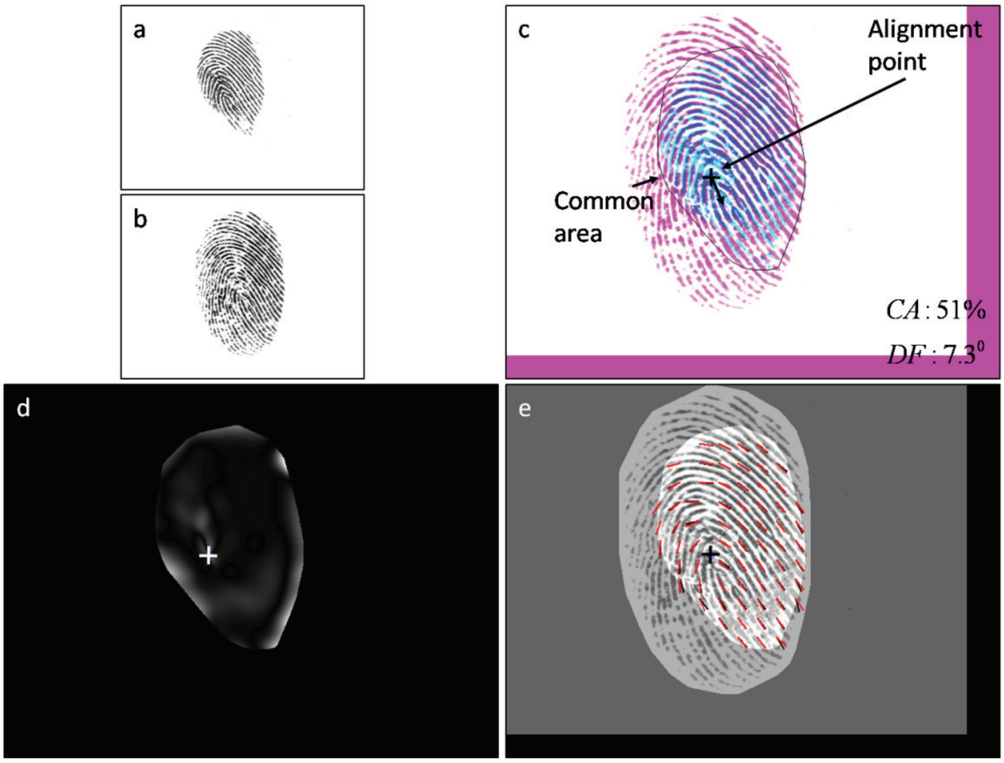


Figure 14 — Example of small common area but low relative deformation: (a) and (b) are a mated pair, (c) common area, pixel-level orientation difference (dark-small, light-large), (e) block-wise orientation difference

4.3.3 Thin plate spline-based measurement

The thin plate spline-based method is one approach to measure the deformation between a pair of fingerprints. In order to measure deformation of mated fingerprint pairs using the thin plate spline, the corresponding minutia sets should be detected robustly. Bazen, et al.[3] proposed the detection of the corresponding minutia sets based on triangular local structure, which is called *minutia neighbourhood*, because the local structures which are originated from only a small area in a fingerprint are unlikely to be seriously deformed by plastic distortions. Since the list of possibly corresponding minutia neighbourhood detected by this local comparison algorithm may contain spurious pairs, the correctness of each pair needs to be verified further using the Shape Context scheme[4] and the RANSAC technique.[5] After the detection of correctly corresponding minutia pairs, the thin plate spline can be applied to compute the bending energy which can be used as the measurement of deformation between the fingerprint pair.

4.3.4 Relationship between orientation field-based deformation and similarity score

Figure 15 shows the scatter plots of the relative deformation (computed from the orientation field) versus the similarity score (obtained by a commercial fingerprint comparison algorithm) for mated pairs over FVC 2000 datasets.[6] It can be carefully said that the relative deformation and the similarity score have an inversely proportional relation. This relation is not so strong because the similarity score is also influenced by other factors such as common area and sample quality. In Figure 15, mated pairs with high similarity scores tend to have low relative deformation while mated pairs with low similarity scores do not necessarily have high relative deformation. Furthermore, mated pairs with high relative deformation tend to produce low similarity scores while mated pairs with low relative deformation do not necessarily produce high similarity scores.

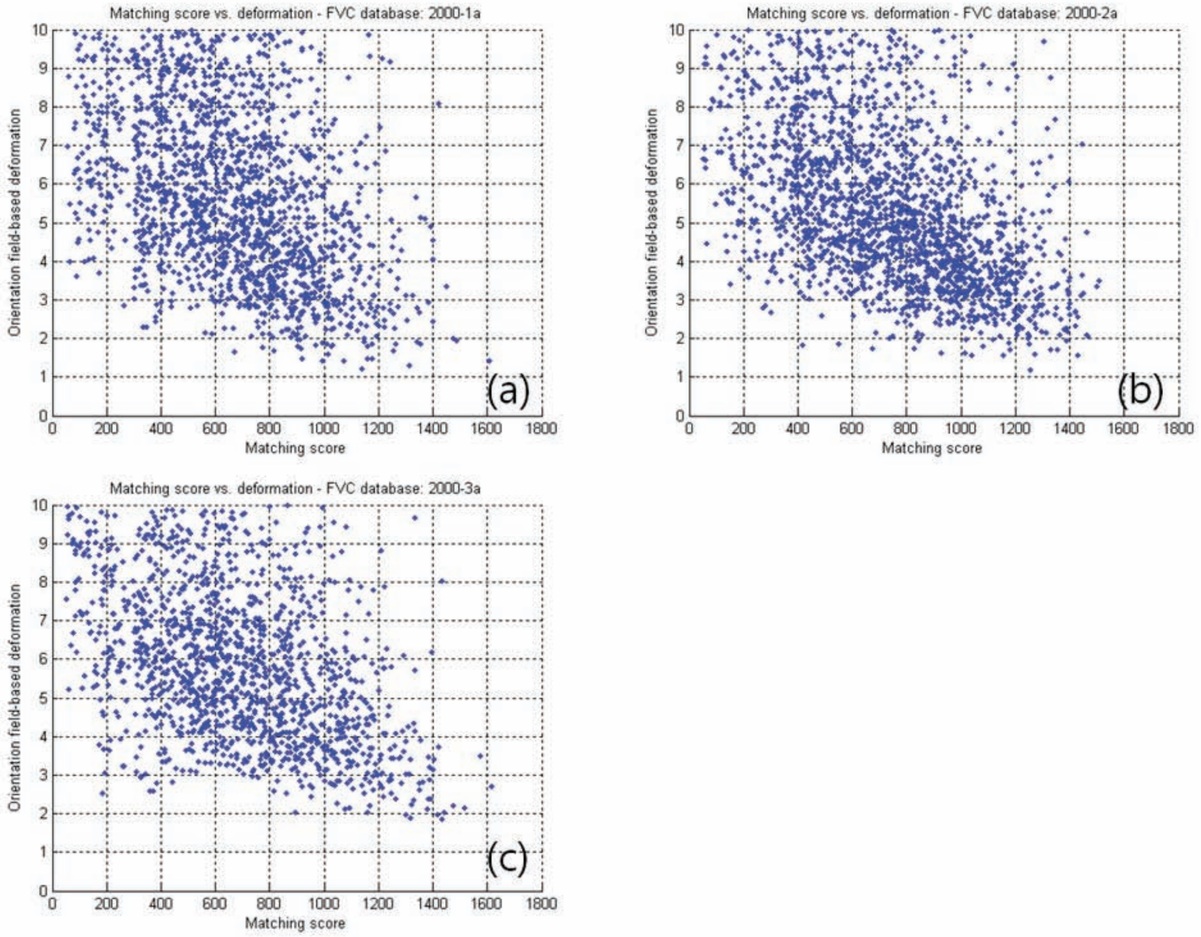


Figure 15 — Scatter plots of relative deformation versus similarity score[9] over FVC 2000 DBs; (a) 2000-1a, (b) 2000-2a, (c) 2000-3a

