Exploiting Samples Quality in Evaluating and Improving Performance of Biometric Systems

Hisham Al-Assam, Ali Abboud, Sabah Jassim
Department of Applied Computing, University of Buckingham, United Kingdom

Abstract

Many approaches have been proposed to enhance the performance of biometric-based recognition when using poor quality biometric samples. It has been shown that reporting average accuracy, to cover a wide range of quality of biometric samples, is not enough to reflect the actual performance. This raises the need to evaluate biometric systems at each quality level separately. Therefore, challenging biometric databases have been recorded with variety of quality conditions and made publically available. This paper highlights the importance of using Adaptive Quality-Based Thresholding (AQBT) when evaluating the performance of biometric systems under different quality conditions. Furthermore, it shows that, in many cases, recognition accuracy evaluation reported in the literature under different quality conditions has two problems. First, the performance is evaluated under hidden assumption, which is AQBT. Second, the reported results cannot be achieved in real-life applications. In order to remedy the two problems, two requirements are to be imposed: 1) the matching criteria should be based on AQBT and 2) at the verification stage the quality level of an input biometric sample should be determined and classified into one of a non-overlapping predefined quality levels. Finally, we propose the use of adaptive quality-based feature extraction to enhance the accuracy of biometric systems. Although this paper focuses on face biometric as a case study, the discussion can be equally applied to other biometric treats. We illustrate our ideas by experiments conducted on the extended Yale B face benchmark dataset.

1. Introduction

Automatic face recognition has been investigated extensively in recent decades [1-2], and has many applications in our daily life activities such as identity management cards (e.g., passports, driving license cards). Face recognition remains a challenging task due to factors such as variations in recording condition, pose, and age. Many algorithms have been developed to automate face recognition such as principal component analysis (PCA) [3], and linear discriminate analysis (LDA) [1]. Most of these algorithms achieve significant recognition accuracy (>90%) in controlled environments but their accuracy drops dramatically in uncontrolled environments [4-6]. The reasons for such dramatic decrease in that recognition accuracy vary from extreme variation in illumination to the uncooperative behavior of individuals. In fact, the quality of biometric data has huge impact on the performance of the biometric systems, and therefore has become the focus of research into improving the robustness and efficiency of biometric systems. Quality measures have been used to adapt feature normalization procedures, feature extraction schemes, fusion schemes and template selection schemes. Quality-adaptive approaches help mitigating the effects of low quality biometric samples, and help improving recognition accuracy. However, the way the performance of quality-based adaptive schemes are measured, in terms of average Equal Error Rate (EER) and average Detection Error Trade-off (DET), may give rise to unexpected consequences due to inconsistency in selecting operation point (i.e. threshold). In most biometric applications, e.g. video-based face recognition systems, there are many poor quality biometric samples that have less chance to be accepted by the system [7]. Hence one can argue that using the same threshold for biometric samples of different quality conditions increases FRR of the biometric system as it will be detailed in section 3. On the other hand, Grother and Tabassi [8] raised some problems in measuring biometric accuracy under different quality conditions. They argued that evaluation measure such as DET does not show the effect of the biometric sample quality on the FAR and FRR separately, and does not show the biometric system performance at different quality levels. Therefore, they suggested using quality ranked evaluation to remedy earlier problems of accuracy measures. In fact, it has been shown that assessing the performance of biometric systems at separate quality levels gives more insight into behavior of these systems under different quality conditions [8].

Moreover, such assessment highlights the strength and weakness of different feature extraction and matching algorithms at each quality level. This will eventually lead to building more robust biometric systems by adaptive selection of biometric algorithms.

In this paper, we argue that the high performances reported in the literature for face recognition under different quality conditions have relied on a hidden assumption, and cannot be achieved in practice, unless Adaptive Quality-Based Thresholding (AQBT) at the matching stage and quality ranked evaluation are adopted. The rest of this paper is organized as follows. Existing work is given in section2. The biometric benchmark database and experimental protocol are described in section 3. Section 4 is dedicated for face recognition evaluation under different quality conditions. AQBT is presented in section 5, and section 6 is devoted to conclusions and future work.

