A Hierarchical Model for the Evaluation of Biometric Sample Quality

Qian He, Zhenan Sun, Tieniu Tan, Yong Zou
Center for Biometrics Authentication and Testing
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, P.O. Box 2728, Beijing, 100190, P.R. China
{qhe, znsun, tnt}@nlpr.ia.ac.cn; zou1020@hotmail.com

Abstract

The evaluation of biometric sample quality is of great importance in the evaluation of biometric algorithms. In this paper, we propose a novel hierarchical model to compute the sample quality on three levels. This model is developed on the basis of three types of influencing factors: global factors, subjective factors and variable factors. We adopt different strategies to compute the corresponding three level qualities: database level quality, class level quality and image level quality. The database level quality is estimated by experience. Then, we compute the mean value of variable number of normalize genuine scores, the quantiles of which are used to determine the class level quality. On the image level quality evaluation, a novel concept of subset frequency is proposed.

1. Introduction

As all kinds of biometric recognition algorithms spring up, an effective method of evaluation becomes much needed. In the previous work of this area, Phillips et al. [8] defined three basic types of evaluation for biometric systems: technology evaluation, scenario evaluation and operational evaluation. However, no matter what kind of evaluation is to be used, the construction of a proper biometric data set comes first in our consideration.

The data may be collected with different sensors, in different circumstances or by different subjects[2]. The results obtained on different data sets need a benchmark to be compared and thus the study of the data set (e.g. data quality and data size) is necessary. Many efforts have been made on addressing the problem of assessing biometric image quality. Kalka et al.[6] studied the impact of various factors of iris image on performance, but they didn't give an exact method to describe the quality. Yi Chen et al.[3] used 2-D wavelets to compute the local quality of iris image. However, the effectiveness of their method depended on the segmentation performance of the image and was only suitable for image preprocessing but sample quality assessment. Shen et al.[9] applied Gabor filter to identify blocks with clear ridge and valley patterns as good quality block. All the methods mentioned above decided the image quality by the characters of the image. Tabassi et al.[4][10] proposed a novel method based on the measurement of the matching scores, on the definition that the image quality should have some prediction of the performance of the recognition algorithms, to assess fingerprint quality. Fingerprint image quality was defined to be five classes, according to the quartiles of genuine matching score distribution. And then artificial neural network was used to train and classify the images. In[5], Patrick et al. proved that quality measurement was predictable of matching performance and gave Tabassi's method a theoretical support.

Previous quality evaluation algorithms treated every single biometric image in the same way, without considering the hierarchical quality distribution of a database. In our model, we first categorize the influencing factors into three types: global factors, subjective factors and variable factors, by which three levels of qualities are presented. The first level quality is measured by experience while the other two levels of qualities are computed through measurement of genuine scores. Here, we use a novel measurement of the scores and import subset frequency to decide the third level quality for the first time. Our model is tested on fingerprint images and the effects are indicated by the decrease of the error rates and the change of the score distributions.

The remainder of this paper is organized as follows. Section 2 first describes three categories of factors, and then the corresponding three levels of qualities are calculated. In this section, we use novel methods to compute each level qualities. Section 3 provides experimental results and discussions. Section 4 gives our conclusion.

2. Hierarchical Quality Model

The three kinds of factors and corresponding level of qualities of our hierarchical model stated above are depicted in Fig 1.

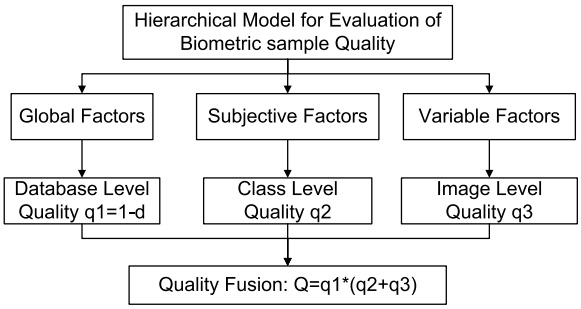


Figure 1. Hierarchical Quality Model

2.1. Influence Factors

According to how much influence it has made, the factors are categorized into three groups, namely, global factors, subjective factors and variable factors. We take the fingerprint images as examples to show the influence of these factors, as is shown in Fig 2 .

