A Bayesian Model for Predicting Face Recognition Performance Using Image Quality

Abhishek Dutta Raymond Veldhuis Luuk Spreeuwers
University of Twente, Netherlands
{a.dutta,r.n.j.veldhuis,l.j.spreeuwers}@utwente.nl

Abstract

Quality of a pair of facial images is a strong indicator of the uncertainty in decision about identity based on that image pair. In this paper, we describe a Bayesian approach to model the relation between image quality (like pose, illumination, noise, sharpness, etc) and corresponding face recognition performance. Experiment results based on the MultiPIE data set show that our model can accurately aggregate verification samples into groups for which the verification performance varies fairly consistently. Our model does not require similarity scores and can predict face recognition performance using only image quality information. Such a model has many applications. As an illustrative application, we show improved verification performance when the decision threshold automatically adapts according to the quality of facial images.

1. Introduction

A face recognition system can make a verification decision to indicate if the subjects contained in a pair of facial images have same (genuine or match) or different (imposter or non-match) identity. For practical applications, we are not only interested in the verification decision but also in the uncertainty associated with the decision about identity. In this paper, we present a Bayesian model to quantify the uncertainty in verification decision.

In addition to the inherent limitations of a face recognition system, there are two major factors that contribute to uncertainty in decision about identity: a) inherent property of some identities which makes verification difficult (as described in [6]); b) the quality (like pose, illumination, noise, etc) of facial image pair. Our model only considers the role of image quality because it has a very strong contribution towards uncertainty in the decision about the identity. For example, a verification decision made using a non-frontal image with uneven lighting entails more uncertainty than a verification decision carried out on frontal mugshots captured under studio conditions. Therefore, our model relies on information about facial image quality to predict performance and to quantify the uncertainty in verification decision.

We propose to use a data driven model to capture the relationship between image quality and corresponding verification performance. We automatically assess facial image quality (like pose and illumination) of facial images and train our model on real verification performance data to find regions in the quality space where the recognition performance varies fairly consistently. Many such models explored in the past require similarity scores to predict recognition performance. Our model can make performance predictions even before the actual recognition has taken place because our model is based solely on the quality of probe (or query) and gallery (or enrollment) image pair.

There are many applications of such models that predict recognition performance: a) verification decision threshold that adapts according to sample quality; b) fusion of results from multiple algorithms; c) facilitate capture of “best” enrollment images by giving feedback to operator about the quality of acquired samples; d) in forensic cases involving a large amount of CCTV footage, such models can help investigators focus their attention on only the “best” quality video frames that entail higher evidential value. As an illustrative example, we apply our model to adaptively vary decision threshold and show that it helps improve verification performance.

This paper is organized as follows: In Section 2, we review some previous work in this area. We describe our Bayesian model in Section 3 and discuss its performance evaluation methodology in Section 4. In Section 5, we describe the experiments designed to train our model and evaluate its performance. We discuss our experiment results in Section 6 and then finally present our conclusions in Section 7.
2. Related Work

The systems that predict performance of a biometric system can be generally classified into two groups. The first group of methods utilizes the similarity score and prior knowledge about the genuine and impostor scores distributions to predict the performance. The second group of performance prediction systems assesses biometric sample quality and uses this information to predict performance—poorer sample quality entails more uncertainty in decisions about identity.

Performance prediction systems based solely on the similarity score first create some feature from the similarity scores and then apply machine learning to model the relationship between these features and the corresponding recognition performance. For instance, [12] computes three features from a set of sorted similarity scores while [11] uses features based on similarity scores that quantify the intrinsic factors (properties of algorithm, gallery set, etc) and extrinsic factors (quality of probe images). Both then use SVM to learn the relationship between these similarity score based features and their corresponding recognition performance. In [10], they compute a feature from impostor score distribution to quantify the “facial uniqueness” and then used Kernel Density Estimation to model a uniqueness based match (genuine) and non-match (impostor) score distributions. The uncertainty in decision about identity is higher in regions of overlapping tails of genuine and impostor score distribution. Therefore, a better model of tails of score distribution is essential for accurate prediction of recognition performance. Following this line of thought, [13] and [14] directly model the tails of similarity score distributions. In [13], the tail of impostor score distribution is modeled as a Weibull distribution. To predict the outcome of a verification decision, they check if the new verification score is an outlier with respect to the model of the tail of impostor score distribution. In [14], the tails of both genuine and impostor score distribution are modeled as a General Pareto Distribution. The normalized distance of a similarity score from the impostor score distribution is used as a performance predicting feature in [12]. Using a Probabilistic Graphical model, they model the joint density of similarity score and these performance predicting features. This allows them to predict the recognition performance.