2. Existing work

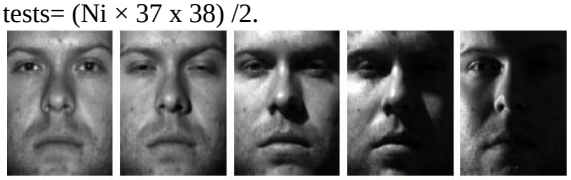
This section focuses on the existing research work in the literature that has been done in the area of face recognition evaluation under different quality conditions. Xiaohua Xie et al [9], evaluated the performance of their developed illumination pre-processing algorithm for CMU PIE and Extended Yale B face databases, and reported (99.8%), (79.4%), (76.1%) and (78.3%) recognition rates for subset2, subset3, subset4 and subset5 of extended Yale B face database respectively. Krizaj et al. [10] developed an enhanced face recognition method based on Scale Invariant Feature Transformation (SIFT) features, and reported recognition rates (100%), (100%), (83.2%) and (82.3%) for subset2, subset3, subset4 and subset5 respectively for the extended Yale B face database. Moreover, they reported different results using different feature extraction methods for the same subsets of the database. Based on LDA, they reported (100%), (99.8%), (56.3%) and (51%) whereas based on the PCA, they reported (93.6%), (55%), (16.7%) and (22%). Many other approaches in the biometric literature [11, 12, 13, 14, 15, and 16] have evaluated the performance of the biometric recognition system in a similar way. In practice, such performances are difficult to achieve because one cannot reproduce the recording conditions associated with these subsets.

Recently some approaches have used quantitative biometric quality measures to adapt biometric recognition schemes whereby the decision threshold is changed dynamically based on the quality of the biometric samples. Dong et al. [7] developed an adaptive approach to dynamically change the decision threshold of iris recognition system based on the quality of iris images to improve the recognition accuracy. Guerra et al. [17] developed a method that adaptively set a threshold for each individual speaker for identification to improve the recognition accuracy. Peter et al. [18], developed a method based on dynamic decision threshold to enhance the performance of fingerprint recognition systems. Although these approaches relied on a dynamic threshold to improve the recognition rate, they did not implement the quality ranked evaluation described in [8] for practical evaluation across different quality levels (section4 highlights the practical importance of quality rank).

3. Experimental protocol and dataset

The extended Yale B face database provides an excellent testing platform for extreme variation in illumination [19]. It consists of still images, in frontal pose, for 38 subjects, each having 64 images captured under different illumination conditions. The total number of images in the database is 2414 images. In addition to these frontal pose images, the database includes an ambient illumination image for each subject but these were not used in our experiments. The database contains five subsets according to the direction of the light-source from the camera axis as shown in Table 1. Samples of images taken from the database are shown in Fig.3, below.

In all experiments, the first three images per user from subset 1 (quality level1 as reported in Table 4 and 5) were selected as reference images to form the gallery set and all the remaining images were used for matching which is based on the Euclidean distances. For a subset (or a quality level) i, if Ni is the number of testing images, then the number of client tests= Ni × 38, and the number of imposter tests= (Ni × 37 x 38) /2.



(a) Subset 1 (b) Subset 2 (c) Subset 3 (d) Subset 4 (e) Subset 5

Figure 1. Sample of images for the same person in different illumination subset

Table 1. Different illumination sets in the extended Yale B database

Subsets	Angles	Number
1	$\theta < 12$	263
2	$20 < \theta < 25$	456
3	$35 < \theta < 50$	455
4	$60 < \theta < 77$	526
5	$85 < \theta < 130$	714

4. Face recognition evaluation under different quality conditions

In this section, we present two sets of experiments. The first set aims to show that the typically reported results of face recognition evaluation based on different illumination subsets are practically inaccurate, and clarify the hidden assumption behind the reported results. It also presents the actual performance of face recognition without AQBT (i.e. fixed decision threshold), and compares it with the reported results by using different illumination subsets. The partitioning is based on arrangement of cameras and lighting sources rather than on a quantitative measure that can be determined from the resulting images. Hence, evaluating the performance of face recognition in each of the five illumination subset separately is laboratory-based evaluation, and cannot reflect reality. The commonly used image illumination quality measure LQI (described later) provides a reasonable base for partitioning the database and practical evaluation of accuracy rate. In the second group of experiments we partition the biometric samples into a number of non-overlapped quality levels and test the performance of the corresponding AQBT. It can be argued that this group of experiments is more realistic and of practical relevance compared to the first set.

