A Method for Selecting and Ranking Quality Metrics for Optimization of Biometric Recognition Systems

Natalia A. Schmid and Francesco Nicolo
Department of Computer Science and Electrical Engineering
West Virginia University, Morgantown, WV 26506
Natalia.Schmid@mail.wvu.edu and fnicolo@mix.wvu.edu

Abstract

In the field of biometrics evaluation of quality of biometric samples has a number of important applications. The main applications include (1) to reject poor quality images during acquisition, (2) to use as enhancement metric, and (3) to apply as a weighting factor in fusion schemes. Since a biometric-based recognition system relies on measures of performance such as matching scores and recognition probability of error, it becomes intuitive that the metrics evaluating biometric sample quality have to be linked to the recognition performance of the system. The goal of this work is to design a method for evaluating and ranking various quality metrics applied to biometric images or signals based on their ability to predict recognition performance of a biometric recognition system. The proposed method involves: (1) Preprocessing algorithm operating on pairs of quality scores and generating relative scores, (2) Adaptive multivariate mapping relating quality scores and measures of recognition performance and (3) Ranking algorithm that selects the best combinations of quality measures. The performance of the method is demonstrated on face and iris biometric data.

1. Introduction

In the field of image and video processing evaluation of quality of images has a number of important applications. These include image acquisition, enhancement, reconstruction, and compression. Image quality metrics designed for these applications are used as figures of merit to quantify degradations or improvements in the images due to various image processing operations [1], [10]. Today it is well understood that (1) selection of appropriate image quality metrics depends on specific applications and (2) image quality measures "should be instrumental in predicting the performance of vision-based algorithms such as feature extraction, image-based measurements, segmentation tasks, etc." (see [1] for details). Rohaly et al. [9] and Corriveau et al. [2] developed a set of attributes that any good objective image (video) quality metric is expected to posses. The main three attributes are prediction accuracy, monotonicity, and consistency.

In all these applications, however, the data after processing, enhancement or compression take the form of images. Therefore, to find a set of image metrics that satisfy the attributes and display reasonable performance is not that hard.

In recent years biometric community began to pay considerable attention to evaluating quality of biometric samples. Biometric images are typical biometric samples. Biometric systems are a type of pattern recognition systems. A typical biometric recognition system operates in two modes, enrollment and authentication or recognition. During the enrollment mode, biometric samples representing different biometric classes are enhanced, processed, encoded and stored in a biometric database. During the recognition mode, a new sample submitted for recognition is preprocessed, enhanced and encoded following the list of procedures used during the enrollment mode. Then, the query encoded data are compared against each entry in the database by involving matching metrics.

The main applications of sample quality in biometric-based recognition systems are (1) to reject poor quality images during acquisition, (2) to use as enhancement metric, and (3) to apply as a weighting factor in fusion schemes. Since a biometric-based recognition system relies on measures of performance such as matching scores and recognition probability of error, it becomes intuitive that the metrics evaluating biometric sample quality have to be linked to the recognition performance of the system. They should be able to predict performance of recognition systems.

In the past, research on biometric sample quality was focused on defining the biometric quality itself. National Institute of Standard and Technology (NIST) organized a number of workshops exclusively devoted to biometric sample quality such as Biometric Quality Workshops (BQW) I and II.

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

The state-of-the art in associating quality of images and signals with the recognition performance of a recognition system (biometric system in particular) does not exist in the form of publications. A few latest ideas were summarized in presentations given at the Multiple Biometric Grand Challenge (MBGC) kick off meeting held in spring 2008. These presentations suggested the use of covariate analysis to relate image/signal quality and recognition performance. These techniques evaluate the covariance or correlation between values of a quality metric and the performance values. Since covariance and correlation are known as second order statistics, they provide only partial (limited) characterization of the relationship between quality metrics and performance.

The absence of a reliable method and a tool for evaluating the effect of a quality metric in predicting performance of a recognition system creates a gap in the base of knowledge on how to use quality metrics and what metrics to use in recognition systems. If a reliable method were available, we would be able to select the best metric among a list of alternatives; rank quality factors in a vector of quality vectors; and use an appropriate quality measure in data fusion schemes.