It is also possible to predict recognition performance based on information about the biometric sample quality. One of the earliest works in predicting performance of a biometric system was presented by [15]. They first show that the normalized match score—which denotes the distance of match score from non-match score distribution—is an indicator of recognition performance. Using an Artificial Neural Network (ANN), they learn the non-linear relationship between fingerprint quality (like clarity of ridges and valleys, number and quality of minutiae, size of image, etc) and corresponding normalized match score. This model of quality and recognition performance (i.e. normalized score) is used to predict the performance of previously unseen fingerprint samples. Using externally assessed fingerprint quality, [17] model the genuine and impostor score distributions using gamma and log normal distributions respectively. This model of score distributions is then used to adaptively select the decision threshold based on quality information. The authors of [1] apply Multi-Dimensional Scaling (MDS) to learn the relationship between image quality features and similarity scores. Using regression, the authors of [2] model the relationship between quality partition (good, bad and ugly) and image-specific (sharpness, hue, etc.) and face-specific (facial expression) properties of a facial image.

Our work most closely relates to the work of [3] which uses a Generalized Linear Mixed Model (GLMM) to model the relationship between image quality (like focus, head tilt, etc.) and the outcome of verification decision. Their analysis shows that some quality metrics are strong indicator of recognition performance. In this paper, we propose a Bayesian framework for modeling the relation between face recognition performance and image quality. We use a probability density function to model this relationship.

3. Model of Image Quality and Recognition Performance

Let \( q = [q_1^p, q_1^g, \cdots, q_n^p, q_n^g] \in \mathbb{R}^{2m} \) denote image quality parameters (like pose, illumination direction, noise, etc.) of a probe and gallery image pair. Throughout this paper, the term image quality refers to any measurable property of facial images that affects the performance of face recognition systems. For a particular face recognition system \( j \), let \( r(j) = [r_1, \cdots, r_n] \in \mathbb{R}^n \) denote the face recognition performance corresponding to a sufficiently large set of different image pairs each having same quality \( q \). Here, we assume that the recognition performance of system \( j \) is not affected by variations in identity [6] and that vector \( q \) is sufficient to capture all the relevant quality variations possible in a facial image pair. Different face recognition systems have varying level of tolerance to image quality degradations and therefore we denote vector \( r(j) \) as a function of a particular face recognition system. To simplify the notation, we simply use \( r \).

We want to model the interaction between image quality \( q \) and recognition performance \( r \) using a probability density function (PDF) \( P(q, r) \). In this paper, we propose a data driven model of \( P(q, r) \) which is trained by gathering recognition performance data \( r \) for the most common types of quality variations \( q \) in probe and gallery image pairs. Once we have trained this model, we can predict recognition performance for a new probe and gallery pair with quality \( q \).
as follows:

\[ r^* = \arg \max_r P(r|q), \]  

where \( r^* \) denotes the most probable estimate of face recognition performance.

The recognition performance prediction \( r^* \) based on our model can be done even before the actual recognition task because our model relies only on the quality of facial images. Many such models, explored in the past, also use the similarity score as a feature for performance prediction. The impostor (or, non-match) score is influenced by both identity and quality of facial images [7]. Hence, it is not possible to differentiate if an extremely low similarity score is due to mismatched identity or comparison of extremely poor facial image pair. Therefore, we avoid using similarity score as a feature in our model. This design decision not only avoids the issues associated with using similarity score as a feature but also allows our model to predict performance even before the actual facial comparison has taken place.

In this paper, we express \( P(q,r) \) using a mixture of \( K \) multivariate Gaussian (MOG):

\[ P(q,r) = \sum_{k=1}^{K} \pi_k \mathcal{N}([q,r]; \mu_k, \Sigma_k), \]  

where, \( \pi_k \) are the mixture coefficients such that \( 0 \leq \pi_k \leq 1; \sum_k \pi_k = 1 \), and \( \mu_k \in \mathbb{R}^{2m+n}, \Sigma_k \) are the mean and covariance matrix of the \( k \)th mixture components. We apply the Expectation Maximization (EM) algorithm to learn the parameters of the MOG in (2).

