Assessing the level of difficulty of fingerprint datasets based on relative quality measures

Shengzhe Li\textsuperscript{a}, Hakil Kim\textsuperscript{a,}\textsuperscript{*}, Changlong Jin\textsuperscript{b}, Stephen Elliott\textsuperscript{c}, Mingjie Ma\textsuperscript{d}

\textsuperscript{a}Graduate School of Information and Communication Engineering, Inha University, Incheon, Republic of Korea
\textsuperscript{b}Department of Computer Science, Shandong University at Weihai, China
\textsuperscript{c}Department of Technology, Leadership and Innovation, Purdue University, West Lafayette, USA
\textsuperscript{d}Department of Statistics, Inha University, Incheon, Republic of Korea

\textbf{A R T I C L E   I N F O}

Article history:
Available online 10 June 2013

Keywords:
Level of difficulty
Fingerprint dataset
Biometric similarity score
Sample quality
Relative sample quality
Common area

\textbf{A B S T R A C T}

Understanding the difficulty of a dataset is of primary importance when it comes to testing and evaluating fingerprint recognition systems or algorithms because the evaluation result is dependent on the dataset. The difficulty exhibited in this paper represents how difficult it is to achieve better recognition accuracy within the specific dataset. Proposed in this paper is a general framework for assessing the level of difficulty of fingerprint datasets based on quantitative measurements of not only the sample quality of individual fingerprints but also the relative differences between genuine pairs, such as common area and deformation. The experimental results over various datasets demonstrate that the proposed method can predict the level of difficulty of fingerprint datasets which coincide with the equal error rates produced by four comparison algorithms. The proposed method is independent of comparison algorithms and can be performed automatically.

\textcopyright 2013 Elsevier Inc. All rights reserved.

1. Introduction

As the worldwide deployment of fingerprint recognition systems has been increasing, the demand for evaluating their performance is also growing rapidly. Many public and private organizations, including academia, have conducted technology evaluations of fingerprint recognition systems with their own datasets. Because these datasets are collected by different organizations and there is no specific method for measuring the difficulty of a dataset, the evaluation results over various datasets cannot be compared. The difficulty exhibited in this paper represents how difficult it is to achieve better recognition accuracy within the specific dataset.

To test the performance of a fingerprint recognition system, the dataset should be on a standardized corpus, ideally collected by a “universal” sensor (i.e., a sensor that collects samples equally suitable for all algorithms tested) [1]. Nonetheless, performance against this corpus will depend on both the environment and the population in which data is collected. Furthermore, building a standardized corpus with a “universal” sensor is also impractical. Therefore, there are inevitable differences in difficulty between datasets. Because of this, there have been other approaches to quantify performance of the biometric system. One method is that of the biometric zoo or menagerie, which examines the performance of the individual. The most popular zoo examples are that of Doddington et al. [14] and Yager et al. [29].

Doddington et al. [14] originally suggested the biometric zoo in the speech and speaker recognition systems and stated that one major factor affecting performance of these systems is inherent differences in the recognizability of different

\* Corresponding author. Tel.: +82 10 6336 7385.
E-mail addresses: szli@vision.inha.ac.kr (S. Li), hikim@inha.ac.kr (H. Kim), cljin@sdu.edu.cn (C. Jin), elliott@purdue.edu (S. Elliott), mmj810611@hotmail.com (M. Ma).

0020-2055/ - see front matter \textcopyright 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.05.025
speakers. Yager et al. [29] extended Doddington’s zoo and defined additional four types of animals in terms of a relationship between a user’s genuine and impostor comparison results. The biometric zoo focuses on the results of an algorithm on a dataset. Thus, the results of zoo analysis are algorithm-dependent and difficult to interpret across the datasets. Some studies take account of the user’s characteristic and developed many score normalization methods in order to predict the performance of the biometric systems [23,24,26,27]. Poh et al. [26] suggested the biometric menagerie index (BMI) that defined as the ratio of the between-client variance and the expectation of the total variance of user’s genuine and impostor comparison. Teli et al. [27] proposed a generalization method for the biometric zoo based on a hierarchical framework. However, these methods still based on the specific algorithms and the difficulty of a dataset cannot directly be represented. O’Connor has proposed a stability score index, which provides insight into particular users who perform poorly or exceptionally well in a particular dataset [25].