The goal of this work is to design a method for evaluating and ranking various quality metrics applied to biometric images or signals based on their ability to predict recognition performance of a biometric recognition system.

The main contribution of the paper are three-fold: (1) compared to all previous works it involves pairs of quality measures assigned to a query biometric sample and to a biometric sample from a biometric dataset; (2) it implements multivariate nonlinear mappings to relate vectors of quality pairs, input variables, and the values of the matching metrics (verification and recognition performance), output variables. Two adaptive multivariate mappings, a Feed-Forward Neural Network (FFNN) and Multivariate Regression Analysis Splines (MARS), are used to model the relationship. These models are optimized with respect to the number of nodes, number of hidden layers, and the number and frequency of "Hockey stick" spline functions; (3) prior to performing the nonlinear mapping, in some scenarios it combines each pair of quality measures into a single score. This is a relative quality score. In some scenarios, the relative quality scores are used as additional inputs to nonlinear mappings.

The designed method was tested on iris and face data. Two sets of quality measures were evaluated and ranked in terms of their ability to predict verification performance of two biometric recognition systems. The systems are (1) an iris recognition system implementing Gabor filter-based encoding [8] and relying on iris image quality designed by Kalka et al. [5] and (2) a commercial face recognition algorithm, called FaceIt G6 matching algorithm provided by Identix Inc. and a face quality package FaceIt G6 quality module.

2. Proposed Method

The importance of quality metrics is evaluated based on their ability to predict recognition performance. In many recognition systems, neither physical nor mathematical relationship between image/signal quality measures and measures of recognition performance can be established. In this case, engineers appeal to so-called "black-box approach." A black-box approach does not assume any specific relationship between two or more sets and does not support any physical model describing a relationship.

The block-diagram describing the implementation of the proposed prediction method is displayed in Figure 1. The diagram links the values of the quality measures and the values of the matching metrics of a recognition system. This section will carefully describe the operation of the three blocks displayed in Figure 1.

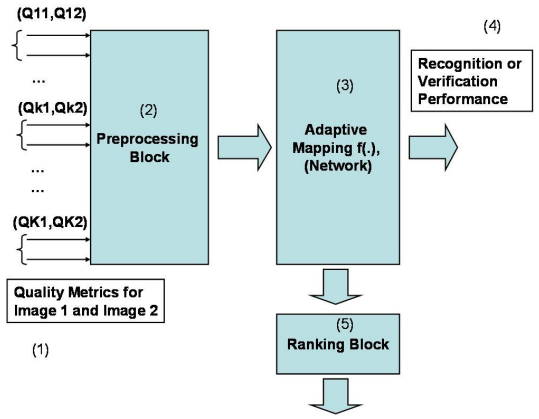

Figure 1. Block-diagram describing the proposed method.

2.1. Input preprocessing block

A distinctive feature of the proposed method is that quality vectors or individual factors come in pairs, (Quality of Image A, Quality of Image B). The goal of Preprocessing Block in Figure 1 is to form a set of relative quality measures in addition to the original quality measures. Preprocessing of quality measures may assume a nonlinear mapping that maps an input quality pair into a single combined or relative quality measure. These combined or relative quality measures can be involved as additional inputs to the multivariate adaptive mapping.

In this work Preprocessing Block operates according to the following three scenarios: (1) Input to Adaptive Mapping is a $2K$ dimensional vector composed of K quality

metrics characterizing image A and K quality metrics characterizing image B. (2) Input to Adaptive Mapping is a K-dimensional vector. Each component is a nonlinear function of a pair (i-th quality measure of biometric sample A, i-th quality metric of biometric sample B). (3) Input to the nonlinear mapping is a $3K$ dimensional vector. The entries include $2K$ dimensional vector from scenario 1 and K dimensional vector from scenario 2.

