On the Stability of Ranks to Low Image Quality in Biometric Identification Systems

Emanuela Marasco[1], Ayman Abaza[2,3]

[1] Lane Department of Computer Science and Electrical Engineering,
West Virginia University
PO Box 6109 Morgantown, WV, USA
emanuela.marasco@mail.wvu.edu
[2] West Virginia High Technology Consortium Foundation
Fairmont, WV 26554, USA
[3] Biomedical Engineering and Systems,
Cairo University, Egypt

Abstract. The goal of a biometric identification system is to determine the identity of the input biometric probe. This is accomplished using a matcher which compares the input probe data against each labeled biometric data present in the gallery database. The output is a set of similarity scores that are ranked in decreasing order. The identity of the gallery entry corresponding to the highest similarity score (i.e., rank 1) is associated with that of the probe. In multibiometric systems, the outputs of multiple biometric matchers are combined. Such a combination, or fusion, can be accomplished at the score level or rank level (apart from other levels of fusion). In the literature, rank is believed to be a stable statistic. However, this belief has not been experimentally demonstrated. The contribution of this paper is to investigate the stability of ranks to the image quality degradation in both unimodal and multimodal scenarios. Experiments were carried out using two databases: 1) West Virginia University (WVU) dataset, composed of four fingerprints per subject for 240 subjects, 2) Face and Ocular Challenge Series (FOCS) collection, composed of three frontal faces per subject for 407 subjects. Experimental results show that, in a unimodal scenario when dealing with low quality data, ranks are more stable than scores. However, such a rank stability is not verified when fusing multiple matchers. Experiments demonstrate that, in the presence of low quality data, performance achieved by score-level fusion is better than that one achieved by rank-level fusion.

1 Introduction

In a generic biometric system, operating in identification mode, the input probe (e.g., a fingerprint image) is compared to the labeled biometric data in the gallery database (e.g., fingerprint database) and a set of similarity scores is generated. Scores are sorted in decreasing order and based on this ordering a set of integer values or *ranks* is assigned to these retrieved identities. The lowest rank indicates the best match; hence the corresponding identity is associated with that of the input probe. The identity of the gallery that corresponds to the true identity of the probe is known as the genuine identity; otherwise it is called impostor one.

The recognition accuracy of a biometric system generally decreases in the presence of low quality biometric data wherein the similarity between the probe and associated gallery image may be reduced [1] [2]. It has been observed that, in such a critical scenario, consolidating the evidence provided by multiple biometric sources can increase the recognition accuracy [3] [4]. Evidence can be integrated *before matching*, at sensor or feature level; or, *after matching* at decision, rank or score level [5] [6]. While the amount of information to integrate progressively decreases from the sensor level to the decision level, the degree of noise also decreases [7] [8]. Since match scores are easy to access and combine, score-level fusion has been widely used.

Recent research [9] [10] [11] has established the benefits of rank-level fusion in identification systems. Ranks only carry information about the relative ordering of the different identities in the gallery. However, there are cases where ranks are useful. First, when the output of commercial systems is only a list of candidate identities and no match scores are given [12]. Second, when conducting statistical parametric tests where distributions are assumed to be normal [13]. These tests may fail when considering match scores whose distributions are not normal. Using ranks instead of match scores, can lead to more robust results. Third, when applying monotonous transformation to match scores, the corresponding ranks are kept unchanged. Ranks do not change when the scale on which the corresponding numerical measurements changes [14]. Further, when combining multiple modalities, fusing ranks does not require a normalization phase as typically needed with heterogeneous match scores [15]. Ranks provided by multiple biometric matchers are consolidated and, for each identity in the gallery, a consensus rank is determined [16].