Many studies have shown that the sample quality (SQ) of a fingerprint strongly affects the performance of a recognition system [9,11,13,15–17,21]. The SQ of a fingerprint is considered to be the reliability of the features that are extracted from the fingerprint, and it can be adopted as a certain weight to reflect how well it can provide information for comparison algorithms to improve recognition performance. However, the SQ of a single fingerprint cannot represent the relative rotation and deformation (DF) or common area (CA) between the reference and the probe images to be compared. Image quality metrics only can quantify the quality of the reference and probe images separately.

The purpose of this paper is to provide a statistical method for assessing the level of difficulty (LOD) of a given fingerprint dataset by measuring the CA, DF and relative sample quality (RSQ) of mated pairs. These factors are combining into a single LOD score by several methods, such as, linear regression, polynomial regression and neural network. The Kruskal–Wallis test is employed to test the statistical significance of the difference among the resulting LODs for datasets [28]. The proposed method can be applied for characterizing and measuring the level of difficulty of fingerprint datasets used in technology evaluation.

The preliminary version of this paper was previously reported in a conference paper [20]. The present paper provides a complete description of the model selection of the LOD and a comparison with the biometric zoo.

2. Definition and model of LOD

2.1. Definition of LOD

The LOD is defined as a relative measure of a fingerprint dataset that represents how “challenging” or “stressing” the fingerprint dataset is for recognition compared to other datasets [2]. There are several influential factors in the performance of fingerprint recognition, such as 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.), cross sensor/cross impression type comparisons, template aging, and subject cooperation [3]. However, some of these factors are impossible to be quantitatively measured from the fingerprint images. In this paper, CA, DF and relative sample quality (RSQ) are considered to be the most influential factors for representing the LOD.

2.2. Common area

Regardless of the comparison algorithms, the CA of a mated pair is one of the major factors determining the similarity score. In general, a larger CA results in a higher similarity score, and conversely, a smaller common area leads to a lower similarity score. Therefore, the similarity score of the mated pair can be considered to be proportional to the CA between the pair. Considering the difference in image size and resolution between datasets, it is more appropriate to define the measure of the CA as the ratio of the intersection between two images to their union rather than the actual common area, as follows:

$$CA = \frac{F_r \cap F_p}{F_r \cup F_p}$$

where $F_r$ and $F_p$ denote the fingerprint foreground in the reference and the probe fingerprints, respectively. CA ranges from 0 to 1, where a larger CA indicates a larger ratio of the common area between the mated pair. The main steps for computing the CA are summarized as follows:

1. Segment the input fingerprint pair.
2. Detect the alignment point of the two fingerprints.
   a. for non-arch fingerprints, the mated singular points are selected as the alignment points,
   b. for arch fingerprints, the maximum point pair in the angular difference and the orientation certainty level along the symmetry line are selected as alignment points [19].
3. Translate, rotate, and align the probe to the reference.
4. Count the number of pixels in the overlapping region and compute CA between them.
Alignment point detection algorithm is used to automatically align two genuine fingerprints based on the singular points (for non-arch type fingerprints) or symmetry lines (for arch type fingerprints) in pixel-level [18]. Fig. 1e illustrates the CA of a pair of mated fingerprints. When there are multiple pairs of alignment points, the one that has the minimum deformation within the common area should be selected.

2.3. Deformation

Due to the impression pressure and the softness of the finger, there is always a deformation in a mated fingerprint pair. Severe deformation generally causes a low similarity score even with a large CA. A traditional approach to the quantification of DF is the measurement of the geometric distortion among pairs of corresponding landmarks in the mated pair [9]. However, it is difficult to extract the corresponding landmarks robustly.