Denote by $\mathbf{Q}(A) = [Q_1(A), Q_2(A), \ldots, Q_K(A)]^T$ and $\mathbf{Q}(B) = [Q_1(B), Q_2(B), \ldots, Q_K(B)]^T$ two K-dimensional vectors of quality measures for two biometric samples A and B. Two specific nonlinear functions used in scenario 2 include:

$$Q_i(relative) = \tanh\left[\pi \frac{|Q_i(A) - Q_i(B)|}{Q_{\max} - Q_{\min}}\right] \quad (1)$$

and

$$Q(relative)_i = \frac{2}{1 + \exp\left[-\alpha \frac{|Q_i(A) - Q_i(B)|}{Q_{\max} - Q_{\min}}\right]} - \frac{1}{2} \quad (2)$$

where $Q_{\max} - Q_{\min}$ is the range of the values that a quality measure takes. We have explored other relative mappings such as $min(Q_i(A), Q_i(B))$, $|Q_i(A) - Q_i(B)|$, $Q_i(A) + Q_i(B)$, $Q_i(A) \times Q_i(B)$ and various compositions of these functions with other non linear mappings. Our findings are such that functions involving the difference $|Q_iA) - Q_i(B)|$ consistently improve prediction performance. For simplicity of implementation Preprocessing Block applies the same transformation to every pair of quality scores.

2.2. Multivariate adaptive mapping

The adaptive mapping shown as the second block in Figure 1 requires training and testing. The training of the mapping block is reduced to estimating a multivariate function relating quality metrics and recognition performance. Recognition performance is in the form of a distance measure or similarity measure defined on a pair of templates, processed and encoded biometric data. If two templates characterize the same biometric class, the pair is genuine and the matching score is genuine. If two templates characterize two different biometric classes, the pair is imposter and the matching score is imposter.

After the adaptive mapping is trained, that is, the mapping function is estimated, the testing of the mapping block is performed by feeding testing data (quality measures) in the nonlinear mapping block and predicting matching scores. Note that training and testing data do not overlap. Also note that the sets of imposter and genuine pairs have to be processed separately. If these two types of data are combined, the estimated mapping looses structure.

Mathematically, the modeling problem is stated as a multivariate regression problem. A general model can be expressed as

$$\hat{Y} = f(\mathbf{X}), \quad (3)$$

where $\hat{Y}$ is the dependent variable, in this case the matching score (recognition performance), and "hat" stands for an approximation to the output Y, $\mathbf{X}$ is a vector of predictive variables, such as pairs of quality measures for biometric samples A and B.

2.2.1 Multivariate Adaptive Regression Splines (MARS)

The MARS model employs a special set of spline functions called "hockey stick" basis functions. These two-sided truncated functions map variable X to a new variable X^* according to $X^* = (X - t)_+$ or $X^* = (t - X)_+$ (a flipped copy), where t is a knot of the basis function.

Suppose that for each input variable X_i, $i = 1, \ldots, K$ there are N observed values $\{x_{i,j} | j = 1, \ldots, N\}$. A pair of basis functions is knotted at each of the observed values and linked as a reflected pair. These $2NK$ basis functions form an initial collection of basis functions $\mathcal{B}$,

$$\mathcal{B} = \{(X_i - t)_+, (t - X_i)_+ | t \in \{x_{i,1}, \ldots, x_{i,N}\}; i = 1, \ldots, K\}.$$

MARS uses the combination of basis functions to approximate model (3). Let $\hat{f}_M(\cdot)$ be a MARS approximation to (3), that is,

$$\hat{f}_M(\mathbf{X}) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(\mathbf{X}), \quad (4)$$

where M is the number of basis functions, $B_m(\mathbf{X})$ is the mth basis function which is either a function in the collection $\mathcal{B}$ or a product of two or more such functions. Given a choice for $B_m(\mathbf{X})$, the coefficients $\{\beta_m | m = 0, \ldots, M\}$ are estimated by minimizing the sum of squared residuals. Basis functions can be highly nonlinear functions of $\mathbf{X}$, but the mapping $\hat{Y} = \hat{f}(\mathbf{X})$ is a linear function of the basis functions. By analogy with $\hat{f}_M(\cdot)$, $\hat{Y}$ is an approximation to the output Y.