To gain more insight into face recognition evaluation, we used in our experiments three Features Extraction Techniques (FETs), namely PCA, LDA, and Discrete Wavelet Transform (DWT). For DWT, we selected the pyramid scheme , which decomposes an image into 3k + 1 subbands (LL_k;HL_k;LH_k;HH_k; : : : ;HL_1;LH_1;HH_1) at a resolution level of k where The LL_k subband is considered as the k-level approximation of I, while HL_k, LH_k, and HH_k captures vertical, horizontal and diagonal features of the image. In all experiments, LH_3 (LH at the third level of decomposition) is used for the DWT.

4.1. Experiment1: Laboratory-based face recognition

Figures (2, 3 and 4) in addition to Tables (2 and 3) show the results of first group of experiments. Figures (2, 3 and 4) show the actual (i.e. practically) authentication accuracy in terms of FAR and FRR based on a fixed decision threshold (Operating Point (OP)) for the three FETs respectively. The OP that gives the EER (FAR=FRR) using subset1 images is selected as a fixed OP, and applied consistently to other subsets (i.e. from subset 2 to subset 5). The three figures illustrate how the FRR increases significantly based on the fixed OP when moving from subset1 to the other subsets (2, 3, 4, and 5). Figure 2, for example, reports FAR= {0, 28.16, 81.34, 99.62, 100} % for the subsets 1,2,3,4, and 5 respectively based on PCA as FET using fixed OP. For authentication purposes, these high FRRs indicate that the possibility of rejecting a genuine individual of low or average quality face images (i.e. subsets (3, 4, and 5)) is high. Nevertheless, for biometric identification applications, such usage of fixed OP in different quality conditions can be exploited by criminals to fool the face recognition systems by manipulating the quality of facial images [20]. The earlier results show that applying the same threshold (i.e. fixed threshold) for different quality conditions is impractical due to the signification increase of FRR when using average and low quality face images.

Table2 presents the actual (i.e. practical) face authentication accuracy in terms of FAR % and FRR % using the three FETs and the five subsets without using AQBT (i.e. fixed OP). It can be seen that the high FRR such as (81%, 99%, and 100 %) for the PCA in the subsets 3, 4, and 5 respectively makes face recognition system impractical using fixed OP, and shows the need for AQBT.

Table 2. Actual results without AQBT (practically accurate) based on the five subsets and the three FETs

FET	subset1		subset2		subset3		subset4		subset5	
	FAR	**FRR**	**FAR**	**FRR**	**FAR**	**FRR**	**FAR**	**FRR**	**FAR**	**FRR**
PCA	0.00	0.0	0.2	28.16	0.03	81.34	0.00	99.62	0.00	100
LDA	0.00	0.0	0.0	1.58	0.00	32.78	0.00	97.56	0.00	99.6
DWT (LH3)	0.00	0.0	0.0	15.79	0.00	58.37	0.00	99.25	0.00	99.5

Table 3 shows experimental evidence on how researchers in the literature typically evaluated the performance of the biometric. It shows authentication accuracy in terms of FAR and FRR evaluated at the five subsets of the extended Yale B face dataset for the three FETs. We argue that the reported results based on non-adaptive OPs, are practically inaccurate. In fact, we argue that these results rely on a hidden assumption. That is presuming the use of AQBT or any other accurate systemic mechanism to change decision threshold adaptively based on the quality of face images. In other words, the decision threshold has been changed implicitly based on the default five illumination subsets. We also show in the second set of experiments that the mentioned assumption is practically inaccurate due to the quality overlap in the five subsets.

Our experimental results indicate that if AQBT is not adopted, false rejection rates becomes very high (always reject) when using low quality face images. In fact this problem could be a source of attack that interferes with the verification through manipulating the recording condition.