Monwar and Gavrilova presented a Markov chain approach for combining ranks from face, ear and iris [10]. Their experiments showed the superiority in accuracy and reliability over other biometric rank aggregation methods. They reported a rank-1 multimodal identification accuracy of 98.5% compared to the unimodal accuracies of 87%, 92% and 94% for ear, face and iris respectively. However, this improvement may be due to the presence of the iris modality. Abaza and Ross proposed a quality-based Borda Count scheme that is able to increase the robustness of the traditional Borda Count in the presence of low quality images without requiring a training phase [17]. Marasco *et al.* proposed a predictor-based approach to perform a reliable fusion at rank level. The predictor (classifier) was trained using both ranks and match scores and designed to operate before fusion [18]. Results demonstrated its effectiveness in detecting potential unimodal identification errors. An interesting analysis was conducted by Lee [19], who investigated the effect of using rank instead of similarity values when combining multiple evidence by Lee [19]. The study focuses on generating rank-similarity curves where the rank-similarity was computed by applying the function γ to the rank of the i^{th} subject following:

$$\gamma_{(r_i)} = [1 - (r_i - 1)/N] \tag{1}$$

$i = 1 \ldots N$, where N indicates the number of enrolled subjects. The resulting value is used as the similarity value of the subject.

In the presence of low quality biometric data, the genuine match score is claimed to be low and it is expected to be an unreliable individual output, able to confuse a score level fusion algorithm and result in a potential identification error. Conversely, the rank assigned to that genuine identity is expected to remain stable even when using low quality biometric data [20]. However, the stability of ranks has been argued but not experimentally demonstrated. The contribution of this paper is to investigate the stability of ranks in the presence of low quality probes in both unimodal and multimodal scenarios. This paper is organized as follows: Section 2 defines the rank stability. Section 3 presents the approaches for fusion at rank level used to conduct this study. Section 4 describes the technique adopted to synthetically degrade the quality of the fingerprint images and the actual low quality face samples used in our experiments. Section 5 reports results and Section 6 summarizes the conclusions of this work.

2 Rank (and Score) Stability

The concept of stability is introduced here in order to have a method to evaluate the robustness of ranks and scores to low quality data. A biometric system is considered stable when small perturbations of its inputs do not alter its outputs [21]. The stability of ranks can be measured as a function of the difference between the rank assigned to the genuine identity using various low quality probes. A rank difference close to zero indicates that the system is rank stable. In this case the variation of the rank assigned to the genuine identity in the gallery when reducing the quality of the probe is limited. Similarly, the stability of match scores is based on the difference between the score assigned to the genuine identity using a high quality and a low quality probe, respectively. In this work, we consider sources of noisy input data that may arise during the image capture where the image quality can be impacted for example by an incorrect presentation of the biometric sample to the system.

Let $\mathbf{G} = [G_1, G_2, ...G_N]$ be the gallery set, composed by N biometric samples belonging to N different subjects. Given a single probe image, N comparisons of the probe against the gallery are performed and N similarity scores are generated. Let P denote a high quality probe image. Let P' denote the same probe image under degradation. Let s_i and s'_i indicate the score output by the matcher after comparing P and the i^{th} gallery, and P' and the i^{th} gallery, respectively. Let $r[s_i]$ and $r[s'_i]$ indicate the rank assigned to the scores s_i and s'_i, respectively. The score-stability statistic τ_S and rank-stability statistic τ_R can be measured as described in Eqn. (2) and Eqn. (3), respectively.

$$\tau_S = \mathbf{f}(s_i - s'_i) \tag{2}$$

$$\tau_R = \mathbf{f}(r[s_i] - r[s'_j]) \tag{3}$$

Ranks (scores) are stable if the rank (score) assigned to the genuine identity would not change with respect to the probe quality. In other words, a small difference in ranks (scores) between using high and low quality probes indicates high stability, and viceversa. A biometric measure which measures the distance between two distributions is the relative entropy. The entropy measures the amount of information required to describe a random variable; however, as a functional of the distribution of the random variable, it does not depend on the actual values assumed by the random variable but only on the probabilities [22]. In order to measure the stability of ranks, it is important to keep the rank value. For the unimodal case, we develop statistical tests as non-parametric measure to estimate rank (and score) stability over low quality images. Tests based on an assumption of normality, like t-test, are not suitable to measure stability. We return this to the fact that these tests poorly approximates data under study. Also, the Wilcoxon and the Sign tests are not suitable to measure stability, since they assume that the distribution under the Null hypothesis is a standard normal. In this paper, we used the Kendall and the Spearman's rank correlation coefficient, whose inputs are two vectors composed by the ranks assigned to the genuine identity with high and low quality probes. For the later tests, higher correlation coefficient value indicates higher stability.