4.4 Relative sample quality

4.4.1 Introduction

The sample quality of fingerprints is known as one of the most decisive factors which influence the matching performance of fingerprint recognition systems. Thus, the distribution of sample quality of a certain fingerprint dataset becomes an indicator of LOD of the dataset. In the technology evaluation where a reference sample can be of low quality, the quality of both samples, not of only the probe sample, in a mated pair must be considered.

4.4.2 Measurement of relative sample quality

From the aspects of relative sample quality of a mated pair, there are four cases in comparison:

(Case 1) high quality of the reference vs. high quality of the probe

(Case 2) high quality of the reference vs. low quality of the probe

(Case 3) low quality of the reference vs. high quality of the probe

(Case 4) low quality of the reference vs. low quality of the probe

Assuming that the influence of the other factors, common area and relative deformation, to the matching performance are negligible, the similarity scores of the above cases, in general, are ordered as:

$$\text{Case 1} > \text{Case 2} \cong \text{Case 3} > \text{Case 4}.$$

Hence, given a mated pair, the measurement of relative sample quality can be defined by any kind of mean, arithmetic, geometric, or harmonic, of individual sample quality values produced by a fingerprint quality metric described in ISO/IEC TR 29794-4.[16]

4.5 Calculating LOD of a dataset

4.5.1 Introduction

Considering that common area (*CA*), relative deformation (*DF*), and relative sample quality (*RSQ*) between a mated pair of fingerprints are major factors influencing the performance of a comparison algorithm, the similarity score of the mated pair will increase as *CA* and *RSQ* increase while *DF* decreases. For a single mated pair, in general, the level of difficulty is proportional to the similarity score and is a function of the influential factors:

$$LOD_\mathrm{p} = f(CA, RSQ, DF^{-1}, \nu) \propto \text{Similarity score}$$

where LOD_p is the level of difficulty for a single pair of fingerprints, ν represents unknown factors, and DF^{-1} indicates an inversely proportional relation between the relative deformation and the level of difficulty.

4.5.2 Measuring LOD of individual pairs

In order to measure LOD_p, the LOD of a single mated pair of fingerprints, from the multiple factors, it is modelled that LOD_p has a multiple nonlinear regression relationship with *CA*, *RSQ*, and DF^{-1} as:

$$LOD_\mathrm{p} = \beta_{11}CA + \beta_{12}CA^2 + \beta_{21}RSQ + \beta_{22}RSQ^2 + \beta_{31}DF^{-1} + \beta_{32}(DF^{-1})^2$$

where β_{ij} (i = 1, 2, 3 and j = 1, 2) are coefficients to be estimated experimentally from a given training dataset. In practice, since LOD is unknown, LOD_p is replaced with the similarity score of each mated pair using a comparison algorithm at hand. After being estimated by multiple nonlinear regression analysis, β_{ij}'s are used in the above model to calculate the LOD distribution of an unknown dataset under evaluation. In applying the coefficients β_{ij}'s obtained from a training dataset to the LOD calculation model for an unknown test dataset, the underlying assumption is that each factor has a similar amount of influence on matching error rates.

Since LOD_p has a proportional relationship with similarity scores, a linear function can be applied for normalization of LOD_p so that LOD_p has an inverse relationship with similarity scores. The normalized LOD_p, $N\ LOD_\mathrm{p}$, is defined as:

$$NLOD_\mathrm{p} = \frac{100 \times (LOD_\mathrm{max} - LOD_\mathrm{p})}{LOD_\mathrm{max} - LOD_\mathrm{min}}$$

where LOD_max and LOD_min are the maximum and minimum of LOD_p, respectively.

Nine non-synthetic datasets from FVC 2000, 2002 and 2004 are used to demonstrate the validity of the above model. Each dataset contains 800 fingerprints captured from 100 fingers. The LOD is measured only for the genuine mated pairs. Figure 16 compares the distributions of CA, DF, and RSQ, respectively, calculated by the methods described above for three FVC datasets (2000-DB2, 2004-DB1, 2004-DB3), and Figure 17 shows the distribution of LOD for individual pairs from the datasets.

