Evaluating and predicting the performance of an identification face recognition system in an operational setting

Marcus A. Butavicius [a]

Abstract

Techniques for evaluating and predicting the performance of identification face recognition systems from trial data are presented. Using template capture and matching information that reflects the influence of situational variables, predictions can be made across several variables of interest (*e.g.*, alarm threshold, gallery size, number of people viewed). This includes predictions for identifying individuals and groups as well as the number of false alarms for both enrolees and non-enrolees. The function relating the probability of identification and gallery size is dependent on alarm threshold such that changes in threshold exert a greater effect on this probability for smaller gallery sizes.

Keywords: Face recognition, Performance evaluation, Operational setting, Receiver operating characteristic curve

Introduction

Of all the biometric technologies, face recognition (FR) systems offer the promise of a method of identification that is minimally intrusive (Blackburn *et al.* 2003, Wayman 1997a, 1997b). However, rigorous testing and evaluation of such systems in operational environments is required to determine their real-world effectiveness.

Traditionally, face recognition systems have been evaluated in at least one of three ways (Phillips *et al.* 2000a, Bone *et al.* 2001). The first is a *technology evaluation*, which involves investigating a system's performance on a standardised database of images (NIST 2002, Phillips *et al*. 2000b, Rizvi *et al.* 1998). The second method is the *scenario evaluation* (Bone and Blackburn 2002; Holmes *et al.* 1991). This involves testing the entire biometric system in a prototype scenario that is similar to the actual operational setting. The third method is conventionally known as an *operational evaluation*, where the system is set up in the operational environment for a period of time and its performance recorded.

However, these techniques are limited in their usefulness for predicting real world performance. On one hand, the technology evaluation fails to incorporate the influence of real world variables and the simple probabilities conventionally produced using this approach do not directly relate to system output, *i.e.*, the probabilities of correctly or incorrectly identifying a person in an operational scenario. On the other hand, the scenario and operational evaluations only provide descriptive statistics on system performance and do not allow predictions to be made across changes in important system parameters and environmental variables. This paper outlines new techniques for evaluating biometric systems that incorporate both the influence of operational variables from the operational evaluation and the statistical predictability of the technology evaluation.

The data collection methodology required for these statistical techniques is outlined in Sunde *et al.* (2003). Briefly, it involves two sets of controlled trials investigating face capture and face matching with both known and unknown viewed persons. The resulting data consists of:

1) *Face capture probabilities* – the proportion of faces captured at least once by the system in each area / camera combination.
2) *Gallery images* – the sets of face images of known persons of the format likely to be used in the system's gallery (*e.g.*, id photos).

[a] Defence Science and Technology Organisation, Edinburgh, South Australia, 5111, Australia (Email: Marcus.Butavicius@dsto.defence.gov.au)

3) *Captured images* – the sets of face images captured by the system consisting of images of the same people from 2) taken from the face capture trials consisting of at least one image of each person from each operational area / camera combination.

Note that by using face captures and face capture probabilities from actual operational trials, the effects of situated variables such as lighting conditions, movement and pose will be reflected in the statistical predictions (Sunde *et al.* 2003).

Preliminary data analysis

The face matching algorithm in the biometric system is used to make all the pairwise comparisons within a set of images consisting of both *Gallery* and *Captured images*. In systems where score normalisation occurs (*i.e.*, the matching process is adjusted so as to maximise inter-individual differences within the gallery), multiple comparisons involving dummy images may need to be performed to acquire these values. The output of these comparisons is a similarity matrix, *i.e.*, all the pairwise comparisons in the set, where the dependent variable is the final similarity measure used by the system to determine whether an alarm is generated. To demonstrate the concepts in this paper, an artificial data set has been generated where the similarity measure has a range of 0 through 100 where 100 indicates a perfect match.

Depending on the number of different types of photos included in the original gallery, several analyses can be performed using the data from subsets of the matrix. Each analysis involves comparing captured images from one operational area with one type of gallery photo. The statistics produced by the analyses can be used to compare different combinations of area and image types as well as to identify problems in specific areas or photo sets (McLindin *et al.* 2003).

For each analysis, all the similarity measures between two different images of the same person (*e.g.*, passport and live capture) constitute a "same person" distribution while all the similarity measures between the same two image types of different people constitute a "different person" distribution. Ideally, the probability density curves for these two distributions should be well separated - the more they overlap, the greater the trade-off between match and false match probabilities at different thresholds.

One measure of the separability of these two distributions derived from the signal detection theory literature, and used in the evaluation of biometric systems, is *d′* (Daugman and Williams 1996, Tanner and Swets 1954, Bolle *et al.* 2000). This is given by:

$$d' = \frac{\mu_1 - \mu_2}{\sigma}, \qquad (1)$$

where μ_1 and μ_2 are the means of the "same person" and "different person" distributions respectively and σ is the shared standard deviation of these distributions. Here *d′* is a measure of the separability of two equivalent normal distributions (in this case the "same person" and "different person" distributions) and may be considered a measure of robustness of the matching system. The parameter *d′* assumes that the "same" and "different" distributions are normally distributed.