In order to overcome this problem, this study adopts the measurement of orientation difference between all the corresponding pixels in the common region. The orientation of each pixel is computed from the pixel-level multi-scale Gaussian filtered orientation field [18]. If there is no relative deformation in a mated pair after alignment, two orientation fields will be exactly matched at the corresponding pixels. Otherwise, there will be angle differences. Therefore, the DF can be quantified by averaging the angle difference over the aligned common region of the orientation field. The DF between the reference and the probe fingerprints is defined as:

\[
DF = \frac{1}{n} \sum_{i \in F_r \cap F_p} \frac{\arccos \left( \cos \theta_i \cos \theta'_i + \sin \theta_i \sin \theta'_i \right)}{2}
\]

where \( \theta_i \) and \( \theta'_i \) are doubled angles, which range from 0° to 360°, at the coupled pixel \( i \) in the common region of the reference and the probe, respectively, and \( n \) is the total number of pixels located in the overlapping region. The maximum angle difference between the two doubled angles, \( \theta_i \) and \( \theta'_i \), is 180°, thus, DF ranges from 0° to 90°.

The process of measuring relative deformation can be summarized as follows:

1. Compute the pixel-level orientation fields from each of a mated fingerprint pair [18].
2. Align the two orientation fields of the fingerprints and extract the common region. (refer to steps 1–3 in computing the CA)
3. Compute the average angle difference between the two orientation fields.

Fig. 1f illustrates the DF between mated fingerprints with different levels of deformation in the overlapped region.

2.4. Relative sample quality

The SQ of a fingerprint is known as the most decisive factor in the performance of a fingerprint recognition system. In fingerprint comparison, however, the SQ of both the reference and the probe should be considered. Thus, the relative sample quality (RSQ) can be defined by the geometric mean of two SQs as

\[
RSQ = \sqrt{SQ_r \cdot SQ_p}
\]
where SQ_r and SQ_p are the sample quality of the reference and the probe fingerprints. There are some studies shown that the orientation certainty level (OCL) based quality assessment can well predict the recognition accuracy, therefore the OCL is adopted as SQ\[17,21,30\].

2.5. LOD modeling

Fig. 2 demonstrates that the above relative measures and the similarity scores have different degrees of positive or negative proportional relations. However, these factors do not solely affect the similarity scores. For those mated pairs that have relatively lower similarity scores, the influential factors might come from many aspects. An effective approach is to combine those factors as a single LOD measure.

The CA, DF, and RSQ in a mated pair of fingerprints are major factors influencing the performance of a comparison algorithm. For a single mated pair, in general, the LOD is proportional to the similarity score and is a function of the influential factors:

\[
\text{LOD} = f(\text{CA}_i, \text{DF}_i, \text{RSQ}_i, \xi_i) \propto \text{Similarity score}
\]

where CA_i, DF_i, and RSQ_i are measured from ith mated pair by (1)–(3), and \(\xi_i\) refers to unknown factors. Assuming that the CA, DF and RSQ are linear or polynomial to the LOD, the difficulty of the mated pair will increase as the CA and RSQ increase while the DF decreases. However, defining a specific model for the LOD is a difficult problem. In this paper, three types of models are assumed and compared; they are the linear regression model, the polynomial regression model and the non-linear neural network model.

A linear regression model is the simplest and straightforward way to model the LOD. Assuming the LOD is a linear combination of the CA, DF and RSQ for each individual mated pair. The linear LOD model is defined as

\[
\text{LOD}_i = \beta_0 + \beta_1 \text{CA}_i + \beta_2 \text{DF}_i + \beta_3 \text{RSQ}_i + \epsilon_i
\]

Fig. 2. Scatter plot of similarity score versus three relative measures. (a) Similarity score versus common area. The larger common areas tend to produce higher similarity scores. (b) Similarity score versus deformation. The smaller deformations tend to produce higher similarity scores. (c) Similarity score versus relative sample quality. The higher relative sample qualities tend to have higher similarity scores.
where $b_i$ is the regression coefficient and $\varepsilon_i$ is a random error. The coefficient $b_i$ can be estimated using the least square method. In practice, since $LOD_i$ is unknown, it is replaced with the similarity score of each mated pair using a comparator at hand.