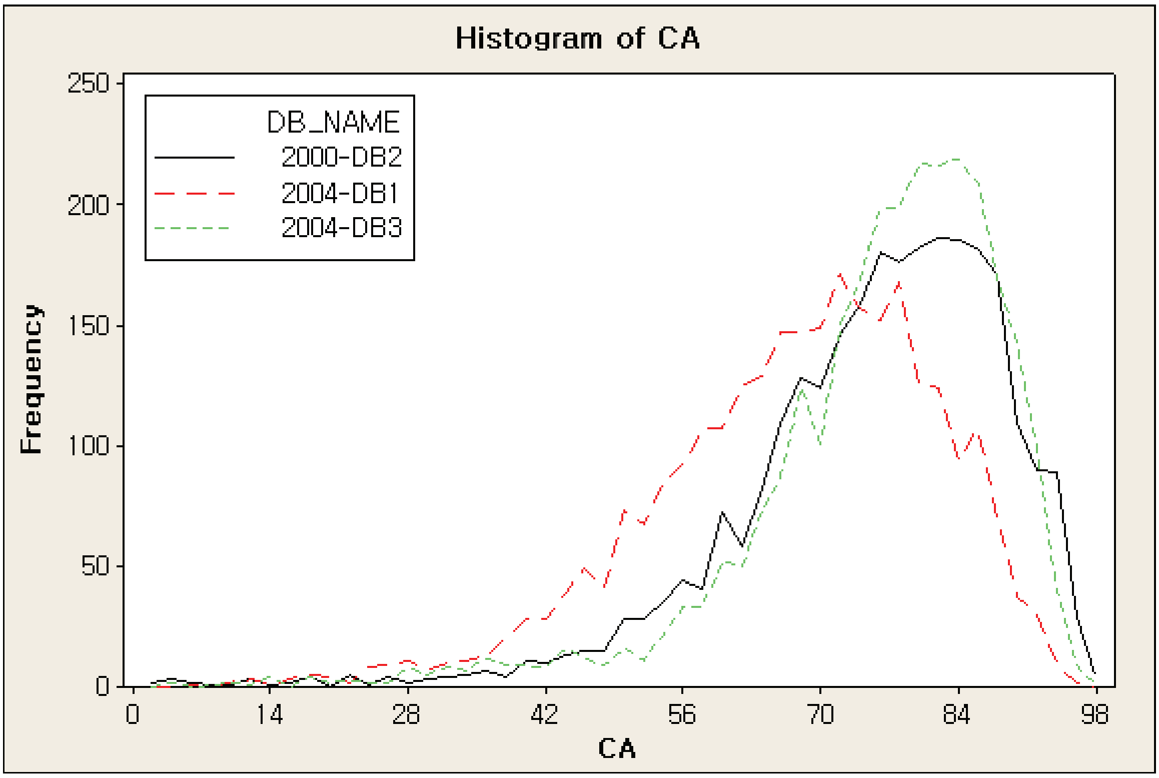


Figure 16 — Histograms of CA of 3 FVC datasets, 2000-DB2, 2004-DB1, and 2004-DB3

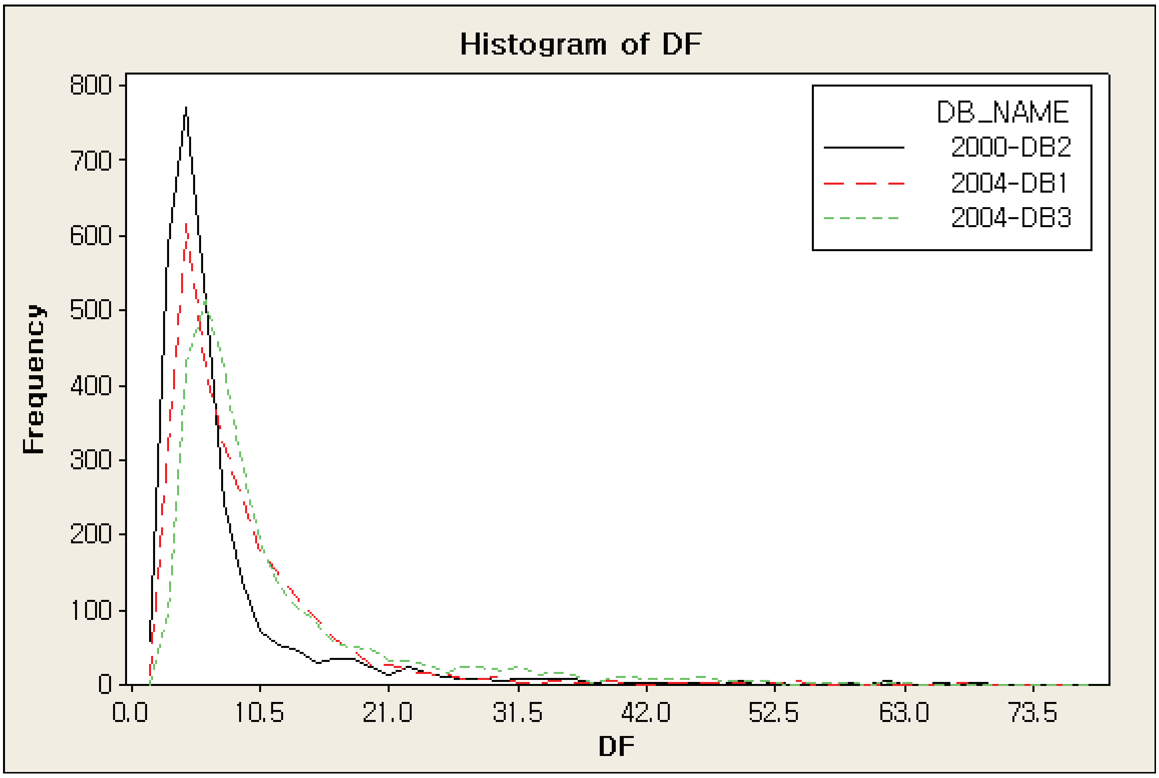


Figure 17 — Histograms of DF of 3 FVC datasets, 2000-DB2, 2004-DB1, and 2004-DB3

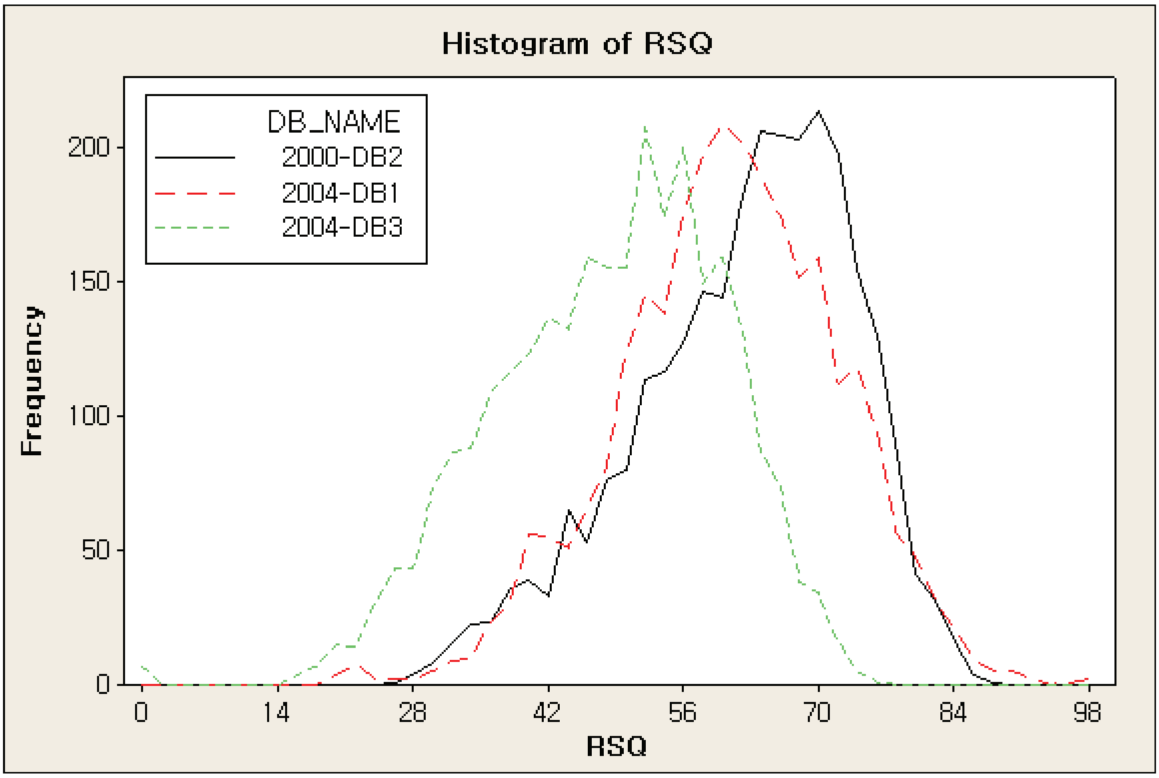


Figure 18 — Histograms of RSQ of 3 FVC datasets, 2000-DB2, 2004-DB1, and 2004-DB3