The global factors make global influence on the database quality, such as sensor and illumination, as is shown in Fig 2(a) and Fig 2(b).

The subjective factors have close connection with the subjects, which mainly refer to the intrinsic factors[1], as is shown in Fig 2(c) and Fig 2(d).

The relative quality of images collected by one biometric sample is determined by the variable factors. For example, at the beginning of the collection, the subjects are not familiar with the equipment or the collection requirements, so the first three images of each sample usually have bad quality. The images affected by this type of factor are shown in Fig 2(e) and Fig 2(f).

2.2. Database Level Quality

Based on the factors mentioned above, our quality model defines three level qualities, as shown in Fig 1.

The top of the hierarchical model is a difficulty coefficient which is determined by the global factors and its value is set by experience. The global difficulty coefficient is denoted by d on the range [0,1]. Higher value of d means more difficult and lower in quality. We define $q_1(q_1 = 1 - d)$ as the quality of this level.

2.3. Class Level Quality

The middle of the hierarchical model is class level quality which represents the quality differences among classes.

The second level quality will be computed according to the following steps. Here, the database is assumed to have N classes with M images in each class. k is the number of the classes, i and j are the images to be compared, $S^1_{k,i,j}$ and $S^0_{k,i,j}$ denote true and false matching scores respectively.

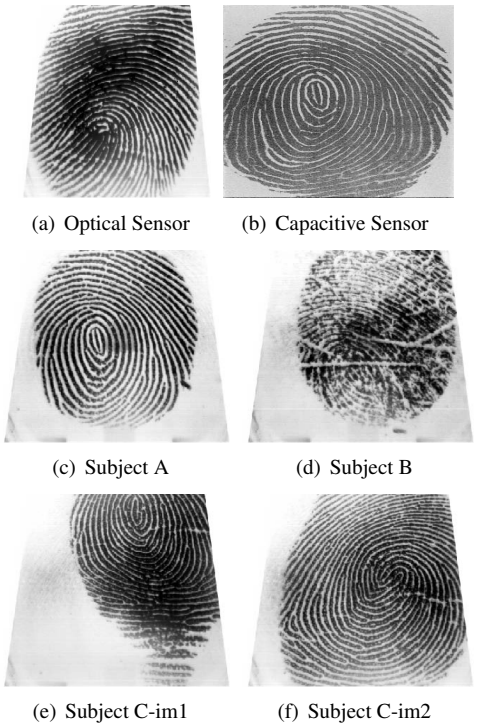


Figure 2. Examples of the images associated with three kinds of factors. (a)(b): images collected by different sensors; (c)(d): images of different fingers collected by the same sensor; (e)(f): images of a single finger collected by the same sensor but in different time periods.

1. Compare every two images in one class to get a genuine score $S^1_{k,i,j}$.

2. For every class, compute the sample mean and standard deviation of its associated U imposter scores. The imposter scores are computed by comparing the first image of the class with the first image of other classes.

$$m_k = U^{-1} \sum_{p=1,p\neq k}^{N} S^0_{p,i,j} \quad (1)$$

$$\sigma_k = (U-1)^{-1} \sum_{p=1,p\neq k}^{N} (S^0_{p,i,j} - m_k)^2 \quad (2)$$

3. Use the statistics computed above to normalize the genuine scores

$$z_{k,i,j} = (S^1_{k,i,j} - m_k)/\sigma_k \quad (3)$$

4. Rank the normalize genuine scores and get the new ones as $\tilde{S}_{k,p}$, where k is the number of the classes, p represents the order on the rank list. Then, compute

the mean of the first R scores on the list.

$$\tilde{\mu}_k = \sum_{p=1}^{R} \tilde{S}_{k,p} \tag{4}$$

5. Compute the empirical cumulative distribution function for $\tilde{\mu}_k$.

$$F(\tilde{\mu}) = \frac{|\tilde{\mu}_k : \tilde{\mu}_k \leq \mu|}{|\tilde{\mu}_k : \tilde{\mu}_k \leq -\infty|} \tag{5}$$

6. Compute the 20 percent, 80 percent quantiles of the distribution of $\tilde{\mu}_k$.

7. Bin $\tilde{\mu}_k$ into three bins based on quantiles of its distribution, as is shown in table 1. This level quality score is denoted by q_2.