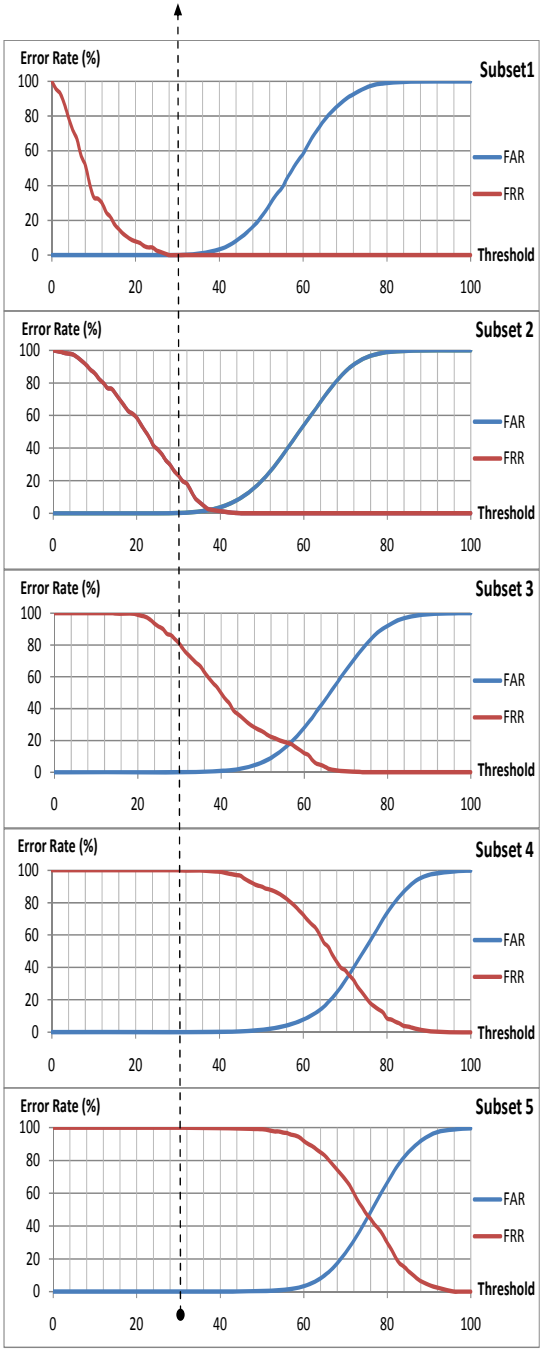


Figure 2. Actual authentication accuracy in terms of FAR and FRR of the five subsets based on PCA