Similarly, a polynomial regression model (which can be considered as a special case of the more general multiple linear regression model) for the LOD is defined as

$$LOD_i = b_0 + b_1 CA_i^2 + b_2 CA_i + b_3 DF_i + b_4 DF_i^2 + b_5 RSQ_i + \varepsilon_i$$

(6)

where $b_i$ is the regression coefficient and $\varepsilon_i$ is a random error.

Since there is no assurance that all of the predictors affect the quality of the model, the most adequate model can be determined by a statistical model selection, such as forward selection, backward elimination and stepwise selection.

If the LOD is considered to have a complex non-linear relationship between the CA, DF and RSQ, a neural network based model can be applied. For training the network, the CA, DF and RSQ of mated pairs are used as the inputs and the corresponding similarity scores are used as the targets.

Remapping the LOD score as an increase in the difficulty of mated pair or dataset, a normalized LOD (NLOD) is defined as

$$NLOD_i = 100 \frac{LOD_{\text{max}} - LOD_i}{LOD_{\text{max}} - LOD_{\text{min}}}$$

(7)

where $LOD_{\text{max}}$ and $LOD_{\text{min}}$ are the maximum and minimum LOD scores over whole datasets, NLOD ranges from 0 to 100, while 0 indicates the easiest and 100 indicates the most difficult.

3. Experimental results

3.1. Datasets and experimental environment

Twelve datasets from Fingerprint Verification Competition (FVC) and two High Resolution Fingerprint (HRF) datasets from the Hong Kong Polytechnic University are used to demonstrate the validity of the LOD models (synthetic fingerprints are excluded) [4–8]. The number of genuine mated pairs in a dataset is $N_{n_i}$, where $N$ is the number of fingerprints in the dataset and $n$ is the number of images of each fingerprint. There are 2800 mated pairs in each dataset of FVC2000–2004, 9240 mated pairs in each of FVC2006, and 6660 mated pairs in each of HRF. The LOD calculation is performed for all of those mated pairs. Four comparison algorithms are used in this experiment: VeriFinger 5.0 (VF) from Neurotechnology, Bozorth3 (B3) from National Institute of Standards and Technology (NIST), US, Fingerprint SDK 2009 form Griaule (GR), and IZZIX SDK from Digent (DG). Both algorithms are invariant to sensor type and image sizes.

3.2. Validation of LOD model

The linear regression/polynomial regression/neural network based LOD models were tested on various datasets/comparators to justify the validity of the proposed LOD model. First, the LOD model is trained by a single dataset and the results showed that the linear regression was the most adequate. Then the LOD model is trained using multiple datasets and the results showed that the neural network becomes more adequate. That is, when the training samples are large enough to represent real world samples, the neural network based method performs very well; otherwise, when the training data is not large enough, the linear regression is recommended as a reasonable replacement. Fig. 3 shows the LOD trained using three models versus equal error rates (EER). Except for FVC2006–1 (a very low resolution sensor) the LODs mutually coincide with the EERs.

To quantitatively determine the goodness-of-fit of the three types of models over four comparators, adjust-$R^2$ [28] of each model–comparator’s combination is computed. This value represents how well the LOD model trained by comparator $A$ predicts the EER of comparator $B$. Table 1 is the cross-validation of the single dataset trained LOD and EER over three models and four comparison algorithms. Table 2 shows another scenario when the LOD is trained using multiple datasets. Overall, considering the trade-off between data dependency and prediction performance, the linear LOD model is adopted in the following experiment.