It is not uncommon for the variability of the "same person" distribution to be less than that of the "different person" distribution. In cases where the distributions are normal but with unequal variance, $d^{1/2}$ should be used (Green and Swets 1966, Swets 1964). This is given by:

$$d^{1/2} = \frac{\mu_1 - \mu_2}{\left(\sigma_1^2 + \sigma_2^2\right)^{0.5}}, \qquad (2)$$

where σ_1 and σ_2 are the standard deviations of the "same" and "different person" distributions respectively. A comparison of the magnitude of *d′* and $d^{1/2}$ in different environment and camera combinations may be useful in determining relative performance advantages. However, their use is limited in biometric evaluation because they are descriptive in nature, their assumptions about the underlying distributions are often unwarranted and because of the dimensionality of these measures.

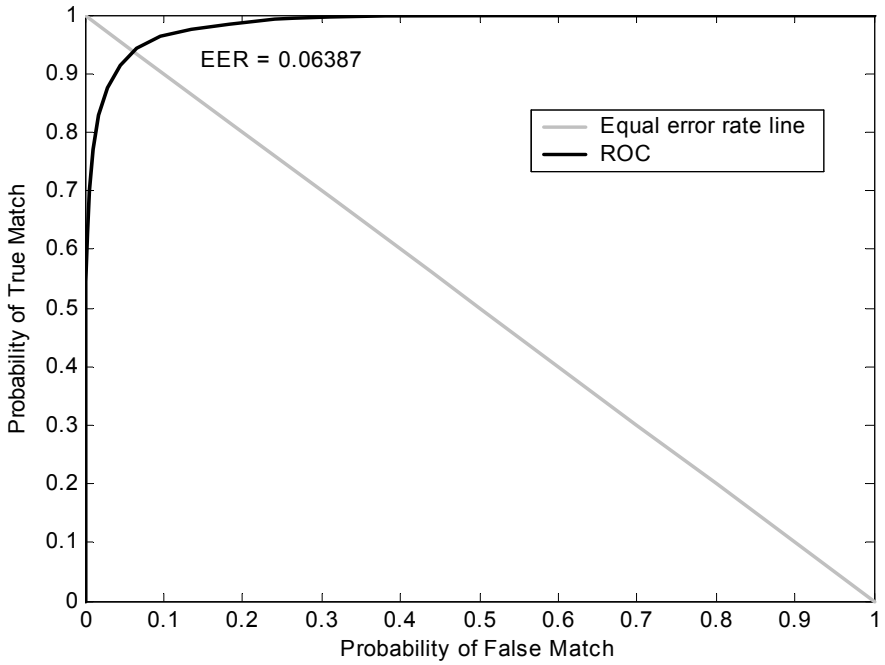

Fig. 1: The Receiver Operating Characteristic (ROC) Curve.

A more detailed analysis requires consideration of basic match probabilities. These can be calculated from the two probability density curves for "same person" and "different person" distributions. Each curve is integrated from the lower bound of threshold values (in this case 0) to the upper bound (in this case 100). For the "same person" distribution this provides the *probability of false nonmatch* ($P_{fnm}(\tau)$), *i.e.*, the probability that the score between a template of an individual and an image of themselves enrolled in the gallery will be below the alarm threshold (τ). The complement of this probability (*i.e.*, 1-$P_{fnm}(\tau)$) provides the *probability of true match* ($P_{tm}(\tau)$). This is the probability that the score between a template of an individual and an image of themselves enrolled in the gallery will be above the alarm threshold.

Integrating the "different person" density function in the same manner provides the *probability of true nonmatch* ($P_{tnm}(\tau)$), *i.e.*, the probability that the score between a template of an individual and an image of a different person enrolled in the gallery will be below τ. The complement of this probability is the *probability of false match* ($P_{fm}(\tau)$). This is the probability that the score between a template of an individual and an image of a different person enrolled in the gallery will be above the threshold.

The effective receiver operating characteristic (EROC) curve

One method of depicting these simple match probabilities is the *Receiver Operating Characteristic* (ROC) curve as displayed in Figure 1. The ROC curve plots the $P_{tm}(\tau)$ on the ordinate against $P_{fm}(\tau)$ on the abscissa. The point at which $P_{tm}(\tau)$ and $P_{fm}(\tau)$ are equal, known as the *Equal Error Rate* (EER), is often used to judge system performance. Generally speaking, the lower this value, the better the system's performance. However, this value is only one indicator of performance and an evaluation of the graph as a whole, especially the function near the ordinate, is necessary for a proper appraisal of system performance. A similar curve also used in the assessment of such systems is the *Detection Error Tradeoff* (DET) curve which plots $P_{fm}(\tau)$ and $P_{fnm}(\tau)$ on log scale axes to highlight performance at higher thresholds.

An additional form of graphical display that combines the conventional ROC curve with the face capture probability, P_c, is proposed. Information from face capture and face match information, where both are linked to

a specific operational scenario, are combined for a more complete overview of system performance. This new graph, called an *Effective Receiver Operating Characteristic* (EROC) curve, is shown in Figure 2.

The probabilities of interest are joint probabilities – under the assumption of independence, these are match probabilities multiplied by the probability of capture. Note that we have assumed independence where the relationship between these probabilities is unknown. This assumption, and the other assumptions of independence in this paper, is worthy of empirical investigation. The *probability of effective true match* ($P_{etm}(\tau)$) is the probability that an individual's template will be captured and that the score between this template and an image of themselves enrolled in the gallery will be above the alarm threshold. The *probability of effective false match* ($P_{efm}(\tau)$) is the probability that an individual's template will be captured and that the score between this template and an image of a different person enrolled in the gallery will be above the threshold. It therefore better captures system performance by incorporating the ability of the system not only to accurately compare images of people to the gallery but also to capture them in the first place.

