A New Performance Evaluation Method for Face Identification - Regression Analysis of Misidentification Risk

Wai Han Ho and Paul Watters
Macquarie University
NSW 2109 Australia
sharonho@ics.mq.edu.au, pwatters@ics.mq.edu.au

Abstract

The performance of a face identification system varies with its enrollment size. However, most experiments evaluated the performance of algorithms at only one enrollment size with the rank-1 identification rate. The current practice does not demonstrate the usability of algorithms thoroughly. But the problem is, in order to measure identification performance at different sizes, experimenters have to repeat the evaluation with samples of those sizes, which is almost impossible when they are large. Approaches using the Binomial theorem with match and non-match scores have been proposed to estimate performance at different sizes, but as a separate process from the evaluation itself. This paper presents a new way of evaluating identification algorithms that allows the estimating and comparing of performance at different sizes, using the regression analysis of Misidentification Risk.

1. Introduction

Face recognition can be broadly classified into verification and identification. Verification deals with the validation of identity claims while identification caters for the identifying of unknowns or validation of negative claims. This paper presents a new performance evaluation method for identification, taking into account the scalability issue.

Identification operations involve one to many comparisons. The identities of incoming unknown images are found by comparing the images with all the enrolled templates and getting the identities of the ones with the highest similarity score. The accuracy rate of a running system is the percentage of incoming unknowns having the right enrolled template located. The approach was adopted by almost all experiments in algorithm evaluation for identification. The performance of an algorithm was determined by finding the rank-1 identification rate with a sample, which is the percentage of probe images in the sample having the right gallery templates located. A gallery set represents the set of enrolled templates, and a probe set acts as the set of incoming unknown images. The approach measures the identification performance of an algorithm at an enrollment size that is equal to the sample size. However, systems designed for identification are likely to be deployed to environments enrolled with different number of people. The chance of error increases with enrollment size, as the probability of having more similar templates within the enrolled increases. Therefore, evaluating algorithms at only one enrollment size has not fully demonstrated the algorithms' performance; their scalability should also be considered. Algorithms should be compared at different enrollment sizes as one may degrade more gracefully than others. But with the current approach, this can only be done by repeating the evaluation with databases of different sizes, which is almost impossible for large sizes.

There were suggestions to use the binomial theorem with match and non-match score distributions in estimating performance at different sizes. Daugman, Maeda *et al*. and Wayman [4, 8, 16] suggested viewing each identification problem at a size of $G + 1$ as making $G + 1$ verifications. Comparisons between image pairs of the same subjects were considered as 'valid claims', and the corresponding similarity scores were termed as match scores. Comparisons between images of different subjects were 'impostor claims', and the corresponding similarity scores were non-match scores.

Phillips *et al*. [12] specified the relationship between 'valid' and 'impostor' claims, and that between match and non-match scores in estimating identification performance. An algorithm could locate the right enrolled template at rank-n if n of the G 'impostor claims' were rejected at a threshold equals to the match score, which meant the non-match scores were smaller than the match score. The probability of an algorithm correctly rejecting an 'impostor claim' at a match score of s was taken as $N(s)$, with N being the cumulative distribution of all the non-match scores. The probability of having n of the G 'impostor

1-4244-1180-7/07/$25.00 ©2007 IEEE

claims' rejected was calculated based on the Binomial theorem. The probability of a match score being at a value of s was $m(s)$, with m being the probability density of all the match scores. The rank-1 identification rate of the algorithm was the integration of: the probability of rejecting all G 'impostor claims' at each possible value s of a match score, $N(s)^G$, multiplied by $m(s)$, the probability of it being at that value. According to [12], the model underestimated identification performance by about 10% at a size of 1,000 and up to 16% for sizes close to 40,000, when tested with 6,000 match scores and 18 million non-match scores without normalization.

Grother *et al*. [5] extended the approach to cater for the open-universe identification scenario, and derived the integration using the linear interpolation of the two distributions and the Monte Carlo sampling method.

Rong *et al*. [15] adapted the method to handle the distortion problem in the real world explicitly, and used the expectation-maximum algorithm and a learning model to train the match score and non-match score distributions that were represented as Gaussian mixtures.