Given the quality \( q \) of previously unseen verification instance, we can apply Bayes’ theorem to (2) and obtain the posterior distribution of recognition performance \( r \) as

\[ P(r|q) = \frac{P(q,r)}{P(q)}, \]  

Since the denominator of (3) does not depend on \( r \), the corresponding most probable estimate of \( r \) for a given quality \( q \) is given by

\[ r^* = \arg \max_r P(q,r), \]  

Substituting \( r^* \) in (3), we can obtain \( P(r^*|q) \) which defines the probability of most probable recognition performance \( r^* \).

4. Performance Prediction Error

Using the PDF of (3), we can obtain the posterior distribution of recognition performance \( r \) for any given point \( q \) in the quality space. Our test data set does not contain sufficient number of verification instances at each point in the quality space. Therefore, even though our model can predict performance at each point in the quality space, we do not have enough test data to evaluate the error in those model predictions. Hence, we evaluate the performance of our model by adopting an alternative view of MOG decomposition which presents the mixture components as clusters.

Recall that, the MOG decomposition of (2) can alternatively be also viewed as partitioning the \([q,r]\) space into \( K \) clusters. We partition all the verification instances in the test data set into a set of \( K \) clusters (or, mixture component). For a previously unseen verification instance in the test data set with quality \( q \), we first compute the most probable estimate of performance \( r^* \) using (4) and then assign it to the cluster \( k^* \) such that

\[ k^* = \arg \max_k \pi_k \mathcal{N}([q,r^*]; \mu_k, \Sigma_k), \]  

where, \( k^* \in [1, \cdots, K] \). Based on these cluster specific verification instances, we compute true verification performance and its credible interval using a Bayesian approach as discussed in Section 4.1. Given a new instance \( q \), the cost of performance prediction is \( O(a^n) \) where \( a \) is the number of levels in each dimension of \( r \) and \( n \) is the dimensionality of \( r \).

We compare these cluster specific true verification performance with our model’s prediction of verification performance at each cluster center. The most probable estimate of verification performance \( r^*_k \) evaluated at the center (i.e. mean) of cluster \( k \) has a credible interval \((c,d)\) of size \( (1-\alpha) \) such that

\[ \int_c^d P(r|q = \mu_k)dr = 1-\alpha, \]  

where, \( \mu_k \) denotes the quality component of \( k \)th mixture component mean \( \mu_k \).

Note that we employ this strategy of evaluating model performance because of the limited nature of our testing data set. Given sufficient test data, our model’s performance can be evaluated at each point in the quality space. As a very rough estimate, we need 100 genuine samples at each point in the quality space for reliable measurements of FRR= 0.01.

4.1. Credible Interval for Computed FRR

We describe a Bayesian approach for computing the credible interval for the cluster specific FRR computed from test data set. Let \( G_k \) and \( I_k \) denote the set of genuine and impostor scores corresponding to cluster \( k \). Given the desired operating point \( \text{FAR}_{\text{desired}} \), we can obtain a decision threshold \( t_k \) by solving the following equation:

\[ \text{FAR}_{\text{desired}} = \frac{n(I_k^+: I_k > t_k)}{n(I_k)}, \]  

where \( n(I_k) \) denotes the cardinality of set \( I_k \). Now, for all instances in \( G_k \), we can make a verification decision \( w \in \{"
\{0, 1\} based on this decision threshold \(t_k\) as follows:

\[
u^{(i)}(t) = \begin{cases} 
1 & \text{if } G^{(i)}_k < t_k, \\
0 & \text{otherwise.}
\end{cases}
\]

Therefore, each verification decision can be thought of as the outcome of a Bernoulli trial. Let \(m\) be a random variable indicating the number of \(v^{(i)}(t) = 1\) observations out of total \(N = |G_k|\) verification decisions. The probability of getting \(m\) success in \(N\) trials follows a Binomial distribution \(\text{Bin}(m|N, \mu)\), where \(P(w = 1|\mu) = \mu\). We are interested in the posterior distribution of \(\mu\) which in turn defines the distribution of FRR given by:

\[
\text{FRR} = \frac{m}{N}.
\]