3.3. Distribution of CA, DF, RSQ and NLOD

The histograms of the CA, DF, RSQ and NLOD for three datasets are shown as examples in Fig. 4. The different characteristics of each dataset can be observed from its distribution. Note that none of the measures follows a normal distribution. The NLOD appears as nearly normal, but this hypothesis is rejected using an Anderson–Darling normality test. Thus, the median might be a more appropriate statistic for comparing the LOD between datasets. FVC2000-2 shows the highest CA, RSQ, the lowest DF, NLOD. On the other hand, FVC2004-1 shows the lowest CA, RSQ, the highest DF, NLOD. Therefore an argument could be made that, among these three datasets, FVC2000-2 is relatively the easiest, and FVC2004-1 is the most difficult.

To compare the NLOD of two or more independent datasets, the Kruskal–Wallis test is employed [28]. The Kruskal–Wallis test is a nonparametric alternative to analysis of variance. The test does not require the data to be normal, but instead uses
the rank of the data values rather than the actual data values for the analysis. Fig. 5 shows the confidence interval of the median of proposed measures versus the datasets.

3.4. Comparing LOD with EER

The LOD is compared with the EER. Four comparison algorithms are utilized as the universal comparators for all datasets. Table 3 compares the median of the NLODs and EERs obtained by the four comparators across the corresponding datasets. Table 3 shows that the NLOD is indeed coincident with the EER. Fig. 6 illustrates a linear relationship between the NLOD and EER. From Table 3, one might notice the EERs of FVC2006-1 and HRF-DBI are exceptionally high for some comparators. For instance, EER of Bzorth3 in FVC2006-1 is near 0.5, which means the resolution and image size of this dataset is not enough for the comparator to process. These cases are excluded in Fig. 6.

According to the description on FVC 2004, the datasets were collected with the aim of creating a benchmark more difficult than FVC 2002, in which the top algorithms achieved accuracies close to 100% [12,22]. To this end, more intra-class variation was introduced, with particular emphasis on skin distortion, a well-known difficulty in fingerprint recognition [10]. This fact also coincides with the rank of the NLOD in Fig. 6, where the datasets from the FVC 2004 were ranked very high, indicating that FVC 2004 datasets are relatively difficult compared to the other FVC datasets.

According to Zhao et al., the maximal translations and rotations are, respectively, about 200 pixels and 20°. Hence, the minimal overlap between a finger’s different impressions is about one fourth of the fingerprint image area [31]. The experimental results demonstrated the proposed LOD is applicable in the high resolution fingerprint datasets and the resulting LOD indicates that the difficulty of HRF DBI is much higher than that of HRF DBII. Therefore, the result coincides with Zhao’s description on this dataset.

Fig. 3. The EER versus three types of LOD model over the datasets. (a) The linear LOD model, (b) the polynomial LOD model and (c) the neural network based LOD model.

The LOD is compared with the EER. Four comparison algorithms are utilized as the universal comparators for all datasets. Table 3 compares the median of the NLODs and EERs obtained by the four comparators across the corresponding datasets. Table 3 shows that the NLOD is indeed coincident with the EER. Fig. 6 illustrates a linear relationship between the NLOD and EER. From Table 3, one might notice the EERs of FVC2006-1 and HRF-DBI are exceptionally high for some comparators. For instance, EER of Bzorth3 in FVC2006-1 is near 0.5, which means the resolution and image size of this dataset is not enough for the comparator to process. These cases are excluded in Fig. 6.

According to the description on FVC 2004, the datasets were collected with the aim of creating a benchmark more difficult than FVC 2002, in which the top algorithms achieved accuracies close to 100% [12,22]. To this end, more intra-class variation was introduced, with particular emphasis on skin distortion, a well-known difficulty in fingerprint recognition [10]. This fact also coincides with the rank of the NLOD in Fig. 6, where the datasets from the FVC 2004 were ranked very high, indicating that FVC 2004 datasets are relatively difficult compared to the other FVC datasets.

