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

Shengzhe Li, Changlong Jin and Hakil Kim
School of Information and Communication Engineering
Inha University
Incheon, Korea
{szli, cljin}@vision.inha.ac.kr, hikim@inha.ac.kr

Stephen Elliott
Department of Technology, Leadership and Innovation
Purdue University
West Lafayette, USA
elliott@purdue.edu

Abstract—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. Proposed in this paper is a general framework of assessing the level of difficulty of fingerprint datasets based on quantitative measurements of not only the sample quality of individual fingerprints but also relative differences between genuine pairs, such as common area and deformation. The experimental results over multi-year FVC datasets demonstrate that the proposed method can predict the relative difficulty levels of the fingerprint datasets which coincide with the equal error rates produced by two matching algorithms. The proposed framework is independent of matching algorithms and can be performed automatically.

Keywords: Level of difficulty, fingerprint dataset, sample quality, relative quality, common area, deformation

I. 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 evaluation 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.

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.

Many studies have shown that the sample quality (SQ) of a fingerprint strongly impacts on the performance of recognition system [2-6]. The SQ of a fingerprint is considered as 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 matching algorithms to improve recognition performance. However, the quality score 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. Further, the SQ of fingerprint is influenced by the type of a sensor. Therefore, the SQ alone cannot fully classify the difficulty level of a dataset.

As pointed out by Hicklin and Reedy [6], the ability to match fingerprints is dependent on three characteristics: (i) number of fingers, (ii) correspondence between the reference and probe images, and (iii) the quality of both the 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 the inherent minutiae content. Image quality metrics can be used to quantify the quality of the reference and probe images separately. However, the similarity of the two fingerprints is what determines the performance of the matcher. Therefore, characterizing the relative difficulty level of a given dataset needs to include the SQ, CA, and DF of mated pairs in a dataset.

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 SQ, CA and DF of mated pairs and combining these factors into a single score by a multiple linear regression. The analysis of variance (ANOVA) is used to compare the difference of difficulty for datasets, and the Tukey’s HSD method is adopted in conjunction with ANOVA to test statistical significance in difference among the resulted LOD for datasets. The proposed method can be applied for characterizing and measuring the relative difficulty levels of fingerprint datasets used in technology evaluation.

II. THE DEFINITION OF LOD AND ITS MODELLING

The LOD is defined as a relative measure of fingerprint dataset that represents how “challenging” or “stressing” the fingerprint dataset is for recognition compared to other datasets [7]. There are several influential factors [8] 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 ageing and subject cooperation, etc.
However, these factors are difficult to measure and the most of datasets used for performance evaluation are homogeneous in the aspects of sensor type and impression type. Therefore, the rest of the properties can be represented and quantified by a fingerprint SQ score. As aforementioned, the CA and DF are more influential factors when matching two fingerprints. Therefore, the quantitative measuring of LOD is defined as follows:

\[ \text{LOD} \propto f(CA, DF, SQ, \xi) \]  

where \( \xi \) refers to unknown factors. The measurement of relative quality of a mated pair, such as CA and DF as well as relative SQ, and the LOD modeling are defined in the following sub-sections.

A. Common area

Regardless of the matching algorithms, the CA of the mated pair is one of the major factors determining the matching score. In general, a larger CA results in a higher matching score, and conversely, a smaller common area leads to a lower matching score. Therefore, the matching score of the mated pair can be considered to be proportional to the CA between the pair.

Considering the difference of the 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 the union of them rather than the actual common area, as follows:

\[ c = 100 \times \frac{(F_r \cap F_p)}{(F_r \cup F_p)} \]  

where \( F_r \) and \( F_p \) denote the area of the foreground in the reference and the probe fingerprints, respectively. \( c \) ranges from 0 to 100 in units of percentage.