The advantage of MARS is in its ability to estimate in an adaptive fashion the location and number of basis functions to guarantee local and global fit of approximation function $\hat{f}_M(\cdot)$ into a set of output measurements. The output points that undergo abrupt changes in their values are described by a large number of closely spaced basis functions to achieve good fit. Conversely, if the output function is smooth and slowly varies within some regions of support, the number of supporting basis functions selected by the MARS is small and sparsely located. In general, MARS trades off complexity of the model and the accuracy of representation, which makes the approach economical and more robust.

MARS operates in two steps: forward selection and backward deletion. In the first step, MARS selects a pair of basis functions which fit the model best at the current stage. To prevent the final model from being overfitted, a backward deletion step is processed to prune basis functions. A modified form of the generalized cross validation criterion is used as the lack-of-fit criterion. As a result of these operations, MARS automatically determines the most important independent variables as well as the most significant interactions among them. Further details on MARS modeling are given in [3].

2.2.2 Neural Network

A Feed Forward Neural Network (FFNN) is selected as a nonlinear mapping [4, 6]. The topology of the network is established by adopting the classical trial and error approach. Starting from a small size network parameterized by a few links and neurons, the network is allowed to grow until a desired value of the Mean Square Error (cost function) is attained. This operation is performed using training data.

The final design is achieved by trading off the complexity and the performance of the network. The analysis of the designed FFNN using different biometric data have shown that a single hidden layer is sufficient to describe the non linear relationship between quality scores and matching scores. Also, for these data, the design based on sigmoidal neurons achieve better performance than the design based on non linear $tanh$ functions. The number of neurons required to approximate the nonlinear mapping between input and output biometric data depends on the type of biometric data and is also distinct for genuine and imposter cases.

The optimization of FFNN is based on Levenberg-Marquardt back-propagation algorithm [7] used to train the adaptive network and to obtain the network weights. The validation process is repeated L times, each time generating a different surrogate of the data by permutating the order of data. An ensemble of L models is obtained, and the final non linear model is selected as a weighted average of L individual functions. This procedure improves validation results for all considered biometric data. The parameter L is optimized by repeating the procedure several times. The L averaging weights are chosen according to $w_i = R_i^2 / \sum_{j=1}^{L} R_j^2, \;\; i = 1, \ldots, L$, where R_i^2 is square correlation between predicted values and actual values in the training set and $\sum_{i=1}^{L} w_i = 1$.

The drawback of this method is in the requirement of long training time. However, in all experiments a single validation stopped at about 30 epochs. This makes the actual training process computationally feasible.

2.3. Performance measures

The goodness of fit between the predicted matching scores and the measured matching scores was evaluated using three criteria (1) the mean square error (MSE), (2) the square of correlation coefficients (R^2), and (3) F statistic. The two equations below are the mathematical definitions of the first and second criteria:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}_i - y_i \right)^2, \tag{5}$$

$$R^2 = \left(\frac{\sum_{i=1}^{N} \left(y_i - \bar{y} \right) \left(\hat{y}_i - \bar{\hat{y}} \right)}{(N-1) s_y s_{\hat{y}}} \right)^2, \tag{6}$$

where $\hat{y}_i$ is the estimated matching score and $\bar{\hat{y}}$ is the sample mean of the estimated matching scores, s_y and $s_{\hat{y}}$ are the sample standard deviations of the measured matching scores and the estimated matching scores, and N is the number of samples.

The F statistic evaluates the ratio between the mean square error of the linear model and the mean square error of the prediction error. In this paper the F statistic is used as an absolute measure.

There are many other statistical measures (such as receiver operating characteristic, Kullback-Leibler distance, etc.) and goodness of fit tests that can be applied to analyze the data. However, the authors found the traditional measures such as R^2, MSE, or F-statistic to be quite informative and intuitive. Apart from this, they are easy to compute.

2.4. Ranking block

The current version of the ranking block performs exhaustive search of the best possible combination of quality measures submitted as an input to Adaptive Mapping. The ranking is based on the quality measures described in the previous section. If K is the length of the vector of quality measures, the number of possible combinations is $(2^K - 1)$. If the vector of quality measures is long, then a more efficient algorithm (for instance, brunch and bound) can be designed to rank individual and combinations of quality measures.