The EROC curve is bounded by a *capture window* (the dotted grey line). This window reflects the ability of the system to capture a face – the smaller the window, the worse the face capture performance. More specifically, it signifies the upper and lower bounds of overall system performance as defined by P_c. The effective error rate (EEER) is the point where $P_{etm}(\tau)$ and $P_{efm}(\tau)$ are equal. As with the EER, the lower this value, the better the system's performance is in general terms. If the system displays the unlikely face capture probability of 1 then the EROC and ROC curves are equivalent. For any P_c less than 1 the two curves will differ and the EEER will be greater than the EER. Importantly, at any operational threshold where the probability of false alarm is effectively zero, the *probability of effective true match* will be lower than the *probability of true match* predicted by the conventional ROC curve. In other words, using the conventional ROC curve can often provide an overly optimistic view of system performance of an operational biometric in an identification application.

In summary, the EROC curve allows a quick visual assessment of system performance that takes into account match and capture performance. Performance can be analysed in a format similar to the ROC curve and in a manner that distinguishes between the contributions of capture rate (*i.e.*, the size of the window) and effectiveness of the matching algorithm (*i.e.*, the shape of the curve). Breaking performance down into separate components allows the researcher to determine why a system is performing poorly (*i.e.*, whether it is having difficulty capturing faces or matching them) and also to better compare system performance between different scenarios. The EROC is particularly useful when the number of conditions to be examined is large.

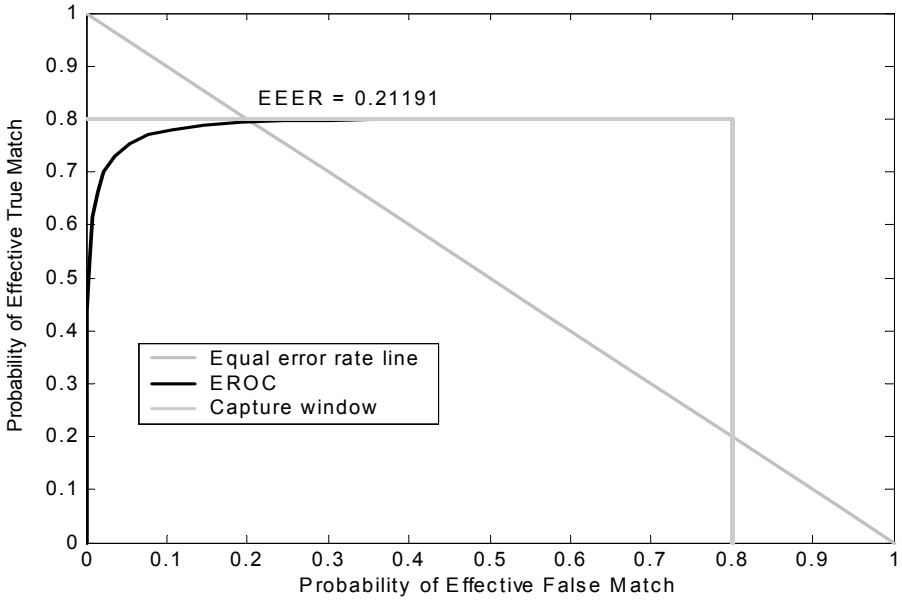

Fig. 2: The Effective Receiver Operating Characteristic (EROC) Curve (P_c = 0.8).

The EROC curve also provides a simple decision heuristic for classifying poor performance, *i.e.*, in cases where an EEER cannot be calculated, system performance is unacceptable. If the capture window does not intersect the equal error rate line then system performance is so poor that even at the most liberal threshold the system does not have a one in two chance of an effective true match. In fact, with a $P_{etm}(\tau)$ of 0.5 the probability of correctly identifying a person of interest, which is the more useful operational metric, is less than 0.5 unless they are the sole enrolee (for further details see the section on *Persons enrolled*).

Higher level measures

While the techniques and measures mentioned above are useful for determining relative performance between competing FR systems and different operational area / camera combinations they do not provide predictions for overall system performance. This section presents methods for providing predictions of higher-level system performance based on the probabilities outlined above. These higher-level measures (*e.g.*, the number of true and false alarms) are more useful in operational terms than the simpler, and often misunderstood, probabilities (*i.e.*, $P_{tm}(\tau)$ and $P_{fm}(\tau)$) that are normally quoted in the biometrics literature. This will involve discussing the formal relationship between these simpler probability estimates and higher-level measures as well as determining confidence bounds for these estimates. This also involves differentiating between equations for probabilities associated with people enrolled in the gallery (signified by r) and those not enrolled (signified by $\neg r$).