In our experiment, 20 percent and 80 percent quantiles are chosen, because middle quality images have the largest number in our database. R equals to $M/2$. All these values can be changed according to the general quality distribution of the database. For example, if more images of bad quality appear in one class, a smaller value should be chosen for R, and if most of the samples are hardly to be recognized, the subset of small matching scores should be much larger.

Table 1. Category of Quality

Quality value	Description	Range of $\tilde{\mu}_k$
0	fair	$-\infty \leq \tilde{\mu}_k \leq F^{-1}(0.2)$
1	good	$F^{-1}(0.2) \leq \tilde{\mu}_k \leq F^{-1}(0.8)$
2	excellent	$F^{-1}(0.8) \leq \tilde{\mu}_k$

2.4. Image Level Quality

The quality of this level is at the bottom of the hierarchical model and is the most important one. It describes the relative quality of the images in one class. As the collection of biometric data is usually under control, most images in the database are of good quality and similar to each other while images of low quality are usually much different from each other. Thus, high matching scores are obtained between two good quality images while low matching scores are obtained in two situations: image of low quality matching with image of low quality; image of low quality matching with image of good quality. A novel method of subset frequency is adopted here to measure the quality of this level. The genuine scores will be binned into three sets, and the image will be thrown to the corresponding bins according to its associated genuine scores. Then, the image will have the quality according to the bin in which it appears the most frequent. The algorithm to compute the quality of this level can be used independently to assign quality to each image in one class. The quantiles can be changed to obtain different range of quality scores.

The specific steps are listed below. We still use the assumption that the database has N classes and each class has M figures. In order to range the final quality into [1,5], three quartiles will be used in our algorithm.

1. For every class compute the genuine scores, $S^1_{k,i,j}$.

2. compute the quartiles of the the genuine scores, qu_1, qu_2, qu_3.

3. Put i, j into one set respectively according to table 2.

4. Each image has the quality of the set which it appears the most frequent. This level quality is denoted by q_3.

Table 2. Three Kinds of Score Sets

Set	Connected Quality	Range of Genuine Scores
B	1	$S^1_{k,i,j} \leq qu_1$
G	2	$qu_1 \leq S^1_{k,i,j} \leq qu_3$
E	3	$qu_3 \leq S^1_{k,i,j}$

2.5. Quality Fusion

Finally, the overall quality is calculated by fusing the three kinds of qualities as follows:

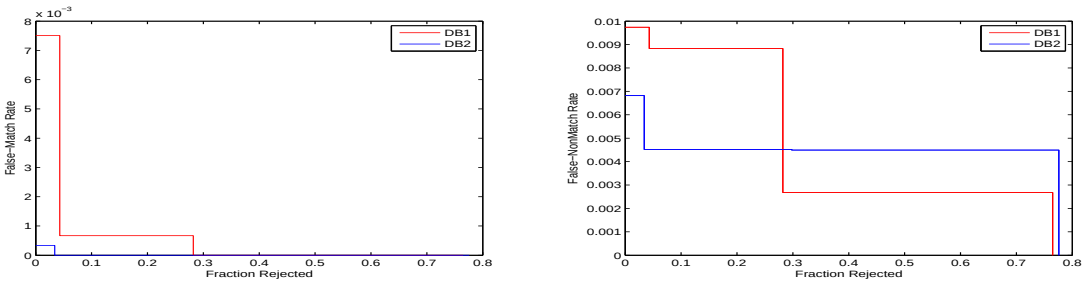
$$Q = q_1 \cdot (q_2 + q_3 = (1 - d) \cdot (q_2 + q_3) \tag{6}$$

Where Q is the overall quality, q_1 represents the database level quality, q_2 represents the class level quality, q_3 represents image level quality, d is the coefficient of database difficulty.