3. Numerical Results

3.1. Iris dataset and experimental results

This section presents the results of analysis performed on West Virginia University (WVU) non-ideal iris dataset. This dataset is composed of 350 classes for a total of 2,413 images. Iris images were processed and encoded using Libor Masek's Matlab adaptation [8] of John Daugman's algorithm. Encoded images in the form of binary IrisCodes

were further pairwise compared by using Hamming distance. Since images were labeled (iris images were assigned iris classes), the matching scores were labeled too. The scores formed using IrisCodes from the same iris class are genuine scores. The scores formed using IrisCodes from two different iris classes are imposter scores. Processing this dataset returned 8,806 genuine matching scores and 2,889,848 imposter scores. Since the number of imposter scores is redundant, the set of imposter scores was subsampled. A representative subset of 109,000 imposter scores was formed.

The computation of the quality of iris images followed the work by Kalka et al. [5]. It suggests six individual quality measures: (1) Motion blur, (2) Defocus, (3) Illumination, (4) Occlusion, (5) Specular Reflection, and (6) Pixel Count and a combined quality measure due to Dempster-Shafer combination rule. Therefore the vectors of quality measures used in this experiment are composed of six components. Since the adaptive multivariate mapping that maps individual and vectors of quality measures into matching scores has to be trained and tested, the authors chose to use half of genuine input-output pairs (4,403 pairs) to train the mappings. The other (nonoverlapping half) was used for validating the performance of the proposed method. For the imposter case, $9,000$ input-output pairs were used for training and $100,000$ for testing.

To perform numerical analysis, the authors used Neural Net Matlab Toolboox and polspline package of R Software. The polspline package provides a non commercial adaptation of Friedman's algorithm (few differences exist).

To evaluate the prediction ability of quality measures and find their best combination, the authors varied the number of quality measures on the input to the adaptive multivariate mapping. Both individual quality measures and their possible combinations were considered. For each combination, a number of performance measures were evaluated (see Sec. 2.3 for details).

Table 1 presents the results of performance evaluation obtained using FFNN. The processing block implemented the prediction scenario 1. The first six rows in the table describe the ability of individual quality measures to predict the performance of the considered iris recognition system. Factor 7 is the combined quality measure due to Dempster-Shafer rule. The following rows in the table present the best combination of two, three, four and five quality measures. The last row describes the case when all quality measures were used.

From the analysis of the data one may conclude that the best individual quality measures are Factor 6 (pixel count) and 4 (occlusion). The R^2 for the two factors are 0.20 and 0.19, respectively. The best pair of quality measures is the pair of Factors 4 and 5 (occlusion and specular reflections) with $R^2 = 0.27$. Note that as the number of involved quality Factors increases, their ability to predict performance of iris recognition system improves. Note that Dempster-Shafer metric (Factor 7) when used as a single factor performs worse than Factor 6 or Factor 4. This result indicates that vectors of quality measures are considerably more informative compared to the Dempster-Shafer score in terms of their ability to predict recognition performance.

A similar set of experiments was performed using MARS as a multivariate adaptive mapping. The results are summarized in Table 2. The results and conclusions are very similar to the results and conclusions above. Note that the performance of MARS is slightly inferior to the performance of the FFNN.

Table 1. Performance of prediction scenario 1 for genuine iris set. The results are obtained with a single hidden layer FFNN composed of 10 neurons.

Factors	R^2	MSE	F
1	0.0708	0.0049	335.40
2	0.0533	0.0050	247.95
3	0.1370	0.0046	698.69
4	0.1956	0.0043	1070.63
5	0.1643	0.0044	865.50
6	0.2046	0.0042	1132.18
7	0.1841	0.0043	993.32
4,5	0.2699	0.0039	1627.62
1,3,4	0.3496	0.0034	2366.55
1,3,4,6	0.3617	0.0034	2494.60
1,3,4,5,6	0.3857	0.0033	2763.62
1,2,3,4,5,6	0.3948	0.0032	2871.72

Table 2. Performance of prediction scenario 1 for genuine iris set. The results are obtained using MARS.