Persons not enrolled

As demonstrated in the EROC curve, $P_{efm}(\tau)$ is the probability that an individual's template will be captured and that the score between this template and an image of another person enrolled in the gallery will be above the alarm threshold. This is simply:

$$P_{efm}(\tau) = P_c * P_{fm}(\tau)\,. \qquad (3)$$

However, galleries normally consist of more than one person. For a person not enrolled in the gallery, a false alarm is generated whenever any of the comparisons the algorithm makes between the face capture of an individual and the face images enrolled in the gallery is above the threshold. In other words, at least one of the comparisons with all of the images in the gallery needs to be above the threshold for a false alarm to be generated. Therefore, extending the work of Wayman (1999), the probability of an effective false alarm for a non-enrolled individual compared to a gallery of more than one person is:

$$P_{ef\neg r}(\tau) = P_c * (1 - P_{tnm}^{n_{gs}}(\tau))\,, \qquad (4)$$

where n_{gs} is the gallery size.

The number of false alarms ($fa'_{\neg r}(\tau)$) associated with non-enrolled persons passing by the system can therefore be estimated by the equation

$$fa'_{\neg r}(\tau) = P_c * (1 - P_{tnm}^{n_{gs}}(\tau)) * n_{\neg r}\,, \qquad (5)$$

where $n_{\neg r}$ is the number of non-enrolled persons viewed by the system in the operational environment.

Figure 3 demonstrates the theoretical relationship between the number of people viewed, the number of people in the gallery and the number of false alarms in a scenario where no enrolled persons are viewed by the system. This depicts predictions of false alarm rate for gallery sizes up to 100 and numbers of population not enrolled up to 1000 with a face capture probability of 80% and with a probability of false match of 0.01.

As can be seen in the graph, when the size of the gallery and the number of people viewed by the system is large the number of estimated false alarms may be unreasonable, *i.e.*, potentially exceeding the number which can be dealt with within an operational setting. One way of determining this is to calculate the estimated number of false alarms based on the number of people viewed per hour and then compare this with the number of responses to such alarms that can be handled by an operator, and any other response systems, in this time (see Kaine 2003). Obviously, in most settings the number of people viewed by the system is fixed such that minimising the false alarm rate can only be achieved by lowering the threshold or keeping the gallery as small as possible.

Persons enrolled

The relationship between probabilities in the ROC and EROC curves and the probabilities of true or false alarms (*i.e.*, an alarm given by the system where either the correct gallery image is highest ranked or an incorrect gallery image is highest ranked) is more complicated. For a person enrolled the probability of an effective true match is given by:

$$P_{etm}(\tau) = P_c * P_{tm}(\tau). \quad (6)$$

This is the probability that the comparison score between an individual's captured template and a template of themselves already enrolled in the gallery is above the match threshold. This is an appropriate indication of performance in cases where, whenever an alarm is generated, an operator checks all potential matches that are above the threshold rather than simply the highest match.

Where this is not the case, this approach overestimates the probability of a true alarm. In most applications, a true alarm is the case where not only is the match score between an individual and their template above the threshold but where no false matches generate a higher-ranked similarity score. In other words, the probability of a true alarm for an enrolled individual is the probability that the comparison of their captured face to their image in the gallery is above the threshold and that the similarity score it produces is higher than that produced by all the other comparisons between the captured image and images of other people in the gallery. In what follows three methods are presented to estimate true alarms.

The first of these methods gives upper and lower bounds on the estimated probability of true alarms. The upper bound represents the best case scenario where, whenever a true match is above the threshold, all of the false matches are below the similarity score of the true match. This upper bound is given simply by $P_{tm}(\tau)$. The lower bound is the worst case scenario where, whenever at least one false match occurs, at least one of these false match similarity values is above the true match value. This lower bound is $P_{tm}(\tau) - (1 - (1 - P_{fm}(\tau))^{ngs-1})$ whenever $1 - (1 - P_{fm}(\tau))^{ngs-1}$ is less than or equal to $P_{tm}(\tau)$ and zero otherwise. However, these bounds are only a crude method for estimating true alarms. For large gallery sizes, relatively high false match probabilities (for example, at low thresholds or for a poorly performing system), or both, the lower bound is zero and therefore uninformative. In addition, the upper bound is unaffected by changes in gallery size. Therefore, in practice the usefulness of these suprema is limited.

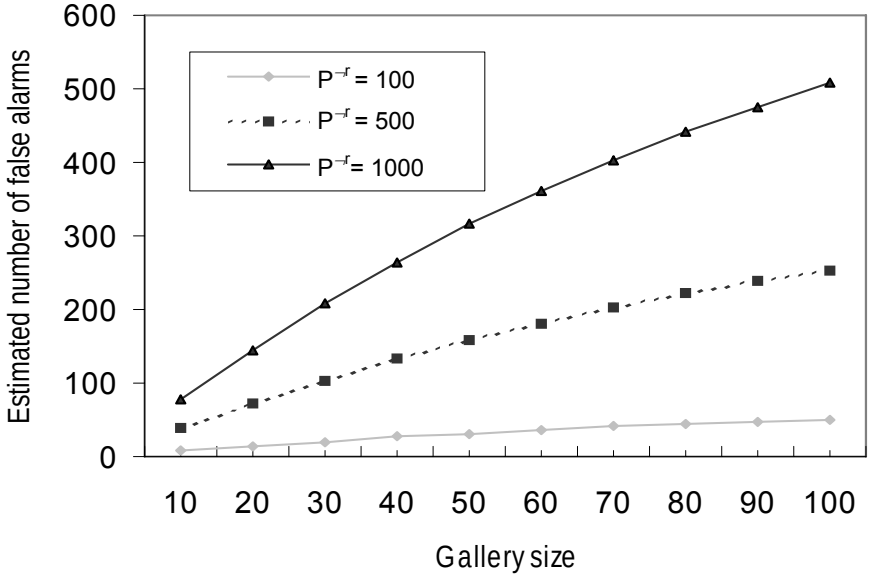

Fig. 3: Estimated number of false alarms as a function of gallery size for three different numbers of non-enrolled people viewed by the system (P_c = 0.8 and $P_{fm}(\tau)$ = 0.01).