3 Experimental Results

This section discusses the used datasets and presents experiments to estimate the stability of ranks and scores for both unimodal and multimodal scenarios.

3.1 Datasets

The performance of the proposed strategy was evaluated using two databases. The first database was collected at West Virginia University (WVU). A subset of this database pertaining to the fingerprint (left thumb [FL1], right thumb [FR1], left index [FL2], right index [FR2]) was used [23]. Fingerprint images were collected using an optical sensor. The entire dataset was divided into five sets: one sample of each identity was used to compose the *gallery* and the remaining four samples of each identity were used as *probes* (P_1, P_2, P_3, P_4). VeriFinger[4] software was used for generating the fingerprint scores. Matching scenarios considered the gallery image of high quality and the probe image degraded to simulate low quality ones. The fingerprint image quality was quantified using the IQF software[24]. Degradation effects are simulated using a gray-scale saturation technique which converts fingerprint pixels corresponding to the ridges into background pixels [17]. The gray-scale saturation level (SL) indicates the gray level value above which pixels are saturated to white (255) (see Fig.1). Figure 2 illustrates the unimodal performance when using various levels of the image quality of probes.

[4] http://www.neurotechnology.com/verifinger.html

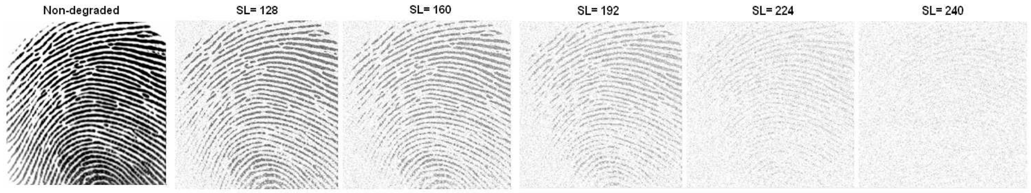


Fig. 1: Examples of low quality fingerprint images artificially degraded by using five different noise saturation levels ST = [128, 160, 192, 224, 240].

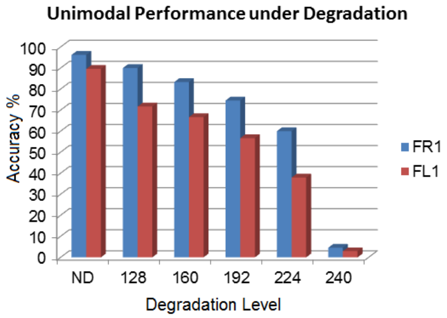


Fig. 2: The unimodal performance decreases when degrading the quality of the probe image at different levels. ND indicates the case with no degradation.

The second database is a subset of the Face and Ocular Challenge Series (FOCS) collection (the Good, Bad and Ugly database) composed by three frontal instances of faces, with two high quality images (from the Good dataset) and one actual low quality image (from the Ugly dataset in which images are taken under uncontrolled illumination, both indoors and outdoors). The partitions of interest are referred to as *Good* and *Ugly*, that have an average identification accuracy of 0.98 and 0.15 respectively [5]. PittPatt[6] software was used for generating the face match scores. Two different matching scenarios were considered: high quality gallery versus high quality probe, referred to as *Good-Good* and high quality gallery versus low quality probe, referred to as *Good-Ugly*. Table 1 provides the details of the database. Fig. 3 shows examples of actual low quality images.

3.2 Results and Discussion

We conducted experiments to estimate the stability of ranks (and match scores) in the presence of low quality input data for both unimodal and multimodal scenarios.

[5] http://www.nist.gov/itl/iad/ig/focs.cfm
[6] http://www.pittpatt.com/

Table 1: Details of the datasets used for the experiments

Database	Biometric	Subjects	Samples	Scores
WVU	Fingerprint (4 fingers)	240	5 per finger	Gen: (1200 × 4) × 4 Imp: (240 × 239 × 25) × 4
FOCS	Face	407	3 per subject	Gen: 407 × 2 Imp: 407 × 406 × 2