Factors	R^2	MSE	F
1	0.0219	0.0052	98.95
2	0.0314	0.0052	142.86
3	0.1188	0.0047	593.67
4	0.1563	0.0045	815.69
5	0.1245	0.0046	626.11
6	0.1663	0.0044	878.16
7	0.1727	0.0044	918.94
4,5	0.2043	0.0042	1130.35
1,3,4	0.2949	0.0037	1841.45
1,3,4,6	0.3084	0.0037	1963.20
1,3,4,5,6	0.3265	0.0036	2134.12
1,2,3,4,5,6	0.3091	0.0036	1969.34

Figures 2-5 show the scatter plots of the predicted matching scores versus the measured matching scores for genuine and imposter scores, separately. In addition to the measures of performance listed in Tables 1 and 2, the scatter plots provide information for subjective (visual) evaluation as well

as objective measures such as bias and slope of a straight line fitted into the data. These measures provide additional information about correlation between predicted and measured matching scores.

Note that the results related to the ability of quality metrics to predict imposter matching scores is less intuitive. Overall, the scatter plots using imposter scores are very compact in the case of iris. Both the range of predicted matching scores and measured matching scores is narrow. They are in a relatively good agreement, apart from a few outliers.

The results summarized in the tables are for prediction scenario 1 only. The authors evaluated both scenario 2 and 3 and concluded that scenario 3 always outperforms scenario 1.

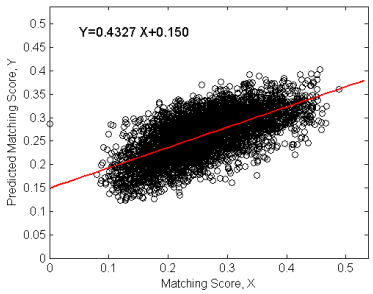

Figure 2. Scatter plot of predicted matching scores versus measured matching scores for the genuine case. The results are obtained with a single hidden layer FFNN composed of 10 neurons. The plot is obtained following prediction scenario 3 and involving all quality measures (six quality measures per iris image) as an input. The equation of regression line (red line) is provided.

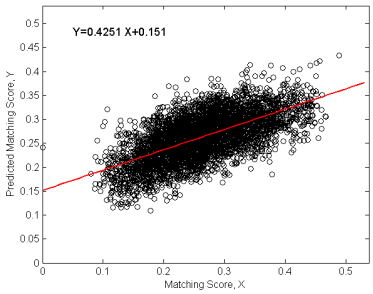

Figure 3. Scatter plot of predicted matching scores versus actual matching scores for the genuine case. The results are obtained using MARS. The plot is obtained following prediction scenario 3 and involving all quality measures per iris image.

3.2. Face dataset and experimental results

WVU Face dataset is composed of 1,745 face images for a total of 270 biometric classes. The FaceIt G6 matching algorithm was applied to process and encode each image in the set. This resulted in 6,074 genuine comparisons and 1,515,566 imposter comparisons. Only a representative

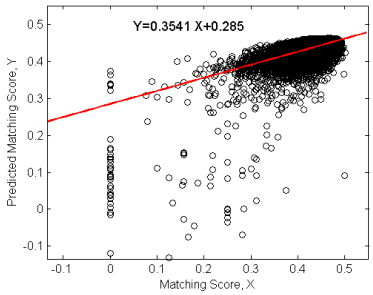

Figure 4. Scatter plot of predicted matching scores versus actual matching scores for the imposter case. The results are obtained with a single hidden layer FFNN composed of 12 neurons and involving quality factors 1,4,6 and prediction scenario 1. The equation of the fitted regression line (red line) is provided.