The second of these methods gives a conservative estimate using probabilities derived directly from the "same" and "different person" distributions (for a similar treatment see Wayman, 1997c). Using this approach, $P_{etar}(\tau)$ (the probability of an effective true alarm for an enrolee) can be given by the joint probability that a) the face is captured by the system, b) the correct image comparison similarity score is above the threshold and that c) none of the incorrect image comparison similarity scores are above the threshold under the assumption that a), b) and c) are independent. This is given by

$$P_{etar}(\tau) = P_c * P_{tm}(\tau) * P_{tnm}^{n_{gs}-1}(\tau). \quad (7)$$

However, this approach to estimating true alarm probability omits a valid case of a true alarm where at least one of the incorrect image comparison scores is above the threshold but where any such scores are lower than the correct image comparison score. This method is conservative because it underestimates the probability of a true alarm in an operational environment. In some cases, the technique may be appropriate given that, especially at the higher thresholds normally used for surveillance applications, the chances of an incorrect comparison score being above the threshold but lower than the correct comparison score are relatively small when the gallery size is kept to a minimum.

This technique can also be used to determine the probability of a false alarm based on enrolled persons. As with true alarms, this technique underestimates the probability of interest (*i.e.*, false alarms) associated with enrolled persons – it omits the valid case of a false alarm where the correct image comparison score is above the threshold yet below the highest ranked comparison score of the other gallery images. Assuming independence, the false alarm probability for enrolled persons is:

$$P_{far}(\tau) = P_c * (1-(1-P_{fm}(\tau))^{n_{gs}-1}) * P_{fnm}(\tau) \quad (8)$$

The third technique for estimating the probability of a true alarm is the most complicated but also the most accurate. As mentioned at the start of this section, the probability of a true alarm can be expressed as the probability that (a) the true match value is above the threshold and (b) that all of the scores from comparisons between the captured face image and the other gallery images are below the true match value. This can be expressed as the intersection of two components. The first component is the probability that the similarity score between a face captured in the live environment and a gallery image of the person is above the match threshold, *i.e.*, $P_{tm}(\tau)$.

The second component is the probability that a score between an individual's image in the live environment and their gallery image is higher than any of the scores between their image in the live environment and the gallery images of any other individual. Put simply, it is the probability that for any given face image from the live environment, a "same" score will be higher than all "different" scores. This can be expressed as the threshold independent probability $P_{s>d}^{n_{gs}-1}$ where n_{gs} denotes the size of the gallery. Assuming independence, the probability of an effective true alarm is the intersection of these two components and is given simply by $P_{tm}(\tau) * P_{s>d}^{n_{gs}-1}$. Assuming independence between these components and the likelihood of face capture, the probability of an effective true alarm with a gallery of two or more images is

$$P_{etar}(\tau) = P_c * P_{tm}(\tau) * P_{s>d}^{n_{gs}-1}. \quad (9)$$

The probability of a false non-alarm needs similar treatment. A false non-alarm can occur when the face image is and is not captured. In the former case, a person is viewed by the system, their face image is captured but an alarm is not generated by either a comparison with the correct gallery image or any of the incorrect images. In other words, it is the joint probability that the score between the captured image and the correct gallery image is below the threshold and that all of the scores between the captured image and all of the other gallery images are also below the threshold. In the latter case, the person is viewed by the system but the face image is not captured. In both cases the system does not generate an alarm. Assuming independence, the equation for an effective false non-alarm with a gallery of two or more images is

$$P_{efnar} = (1 - P_c) + (P_c * P_{fnm}(\tau) * P_{tnm}^{n_{gs}-1}(\tau)) \quad (10)$$

As with the probability of a true alarm, the probability of a false alarm for an enrolee can be broken down into two components. Assuming independence, it is the probability (a) that at least one of the scores between the captured image and the gallery images of other people is above τ multiplied by the probability (b) that the score associated with the correct gallery image is less than the highest of the scores associated with the false images. The first component is the complement of the probability that none of the scores between the captured image and the gallery images of other people is above the match threshold. For a gallery of two or more images this is $(1 - P_{tnm}^{n_{gs}-1}(\tau))$.

The second component is the probability that a "same person" score is lower than at least one "different person" score. This is the complement of the probability that the "same person" score is above all "different person" scores which can be expressed as the threshold independent probability $1 - P_{s>d}^{n_{gs}-1}$. Assuming independence, the probability of an effective false alarm for an enrolee is therefore:

$$P_{efae}(\tau) = P_c * (1 - P_{tnm}^{n_{gs}-1}(\tau)) * \left(1 - P_{s>d}^{n_{gs}-1}\right) \quad (11)$$

Effect of gallery size on true alarm probabilities

This equation for estimating effective true alarms for enrolees allows us to answer a question frequently asked in the analysis of identification biometric systems: how does varying the gallery size impact on the probability of a true alarm? Figure 4 shows the theoretical influence of gallery size on the probability of true match across three different thresholds from the simulated data. Probability of capture is 0.8 and the thresholds of 60, 65 and 70 are associated with $P_{tm}(\tau)$'s of 0.83, 0.46 and 0.12 respectively. Generally speaking, the function relating true alarms to gallery size is dependent on threshold. More specifically, for any reasonable threshold, the higher the threshold the less sensitive the probability of true alarm is to increasing gallery size.