Assuming a Beta distribution as the prior distribution of \(\mu\), the posterior distribution of \(\mu\) is the product of binomial likelihood function \(\text{Bin}(m|N, \mu)\) and beta prior \(\text{Beta}(a, b)\). Based on the property of conjugate prior [4, p.70], we know that the posterior distribution of \(\mu\) is a Beta distribution \(\text{Beta}(m + a, l + b)\), where \(l = N - n\). The FRR given by (9) has a Bayesian credible interval \((c, d)\) of size \(1 - \alpha\) such that

\[
\int_c^d \text{Beta}(\mu; m + a, l + b) d\mu = 1 - \alpha.
\]

Since we do not have any prior knowledge about \(\mu\), we assume a uniform prior \(\text{Beta}(a = 1, b = 1)\).

5. Experiments

![Figure 1. MultiPIE camera and flash positions used in this paper.](image)

We present experiment results to show that the proposed model of (2) can indeed capture the relation between quality \(q\) and performance \(r\). In this study, we use the FaceVACS [5] recognition system and the neutral expression, four session (first recording only) subset of the MultiPIE data set [9]. Of the total 337 subjects, our training set consists of the 129 subjects that are present in all four sessions. The remaining 208 subjects are used for testing.

For both training and testing data set, we use the high quality frontal mugshots for the gallery (or, enrollment) set. Image quality \(\text{(i.e. pose and illumination)}\) variations are only present in the probe (or, query) set. The probe set contains images from the 5 camera and 5 flash positions as depicted in Figure 1. Since the quality of our gallery set remains fixed, in all our experiments the quality vector quantifies the pose and illumination of only the probe image: \(\text{i.e. } q = [q_1, q_2]\). In Section 5.1, we describe the quality vector \(q\) in more detail. Furthermore, the recognition performance vector is a single dimensional quantity \(r = [r_1]\) where, \(r_1\) denotes the False Reject Rate (FRR) (in base 10 log scale). For all results presented in this paper, the False Accept Rate is fixed to 0.001. This experimental design simulates real world verification scenario where the gallery is fixed to a set of high quality frontal mugshots and the variable of interest is the expected face recognition performance \(\text{(i.e. FRR)}\) at some predefined operating point \(\text{(i.e. FAR)}\).

We have designed our experiment such that there is minimal impact of session variation and image alignment on the face recognition performance. We select the high quality gallery image from the same session as the session of the probe image. Furthermore, we disable the automatically detected eye coordinates based image alignment of FaceVACS by supplying manually annotated eye coordinates. This ensures that there is consistency in facial image alignment even for non-frontal view images.

5.1. Image Quality Assessment

Many types of quality variations can degrade the quality of a facial images. There exists a multitude of algorithms in Computer Vision to assess common facial image properties like pose, illumination direction, noise, blur, \textit{etc}. In this paper, we use the Image Quality Assessment (IQA) tool \texttt{dbassess} included with the FaceVACS [5] SDK. This IQA tool measures a large number of image quality parameters. However, we only use the \texttt{DeviationFromFrontalPose} \((q_1)\) and \texttt{DeviationFromUniformLighting} \((q_2)\) parameters because our training and testing data set mainly contains variations in pose and illumination.

In Figure 2, we show the distribution of these two quality parameters \((q_1, q_2)\) for probe images in the training data set. The distribution of \(q_1\) for frontal view images is centered around \(-1.0\) while for non-frontal views, it shifts toward \(+2.0\). Similarly, while keeping the pose fixed to frontal view, we vary the illumination and observe that for frontal illumination the distribution of \(q_2\) is centered around \(-2.0\) while for other illumination conditions it shifts towards values \(\geq 0\). This shows that the two quality parameters have the desired response towards the pose and illumination variations present in our data set.