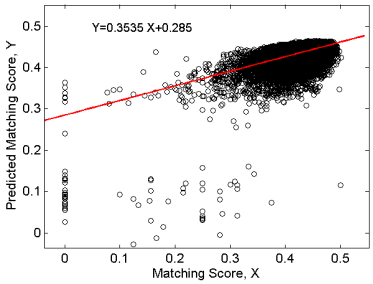

Figure 5. Scatter plot of predicted matching scores versus actual matching scores for the imposter case. The results are obtained using MARS and involving quality factors 1,4,6 and prediction scenario 1.

subset of 109,000 imposter scores is used to evaluate the ability of various face quality measures to predict performance of the face recognition system. The quality of face images are estimated by using FaceIt G6 quality module. The module outputs eleven quality scores per image: (1) Darkness, (2) Brightness, (3) Exposure, (4) Focus, (5) Resolution, (6) Cropping, (7) Glare, (8) Faceness, (9) Contrast, (10) Texture, and (11) Overall quality factors obtained as minimum of a selected combination of quality factors. Half of genuine scores is chosen as a training set and remaining half as a testing set. In the imposter case, we use 9,000 scores for training purpose and remaining 100,000 scores for testing.

The results in Table 3 summarize prediction performance of the FFNN with a single hidden layer composed of 10 neurons applied to genuine testing data. Note that the individual factors such as Factor 3 (exposure) or Factor 5 (Resolution) considerably outperform the Overall metric. Their R^2 are 0.27, 0.22, and 0.18, respectively. The vectors of 9 and 10 scores predict performance of the recognition system based on FaceIt best.

Table 4 presents similar results when FFNN is replaced by MARS.

Figures 6-9 show the scatter plots of the predicted match-

Table 3. Performance of prediction scenario 1 using genuine face matching scores. The results are obtained using FFNN with a single hidden layer composed of 10 neurons.

Factors	R^2	MSE	F
1	0.1973	510.95	746.36
2	0.0834	598.47	276.30
3	0.2751	458.11	1151.95
4	0.0947	574.88	317.70
5	0.2212	493.05	862.12
6	0.0038	631.05	11.69
7	0.0011	632.46	3.46
8	0.1433	546.76	507.81
9	0.0401	606.92	126.93
10	0.0207	621.30	64.16
11	0.1834	521.75	682.00
1,2	0.2951	443.33	1270.82
1,3,8	0.3773	392.08	1839.20
1,2,3,4	0.3564	404.92	1680.77
3,4,5,6,7,8	0.4992	318.89	3026.14
2,5,6,7,8,9,10	0.3946	381.62	1978.24
1 through 9	0.5177	304.48	3258.52
1 through 10	0.5100	310.55	3159.23

Table 4. Performance of prediction scheme (1) for genuine face matching scores. The results are obtained with MARS.

Factors	R^2	MSE	F
1	0.1072	565.63	364.76
2	0.0755	603.35	248.03
3	0.1332	548.69	466.45
4	0.0801	581.93	264.29
5	0.1304	549.27	455.25
6	0.0000	633.11	x
7	0.0000	633.11	x
8	0.0793	581.89	261.67
9	0.0000	633.11	x
10	0.0178	621.90	55.07
11	0.1005	568.54	339.30
1,2	0.1450	550.64	514.81
1,3,8	0.1771	517.59	653.25
1,2,3,4	0.2028	516.41	772.29
3,4,5,6,7,8	0.1653	526.65	601.27
2,5,6,7,8,9,10	0.2034	522.84	775.11
1 through 9	0.3214	426.54	1437.96
1 through 10	0.3255	424.22	1465.26

ing scores as a function of measured matching scores for the face biometrics. Note the compactness of both predicted and measured imposter scores. The range of values they take is very small. They are clustered in the positive quadrant around zero. Therefore, the plots involving imposter matching scores are not informative. Alternatively, the plots based on genuine scores provide additional information that is not included in the tables. Again scenario 3 outperforms scenario 1. Based on the plots and tables, the quality measures due to FaceIT G6 have much higher ability to predict recognition performance compared to the ability of iris quality measures described in [5] to predict performance of iris recognition system based on IrisCode.

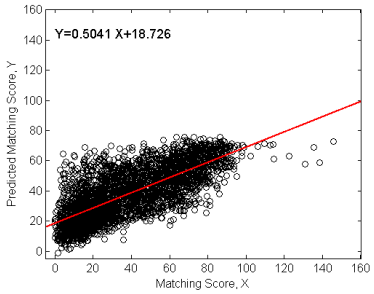

Figure 6. Scatter plot of predicted face matching scores versus actual matching scores in the genuine case. The results are obtained with a one hidden layer FFNN with 10 neurons and involving all quality factors and prediction scheme (3).