Detection of a group

Another important operational measure is the ability of the system to detect the activities of a group. For example, consider a group of interest where images of each member are enrolled in the gallery. Let us also assume that to detect a group we need to detect one or more of its members. In other words, when at least one of the members is detected we have been alerted to the possible presence of the entire group. Assuming independence, the probability that at least one member of the group will be detected (*i.e.*, an effective true alarm for at least one person in the group) is given by

$$P_{etag}(\tau) = 1 - (1 - (P_c * P_{tar}(\tau)))^{n_g}, \quad (12)$$

where n_g is the number of people in the group and $P_{tar}(\tau)$ is the probability that an enrolled person will generate a true match alarm. Figure 5 depicts probabilities for detecting groups of people (1 – 20) for three different thresholds – 60, 65 and 70 (producing $P_{tm}(\tau)$'s of 0.88, 0.55 and 0.17 respectively) – a capture rate of 0.8 and a gallery size of 100 members.

As can be seen in Figure 5, even when the performance of the system for individuals is relatively meagre, the cumulative effects for the detection of groups may result in acceptable overall performance. In other words, while the conventional analysis of system performance (*i.e.*, $P_{tm}(\tau)$ and $P_{fm}(\tau)$) may be pessimistic, the actual system performance in operation may be acceptable with respect to the detection of groups.

Confidence Intervals

In addition to the point estimates depicted in the ROC and EROC curves, 95% confidence intervals can also be provided for $P_{tm}(\tau)$ and $P_{fm}(\tau)$ (Crow *et al.* 1960, Wayman 1998). We can then be 95% confident that the unknown population parameter (in this case the match probabilities) is within these bounds based on the samples we have drawn. For $P_{tm}(\tau)$ from the "same person" distribution, the 95% confidence interval at any given threshold is given by:

Upper bound

$$= \frac{r + y + \sqrt{\left((r+y)^2 - n + z^2\right)\frac{r^2}{n}}}{n + z^2}$$

(13)

Lower bound

$$= \frac{r + y - \sqrt{\left((r+y)^2 - n + z^2\right)\frac{r^2}{n}}}{n + z^2} \quad (14)$$

where r = the number of "same person" scores above the threshold and
n = the total number of "same person" scores,

z = normal deviate and $y = \frac{z^2}{2}$.

For the "different person" distribution, a different approach is required because the cross-comparisons are no longer independent. From Wayman (1998), the

95% confidence limits for $P_{fm}(\tau)$ at a given threshold are:

Upper bound = $P_{fm}(\tau) + 1.96 * \left(\frac{\hat{\sigma}(\tau)}{\sqrt{N}}\right)$ (15)

Lower bound = $P_{fm}(\tau) - 1.96 * \left(\frac{\hat{\sigma}(\tau)}{\sqrt{N}}\right)$ (16)

where

$$\hat{\sigma}^2 = \frac{1}{N(N-1)^2}\sum_{h=1}^{N}\left(\sum_{k\neq h} r(h,k) + \sum_{k\neq h} r(k,h)\right)^2 - 4 * \left(P_{fm}(\tau)\right)^2$$

where $r(h,k)$ is the similarity score between person h and person k in the sample.

In general, confidence bounds are rarely reported in biometric system evaluations. One reason is that the confidence bounds relate only to error in sampling from the test population in the test environment and both the participants and the environment used in testing often differ from the operational scenario (Wayman 1998). In this evaluation methodology we have minimised