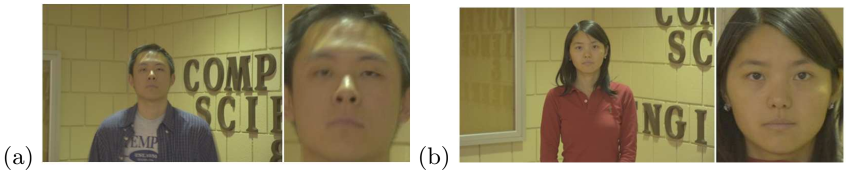
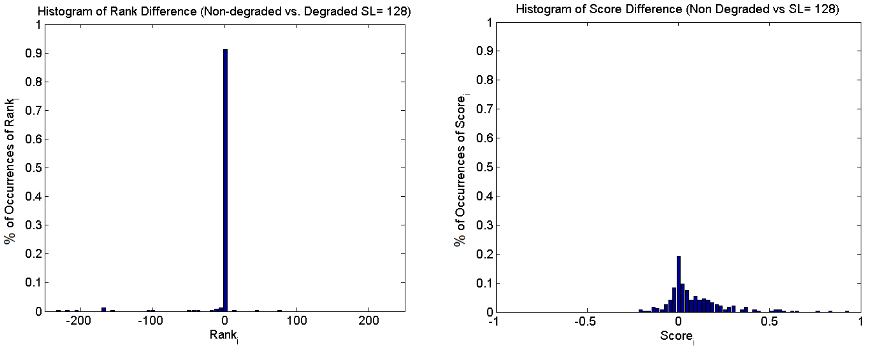


Fig. 3: Examples of face images taken from the Face and Ocular Challenge Series (FOCS) collection: (a) sample image from the Ugly partition; (b) sample image from the Good partition.

Unimodal Rank /Scores Stability. Ranks appears to be more stable than match scores, see Fig. 4 for fingerprints (a similar result is obtained for faces as well). Histograms of the difference between the rank (score) assigned to the genuine identity in the presence of a high and low quality probe image is shown in Fig. 4.

Multimodal Rank /Scores Stability. We integrated ranks in multimodal biometric systems, and compare them to the performance achieved using scores.

Fig. 5 (a) shows the accuracy achieved by rank- and score-level fusion schemes when combining four fingerprints where two are of low quality. In this scenario, the modified highest rank exhibits the best performance among the considered rank level fusion schemes; the achieved rank identification rate decreases from 92.08% to 57.5% when increasing the degradation factor. The score sum performance decreases only from 99.17% to 86.67%. Fig. 5 (b) shows the accuracy achieved when fusing four fingerprints where one is of low quality. The modified highest rank exhibits the best robustness to image quality degradation. It achieves a rank-1 identification rate of 97.08% when the noise saturation level applied to one fingerprint image in every pair is 128 and 85.00% when increasing the noise saturation level to 240. However, the performance of the score sum exceeds that obtained by rank level fusion by achieving a rank-1 identification rate of 99.17% in both non-degraded and degraded conditions. The experiments showed that rank is stable using low quality probes. The correlation coefficient value for ranks is higher than that for match scores, ranks are more stable than scores. When the level of quality degradation is significant, both ranks and scores are not stable (0.418 for ranks and 0.199 for scores).



(a) (b)

Fig. 4: Histograms of the difference between the rank (and the score) assigned to the genuine identity in the gallery before and after degradation of the probe image: (a) Rank difference: Non Degraded vs. degraded with SL= 128; (b) Score difference: Non Degraded vs. degraded with SL= 128.