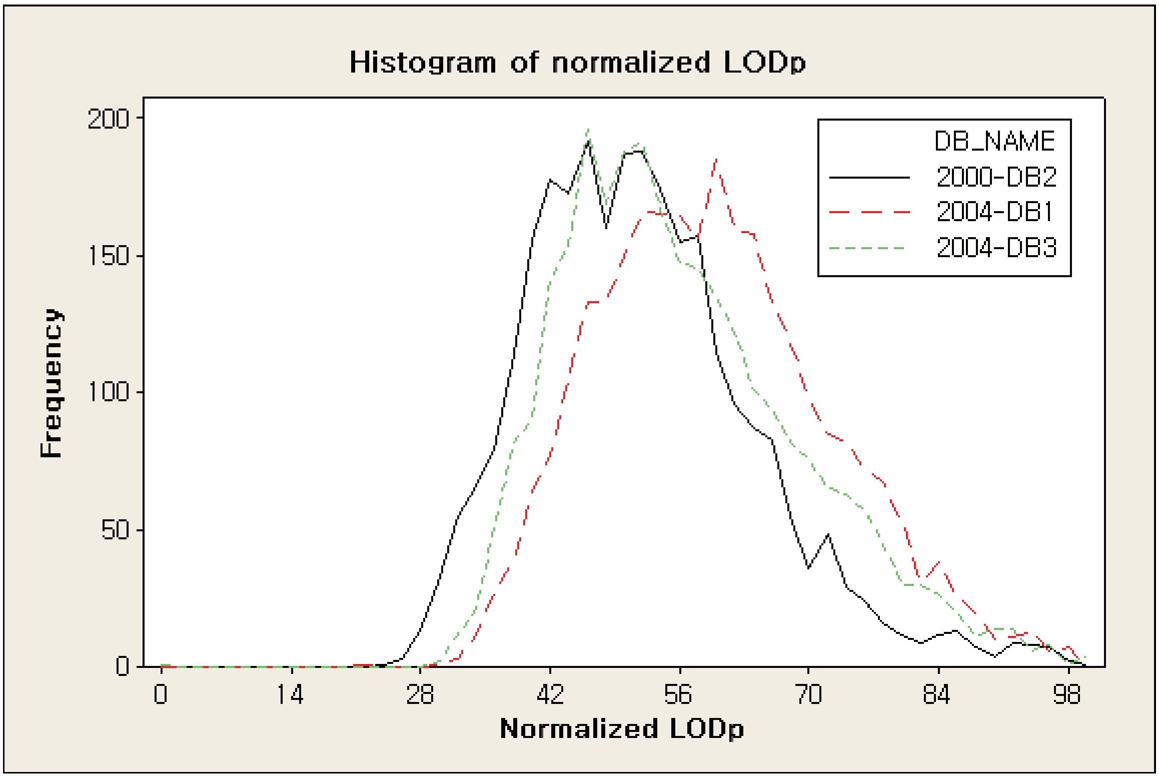


Figure 19 — Histogram of Normalized LOD of 3 datasets, 2000-DB2, 2004-DB1, and 2004-DB3

The one-way ANOVA (ANalysis Of VAriance) and the Tukey's HSD (Honestly Significant Difference) tests are applied to the LOD distributions to examine whether their differences are significant. Figure 18 shows the results of the ANOVA test and grouping of datasets in three difficulty levels using the Tukey's HSD test. In this figure, datasets in different colours are significantly different.

Since the LOD of a dataset is computed based on CA, DF, and RSQ, it is independent of comparison algorithms. However, it is desirable that the LOD has a certain monotonical relationship with the metrics of matching performance, e.g. the equal error rate (EER) of a certain "universal" comparison algorithm. Two widely used comparison algorithms (VeriFinger 5.0[9] and Bozorth3[10]) are utilized as the universal comparison algorithms for all the datasets.

Table 1 also compares the ranked average NLODs against EERs and FRR obtained by the two comparison algorithms across the corresponding datasets. The ranks of LODs are categorized into three classes: easy, medium, and difficult. The table shows that the measured NLOD is almost coincident with actual EERs except the datasets 2000-DB3 and 2002-DB3, which requires further investigation. Figure 20 illustrates a linear relationship between Normalized LOD and EER as measured through Bozorth3.

Table 1 — The Normalized LOD of dataset and corresponding EER computed by comparison algorithms

Class	Dataset	NLOD	EER		FRR(FAR = 0.01 %)
			VeriFinger	Bozorth3	VeriFinger
Easy	2000-DB2	51.85	0.8214	4.28	1.64 %
	2002-DB1	54.46	0.9286	3.38	1.86 %
	2002-DB2	54.88	0.6964	2.51	1.39 %
Medium	2000-DB1	55.59	3.4464	5.98	6.89 %
	2004-DB3	56.14	3.9821	6.64	7.96 %
	2000-DB3	56.29	5.4643	8.0	10.93 %
Difficult	2002-DB3	57.16	2.9821	9.8	5.96 %
	2004-DB2	57.48	5.403	10.75	10.79 %
	2004-DB1	59.68	6.625	13.71	13.25 %

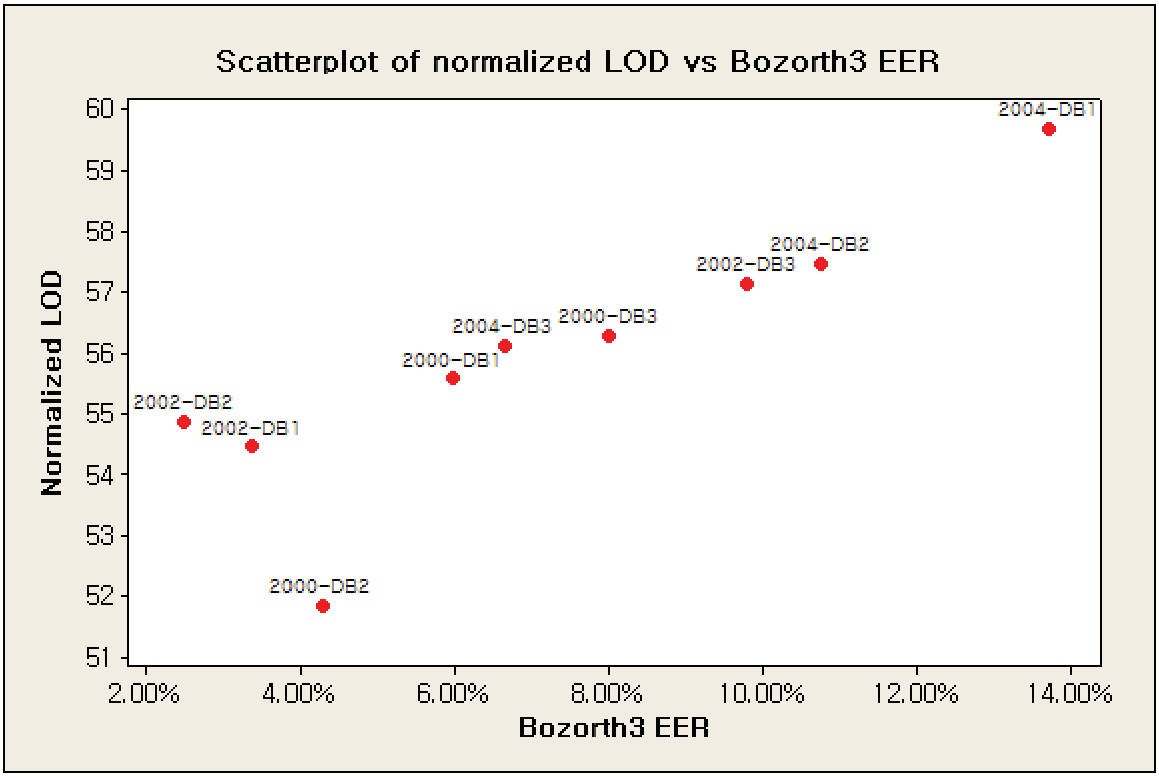


Figure 20 — Example of relationship between Normalized LOD and Bozorth3 EER

5 Analysis of mated pair data characteristics based on comparison results

5.1 General

By using comparison scores, and not image quality values, a technology testing dataset generation methodology that extracts and organizes meaningful data for practical accuracy evaluation can be possible. Furthermore, this methodology is very straightforward. The following observations can be made:

— Generally, each comparison algorithm has different scoring characteristics, and the score values of different algorithms will not be the same, even for mated sample data with same image quality values.

— It is possible to measure "matchability" based on the results of multiple comparison algorithms.

5.2 Matchability

5.2.1 Concept of matchability

The concept of matchability includes the following:

— For a single mated pair data, the matchability is determined for each vendor algorithm.

— Mated pair data that are labelled as capable of being matched are referred to as "matchable".

5.2.2 Criteria for determining matchability

Matchability is a function of the proportion of algorithms through which mated pairs match. Criteria for determining matchability are as follows:

— A mated pair that matches through a large proportion of algorithms is more matchable than a mated pair that matches through a small proportion of algorithms.

— In the aggregate, a dataset whose mated pairs match through a large proportion of algorithms is more matchable than a dataset whose mated pairs match through a small proportion of algorithms.

— If more algorithms with different characteristics can be used, a more universal matchability can be expected to become available.

— The matchability label of each mated pair data is assigned for each comparison algorithm of all vendor software.

— This TR does not provide guidance on evaluating the impact of different sensor types or sensing technologies on matchability. However, tests can be designed that examine the impact of sensor variation – e.g. use of different sensors to collect probe and gallery data – on matchability.

5.2.3 Decision of matchability

Mated pair comparisons may result in match / no match decisions or in comparison scores, depending on the algorithm. For the purposes of matchability assessment, access to comparison scores is strongly preferred. Matcher-specific decision thresholds can be used to determine whether a given mated pair comparison is declared a match. The benefit of this approach is that a relatively small number of mated pairs can be used to characterize a dataset's LOD.

Threshold determination may be based on testing organizations' previous experience with a given algorithm. An understanding of algorithm-specific comparison score distributions will typically simplify matchability-based LOD assessments. Such an understanding will also improve inter-organizational collaboration: sharing score-based decision criteria is more useful than sharing opaque, rank-based results. To sufficiently understand thresholds, execution of substantial non-mated comparisons is typically required.

EXAMPLE A testing organization may have previously established that for comparison algorithm B, a comparison score of 100 typically corresponds to a false match rate of 0.01 %, such that 100 is a reasonable operating point. For the purposes of matchability determination, mated pairs that score below (weaker than) 100 can be considered non-matches, and mated pairs that score 100 or higher can be considered matches.

In some cases, it may be necessary or useful to perform a rank-based analysis of mated pairs, such as when testing an algorithm that only functions in identification mode. In this case, the match decision is based on whether the correct reference matches at Rank 1 against a given probe.

Comparison algorithms used in matchability determinations should not generate heavily quantized comparison scores. Quantization reduces insight into comparison scores; such insight may be necessary to differentiate between subtle differences in mated pair comparison scores. For example, some comparison algorithms only return scores when comparisons are successful, returning null or failure results in the case of failed comparisons. This behaviour is undesirable in that it reduces visibility into borderline comparisons which may be particularly relevant to matchability and to LOD in general.

Table 2 provides an example of deciding matchability using multiple comparison algorithms.

Table 2 — Example of matchability table

Mate Pairs	Algorithm1	Algorithm2	Algorithm3	Matchability
pair1	Match	Match	Match	High
pair2	Match	Match	Nonmatch	Medium
pair3	Nonmatch	Nonmatch	Nonmatch	Low
...	...	...	...	...
pairN	Nonmatch	Nonmatch	Nonmatch	Nonmatchable

The LOD can be associated with a certain performance test and discriminated by the test name, such as “NIST’s MINEX 2004”[7] and “MTIT”[8]. Comparison algorithms can be collected by the organization which held the performance test.

Suppose that we have M algorithms available to assess the LOD of the dataset. Let the number of mated pairs in the dataset be N. For a given mated pair, an algorithm produces a score of +1 if it (1) generates a comparison score stronger than the declared threshold or (2) identifies the correct mate as a rank one match. The algorithm produces a score of −1 otherwise. In this way, we can produce a score for each comparison algorithm, for each probe, which we shall write as $s_{n,i}$ where n refers to the comparison and i refers to the algorithm. We can thus construct the following matrix system:

$$\begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,M} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N,1} & s_{N,2} & \cdots & s_{N,M} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{pmatrix} = \begin{pmatrix} M \\ M \\ \vdots \\ M \end{pmatrix}$$

where the matrix on the left is written $\boldsymbol{S}$, the vector of unknown weights is $\boldsymbol{w}$ and the target on the right is $\boldsymbol{t}$. Note that the matrix $\boldsymbol{S}$ will consist entirely of +1s and −1s. Suppose that we solve the above system for $\boldsymbol{w}$ in a least-squares sense and that we find the minimum-norm solution for $\boldsymbol{w}$, i.e. we find the value of $\boldsymbol{w}$ with the smallest $||\boldsymbol{w}||$ that minimises $||\boldsymbol{Sw} - \boldsymbol{t}||$. This solution is revealing about how easy the data are to match, consider the following extreme cases:

- If the data are all easy to match, then all the comparison algorithms will get the answer correct and the elements of $\boldsymbol{w}$ will all be 1.
- If the data are impossible to match, then all the comparison algorithms will get the answer incorrect and the elements of $\boldsymbol{w}$ will all be −1.

We therefore propose to use the following metric as a measure of dataset matchability:

$$\text{Database matchability} = \|1 - \boldsymbol{w}\|_1 = \sum_{i=1}^{M} |1 - w_i|$$

The minimum value of this quantity is zero, i.e. all the comparison algorithms get all the comparisons correct. The maximum value of this quantity is twice the number of algorithms used, $2M$. This metric can be normalized to [0, 1] by dividing with $2M$.

Let us consider some more examples. Suppose that we have three comparison algorithms, among which algorithm 1 gets most comparisons correct and algorithms 2 and 3 get them mostly wrong. Then we would expect the solution vector to be near to [3 0 0], and the dataset difficulty would be (|1-3| + |1-0| + |1-0|) = 4. Now suppose that algorithms 1 and 2 get most comparisons correct, then we would expect the solution vector to be [1.5 1.5 0] which would give a difficulty of (|1-1.5| + |1-1.5| + |1-0|) = 2. This is intuitively the correct result, because if two algorithms get the correct answers then the dataset is easier than if only one algorithm gets it right.