The suggested binomial-based models derived algorithm performance based on the distribution of all the non-match scores of the selected sample. However, an identification operation of an incoming image involves only the non-match scores of that particular image. The models rely on different individuals' non-match score distributions being identically distributed, not only the match and non-match scores being independent. Johnson *et al*. [7] presented a model that predicted performance based on individuals' non-match scores separately, using a count method.

We suggest using a completely different approach, the regression analysis of Misidentification Risk, to represent the performance and scalability of algorithms in one go. The Misidentification Risk of a person associated with an algorithm under certain conditions is the probability of the person being misidentified by others, when compared to all people in the target population under the same condition by the algorithm. It reflects the result of the decidability [3] or separability of the person's match and non-match score distributions. The distribution of Misidentification Risk consolidates the effect of individuals' decidability. The probability of correctly identifying a person at different enrollment sizes can be estimated from the distribution of Misidentification Risk. Experimental results showed that different algorithms' distribution of Misidentification Risk could be represented with either the Beta or Weibull distributions with high goodness of fit, and that point estimates of identification rates at different enrollment sizes from the best-fit curves matched empirical results with high accuracy.

In Section 2, we describe the concept of Misidentification Risk and how it relates to the performance of algorithms. We discuss also the technical issues in applying the concept to represent algorithms' identification performance. Section 3 details the experiments done in determining the goodness of fit of the Beta and Weibull distributions to Misidentification Risk. Section 4 compares the performance estimated using the new approach and the binomial-based approach with empirical results. Section 5 summarizes the discussion.

2. Misidentification Risk

2.1. The relationship between distributions of Misidentification Risk and identification Performance

Looks are diverse, but still there are people who look alike. One may find it easy to identify some people but have difficulty identifying others. The same is true for automatic face identification systems. A person may be identified correctly by one algorithm but incorrectly by other algorithms. The number of people that could possibly be misidentified as the person amongst the same group of people could also be different. A person will have a different risk of being misidentified with different algorithms.

The Misidentification Risk of a person with an algorithm under certain conditions is the percentage of people in the target population that would be misidentified as him by the algorithm under the same conditions. The conditions can be any magnitude change in illumination, pose, expressions and so forth.

The Misidentification Risk of a person with an algorithm directly relates to the probability of the person being correctly identified by the same algorithm amongst a subgroup of the target population. A person having a Misidentification Risk of p under some conditions will have Mp misidentifications on average in independent sets of M people under the same conditions.

Therefore, in order for the person to have an average of < 1 misidentification amongst M people, his Misidentification Risk must satisfy

$$M * p < 1, \ or \ p < \frac{1}{M}. \tag{1}$$

That is, people having a Misidentification Risk $< \frac{1}{M}$ with an algorithm on certain conditions will on average be correctly identified against a group of other M people under the same conditions. This implies that the percentage of people in a population having a Misidentification Risk $< \frac{1}{M}$ will be the percentage of people that can possibly be identified by the algorithm when having to compare to a group of other M people. In other words, if we have different groups of $M + 1$ people randomly taken from the population, the average probability of a person being correctly identified in

these groups is the percentage of people having a Misidentification Risk $< \frac{1}{M}$. $M + 1$ can be approximated by M except when M is small.

This means that an algorithm's cumulative distribution of Misidentification Risk of a target population gives the average probability of it correctly identifying a person for different values of M. Denoting the distribution as $F(x)$, the average rank-1 identification rate $P(M, 1)$ (the probability of correctly identifying a person amongst a group of M people) is

$$P(M, 1) = F(\frac{1}{M}). \quad (2)$$

Based on the same reasoning, the average identification rate $P(M, n)$ at rank n is

$$P(M, n) = F(\frac{n}{M}). \quad (3)$$

As an algorithm's distribution of Misidentification Risk under certain conditions relates to its performance at different enrollment sizes under the same conditions, we suggest assessing an algorithm with its distribution of Misidentification Risk.