According to Zhao et al., the maximal translations and rotations are, respectively, about 200 pixels and 20°. Hence, the minimal overlap between a finger’s different impressions is about one fourth of the fingerprint image area [31]. The experimental results demonstrated the proposed LOD is applicable in the high resolution fingerprint datasets and the resulting LOD indicates that the difficulty of HRF DBI is much higher than that of HRF DBII. Therefore, the result coincides with Zhao’s description on this dataset.

The computation of the NLOD is based on the median statistic, thus it reflects an average level of difficulty of the dataset. However, the NLOD could not map to the EER directly and precisely. Some factors might have not been considered influencing this difficulty. From the observation, one factor may be the sensor type. While the proposed measures are independent of the sensor type, the comparators are severely influenced by the sensor type. In the LOD analysis, FVC2000-2 resulted in lower difficulty than FVC2002-2, while on the contrary, the EER of FVC2000-2 was higher than that of FVC2002-2. In fact, FVC2000-2 was collected using a low-cost capacitive sensor and FVC2002-2 was collected using an optical sensor; however, the proposed measures could not describe this difference. This issue is remained as a future work.
Table 1
The cross-validation of the single dataset trained LOD and EER over three types of models and four comparison algorithms. The adjust-$R^2$ of each model-comparator’s combination is computed.

<table>
<thead>
<tr>
<th>Testing (11)</th>
<th>Training (1)</th>
<th>Linear model LOD $= b_0 + b_1CA + b_2DF + b_3RSQ$</th>
<th>Polynomial model LOD $= b_0 + b_1CA^2 + b_2DF^2 + b_3DF + b_4RSQ$</th>
<th>Neural Network 10 hidden layers LOD-VF (%) LOD-B3 (%) LOD-GR (%) LOD-DG (%) LOD-VF (%) LOD-B3 (%) LOD-GR (%) LOD-DG (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER-VF</td>
<td>53.10</td>
<td>60.80</td>
<td>65.90</td>
<td>68.40</td>
</tr>
<tr>
<td>EER-B3</td>
<td>45.60</td>
<td>59.90</td>
<td>79.20</td>
<td>69.20</td>
</tr>
<tr>
<td>EER-GR</td>
<td>54.60</td>
<td>61.30</td>
<td>64.40</td>
<td>67.40</td>
</tr>
<tr>
<td>EER-DG</td>
<td>39.80</td>
<td>47.10</td>
<td>52.80</td>
<td>53.90</td>
</tr>
<tr>
<td>Avg. R-Sq</td>
<td>60.67</td>
<td>60.80</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2
The cross-validation of the multiple datasets trained LOD and EER over three types of models and four comparison algorithms. The adjust-$R^2$ of each model-comparator’s combination is computed.

<table>
<thead>
<tr>
<th>Testing (11)</th>
<th>Training (7)</th>
<th>Linear model LOD $= b_0 + b_1CA + b_2DF + b_3RSQ$</th>
<th>Polynomial model LOD $= b_0 + b_1CA^2 + b_2DF^2 + b_3DF + b_4RSQ$</th>
<th>Neural Network 10 hidden layers LOD-VF (%) LOD-B3 (%) LOD-GR (%) LOD-DG (%) LOD-VF (%) LOD-B3 (%) LOD-GR (%) LOD-DG (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER-VF</td>
<td>60.50</td>
<td>71.40</td>
<td>72.30</td>
<td>70.00</td>
</tr>
<tr>
<td>EER-B3</td>
<td>43.30</td>
<td>45.50</td>
<td>81.10</td>
<td>48.60</td>
</tr>
<tr>
<td>EER-GR</td>
<td>60.50</td>
<td>69.70</td>
<td>69.70</td>
<td>67.70</td>
</tr>
<tr>
<td>EER-DG</td>
<td>46.50</td>
<td>56.60</td>
<td>58.40</td>
<td>54.40</td>
</tr>
<tr>
<td>Avg. R-Sq</td>
<td>63.43</td>
<td>53.39</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In this paper, the animals are analyzed using the proposed relative measures. Yager et al. [29] defined a biometric zoo as following. Let $G$ and $I$ be the set of average genuine and imposter match scores, respectively. Let $G_{H}$ be the 25% of users with the highest genuine statistics, and $G_{L}$ be the 25% of users with lowest genuine statistics. Similarly, let $I_{H}$ be the 25% of users with the highest imposter statistics, and $I_{L}$ be the 25% of users with the lowest imposter statistics. The definitions of the animals are:

- **Worms**: Worms are the worst conceivable users of a biometric, and belong to set $G_{L} \cap I_{H}$.
- **Chameleons**: Chameleons always appear similar to others, receiving high match scores for all verifications, and belong to set $G_{H} \cap I_{L}$.

**Table 3**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Median of CA</th>
<th>Median of DF</th>
<th>Median of RSQ</th>
<th>Median of NLOD</th>
<th>VeriFinger EER</th>
<th>Bozorth3 EER</th>
<th>Griaule EER</th>
<th>Digent EER</th>
<th>Average EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000-DB1</td>
<td>0.72</td>
<td>5.3</td>
<td>81.0</td>
<td>64.6</td>
<td>0.017</td>
<td>0.053</td>
<td>0.046</td>
<td>0.012</td>
<td>0.032</td>
</tr>
<tr>
<td>2000-DB2</td>
<td>0.77</td>
<td>4.9</td>
<td>85.5</td>
<td>62.2</td>
<td>0.006</td>
<td>0.060</td>
<td>0.048</td>
<td>0.007</td>
<td>0.030</td>
</tr>
<tr>
<td>2000-DB3</td>
<td>0.74</td>
<td>5.8</td>
<td>78.1</td>
<td>65.2</td>
<td>0.042</td>
<td>0.076</td>
<td>0.088</td>
<td>0.051</td>
<td>0.064</td>
</tr>
<tr>
<td>2002-DB1</td>
<td>0.71</td>
<td>4.6</td>
<td>82.8</td>
<td>63.7</td>
<td>0.004</td>
<td>0.030</td>
<td>0.015</td>
<td>0.008</td>
<td>0.014</td>
</tr>
<tr>
<td>2002-DB2</td>
<td>0.73</td>
<td>5.0</td>
<td>84.1</td>
<td>63.8</td>
<td>0.004</td>
<td>0.024</td>
<td>0.023</td>
<td>0.008</td>
<td>0.015</td>
</tr>
<tr>
<td>2002-DB3</td>
<td>0.67</td>
<td>4.4</td>
<td>82.0</td>
<td>64.8</td>
<td>0.013</td>
<td>0.099</td>
<td>0.056</td>
<td>0.024</td>
<td>0.048</td>
</tr>
<tr>
<td>2004-DB1</td>
<td>0.69</td>
<td>6.5</td>
<td>81.5</td>
<td>67.8</td>
<td>0.059</td>
<td>0.131</td>
<td>0.101</td>
<td>0.068</td>
<td>0.090</td>
</tr>
<tr>
<td>2004-DB2</td>
<td>0.69</td>
<td>5.4</td>
<td>77.8</td>
<td>66.0</td>
<td>0.040</td>
<td>0.103</td>
<td>0.099</td>
<td>0.033</td>
<td>0.069</td>
</tr>
<tr>
<td>2004-DB3</td>
<td>0.78</td>
<td>7.4</td>
<td>85.9</td>
<td>65.9</td>
<td>0.033</td>
<td>0.066</td>
<td>0.073</td>
<td>0.033</td>
<td>0.051</td>
</tr>
<tr>
<td>2006-DB1</td>
<td>0.80</td>
<td>8.0</td>
<td>42.7</td>
<td>67.4</td>
<td>0.092</td>
<td>0.463</td>
<td>0.153</td>
<td>0.090</td>
<td>0.199</td>
</tr>
<tr>
<td>2006-DB2</td>
<td>0.82</td>
<td>4.8</td>
<td>81.6</td>
<td>60.5</td>
<td>0.002</td>
<td>0.016</td>
<td>0.019</td>
<td>0.008</td>
<td>0.011</td>
</tr>
<tr>
<td>2006-DB3</td>
<td>0.72</td>
<td>6.4</td>
<td>86.6</td>
<td>66.5</td>
<td>0.029</td>
<td>0.075</td>
<td>0.050</td>
<td>0.038</td>
<td>0.048</td>
</tr>
<tr>
<td>HRF-DB1</td>
<td>0.76</td>
<td>6.0</td>
<td>85.5</td>
<td>65.1</td>
<td>0.154</td>
<td>0.181</td>
<td>0.233</td>
<td>0.252</td>
<td>0.205</td>
</tr>
<tr>
<td>HRF-DBII</td>
<td>0.83</td>
<td>4.0</td>
<td>87.6</td>
<td>58.4</td>
<td>0.002</td>
<td>0.039</td>
<td>0.019</td>
<td>0.020</td>
<td>0.020</td>
</tr>
</tbody>
</table>