5.2. Training

In order to train our data driven model of (2), we would ideally want a very large number of probe images with same
quality to evenly occupy each position in the quality space. However, it is difficult to obtain such a data set. In Figure 4 (left), each point corresponds to a unique probe image in our training data set. We observe that some of the regions in the quality space are sparsely populated by the training data. Therefore, we apply a quality space sampling strategy that adapts according to the nature of available training data. We define sampling points along $q_1$ and $q_2$ based on the $N_q$ quantiles ($=28$) of evenly spaced probabilities in the quality space. At each sampling point $q = [q_1, q_2]$, we select the closest $N_s$ samples ($=250$) around $q$. We aggregate all similarity scores for which the quality of probe corresponds to the closest $N_s$ samples. These aggregated scores define the $r$ vector (i.e. FRR at FAR=0.001) for that particular $q$ vector. To avoid collecting scores from very large distances, we discard the sample point $q$ that do not acquire sufficient scores within certain predefined range. For the training data set, Figure 4 (right) shows the true FRR at each sampling point in the quality space. For some sample points, the FRR=0 and therefore to avoid $-\infty$ in the log scale, we assign all such instances as $r_1=3.0$ (i.e. FRR=0.001).

From the training data set, we have a set of 441 training vectors $[q, r]$, which is used to learn the model parameters for (2). We use the EM algorithm implementation available in the R library mclust [8]. We select the number of clusters $K=5$ because on the training set, this results in most distinct clusters in the quality space. Furthermore, given the limited nature of our training data set, we cannot reliably estimate a model with full covariance matrix. Therefore, we select the VVI (see [8] for details) model parametrization which defines covariance matrix as: $\Sigma_k = \lambda_k A_k$, where $A_k$ is a diagonal matrix whose elements are proportional to the eigenvalues and $\lambda_k$ is an associated constant of proportionality. Here, $\lambda_k$ and $A_k$ govern the volume and shape of $k$-th mixture component.

The projection of resulting six mixture components regions in the quality space is shown in Figure 3. In Figure 5 (left), we show the plot of $(r^*, q_1, q_2)$ where $r^*$ denotes the most probable estimate of recognition performance computed using (4). In Figure 5 (right), we show the corresponding value of probability density function $P(r^*|q_1, q_2)$. Note that the FRR values are shown in log scale in order to
provide better visualization of different regions in the quality space formed by the six mixture components.

5.3. Performance Prediction

Figure 6. Cluster specific verification performance where the error bars indicate 95% credible interval.

As described in Section 4, Figure 6 shows the cluster specific verification performance as predicted by our model (at each cluster center in the quality space) and for the training and testing data set. The error bars indicate 95% credible interval (i.e. $\alpha = 0.05$).

Figure 7. Performance for verification decisions based on our cluster-specific threshold and a naive threshold scheme.

As an illustrative application of our model, we show that adapting the verification decision threshold based on image quality information can improve verification performance. From the training data set, we compute a cluster specific decision threshold for all verification instance without considering the image quality. The selected $FAR = 0.001$ denotes a single operating point and therefore gives us a single point in the ROC curve shown in Figure 7. Therefore, we repeat this procedure for other values of FAR and obtain the full ROC curve shown in Figure 7.

The adaptive decision threshold scheme based on our model achieves a $FRR = 4.72\%$ (at $FAR = 0.00102\%$) while a naive scheme achieves a $FRR = 5.53\%$ (at $FAR = 0.0107\%$) – an improvement of 0.81% in FRR. This improvement in performance, though small, shows the merit of our model in exploiting image quality information for performance prediction. Further performance gain can be achieved by including additional image quality parameters and by using less constrained model parametrization.

6. Discussion

In this paper, we build a data driven model to learn the relation between the quality $q$ of probe and gallery image pair and the corresponding recognition performance $r$. Recall that, although our model is capable of handling quality of both probe and gallery images, we only consider the quality of probe images in our experiments. This is done to simulate real world verification scenario in which the quality of gallery image is fixed to a set of high quality frontal mugshots and only the quality of probe images vary. Our model parametrizes the $[q, r]$ space into a linear combination of multivariate Gaussian. In Figure 3, we show the projection of these Gaussians in the quality space. FaceVACS is fine tuned for optimal verification performance on frontal view images for which its performance remains largely invariant to illumination variation. Mixture Component (MC) 1 captures this property of the system and therefore occupies the region of quality space corresponding to frontal view ($-3 \leq q^p_k \leq 0$) and all possible illumination condition ($-4 \leq q^i_k \leq 0$). MC 3 corresponds to slight non-frontal pose and depicts the property of FaceVACS to be nearly tolerant to a pose variation of $\pm 15^\circ$. All the remaining four MC i.e. $k \in \{2, 4, 5\}$ are located in a region of quality space corresponding to the non-frontal pose (i.e. $1 \leq q^p_k \leq 4$). In this region, there are multiple clusters along the illumination variation axis (i.e. $q^i_k$) which indicates that for non-frontal pose, illumination variation has strong impact on verification performance. Furthermore, MC 2 corresponds to non-frontal pose and illumination images – the worst possible image quality in our data set.