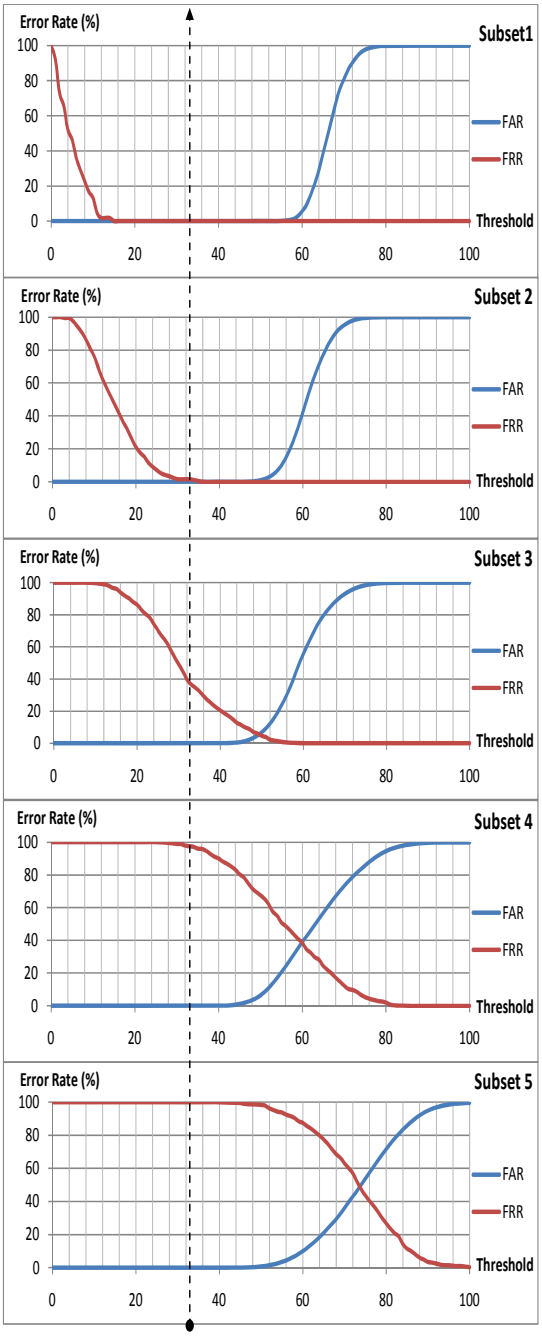


Figure 3. Actual authentication accuracy in terms of FAR and FRR of the five subsets based on LDA

Table 3. Reported results without AQBT (practically accurate) based on the five subsets and the three FETs

FET	subset1		subset2		subset3		subset4		subset5	
	FAR	**FRR**	**FAR**	**FRR**	**FAR**	**FRR**	**FAR**	**FRR**	**FAR**	**FRR**
PCA	0.0	0.0	2.98	2.98	18.2	18.22	33.60	33.6	32.51	32.51
LDA	0.0	0.0	0.00	0.00	3.07	3.07	28.67	28.6	29.25	29.25
DWT (LH3)	0.0	0.0	0.34	0.34	1.65	1.65	7.38	7.3	5.32	5.32

4.2. Experiment 2: Actual Real-life-Based Face Recognition Evaluation

The first group of experiments highlights the importance of using AQBT for achieving good face recognition accuracy across different quality conditions. The main idea of AQBT is to divide face image qualities quality levels so that for each quality level there is a decision threshold (OP) that enable system to achieve the optimal accuracy. In actual real-life face recognition system, at each

authentication or identification attempt, a systemic mechanism is used to quantify the quality of an input face image (quality measure) and then the biometric system assigns this image to a specific quality level, and therefore, the corresponding decision threshold is selected.

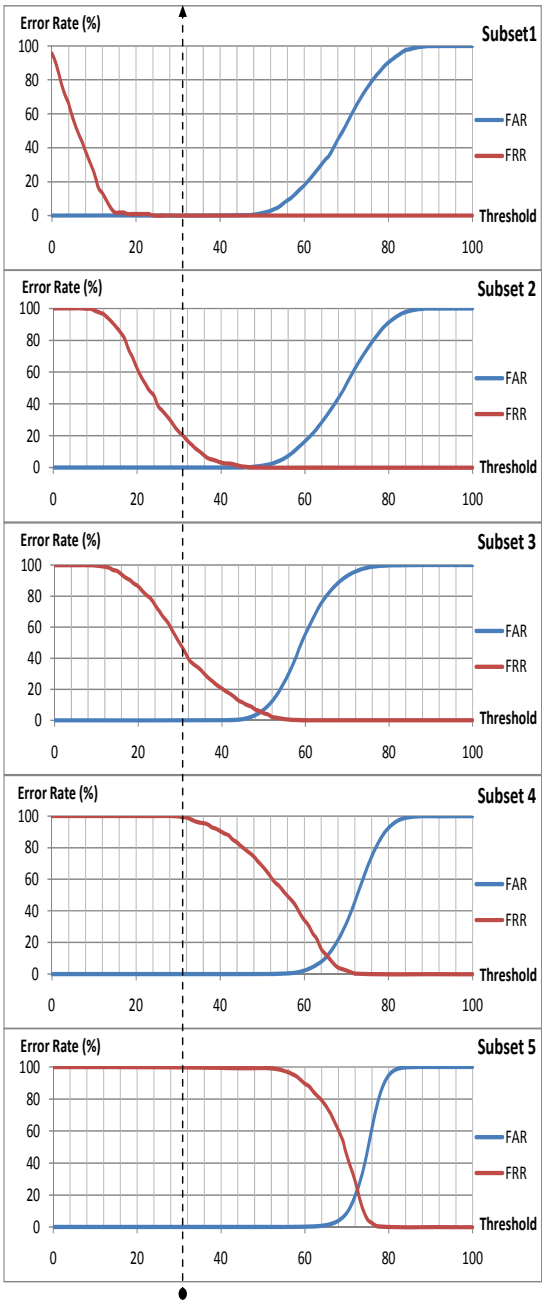


Figure 4. Actual authentication accuracy in terms of FAR and FRR of the five subsets based on DWT (LH_3)