3. Experiment and Analysis

Our quality model is tested on FVC2002[7] DB1 and DB2, which are collected by two types of optical sensor, and each of them contains 8 impressions of 110 fingers with the same image size. The quality indices obtained by our method are tested at the matching stage. In the experiment, d equals to zero. Here, we ignore the global factors, as the two databases have no specific quality differences.

Table 3 shows the the fractions of each level quality in the two databases. We use Kolmogorov Smirnov (KS) test[5] to measure separation of genuine and imposter score distribution. As more images of low quality are removed, a larger KS test statistic is expected. Table 4 shows the KS statistics of the two databases, in which "bad" refers to the original database, "fair" refers to the database without quality 1, "good" refers to the database without quality 1 and 2, and "excellent" refers to the database without quality 1, 2 and 3. Figure 3 shows the decrease of both FMR01 (False Match Rate when False Non-Match Rate is 0.01)



(a) Reduction of FMR01

(b) Reduction of FNMR01

Figure 3. Error versus reject performance for two databases. (a) and (b) show reduction in FNMR01 (FMR=0.01) and FMR01 (FNMR=0.01) as the increase of the fraction of bad quality samples been rejected. The similarity scores come from a commercial matcher.

and FNMR01 (False Non-Match Rate when False Match Rate is 0.01) as more bad quality images are removed. The results show that our model provides an effective quality measurement of biometric samples.

Table 3. Quality Fraction

Quality Value	1	2	3	4	5
DB1	0.043	0.239	0.483	0.207	0.028
DB2	0.034	0.264	0.478	0.198	0.026

Table 4. KS Test for Separation of Genuine and Imposter Scores

samples	bad	fair	good	excellent
DB1	0.989	0.990	0.996	1
DB2	0.992	0.994	0.995	1

4. Conclusion

This paper has presented a hierarchical model to evaluate the quality of biometric images. Our method can be used to evaluate large database directly. The key contribution of our method is that we are the first one to propose a novel hierarchical model to compute the quality. Besides, in terms of the computation of the third level quality, we adopt a new method by using the subset frequency to assess the quality. Each level quality of our model can be used independently. The parameters involved in the model can be adjusted according to situation in order to obtain best quality measurement. Our model can be extended to other modalities as it is based on the measurement of the matching scores instead of specific image characters. Our future work would include exploring exact definition of the score grades for each level that are significantly different from each other and implementing new robust methods to compute each level quality.

Acknowledgement

This work is funded by research grants from the National Basic Research Program of China (2004CB318110), the National Science Foundation (60605014, 60332010, 60335010 and 2004DFA06900). The authors also thank the anonymous reviewers for their valuable comments.

References

[1] Biometric sample quality standard-part1:framework. International Standard ISO/IEC FDIS 19784, 2005.

[2] A.J.Mansfield and J.L.Wayman. Best practices in testing and reporting performance of biometric devices. *NPL Report CMSC*, (14), 2002.

[3] Y. Chen, S. C. Dass, and A. K. Jain. Localized iris image quality using 2-d wavelets. In *International Conference on Biometrics*, pages 373–381, 2006.

[4] E.Tabassi, C.L.Wilson, and C.L.Watson. Fingerprint image quality. In *Nist research report NISTIR 7151*, August 2004.

[5] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Transaction on Pattern Analysis And Machine Intelligence*, 29(4):531–543, 2007.

[6] N. D. Kalka, V. Dorairaj, Y. N. Shah, N. A. Schmid, and B. Cukic. Image quality assessment for iris biometric. In *SPIE Conference on Biometric Technology for Human Identification*, April 2006.

[7] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. Fvc2002: Second fingerprint verification competition. In *International Conference on Pattern Recognition*, pages 811–814, 2002.

[8] P. Phillips, A. Martin, C.L.Wilson, and M. Przybocki. An introdution to evaluating biometric systems. *Computer*, pages 56–63, February 2000.

[9] L. Shen, A. C. Kot, and W. M. Koo. Quality measures of fingerprint images. In *Audio- and Vedio-Based Biometric Person Authentication*, 2001.

[10] E. Tabassi and C. L.Wilson. A novel approach to fingerprint image quality. In *IEEE International Conference on Image Processing*, September 2005.