Based on (4), we compute the most probable estimate of verification performance vector $r^*$ at each point in the quality space and show it in Figure 5 (left). This map clearly marks the boundary of MC $\{1, 2, 5\}$. However, the boundary between MC 3 and 4 is not clearly visible because these two mixture components have very small difference in verification performance. Yet, our model use two mixture components to represent this region because multiple types of
quality variations (corresponding to different regions in the quality space) can have similar impact on verification performance. For example, for a uniformly illuminated non-frontal view facial image (i.e. camera 04,1 and flash 09), the verification performance is degraded to a certain level by the non-frontal pose of the facial images. Whereas, for a poorly illuminated near-frontal view facial image (i.e. camera 05,0 and flash 01), the verification performance is degraded to a similar level but due to poor illumination. Therefore, we expect multiple regions in the quality space to have similar verification performance. This phenomenon is nicely illustrated by the L shaped region formed by MC 3 and 4. The corresponding map of $P(r^*|q)$ is shown in Figure 5 (right) which depicts low confidence in model prediction in the boundary regions of mixture components.

The cluster specific verification performance of Figure 6 shows that for clusters $k = \{1, 2, 4\}$, our model accurately classifies both testing and training verification samples into clusters for which the verification remains fairly consistent. For clusters $k = \{2, 5\}$, we observe a large difference between the true verification performance of the training and testing data set. This indicates that the two quality parameters (pose and illumination) used in our model may not be sufficient to capture all the variations that exist in our data set. For example, some subjects in our data set wear glasses and some others have a large part of their face occluded by facial hair. Another reason for this large variation might have to do with the diagonal model parametrization we use in this paper. While this parametrization reduces the model complexity, it enforces independence constraint between the quality and recognition performance variables.

Figure 6 also shows our model predictions which is very
close to the true verification performance observed on the training and testing data set. For cluster 2, we observe a very large confidence interval for our model’s prediction. Recall that, we evaluate the model’s performance at each cluster center. Since our model training is based on only 250 scores samples at each sample points in the quality space, we observe large variance in the model predictions. For our training and testing data set, we observe small variance in verification performance because the estimates are based on a very large number of samples (> 1000) classified to each cluster. These limitations are common for most data driven model indicating the need for more densely distributed training data in the quality space.

7. Conclusion

In this paper, we propose a data driven model to learn the relation between facial image quality and the corresponding recognition performance. Adopting a Bayesian approach, we model this relationship as a probability density function. For previously unseen verification instance, we predict the verification performance by evaluating the posterior distribution for the given image quality. This posterior distribution also quantifies the uncertainty in decision about identity. A remarkable property of our model is that it relies solely on image quality information and does not require similarity scores to make predictions about recognition performance. For a data set containing pose and illumination variations, we have shown that the proposed model is able to identify regions (i.e. cluster) in the quality space over which the face recognition performance varies fairly consistently. Furthermore, we have also shown an illustrative application of our model where we observe improvement in verification performance by using image quality information to adapt the decision threshold.

A limitation of the proposed data driven model is that it requires sufficiently large number of training samples spread densely in the quality space. Given that we succeed in acquiring sufficient training and testing data, we envisage to extend our model to include additional quality parameters (noise, sharpness, expression, etc.) and more recognition performance parameters (like Area Under ROC, calibrated log-likelihood ratio, more points on ROC, etc. and their combinations).

Acknowledgements

This work was supported by the BBfor2 project which is funded by the EC as a Marie-Curie ITN-project (FP7-PEOPLE-ITN-2008) under Grant Agreement number 238803. We would also like to thank Cognitec Systems GmbH, for supporting our research by providing the FaceVACS software. Results obtained for FaceVACS were produced in experiments conducted by the University of Twente, and should therefore not be construed as a vendors maximum effort full capability result.

References