A subtle point about the above example is that if both algorithms are in agreement (i.e. the system is under-determined), then if we were just minimising $||\boldsymbol{Sw} - \boldsymbol{t}||$ an answer of [3 0 0] would have been acceptable in both cases. However, because we insist on the solution for $\boldsymbol{w}$ that minimises $||\boldsymbol{w}||$ then the solution of [1.5 1.5 0] will be the one that is produced.

It is well known that the minimum norm least squares solution of the proposed matrix solution has an exact solution that can be computed using the Singular Value Decomposition (SVD). Using the SVD also has the advantage that it deals with the case when $\boldsymbol{S}$ is rank deficient, which can occur, for example when all the algorithms produce the same output for all data.

5.2.4 Controlling for the influence of individual factors on overall performance

Assuming that all the influencing factors to the quality of fingerprint samples are reflected in comparison results, the amount of influence of individual factors to the overall performance may be estimated by carefully constructing a dataset for evaluation. In order to analyse the influence of aging, for example, fingerprints should be collected so that the correlation between the aging period and the comparison scores is computed.

The same can be said for sensor types and others. However, in order to reflect the differences of sensor properties into datasets, to the extent possible, samples of the data of the same test subjects acquired by different sensors should be used.

Also, when using fingerprint data of multiple fingers of the same person, it is said that the patterns of the fingers are similar. However, this Technical Report does not specifically handle such data characteristics distinctly. This is because, if it is possible to correctly distinguish the fingers that have high pattern correlativity, it would be suitable for inspecting the performance characteristics of inter-pattern discrimination.

5.3 Building datasets of different levels of difficulty

Datasets of different levels of difficulty can be built as follows:

- An “easy” standard corpus can be built by selecting the intersection of correctly matched mated pair data for all vendor algorithms.
- A “medium“ standard corpus can be built by selecting mated pair data for which a match was achieved with at least two algorithms.
- A “difficult” standard corpus can be built by selecting mated pair data where a match was achieved with at least one algorithm.

The LOD of a corpus can be associated with the distribution of the number of algorithms that correctly match or achieve a 1st ranked match for all the mated pairs.

Rank or score information may also be used for specifying a more challenging LOD. For example, a rank-2 result indicates that a given mated pair was more difficult to a match rank-1 result.

Figure 24 shows the result of an experiment of building fingerprint datasets with different levels of difficulty, where three comparison algorithms are used with 85 mated pairs, and Figure 25 compares the DET curves obtained by a single comparison algorithm over two different levels of difficulty, “easy (green)“ and “difficult (red)“.

NOTE 1 While matchability successfully provides a measure of number of difficult to match mated pairs in the dataset, it does not provide any information about non-matedairs that are easy to be falsely matched.

NOTE 2 Matchability as a measure of LOD of dataset is modality independent and can be applied to dataset of other biometric characteristic such as face.

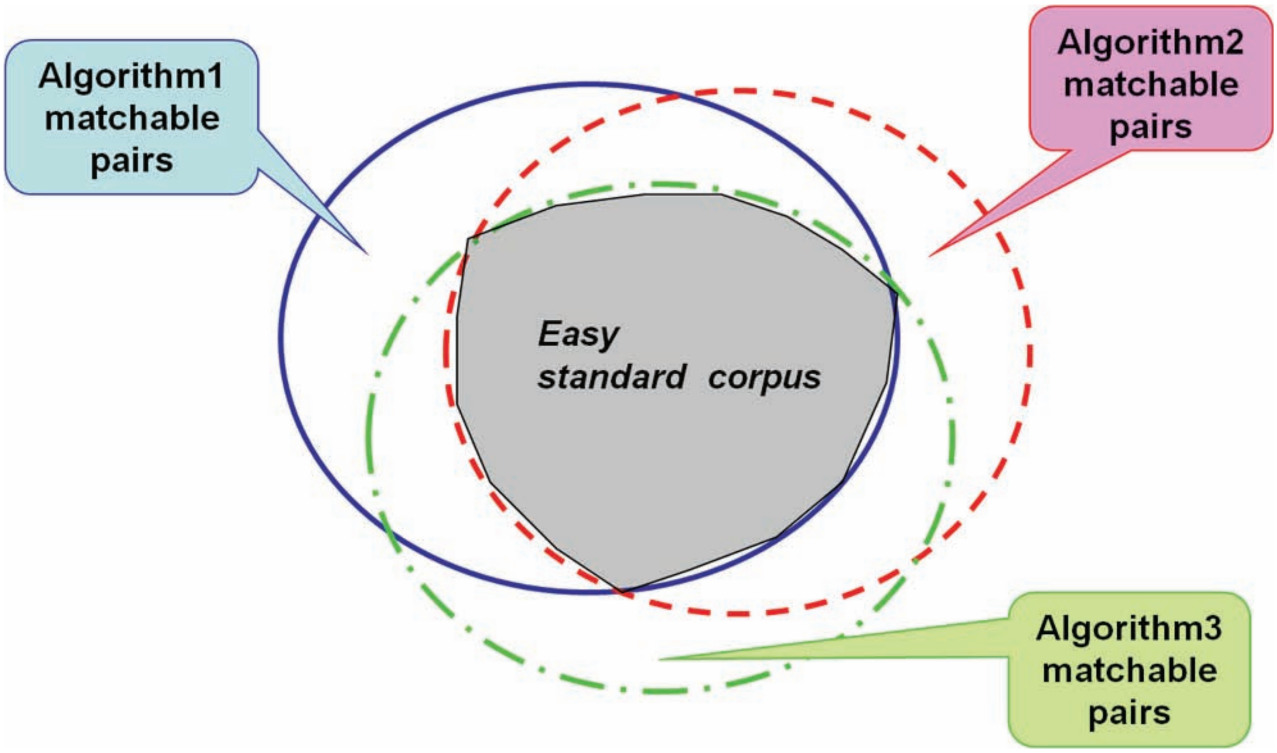

Algorithm1 matchable pairs
Algorithm2 matchable pairs
Easy standard corpus
Algorithm3 matchable pairs