We used the Luminance Quality Index (LQI) [21] to assess the quality of testing images as given in the following equation:

$$LQI = \frac{2\,\bar{x}\,\bar{y}}{(\bar{x})^2 + (\bar{y})^2}$$

Where x, y are the input and reference images respectability, $\bar{x}, \bar{y}$ are the mean of x and y.

In practice, the LQI of an image with respect to another reference image is calculated for each window of size 8x8 pixels in the two images, and the average of these entire blocks defines the LQI of the entire image. Based on Extended Yale B face database, the reference image used to calculate LQI index is the average face image of all 38 individuals (i.e. the average of the frontal pose and in direct illumination image (P00A+000E+00) of each subject) [22].

By measuring the face images quality of the five subsets of the extended Yale using the LQI, we found that the image qualities of the subsets are overlapped and, therefore, it is not feasible to systemically identify the subset of a given image accurately. In other words, AQBT cannot be applied effectively on the predefined subsets of the extended Yale due to the mentioned overlapping. To achieve accurate evaluation of face recognition under different quality conditions, we divided face images of the extended Yale into three well-separated groups (or quality levels) according to their LQIs. The three quality levels are selected because there is a noticeable difference in the performance between these levels compared to the performance using 4 or 5 quality levels [8]. Quality Level 1 (QL1), LQI ∈]0.88 1], represents good quality images, which is roughly similar to subset1 and 2. Quality Level 2 (QL2), LQI∈ [0.75- 0.88], represents medium/average quality images, which is roughly similar to subset 3. And Quality Level 3 (QL3), LQI ∈ [0 0.75[, represents low quality images, which is roughly similar to subset 4 and 5. Table 4 reports the actual authentication accuracy in terms of FAR % and FRR % of the three FETs for each of the three quality levels without using AQBT. It can be noticed that the FRR % is high in QL2 and QL3 due to the use of fixed decision threshold.

Table 4. Actual results without AQBT based the three quality levels and the five subsets and the three FETs

FET	Quality level 1		Quality level 2		Quality level 3	
	FAR	**FRR**	**FAR**	**FRR**	**FAR**	**FRR**
PCA	6.11	6.1	4.52	46.8	0.64	92.2
LDA	0.08	0.1	0.19	31.1	0.05	89.5
DWT (LH3)	2.79	2.8	1.44	9.29	0.21	51.8

Table 5, on the other hand, shows the actual authentication accuracy in terms of FAR % and FRR

% of the three FETs for each of the three quality levels based on AQBT. It can noticed that using AQBR and quality ranks leads to gain better insight into the actual performance of face recognition system under different quality levels (ranks). It is important to mention that in Table 5, the adaptive OP is selected to achieve the EER at each quality level. However, using AQBT and quality ranking offers flexibility in selecting OPs for each quality level, in addition to producing actual and practical results.

Table 5. Actual results using AQBT based the three quality levels and the five subsets and the three FETs

FET	Quality level 1		Quality level 2		Quality level 3	
	FAR	**FRR**	**FAR**	**FRR**	**FAR**	**FRR**
PCA	6.11	6.11	17.86	17.86	34.67	34.67
LDA	0.08	0.08	5.83	5.83	25.58	25.58
DWT (LH3)	2.79	2.79	4.53	4.53	6.98	6.98

It can be noticed from Table 5 that the performance of the three FETs varies across the three quality levels. For example, LDA has the lowest error rates in QL1 while LH_3 outperforms PCA, and LDA in QL2, and QL3.