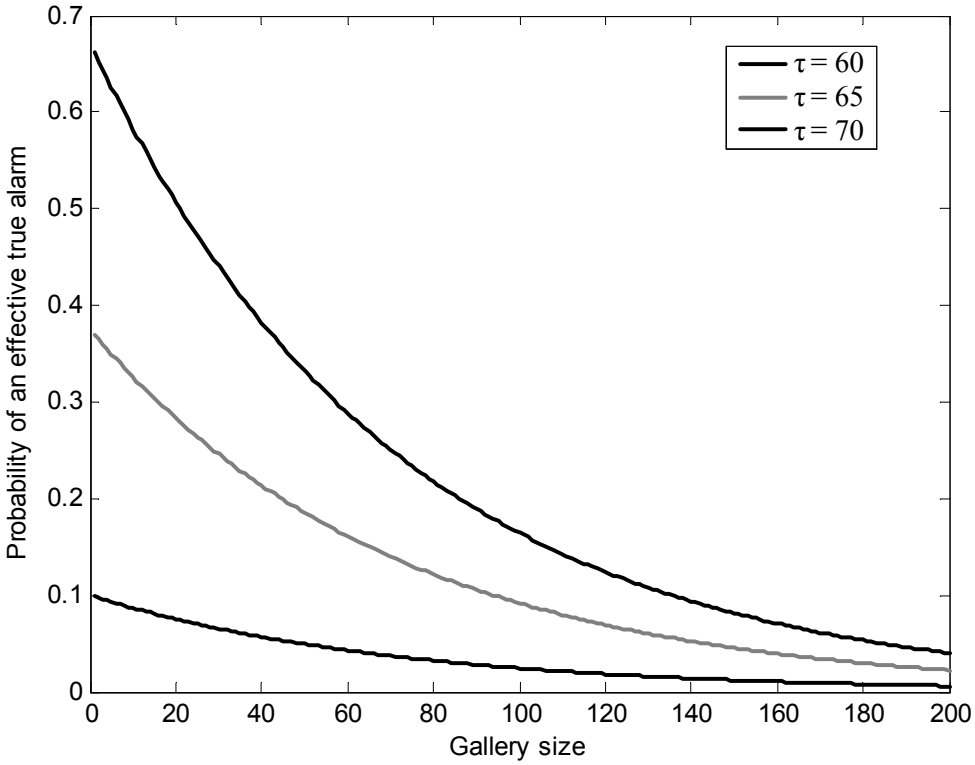

Fig. 4: Probability of an effective true alarm for three different thresholds (55,60 and 65) and for gallery sizes 1 through 200.

environmental differences between testing and deployment (Sunde *et al.* 2003). However, when the test participants in this methodology are not representative of the population that ultimately will be viewed by the system such confidence bounds, and in fact the point estimates themselves, may be of limited benefit.

Conclusion

The new techniques presented for evaluating and predicting the performance of an identification face recognition system in an operational setting have combined elements of signal detection theory and basic probability theory. They allow superior predictions of performance over conventional biometric system evaluations because they a) reflect the influence of situational variables such as lighting, pose and movement and b) allow for variation in a larger set of parameters such as threshold, gallery size, target type (enrolee vs. non-enrolee, group vs. individual) and the number of people viewed by the system. In addition, the EROC curve that has been presented allows a visual assessment of system performance both in terms of face capture and face match in an operational setting. This technique may be of assistance in quickly determining optimal areas and / or system out of a number of possible alternatives as well as identifying causes of poor system performance.

In addition, there are two significant practical implications from this work. Firstly, the ability to detect a group, which is a useful operational measure in certain contexts, may be adequate even when the ability to detect an individual is poor. Secondly, gallery size has a significant impact on system performance. For effective false alarms, the effect of increasing gallery size can paint a pessimistic view of system performance for the parameters used in many operational settings. For effective true alarms, the impact of increasing gallery size is dependent on the threshold setting such that the lower the threshold, the more sensitive the probability is to increasing gallery size.

The techniques provided in this paper are provided with the following caveats:

1. This approach does not specifically address variations in system performance due to multiple captures of a person in an environment or multiple enrolment images of a person. This would require separation of the variance between different people from the

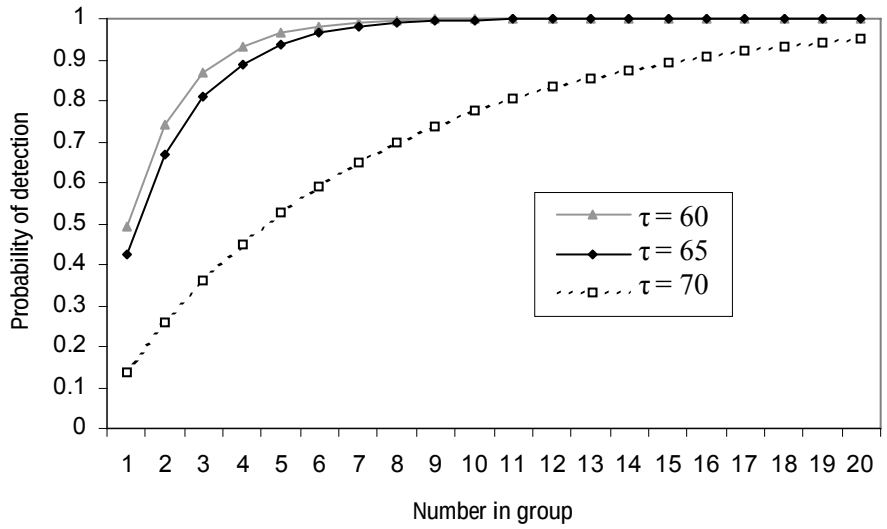

Fig. 5: Probabilities for detecting groups (P_c= 0.8, n_{gs} = 100).

cameras for implementing the biometric variance between images of the

same person such as in an analysis of variance, multivariate analysis of variance or mixed-effects models.

2. It assumes that the system compares the template from the environment to every gallery image in a single pass. In systems where multiple comparison stages are used in the matching process the analysis will need to be more sophisticated.
3. For valid results, this technique requires a relatively large database of images and participants for both face capture and face matching trials. Practically speaking, the requirement of a large number of participants is both time consuming and costly. However the results are more appropriate than those proposed in alternative evaluation methodologies and any such costs may be insignificant in comparison to those caused by the inappropriate implementation of a face recognition system in an operational setting.
4. The participants in the trials need to be representative of the population that will actually be viewed by the system. Where this is not the case, the predictive power of these techniques is limited.
5. The assumptions of independence need to be examined. For example, the relationship between face capture and face matching should be empirically investigated. Given the possible interaction between influencing factors in an operational environment, it should be noted that any correlations may be linked not only to the system itself but the environment and sample upon which the testing was conducted so that any conclusions may be limited in their generality.
6. This approach has neglected an important aspect of the system performance, namely human-machine interaction. The final link in the decision chain is the operator because ultimate judgment on the correctness of an alarm is a human one. This may involve an operator viewing the images presented by the system, face-to-face recognition and trust in the system's output (see Vast and Butavicius 2005 and Lee, Vast and Butavicius, in press). It may also involve the activity of further response systems (Kaine 2003). Therefore, the evaluation methodology presented in this paper only investigates one part of the complete operational system (Sunde *et al.* 2003).