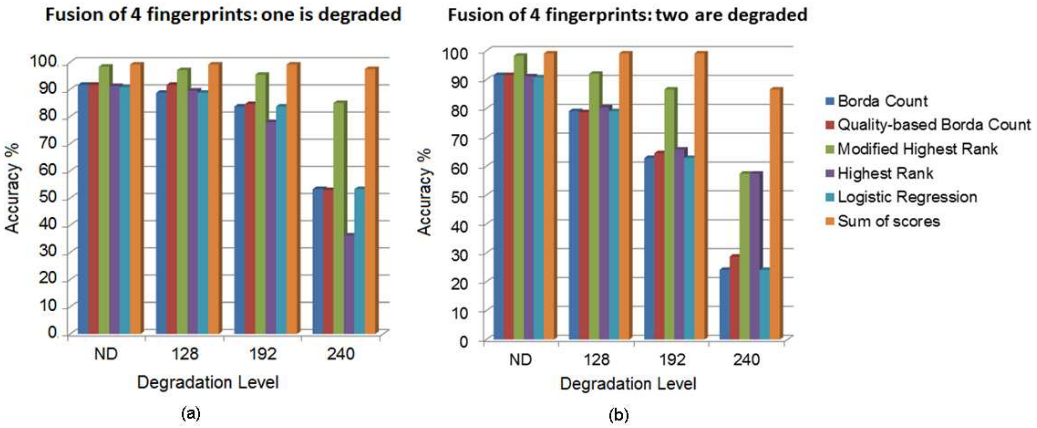


Fig. 5: (a) Fusion of four fingerprints when one of them is degraded: change in performance of different schemes at rank and score under different degradation levels. (b) Fusion of four fingerprints when two of them are degraded: change in performance of different schemes at rank and score under different levels of degradation of two fingerprint probe images.

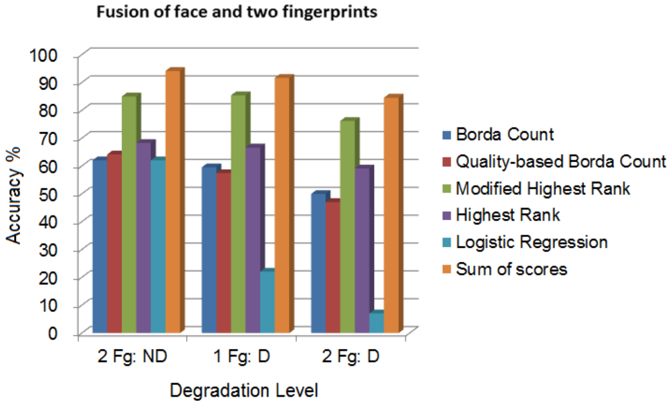


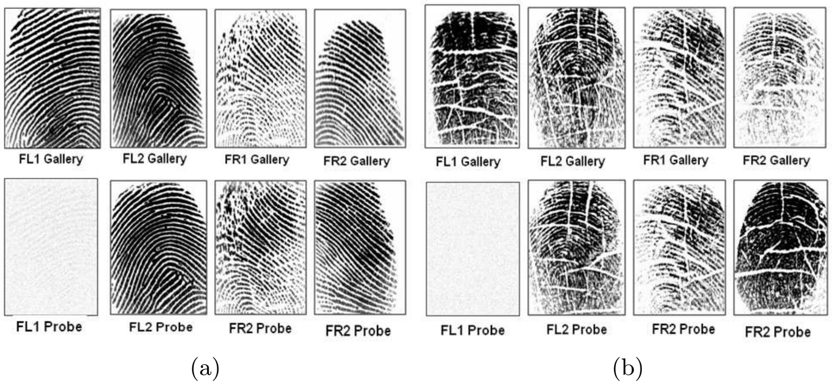
Fig. 6: Fusion of one face and two fingerprints both are of low quality. The face modality is taken from the GBU data set of the FOCS database and both fingerprints from WVU database (FR1 and FL1 fingers).

Fig. 6 illustrates the performance when fusing one face and two fingerprints all of low quality. When combining ranks, the best accuracy is achieved by the Modified Highest Rank. For the traditional Borda Count scheme, the presence of only one incorrect identification where a high rank is assigned to the genuine identity to have a final error. The Highest Rank rule requires that only one of the combined biometric matchers assigns rank-1 to the genuine identity. Errors due to ties are solved with its modified version. When all the combined modalities are all of low quality, the Quality-based Borda Count is the most effective fusion scheme. However, fusion at score level outperforms all the rank-level fusion schemes.

Fig. 7 (a) shows a fusion casework in which the sum of scores and the quality-based Borda Count assign a rank greater than 1, while all the other rank level fusion schemes performs well. In Fig. 7 (b) where only the sum of scores is able to output a correct decision.

