

---

## Method to Evaluate Pose Variability in Automatic Face Recognition Performance

---

### Yednek Asfaw\*

Systems and Computer Engineering, Carleton University,  
Ottawa, Ontario, Canada

E-mail: yasfaw@sce.carleton.ca

\*Corresponding author

### Guy Scott

Citizenship and Immigration Canada,  
Ottawa, Ontario, Canada

E-mail: Guy.Scott@cic.gc.ca

### Paul Pelletier

Citizenship and Immigration Canada,  
Ottawa, Ontario, Canada

E-mail: Paul.Pelletier@cic.gc.ca

### Andy Adler

Systems and Computer Engineering, Carleton University,  
Ottawa, Ontario, Canada

E-mail: adler@sce.carleton.ca

**Abstract:** Automatic face recognition technology allows verification and identification of individuals from photographic face images; this technology has important applications for verification of identity documents. Such technology has improved dramatically over the past decade, to the point where face images may be used for identification in large databases with relatively low error rates. One key concern for such government applications is the extent to which recognition performance degrades as the quality of images decreases. This paper introduces a method to evaluate the impact of variability in face pose on face recognition accuracy. For each pair of images of a given pose difference, a genuine distribution was calculated, while the impostor distribution was calculated from all non-matching images. Confidence intervals were determined using non-parametric bootstrapping. Experiments were conducted with volunteers who were asked to assume specific poses from neutral pose to  $\pm 20$  deg in each of the roll, pitch, and yaw directions. Data were analyzed to determine changes in recognition performance, using three leading commercial face recognition algorithms. Results indicate that roll variations made

a relatively small effect on performance, while pitch and yaw variations produce a large and significant increase in error rates. More recent algorithms show better results at low pose variability, and thus are relatively more sensitive to pose changes.

**Keywords:** Automatic Face Recognition; Receiver Operator Curve; Biometric Sample Quality; Biometric Performance Analysis

## Reference

---

## 1 Introduction

This paper introduces a method to evaluate the effect of variability in face pose on the accuracy of face recognition biometric algorithms. Automatic face recognition (AFR) technology allows verification and identification of individuals from photographic images; a live image of a person may be compared to previously acquired photos, or two photos may be compared to determine if they represent the same person. AFR has seen very active research and development since the early 1990's(16) and has shown dramatic improvements in accuracy over this period (1; 13; 12).

AFR is well suited to the requirements of government ID documents such as passports, visas and driver's licenses. Such documents are printed with an image of the holder's face which is acquired as part of the issuance process. On presentation of the document, the presenter needs to be verified as the legitimate holder by comparing against the image acquired at issuance time. While other biometric features, such as fingerprint and iris images are also commonly stored on identification documents, AFR has especial value because: 1) historic databases of face images are already maintained by governments, and 2) face images form a natural way to perform identity verification with which the general population is already comfortable. A similar conclusion was reached by the international civil aviation organization (ICAO) which recommends that face be used as the primary biometric feature for passport documents(8). Biometric performance of AFR is generally understood to be poorer than that of biometric features from fingerprint or iris images (although recent work(12) suggests that is not true for very high resolution face images). AFR performance (in terms of error rates) is best when care is taken to acquire high quality images with uniform illumination and pose. Thus, in order to maintain good biometric performance, it is important to ensure high quality images are enrolled. Recommendations for photo capture and image quality are part of the ISO standard for AFR(9). Additionally, several countries have established standards and guidelines for passport and visa photo capture, specifically designed to support AFR performance(11). On the other hand, poor quality images dramatically increase AFR error rates(4). Such biometric image quality degradation may be classified in terms of character (inherent features), fidelity (accuracy of features), or utility (predicted biometrics performance)(10). Degradation due to poor capture or camera settings is classified as low fidelity

quality: poor lighting, contrast, glare, low resolution or excessive compression. Variability in the face pose, or due to expression, or changes due to makeup and facial hair are classified as low character quality. The utility quality may then be calculated from analysis of test data from quality changes. Variability in pose changes the geometric relationship between face features; comparison of images of the same face at different poses will thus affect the biometric matching due to these perspective differences. We note that one approach that attempts to address pose variability is 3D AFR, in which special cameras capture range information in addition to the face images, from which the 3D shape of the face may be represented(2). However, since such an approach requires special cameras, it represents a large infrastructure change for identity document issuance, and was not considered feasible at this time.

The study of biometric image quality and biometric performance has mainly focused on developing improved face recognition algorithms (5; 6; 7). Furthermore, literature on biometric performance comparison is mostly between AFR engines and humans(1; 3). This paper develops a method to track and objectively evaluate biometric performance in relation to pose variability using photographs captured for passport and visa documents. The method can be used to compare different versions AFR and re-evaluate the pose angle constraints in place for face capture.

Many AFR systems are designed for a closed identification, open identification (watchlist) or verification. In the verification problem a subject submits a biometric sample or probe, along with a claim of identity. The probe is matched against the claim's corresponding enrollment sample. The claim is accepted if it passes a threshold. The performance of this one-to-one problem is analyzed using the ROC curve. The ROC curve is generated by comparing a set of probes against a gallery with the assumption that the ground truth is known. The resulting ROC curve is independent of the gallery size used because it is a measure of one-to-one performance.

In the identification problem, no claim is submitted with the probe sample; the application must determine the identity of the subject. The identification is defined in two ways: open identification in which a probe does not necessarily have a genuine match in the gallery or closed identification where every probe has a corresponding enrollment sample in the gallery of arbitrary size. Typically, open identification performance is measured using an Alarm curve which is constructed like an ROC curve except that for each probe there is a single imposter contribution to the false alarm score distribution.