Figure 21 — Conceptual diagram of “Easy” standard corpus

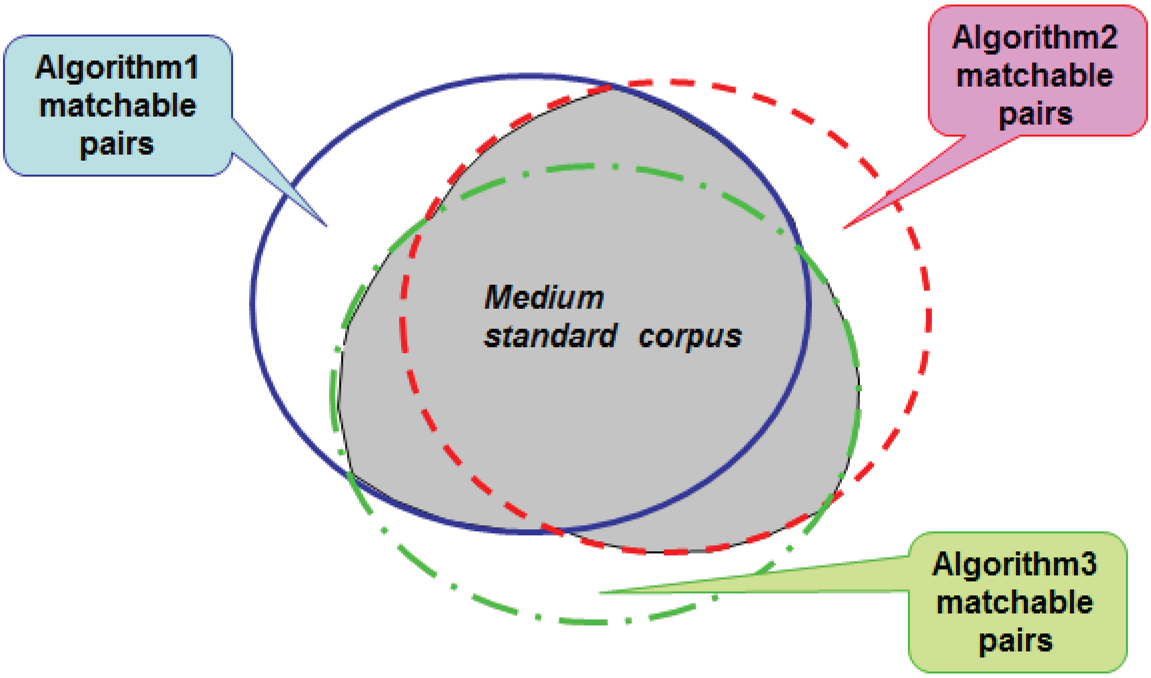

Algorithm1 matchable pairs
Algorithm2 matchable pairs
Medium standard corpus
Algorithm3 matchable pairs

Figure 22 — Conceptual diagram of “Medium” standard curpus

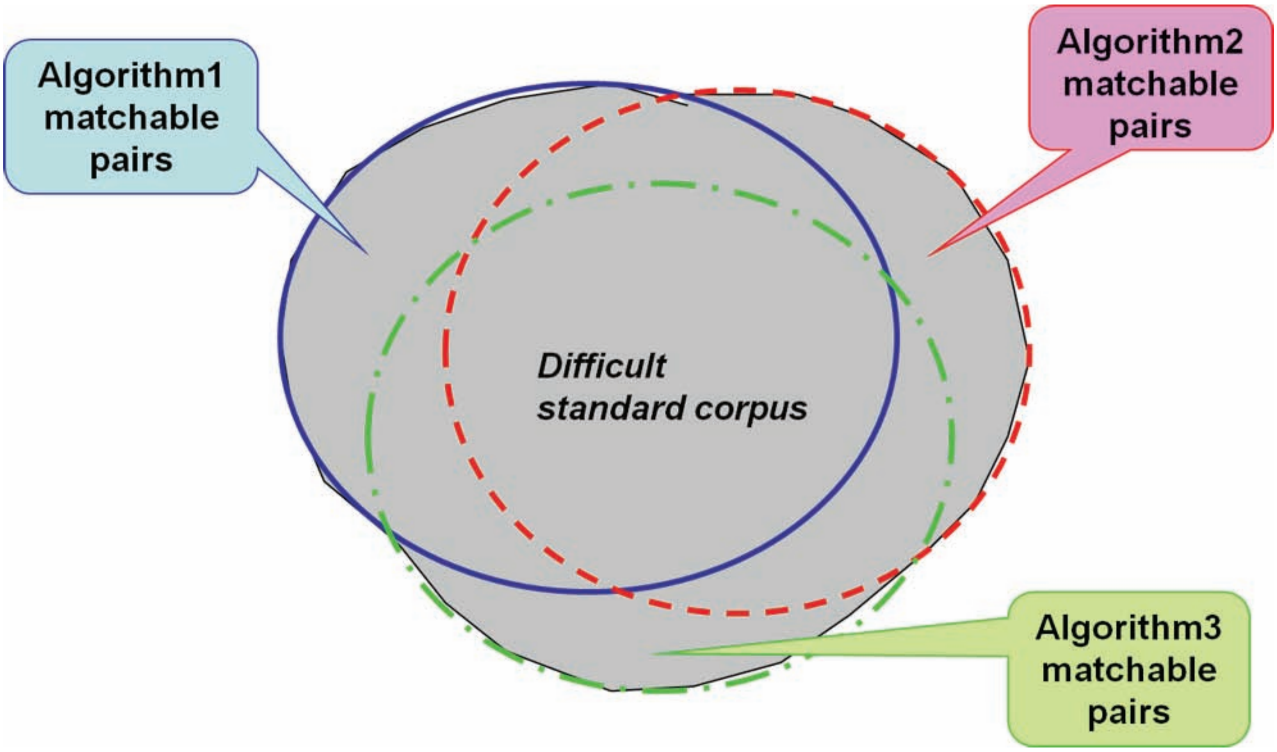
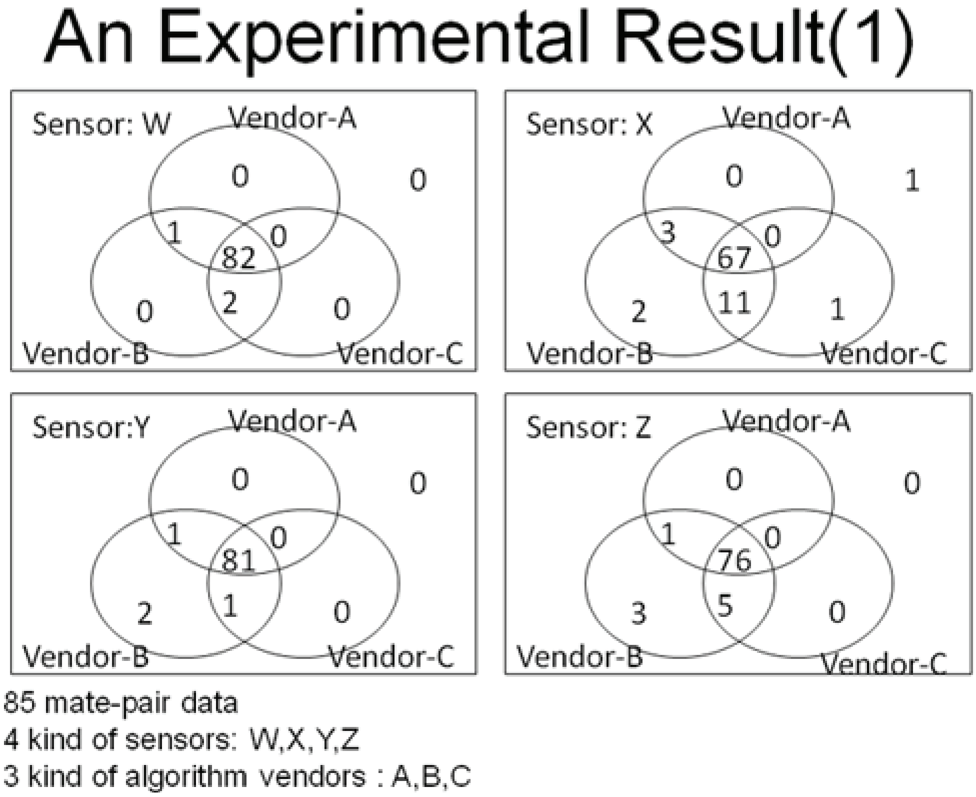