2.2. Technical issues in representing performance with Misidentification Risk

Equation 2 is conceptually easy, but the actual processing would be much more complicated if what we have are discrete values of Misidentification Risk and their accumulative percentages. As target populations are usually large, the possible number of Misidentification Risk values is large and close to continuous. Therefore, if distributions of Misidentification Risk of different algorithms can be represented by a continuous mathematical distribution model, then any difference between algorithms is embraced in the corresponding parameter values, instead of sets of (risk, percentage) values.

Because populations targeted by algorithms are usually large. Their distributions of Misidentification Risk can only be estimated from samples randomly drawn from the populations. When the Misidentification Risk of a population is estimated using a sample of size Z, the values are quantized into discrete values in steps of $\frac{1}{Z}$.

The accuracy of using the distribution of Misidentification Risk measured from a sample to represent an algorithm's performance depends on three issues: (a) Can an algorithm's distribution of Misidentification Risk when measured by a sample be adequately represented by a mathematical distribution model? (b) How closely can we reproduce a population's distribution from the Misidentification Risk measured from a sample? (c) The effectiveness of using the distribution of Misidentification Risk in estimating algorithms' performance at different enrollment sizes.

We address the first issue in Section 3. Experimental results showed that distributions of Misidentification Risk could be closely represented by a continuous probability distribution model. The second and third issues cannot be answered separately due to the same reason - the size of target populations. We address the two issues together in Section 4, where estimated rank-1 identification rates at different enrollment sizes are compared with empirical results.

3. Characteristics of Misidentification Risk distributions

We believe there is a certain regularity underlying the Misidentification Risk generated by different algorithms.

Hypothesis: The same continuous mathematical distribution model may be used to represent the cumulative distribution of Misidentification Risk of different algorithms, with a goodness of fit measured by R^2 greater than 0.95.

The adequacy of the Weibull and Beta distribution models in representing the data was tested

3.1. A brief discussion of the Beta distribution

The Beta distribution is a continuous distribution on the interval of [0 1] parameterized with 2 qualities α and β. It has a probability density function (pdf) f defined as

$$f(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) + \Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}, \quad (4)$$

where Γ is the gamma function.

The cumulative distribution function (cdf) is defined as

$$F(x|\alpha, \beta) = \frac{\int_0^x t^{\alpha-1}(1 - t)^{\beta-1} dt}{(\Gamma(\alpha)\Gamma(\beta))/\Gamma(\alpha + \beta)}. \quad (5)$$

3.2. A brief discussion of the Weibull distribution

The Weibull distribution, published in 1939 by Waloddi Weibull, is one of the most widely used lifetime distributions in reliability engineering [9, 14, 2].

The two-parameter pdf is :

$$f(T|\beta, \gamma) = \frac{\beta}{\eta}(\frac{T}{\eta})^{\beta-1} e^{-(\frac{T}{\eta})^{\beta}}, \quad (6)$$

where η is the scale parameter, and β is the shape parameter. The shape of the pdf takes on a variety of forms depending on β. It reduces to that of the 2-parameter exponential distribution when β = 1, and may approach normal when β lies between 2.6 and 3.7.

The **two-parameter cdf** is given by:

$$F(T|\beta, \gamma) = 1 - e^{-(\frac{T}{\eta})^{\beta}}. \quad (7)$$

3.3. Experiments on Weibull and Beta regression analysis of Misidentification Risk

We tested the hypothesis using images on the fa and fb lists of the FERET face database [11] and the five algorithms implemented by [1]. The images were full frontal face images taken on the same day but with different expressions (the conditions). The five algorithms were the PCA with Euclidean distance as a measure of similarity (PCA Euclidean), PCA with Mahcosine as a measure of similarity (PCA Mahcosine), Linear Discriminant Analysis (LDA), Bayesian ML and Bayesian MAP.

The experiment followed the new Owner-Tester experimental design proposed by Ho *et al*. [6]. The new setup uses an owner and a tester set of images, each from a different group of subjects. The gallery and probe sets of the current setup are made up of images from the same group of subjects. Each owner has two images in the owner set, one acting as the 'known' image and the other as the 'unknown'. Each tester has one image in the tester set as the 'known' image. With the setup, desired identification behaviour of a population is estimated from the owners, whose images are compared with the tester images in the tester set.