The main steps for computing the CA are summarized as follows:

1) Segment the input fingerprint pair
2) Detect the alignment point in two fingerprints
   a) for non-arch fingerprints, the mated singular points are selected as the alignment points,
   b) for arch fingerprint, the maximum point pair in the angular difference and the orientation certainty level along the symmetry line is selected as alignment points.
3) Translate, rotate, and align the probe to the reference.
4) Count the number of pixels in the overlapped region and compute the ratio between them.

Figure 1 illustrates the CA of a pair of mated fingerprints. When there are multiple pairs of alignment points, the one which has the minimum deformation within the common area can be selected.

B. Deformation

Due to the impression pressure and the softness of the finger, there is always relative deformation between a mated fingerprint pair. Severe deformation generally causes a low matching score even with a large CA. A traditional approach to the quantification of DF is the measurement of 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 orientation field in [10]. If there is no relative deformation between a mated pair after alignment, two orientation fields will 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 pair. The DF between the reference and the probe fingerprints is defined as:

\[ d = \frac{1}{n} \sum_{i \in F_R \cap F_P} \left( \arccos \left( \frac{\cos \theta_i \cdot \cos \theta'_i + \sin \theta_i \cdot \sin \theta'_i}{2} \right) \right) \]  

(3)

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

The process of measuring relative deformation can be summarized as follows:
1) Compute the pixel-level orientation fields of the two fingerprints [10].
2) Align the two orientation fields of the fingerprints and extract the common region (refer to the steps 1-3 in computing CA).
3) Compute the average angle difference between the two orientation fields.

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

C. Relative sample quality

SQ of a fingerprint is known as the most decisive factor of the performance of a fingerprint recognition system. In matching, however, 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 [2] as

\[ q = \sqrt{q_r \cdot q_p} \]  

(4)

where \( q_r \) and \( q_p \) are the sample quality measure defined in [2] of the reference and the probe, respectively.

D. LOD modelling

Because the measures of the CA, DF and RSQ are intended to represent certain aspects of the LOD, it is natural to combine the different measures into one measure of the LOD. To predict the LOD from the multiple measurements, it is assumed that the CA, DF and RSQ have a linear relationship with the LOD. Therefore, a multiple regression model for the LOD of an individual mated pair is defined as

\[ l = \beta_1 c + \beta_2 d + \beta_3 q + \varepsilon \]  

(5)

where \( \beta \)'s are coefficients and \( \varepsilon \) is a random error.
The coefficients $\beta$’s can be estimated using the least square method, where $c, d$, and $q$ are given by the Eq. (2), (3), and (4), and the LOD $I$ is replaced with the matching scores of mated pairs produced by VeriFinger 6.2.

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

The non-synthetic datasets from FVC 2000, 2002 and 2004 [11-13], are used to demonstrate the validity of the proposed method. Each dataset contains 800 fingerprints which are captured from 100 fingers. The LOD calculation is performed only for the genuine mated pair, therefore there are totally 2,800 LOD scores for each dataset that can be used are obtained for each dataset.

A. LOD of individual pairs and its distribution

The histograms of the LOD for 3 datasets are shown as examples in Figure 2. The LOD scores of each dataset show a normal distribution. Apparently a higher LOD value means that the mated pairs have larger common areas and smaller deformations, so the dataset is relatively “easier”, and vice versa. Figure 3 also shows that the mean of the dataset 2000-DB2 is the largest and 2004-DB1 is the smallest. In other words, 2000-DB2 is relatively easier than the other datasets, and 2004-DB1 is more difficult.

The one-way ANOVA and the Tukey’s HSD methods are applied to the LOD scores to inspect whether their difference is significant. Figure 3 shows the results of the ANOVA and grouping using the Tukey’s HSD method. As shown in this figure, datasets that do not share a common letter are significantly different. Therefore, the relative difficulty of any of the two datasets can be compared.