5. Adaptive quality-based feature extraction

Figure 5 presented to show the performance of FETs in terms of error rate across different quality levels using half total error rate (HTER) performance measure.

$$\text{HTER} = \frac{\text{FAR} + \text{FRR}}{2}$$

We can notice from these results that the performance of these FETs varies with different quality conditions. At quality level 1 (QL1), all FETs have low error rate but LDA has the lowest error rate among them. Also, the performance of DWT is close to the performance of LDA algorithm while PCA is the worst among them in this quality state. At quality level 2 (QL2), LDA and DWT have almost the same performance and PCA has the highest error rate compared with other FETs. In the worst quality condition (quality level 3), the performance of DWT is far better than other FETs (LDA, PCA). This can be attributed to the ability of frequency transforms to provide valuable tools for signal processing and analysis. Frequency information content conveys richer knowledge about features in signals/images that should be exploited to complement the spatial information. To sum up, these results provide strong evidence about the need to use different FETs adaptively based on the quality of input face images however there is a need to do more investigations to prove such concept.

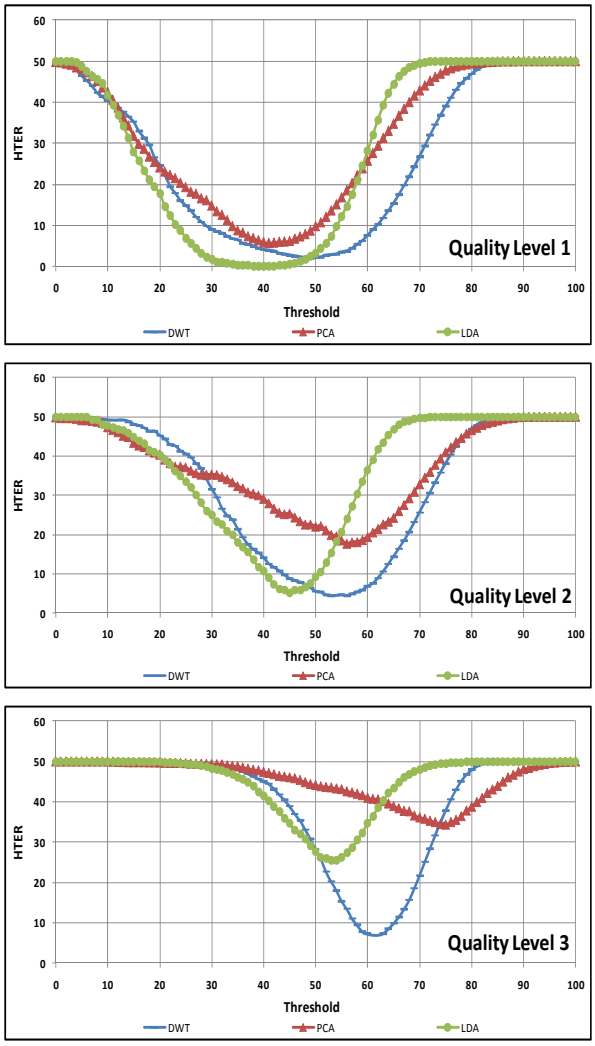


Figure 5. Performance of FETs in terms of error rate (HTER) across different quality levels

6. Conclusion

We have shown that performance evaluations of face recognition under different quality conditions reported in the literature relied on a hidden assumption. we have demonstrated that the presented experimental results of such systems does not have strong relevance in practice and cannot be attained in real-life applications. This could be remedied by imposing two requirements: 1) the system should implement an Adaptive Quality-Based Threshold (AQBT), and 2) the quality level of a fresh input face image should be accurately classified to one of non-overlapped predefined quality levels. It has been shown that the performance of FETs varies across the quality levels using AQBT. Future work will focus on exploiting AQBT applied on different

quality levels to practically enhance the overall performance of biometric systems. This will be done by investigating the fusion of FETs, modeling different facial expression, and adaptively fuse multi-stream biometrics.

7. References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, " Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," in IEEE Trans. PAMI, Special issue on Face Recognition, vol. 19,no. 7, pp. 711–720, 1997.

[2] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, " Face recognition using laplacian faces," in IEEE Trans. PAMI, vol. 27, no. 3, pp. 328–340, March 2005.