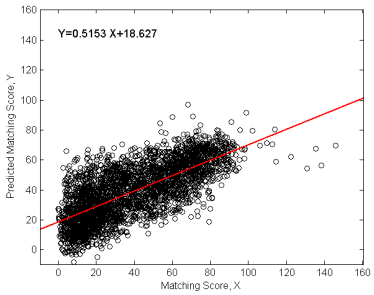

Figure 7. Scatter plot of predicted face matching scores versus actual matching scores in the genuine case. The results are obtained with MARS and involving all quality factors and prediction scheme (3).

4. Conclusions

This paper presents a method for selecting and ranking quality measures designed to operate on raw biometric samples (in the form of images and signals). Since the relationship between quality measures and recognition performance (in the form of matching scores and probability of recognition error) is highly nonlinear, the authors approximate this relationship using adaptive multivariate mappings. MARS and FFNN were selected and optimized for this purpose.

The designed methods operate on pairs of quality values that are produced when a quality measure is applied to a query biometric sample and to a selected enrolled sample. Three scenarios for preprocessing quality pairs are de-

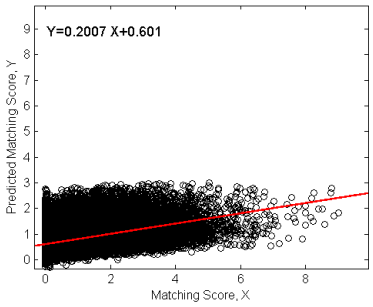

Figure 8. Scatter plot of predicted face matching scores versus actual matching scores in the imposter case. The results are obtained with a one hidden layer FFNN with 12 neurons and involving quality factors 1-9 and prediction scheme (1).

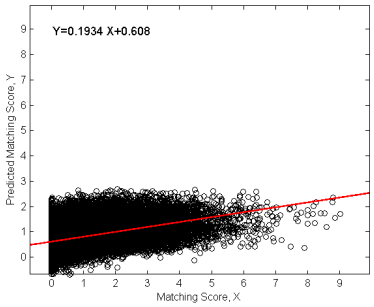

Figure 9. Scatter plot of predicted face matching scores versus actual matching scores in the imposter case. The results are obtained using MARS and involving quality factors 1-9 and prediction scheme (1).

scribed. The scenario involving direct and relative measures outperforms the other two.

The performance of the proposed method for selecting and ranking quality measures was evaluated by comparing predicted matching scores versus measured matching scores. A number of objective statistics such as R^2, F statistic, and MSE were evaluated.

The numerical analysis performed using iris and face biometric data resulted in a number of observations and conclusions. Based on the obtained results: (1) relative quality measures carry additional information compared to the original individual quality measures; (2) vectors of metrics perform considerably better compared to combined metrics, unless a combination rule is found that preserves the information contained in a vector of quality measures.

References

[1] I. Avcibas, B. Sankur, and K. Sayood. Statistical evaluation of image quality measures. *Journal of Electronic Imaging*, 11(2):206–223, 2002.

[2] P. Corriveau and A. Webster. Vqeg evaluation of objective methods of video quality assessment. *SMPTE Journal*, 108:645–648, 1999.

[3] J. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–141, 1991.

[4] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[5] N. D. Kalka, J. Zuo, V. Dorairaj, N. A. Schmid, and B. Cukic. Image quality assessment for iris biometric. *Proc. of SPIE Conference*, 6202:61020D1–62020D11, April 2006.

[6] V. Kurkova. Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, 5:501–506, 1992.

[7] D. W. Marquardt. An algorithm for least square estimation on nonlinear parameters. *Journal of SIAM*, 11:431–441, 1963.

[8] L. Masek and P. Kovesi. Matlab source code for a biometric identification system based on iris patterns. *The School of Computer Science and Software Engineering, The University of Western Australia*.

[9] A. M. Rohaly. Video quality experts group: Current results and future directions. *Proc. of SPIE Conference*, 4067(11):742–753, 2000.

[10] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. on Image Processing*, 15(11):3440–3451, 2006.