The distributions of Misidentification Risk of the algorithms under the conditions represented by the images were derived from the risk of individual owners being misidentified by the testers. The Misidentification Risk of an owner is the number of testers having a smaller similarity score with the owner's unknown image than that between the owner's image pair, divided by the total number of testers.

The regression analysis of the five algorithms with the two models was tested using three owner-tester groups of sizes '(1)' to '(3)' as shown on the second and third columns of Tables 1 and 2. For each case, subjects were randomly chosen as owners or testers. The images on the fa and fb lists of subjects chosen as owners took the role of the 'known' and 'unknown' images respectively. For subjects chosen as testers, only the image on the fa list was taken.

There are many different regression methods suggested in the literature to fit data into a given mathematical model, such as the least square estimation (LSE), relative least square, maximum likelihood, moment estimators, ridge regression, least absolute deviations and robust ridge regression. As the purpose of the experiment was to determine if distributions of Misidentification Risk can be represented by the selected models, and to find the one that better describes face recognition data, not a comparison of parameter estimators, we have chosen the most common one, the LSE, to find the best-fit curves. Following the same reason, although there are different ways of measuring how good a model describes given data, like the total deviation, mean squared errors of residuals and correlation coefficient (R^2), we have chosen R^2 as the metric for testing the goodness of fit of the models. R^2 is a fraction between 0 and 1, with higher values indicating better fits. R^2 was computed as the sum of the squares of the distances of the points from the best-fit curve determined, normalized by the sum of the squares of the distances of the points from a horizontal line through the mean of all Y-axis values. As the two models being evaluated have the same number of parameters, they are of similar complexity. We do not have to consider whether any decrease in sum-of-squares was the result of an increase in complexity [13].

3.4. Results of regression analysis

Tables 1 and 2 give the regression analysis results of the five algorithms using the three datasets, for the Beta and Weibull models respectively. The values of R^2 show that both models fitted the data obtained from all the algorithms and datasets very well, all except one had a R^2 greater than 0.95. This was especially true for the Weibull model which gave relatively better results, with R^2 ranged from 0.969 to 0.997. The R^2 of the Beta model ranged from 0.947 to 0.989. Therefore, the hypothesis is true.

Is the Weibull distribution model adequate in representing distributions of Misidentification Risk? It is generally agreed in the literature that when deciding whether a model is suitable to represent certain data, the following criteria should be considered: Descriptive adequacy - does the model provide a good description of the observed data? Generalizability - does the model predict well the characteristics of data that will be observed in the future? Complexity - does the model capture the phenomenon in the least complex possible manner? The Weibull model has fitted the data with $R^2 > 0.95$, so it should be descriptively adequate for the data. Although the experiment tested only five, not all available algorithms, the results have given us confidence that the mathematical model can likely be generalized to fit data of other algorithms. The 2-parameter Weibull model has only two parameters and a well formed and simple cdf. Therefore, we think the complexity of the Weibull model is low, but will continue to test if any 1-parameter models can also suit the task.

4. Using the Weibull model to estimate performance

As the Weibull model could represent the distributions of Misidentification Risk better than the Beta model, we show here how the Weibull model can be used to estimate the performance of algorithms at different enrollment sizes. Based on Section 2.1, we can estimate the rank-1 identification rate of an algorithm when enrolled with M people by substituting $\frac{1}{M}$ into the distribution function as in Equation 2. Replacing the arbitrary distribution F with the Weibull cdf:

$$P(M,1) = F(\frac{1}{M}) = 1 - e^{-(\frac{1}{M*\eta})^{\beta}}. \qquad (8)$$

Table 1. Regression results from the Beta distribution

Algorithm	Dataset		R^2
	Tester Size	Owner Size	
PCA Euclidean	(1) 199	812	0.955
PCA Mahcosine			0.959
LDA			0.992
Bayesian ML			0.963
Bayesian MAP			0.965
PCA Euclidean	(2) 598	598	0.975
PCA Mahcosine			0.947
LDA			0.985
Bayesian ML			0.980
Bayesian MAP			0.978
PCA Euclidean	(3) 997	199	0.980
PCA Mahcosine			0.967
LDA			0.989
Bayesian ML			0.980
Bayesian MAP			0.970