Acknowledgements

Thanks to Jadranka Sunde, Charles Pearce, Stephen Bourn, Ray Johnson, Ian Graves, Neville Curtis, Brandon Pincombe, Vladimir Ivancevic, Chris Woodruff and Ted Dunstone. A portion of this paper is contained in a draft DSTO report written by M.A. Butavicius, S. Bourn and N. Curtis.

References

[1] T. Blackburn, M.A. Butavicius, I. Graves, D. Hemming, V. Ivancevic, R. Johnson, A. Kaine, B. McLindin, K. Meaney, B. Smith and J. Sunde, "Biometrics technology review 2002," *DSTO General Document Series,* vol. 359, pp. 1-42, 2003.

[2] R.M. Bolle, S. Pankanti and N.K. Ratha, "Evaluation techniques for biometrics-based authentification systems (FRR)," *Proc. of International Conference on Pattern Recognition,* pp. 2831-2837, 2000.

[3] M. Bone and D.M Blackburn, "Face recognition at a chokepoint: scenario evaluation results," *DoD Counterdrug Technology Development Program Office Report,* 2002.

[4] M. Bone, J.L. Wayman and D.M. Blackburn, "Evaluating facial recognition technology for drug control applications," *Proc. of ONDCP International Counterdrug Technology Symposium,* pp. 1-17, 2001.

[5] E. Crow, F.A. Davis and M.W. Maxfield, *Statistics manual,* Dover, New York, 1960.

[6] J.G. Daugman and G.O. Williams, "A proposed standard for biometric decidability," *Proc. of CardTech/SecureTech,* pp. 223-234, 1996.

[7] D.M. Green and J.A. Swets, *Signal detection theory and psychophysics,* Wiley, New York, 1966.

[8] J.P. Holmes, L.J. Wright and R.L. Maxwell, "A performance evaluation of biometric identification devices," *Sandia*

National Laboratories Report, SAND91-0276 (UC – 906), 1991.

[9] A.K. Kaine, "The impact of facial recognition systems on business practices within an operational setting," *Proc. of ITI'03,* 315-320, 2003.

[10] B.A. McLindin, M.A. Butavicius and K. Meaney, "Gallery Image Effects on Facial Recognition Systems," *Proc. of EC-VIP,* pp. 445-460, 2003.

[11] NIST, "Summary of NIST standards for biometric accuracy, tamper resistance, and interoperability," *National Institute of Standards and Technology Report,* 2002.

[12] P. J. Phillips, A. Martin, C. L. Wilson and M. Przybocki, "An introduction to evaluating biometric systems," *IEEE Computer,* vol. 33(2), pp. 56-63, 2000a.

[13] P.J. Phillips, H. Moon, S.A. Rizvi and P.J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22(10), pp. 1090-1104, 2000b.

[14] S. Rizvi, P.J. Phillips and H. Moon, "The feret verification testing protocol for face recognition algorithms," *National Institute of Standards and Technology Report,* NISTIR 6281, 1998.

[15] J. Sunde, M. Butavicius, I. Graves, D. Hemming, V. Ivancevic, R. Johnson, A. Kaine, B. McLindin, K. Meaney and J. Sunde, "A methodology for evaluating the operational effectiveness of facial recognition systems," *Proc. of EC-VIP,* pp. 441-448, 2003.

[16] J.A. Swets, *Signal detection and recognition by human observers,* Wiley, New York, 1964.

[17] W.P. Tanner and J.A. Swets, "A decision-making theory of visual detection," *Psychological Review,* vol. 61, pp. 401-409, 1954.

[18] R.L. Vast and M.A. Butavicius, "A literature review of face recognition for access control: human versus machine solutions," *DSTO Technical Document Series,* vol. 1747, pp. 1-35, 2005.

[19] R.L. Vast, M. D. Lee and M.A. Butavicius, "Face matching under time pressure and task demands," *Proc. Of Cognitive Science,* in press.

[20] J.L. Wayman, "Benchmarking large-scale biometric identity systems," *Proc. Of CardTech/SecurTech,* pp. 314-331, 1997a.

[21] J.L. Wayman, "The science of biometric technologies: classifying, testing, evaluating and selecting," *Proc. of CardTech/SecurTech,* pp. 385-396, 1997b.

[22] J.L. Wayman, "A scientific approach to evaluating biometric systems using a mathematical methodology," *Proc. of CardTech/SecureTech,* pp. 477-492, 1997c.

[23] J.L. Wayman, "Technical testing and evaluation of biometric identification devices," in A. Jain, R. Bolle and S. Pankanti (eds.), *Biometrics: information security in a networked society,* Kluwer, Norwell, pp. 345-368, 1998.

[24] J.L. Wayman, "Error-rate equations for the general biometric system," *IEEE Robotics and Automation Magazine,* vol. 3, pp. 35-48, 1999.