Algorithm1 matchable pairs
Algorithm2 matchable pairs
Difficult standard corpus
Algorithm3 matchable pairs

Figure 23 — Conceptual diagram of “Difficult” standard corpus



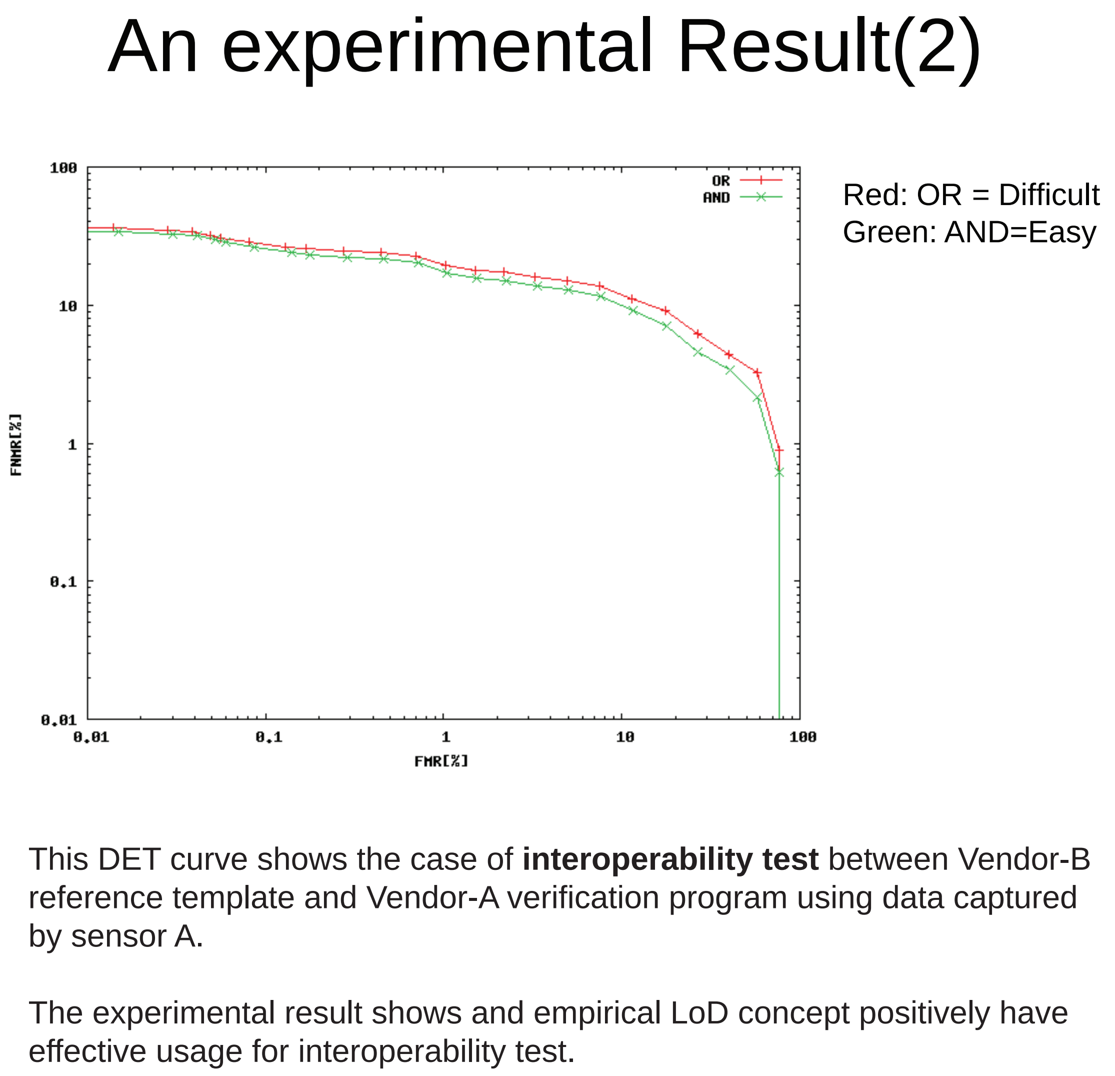
An Experimental Result(1)

85 mate-pair data

4 kind of sensors: W,X,Y,Z

3 kind of algorithm vendors : A,B,C

Figure 24 — An experimental result of building fingerprint datasets with different levels of diffculty



An experimental Result(2)

Red: OR = Difficult
Green: AND=Easy

This DET curve shows the case of **interoperability test** between Vendor-B reference template and Vendor-A verification program using data captured by sensor A.

The experimental result shows and empirical LoD concept positively have effective usage for interoperability test.

Figure 25 — Comparison of DET curves over two different levels of difficulty

Bibliography

[1] Hicklin R.A., & Reedy C.L. Implications of the IDENT/IAFIS Image Quality Study for Visa Fingerprint Processing,“ Technical Report, MitreTek Systems, Inc., Oct. 31, 2002

[2] Jin C., & Kim H. Pixel-level singular point detection from multi-scale Gaussian filtered orientation field. *Pattern Recognit.* 2010, **43** pp. 3879–3890

[3] Bazen A.M., & Gerez S.H. Fingerprint matching by thin-plate spline modelling of elastic deformation. *Pattern Recognit.* 2003, **36** pp. 1859–1867

[4] Belongie S., Malik J., Puzicha J. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* 2002, **24**, pp. 509–522

[5] Fischler M.A., & Bolles R.C. Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography *Commun. ACM*, 1981, **24**, pp. 381–395

[6] http://bias.csr.unibo.it/fvc2000/Downloads/fvc2000_report.pdf

[7] http://www.nist.gov/itl/iad/ig/minex.cfm

[8] Bazin A.I., & Mansfield T. An Investigation of Minutiae Template Interoperability. 2007 IEEE Workshop on Automatic Identification Advanced Technology, pp. 13–18, June 2007

[9] VeriFinger 5.0., Neurotechnology, http://www.neurotechnology.com/verifinger.html

[10] Bozorth3, NIST, http://www.nist.gov/itl/iad/ig/nbis.cfm

[11] ISO/IEC 19795-1, *Information technology — Biometric performance testing and reporting — Part 1: Principles and framework*

[12] ISO/IEC 19795-2, *Information technology — Biometric performance testing and reporting — Part 2: Testing methodologies for technology and scenario evaluation*

[13] ISO/IEC TR 19795-3, *Information technology — Biometric performance testing and reporting — Part 3: Modality-specific testing*

[14] ISO/IEC 19795-4, *Information technology — Biometric performance testing and reporting — Part 4: Interoperability performance testing*

[15] ISO/IEC 29794-1, *Information technology — Biometric sample quality — Part 1: Framework*

[16] ISO/IEC TR 29794-4, *Information technology — Biometric sample quality — Part 4: Finger image*

ICS 35.040

Price based on 28 pages