3.5. Comparing LOD with biometric zoo

In this paper, the animals are analyzed using the proposed relative measures. Yager et al. [29] defined a biometric zoo as following. Let $G$ and $I$ be the set of average genuine and imposter match scores, respectively. Let $G_{H}$ be the 25% of users with the highest genuine statistics, and $G_{L}$ be the 25% of users with lowest genuine statistics. Similarly, let $I_{H}$ be the 25% of users with the highest imposter statistics, and $I_{L}$ be the 25% of users with the lowest imposter statistics. The definitions of the animals are:

- **Worms**: Worms are the worst conceivable users of a biometric, and belong to set $G_{L} \cap I_{H}$.
- **Chameleons**: Chameleons always appear similar to others, receiving high match scores for all verifications, and belong to set $G_{H} \cap I_{L}$. 
Phantoms: Phantoms lead to low match scores regardless of who they are being matched against, themselves or others, and belong to the set \( G_L \cap I_L \).

Doves: Doves are the best possible users in biometric systems. They are defined by the set \( G_H \cap I_L \).

Fig. 6. The relationship between the NLOD and EER. EERs are computed from VeriFinger, Bozorth3, Griaule and Digent.

Fig. 7. The median and confidence interval of NLOD across the animals.
Fig. 7 demonstrates the relation between the proposed measures and a variety of animals. The animals are classified from each dataset (excluding FVC2006-1) utilizing the Bozorth3 algorithm. Sheep is part of Doddington’s zoo [14] which comprises the default population and has been added to the experiment. Considering sheep as a reference, worms and phantoms are relatively difficult, while chameleons and doves are relatively easier. This coincides with the original definition of Yager’s zoo. Furthermore, the animals between different datasets can be compared, which is shown in Fig. 8. For example, the LOD of worms in an easy dataset FVC2000-2 is larger than the LOD of chameleons in a difficult dataset FVC2004-1.

4. Conclusions

This paper presented an effective method for measuring and characterizing the LOD of fingerprint datasets. The proposed method is based on the differential factors between mated pairs, such as the common area, deformation, and relative sample quality. The experimental results showed that a multiple linear regression using the three factors effectively embodies the LOD of fingerprint datasets without actual comparison operations. This method can be used to predict the LOD of a given dataset objectively and independently of the comparison algorithms.

Acknowledgments

This work was supported by the IT R&D program of MOTIE/KEIT. [10039149, Development of Basic Technology of Human Identification and Retrieval at a Distance for Active Video Surveillance Service with Real-time Awareness of Safety Threats].

References