Table 2. Regression results from the Weibull distribution

Algorithm	Dataset		R^2
	Tester Size	Sample Size	
PCA Euclidean	(1) 199	812	0.990
PCA Mahcosine			0.991
LDA			0.970
Bayesian ML			0.992
Bayesian MAP			0.995
PCA Euclidean	(2) 598	598	0.997
PCA Mahcosine			0.989
LDA			0.986
Bayesian ML			0.995
Bayesian MAP			0.994
PCA Euclidean	(3) 997	199	0.982
PCA Mahcosine			0.991
LDA			0.969
Bayesian ML			0.994
Bayesian MAP			0.989

Hypothesis: The suggested concept and the Weibull representation of Misidentification Risk may be used to estimate identification performance at different enrollment sizes.

4.1. Experiments on performance estimation

The same experimental setup and datasets as in Section 3.3 were used to test the approach in estimating performance. We compared the rank-1 identification rates estimated from the Weibull Regression at the following enrollment sizes: 200, 599 and 998, with the estimations from the binomial-based method [12] without normalization and the corresponding empirical results measured.

Tables 3 and 4 tabulate the rank-1 estimations from the

Table 3. A comparison of rank-1 estimations from the Weibull regression method with empirical results at different sizes

Observation from	Estimation /Empirical	Performance at Sizes		
		200	599	998
PCA Euc.	Empirical	0.88	0.79	0.74
	(1)199,812	0.89	0.81	0.77
	(2)598,598	0.87	0.80	0.76
	(3)997,199	0.87	0.78	0.73
PCA Mah.	Empirical	0.92	0.89	0.81
	(1)199,812	0.92	0.87	0.84
	(2)598,598	0.93	0.89	0.87
	(3)997,199	0.91	0.85	0.82
LDA	Empirical	0.78	0.76	0.7
	(1)199,812	0.76	0.68	0.64
	(2)598,598	0.81	0.74	0.71
	(3)997,199	0.78	0.70	0.66
Bayes. ML	Empirical	0.90	0.86	0.77
	(1)199,812	0.91	0.85	0.82
	(2)598,598	0.91	0.86	0.84
	(3)997,199	0.91	0.82	0.78
Bayes. MAP	Empirical	0.90	0.86	0.77
	(1)199,812	0.90	0.84	0.81
	(2)598,598	0.91	0.86	0.84
	(3)997,199	0.91	0.82	0.78

Table 4. A comparison of rank-1 estimations from the binomial-based model with empirical results at different sizes

Observation from	Estimation /Empirical	Performance at Sizes		
		200	599	998
PCA Euc.	Empirical	0.88	0.79	0.74
	(1)199,812	0.79	0.70	0.66
	(2)598,598	0.74	0.65	0.62
	(3)997,199	0.70	0.62	0.58
PCA Mah.	Empirical	0.92	0.89	0.81
	(1)199,812	0.91	0.86	0.83
	(2)598,598	0.90	0.85	0.82
	(3)997,199	0.88	0.81	0.78
LDA	Empirical	0.78	0.76	0.7
	(1)199,812	0.63	0.57	0.54
	(2)598,598	0.61	0.54	0.51
	(3)997,199	0.57	0.50	0.47
Bayes.ML	Empirical	0.90	0.86	0.77
	(1)199,812	0.72	0.65	0.62
	(2)598,598	0.66	0.59	0.56
	(3)997,199	0.64	0.57	0.53
Bayes. MAP	Empirical	0.90	0.86	0.77
	(1)199,812	0.71	0.64	0.61
	(2)598,598	0.66	0.59	0.55
	(3)997,199	0.64	0.56	0.53

Weibull and binomial-based models [12] respectively, with the empirical results included for comparison. The third to fifth columns show the estimated or empirical rank-1 identification rates at the three sizes respectively. The row 'Empirical' for each algorithm shows the rank-1 identification rate directly measured from the three datasets. The rows with '(1)', '(2)' and '(3)' show the estimations from the Weibull or binomial-based model. '(1)' to '(3)' represent the same datasets as in Table 2.