[3] Turk, M., Pentland, A., "Eigenfaces for recognition," in journal of Cognitive Neuroscience ,vol.3, no.1,PP. 71– 86, 1991.

[4] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J.Marques, J. Min, and W.Worek, "Overview of the face recognition grand challenge," in: IEEE workshop on Computer Vision and Pattern Recognition (CVPR), 2005.

[5] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: a database for studying face recognition in unconstrained environments," in Tech. Rep. 07-49,University of Massachusetts, Amherst, Mass, USA, October 2007.

[6] J. Ruiz- del- Solar, R. Verschae, M. Correa, "Recognition of Faces in Unconstrained Environments: A Comparative Study," in EURASIP Journal on Advances in Signal Processing (Recent Advances in Biometric Systems: A Signal Processing Perspective), Vol. 2009, Article ID 184617, 19 pages, 2009.

[7] Wenbo Dong; Zhenan Sun; Tieniu Tan; Zhuoshi Wei, "Quality-based dynamic threshold for iris matching,", 16th international conference on image processing (ICIP), pp. 1949 -1952, 2009.

[8] P. Grother and E. Tabassi, "Performance of biometric quality measures," IEEE Trans. Pattern Anal. Mach.Intell., vol. 29, no. 4, pp. 531–543, Apr. 2007.

[9] Xiaohua Xie, Wei-Shi Zheng, Jianhuang Lai, and P.C. Yuen, " Face illumination normalization on large and small scale features," in IEEE CVPR2008., PP. 1–8, June 2008.

[10] Krizaj, J; Struc, V.; Pavesic, N, 'Adaptation of SIFT features for face recognition under varying illumination',MIPRO,IEEE Proceedings of the 33rd International Convention, pp. 691 - 694,2010.

[11] J. Ruiz-del-Solar and J. Quinteros, "Illumination compensation and normalization in eigenspace-based face recognition: a comparative study of different pre-processing approaches," *Pattern Recognition Letters*, vol. 29, no. 14, pp. 1966–1979, 2008.

[12] S. A. Jassim, H.Sellahewa, "A wavelet-based approach to face verification/recognition,", Proc. SPIE, vol. 5986, 2005.

[13] VUCINI E., GOKMEN M., GROLLER M.E., "Face recognition under varying illumination," The 15th International Conferences in Central Europe on Computer Graphics, Visualization and Computer Vision, pp. 57–64, 2007.

[14] Xie, X., Lam, K.-M., "An efficient illumination normalization method for face recognition" Pattern Recognition Lett. 27, pp. 609–617, 2006.

[15] Wang, H., Li, S., Wang, Y., "Face recognition under varying lighting conditions using self quotient image," In Proc. 6th Internat. Conf. on Face and Gesture Recognition, Seoul, Korea, pp. 819–824, May 2004.

[16] Short, J., Kittler, J., Messer, K., "A comparison of photometric normalisation algorithms for face verification," In Proc. sixth Internat. Conf. on Face and Gesture Recognition, Seoul, Korea, pp. 254–259, 2004.

[17] Castilllo-Guerra E., Diaz-Amador R. and Julian C.-B., "Adaptive threshold estimation for speaker verification system,", in Proceeding of Acoustic'08, Paris, 2008.

[18] Peter Z Lo, "Dynamic thresholding for a fingerprint matching system ," US Patent App. 11/031,835, 2005.

[19] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From Few to Many: Generative Models for Recognition under Variable Pose and Illumination," IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(6), pp. 643–660, 2001.

[20] Wein, L., M. M. Baveja, "Using fingerprint image quality to improve the identification performance of the U.S. visitor and migrant status indicator technology program", Proc. National Acad. Sci. 102 (21), pp. 7772 –7775, 2005.

[21] Alan C. Bovik, Zhou Wang, "A Universal Image Quality Index," IEEE Signal Processing Letters, 9(3), pp. 81- 84, 2002.

[22] H. Sellahewa, S. Jassim, "Image-Quality-Based Adaptive Face Recognition," IEEE Trans. on Instrumentation and Measurement 59(4), pp. 805–813, 2010.