4 Conclusion

This study carried out an investigation regarding the stability of the rank in the context of biometrics. Further, we analyzed different non learning-based rank level fusion schemes in the presence of both synthetically degraded fingerprint images and actual low quality face images. The experiments showed that rank is stable with low quality images, when the level of degradation is not significant; While both ranks and scores are not stable, when the level of degradation is significant. Further, ranks are more stable than scores since they present a higher rank correlation coefficient value. (However, the performed study may be



(a) (b)

Fig. 7: (a)Fusion scenario where gallery is of high quality for all the matchers but one of the probe (FL1) is low quality (obtained with SL= 240). (b) Examples of a high quality (ND) and low quality fingerprint images where the low quality probe has been obtained with SL= 240. For this subject, when using the low quality probe, only the sum of scores is able to correctly classify that probe.

dependent upon the matcher used). Conditions under which it is reasonable to use ranks can be expressed as follows:

- When match scores are not available, fusing ranks by applying the modified highest rank scheme leads to the best identification accuracy.
- When match scores are available, a better identification accuracy can be obtained by employing score level fusion.

5 Acknowledgments

The authors are grateful to West Virginia University and, in particular, Dr. Arun Ross. This material is supported by ONR under Contract No. N00014-09-C-0388 awarded to West Virginia High Technology Consortium Foundation.

References

1. M. Monwar and M. Gavrilova. Multimodal biometric system using rank-level fusion approach. *IEEE Transactions on Systems, Man, and Cybernetics*, 39:867–878, August 2009.
2. P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Transaction On Pattern Analysis and Machine Intelligence*, 29(4):531–543, 2007.
3. J. Kittler, Y. P. Li, J. Matas, and M. U. R. Sanchez. Combining evidence in multimodal personal identity recognition systems. *International Conference on Audio- and Video-based Biometric Person Authentication*, pages 327–334, 1997.
4. J. Fierrez-Aguilar, Y. Chen, J. Ortega-Garcia, and A. K. Jain. Incorporating image quality in multi-algorithm fingerprint verification. *International Conference on Biometrics (ICB)*, pages 213–220, January 2006.

5. A. Ross, K. Nandakumar, and A. Jain. *Handbook of MultiBiometrics*. Springer, 2006.
6. J. Kittler, M. Hatef, R. P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.
7. A. Ross, A. Jain, and K. Nandakumar. *Introduction to Biometrics: A Textbook*. Springer, 2011.
8. A. Ross and A. Jain. Multimodal biometrics: an overview. *12th European Signal Processing Conference (EUSIPCO)*, pages 1221–1224, 2004.
9. O. Melnik, Y. Vardi, and C. Zhang. Mixed group ranks: Preference and confidence in classifier combination. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 26(8):973–981, August 2004.
10. M. Monwar and M. Gavrilova. Markov chain model for multimodal biometric rank fusion. *Signal, Image and Video Processing*, pages 1863–1703, 2011.
11. A. Saranli and M. Demirekler. A statistical unified framework for rank-based multiple classifier decision combination. *Pattern Recognition*, 34:865–884, 2001.
12. S. Labovitz. The assignment of numbers to rank order categories. *American Sociological Review*, 35(3):515–524, June 1970.
13. M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, December 1937.
14. F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, December 1945.
15. K. Nandakumar, A. Jain, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285, 2005.
16. A. Ross and A. Jain. Information fusion in biometrics. *Pattern Recognition Letters 24*, pages 2115–2125, 2003.
17. A. Abaza and A. Ross. Quality-based rank level fusion in biometrics. *Third IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–6, September.
18. E. Marasco, A. Ross, and C.Sansone. Predicting identification errors in a multibiometric system based on ranks and scores. *Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–6, September 2010.
19. J. Lee. Combining multiple evidence from different properties of weighting schemes. *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, (9):180–188, 1995.
20. S. Tulyakov and V. Govindaraju. Combining biometric scores in identification systems. pages 1–34, 2006.
21. R. Lempel and S. Moran. Rank-stability and rank-similarity of link-based web ranking algorithms in authority-connected graphs. *Third IEEE International Conference on Biometrics: Theory, Applications and Systems*, September 2004.
22. Andy A. Adler, R. Youmaran, and S. Loyka. Towards a measure of biometric information. *Canadian Conference on Electrical and Computer Engineering.*, pages 210–213, 2006.
23. S. Crihalmeanu, A. Ross, S. Schuckers, and L. Hornak. A protocol for multibiometric data acquisition, storage and dissemination. *Technical Report, West Virginia University*, 2007.
24. N. Nill. Mitre technical report IQF (Image Quality of Fingerprint) software application. 2007.