The experimental results show that the rank-1 identification rates estimated for the five algorithms using the proposed method had much smaller deviations from the empirical results than those estimated using the binomial-based model. Deviations from the proposed method ranged from 0 to $|0.08|$, with the majority being $|0.01|$. Those from the binomial-based model ranged from $|0.01|$ to $|0.3|$, with more than half greater than $|0.15|$. The estimations from Weibull regression were quite evenly distributed between over and under estimations, with 16 of the former and 23 of the latter, while all except one of the binomial-based estimates were underestimations. The larger deviations obtained in this experiment for the binomial-based model when compared to those presented in [12] might be caused by the sample size of this experiment being much smaller.

The experimental results have shown that the Weibull Regression of Misidentification Risk method can be used in estimating identification performance at different sizes. It has given high accuracy in performance estimation.

5. Summary and future work

It is important to estimate algorithm performance at different enrollment sizes, but current approaches have not provided us with an easy and accurate way to do so. We showed in this paper how performance at different sizes could be easily estimated using our proposed regression analysis model of Misidentification Risk. Preliminary results have shown that the Weibull model can accurately represent the distributions of Misidentification Risk from different algorithms and samples, and the approach have given performance estimates close to the empirical measurements. We therefore think that the approach has pointed us to a new way of estimating identification performance.

The experimental results showed only point estimates using a database of around 1200 people. We will continue the research using larger databases to find out whether the regression approach can be applied in general, especially for larger enrollment sizes and enrollments with a completely different group of people. We will also continue to test the robustness of the approach by measuring the width and coverage of confidence intervals.

Acknowledgement

Portions of the research in this paper use the FERET database [11, 10] of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office

References

[1] The csu face identification evaluation system, version 5.0. Technical report, Colorado State University.

[2] W. R. Blischke and D. P. Murthy. *Reliability: modeling, prediction, and optimization*. A Wiley Interscience Publication, John Wiley & Sons Inc, 2000.

[3] J. Daugman. Biometric decision landscapes. technical report. Technical Report TR482, University of Cambridge Computer Laboratory, 2000.

[4] J. Daugman. The importance of being random: statistical principles of iris recognition. *PR*, 36:279–291, 2003.

[5] P. Grother and P. Phillips. Models of large population recognition performance. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 68–94, 2004.

[6] W. H. Ho, P. Watters, and D. Verity. Using an owner-tester setup for face identification/biometrics evaluation. In *(submitted to) proceedings of the 2nd International Conference on Biometrics*, 2007.

[7] A. Johnson, J. Sun, and A. Boick. Using similarity score from a small gallery to estimate recognition performance for large galleries. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 100–103, 2003.

[8] T. Maeda, M. Matsushita, and K. Sasakawa. A performance model for biometrics identification systems. *Systems and Computers in Japan*, 36(11), 2005.

[9] D. P. Murthy, M. Xie, and R. Jiang. *Weibull models*. Wiley Interscience, 2004.

[10] P. Phillips, H. Moon, S. Rizvi, and P. J. Rauss. The feret evaluation methodology for face recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1090–1104, 2000.

[11] P. Phillips, H. Wechsler, J. Huang, and P. Rauss. The feret database and evaluation procedure for face recognition algorithms. *Image and Vision Computing Journal*, 16(5):295–306, 1998.

[12] P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002 evaluation report. Technical report, FRVT, 2003.

[13] Prism. A complete guide to nonlinear regression, graph pad prism. Online Graph Pad Prism, Read 6 Nov 2006.

[14] P. J. Smith. *Analysis of Failure & Survival Data*. Chapman & Hall /CRC, 2000.

[15] R. Wang and B. Bhanu. Learning models for predicting recognition performance. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, 2005.

[16] J. L. Wayman. Error-rate equations for the general biometric system. *IEEE Robotics and Automation Magazine*, 6(1):35–48, 1999.