B. Comparing the LOD with the EER

Since the LOD of a dataset is computed from the CA, DF, and RSQ, it is not influenced by the matching algorithms. However, it is desirable that the LOD has a certain monotonical relationship with the matching performance. Therefore, in this experiment, the LOD is compared with the equal error rates (EER) of certain “universal” matching algorithms. Although the EER results of various matching algorithms over FVC datasets have been reported, the performance evaluation has been conducted over multiple years and the participating algorithms are not the same for each competition, thus the reported EERs in different years cannot be compared. Therefore, two well known matching algorithms (VeriFinger 5.0 and Bozorth3) are utilized as the universal matchers for all datasets.

Table 1 compares the ranked average LODs and EERs obtained by the two matchers across the corresponding datasets. The ranks of LODs are categorized into three classes: easy (blue), medium (green), and difficult (yellow). The table shows that the LOD is indeed coincident with EER except the datasets 2000-DB3 and 2002-DB3, which requires further investigation. Figure 4 illustrates a linear relationship between LOD and EER when the LOD is normalized from 0 to 10.

![Figure 3: One-way ANOVA: LOD versus DB_NAME.](image-url)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LOD</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>VeriFinger</td>
</tr>
<tr>
<td>2000-DB2</td>
<td>583.6</td>
<td>0.8214</td>
</tr>
<tr>
<td>2002-DB1</td>
<td>546</td>
<td>0.9286</td>
</tr>
<tr>
<td>2002-DB2</td>
<td>540.1</td>
<td>0.6964</td>
</tr>
<tr>
<td>2000-DB1</td>
<td>529.8</td>
<td>3.4464</td>
</tr>
<tr>
<td>2004-DB3</td>
<td>519.7</td>
<td>3.9821</td>
</tr>
<tr>
<td>2000-DB3</td>
<td>519.7</td>
<td>5.4643</td>
</tr>
<tr>
<td>2002-DB3</td>
<td>507.2</td>
<td>2.9821</td>
</tr>
<tr>
<td>2004-DB2</td>
<td>502.7</td>
<td>5.403</td>
</tr>
<tr>
<td>2004-DB1</td>
<td>471.1</td>
<td>6.625</td>
</tr>
</tbody>
</table>

The color indicates the relative difficulty group which determined by the LOD or EER. Blue: easy group. Green: normal group. Yellow: difficult group.

TABLE I. THE LOD OF DATASET AND CORRESPONDING EER COMPUTED BY MATCHERS

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LOD</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>VeriFinger</td>
</tr>
<tr>
<td>2000-DB2</td>
<td>583.6</td>
<td>0.8214</td>
</tr>
<tr>
<td>2002-DB1</td>
<td>546</td>
<td>0.9286</td>
</tr>
<tr>
<td>2002-DB2</td>
<td>540.1</td>
<td>0.6964</td>
</tr>
<tr>
<td>2000-DB1</td>
<td>529.8</td>
<td>3.4464</td>
</tr>
<tr>
<td>2004-DB3</td>
<td>519.7</td>
<td>3.9821</td>
</tr>
<tr>
<td>2000-DB3</td>
<td>519.7</td>
<td>5.4643</td>
</tr>
<tr>
<td>2002-DB3</td>
<td>507.2</td>
<td>2.9821</td>
</tr>
<tr>
<td>2004-DB2</td>
<td>502.7</td>
<td>5.403</td>
</tr>
<tr>
<td>2004-DB1</td>
<td>471.1</td>
<td>6.625</td>
</tr>
</tbody>
</table>
FVC 2004 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 percent. To this end, more intra-class variation was introduced, with particular emphasis on skin distortion, a well-known difficulty in fingerprint recognition [14]. This fact also coincides with the rank of the LOD in figure 4, where the three datasets from the FVC 2004 have ranked the 5th, 8th, and 9th, indicating that these datasets are relatively difficult compared to the others.

IV. Conclusions

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

References

[7] "ISO/IEC NP 29198 - Biometrics - Characterization and measurement of difficulty for fingerprint databases for technology evaluation ".