Hube (17) has shown that Alarm curves for arbitrary size can be estimated to first order by an ROC curve. Based on this assumption, the method presented in this paper will be demonstrated using the ROC curve which can be extended to Alarm curves. Furthermore, the method is applied to a sample set that is representative of the target population where each sample has a unique biometric template and each probe has a set of genuine matches in the gallery.

This work was conducted as part of the development of the Canadian contribution to the ISO face recognition standard(9), in which a need was identified to specify the acceptable limits for pose variability in photographs captured for passport and visa documents. In the remainder of this paper, we describe the method (sec. 2 and sec. 3), photo capture protocol(sec. 4), and the results obtained(sec. 5).

## 2 Face Recognition

The AFR is designed to handle all application modes: verification, identification and watchlist. However, the method proposed in this paper deals with evaluation of pose variability in AFR using ROC curves; therefore the AFR is used in verification application mode. The method is applied on a face image database of size  $N$ . The database contains images of several subjects each with different pose angles. Each image is analyzed to determine the position and size of the face followed by locating the center of the eyes. The image is then analyzed to determine if the quality is sufficient for AFR. If the image passes the quality check, it is then normalized to create a token image. The token image is then preprocessed using histogram equalization and intensity normalization. From the preprocessed token image relevant features are extracted. These extracted features constitute the biometric template. Using each AFR algorithm, a biometric template  $\mathbf{T}_i^A$  was generated for algorithm,  $A$  and image,  $i$ . Based on these templates, the complete similarity score comparison matrix,  $\mathbf{S}^A$  of size  $N \times N$ , was generated where  $\mathbf{S}_{i,j}^A$  represents the similarity score generated by algorithm  $A$  between  $\mathbf{T}_i^A$  and  $\mathbf{T}_j^A$ . In order to study the biometric performance of algorithm  $A$  for a specified pose variability, the following procedure was used from  $\mathbf{S}^A$  to generate the genuine and impostor distributions from which ROC curves could be calculated.

### 2.1 Impostor Distribution

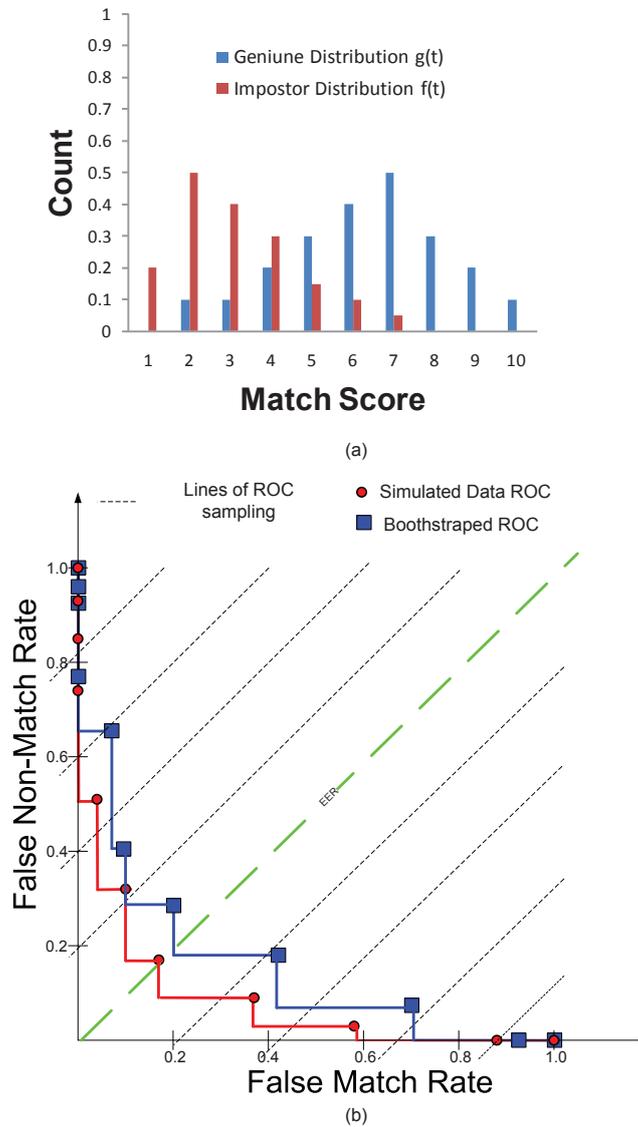
The impostor distribution,  $F^A$ , was taken to be function of the AFR algorithm only, and not the pose variability.  $F^A$  represents the distribution of all similarity scores  $\mathbf{S}_{i,j}^A$ , in which templates  $\mathbf{T}_i^A$  and  $\mathbf{T}_j^A$  are not from the same person. If we define  $\mathbf{X}$  for the match scores from the mated pairs with length  $P$ ,  $F^A$  is expressed as:

$$F^A(t) = \frac{1}{P} \sum_{p=1}^P 1(X_p \leq t) \quad (1)$$

### 2.2 Genuine Distributions

For each subject  $s$ , the template from image with either roll, yaw, and pitch angle  $\theta$  was defined as  $R_s(\theta)$ ,  $Y_s(\theta)$ ,  $P_s(\theta)$ , respectively. Therefore, a genuine distribution,  $G_{\Delta}^A$ , was calculated for each AFR algorithm and pose difference, where  $\Delta$  is pose angle difference. Thus,  $G_{\Delta R=10 \text{ deg}}^A$  represents the distribution of similarity scores in which the *roll* angle differs by 10 deg for the algorithm.  $G_{\Delta}^A$  is calculated from all similarity scores  $\mathbf{S}_{i,j}^A$  for which images  $i$  and  $j$  represent the same person *and* the pose angle difference is  $\Delta$ . For example, for  $\Delta R = 10$  deg, the comparison between images  $R_s(-10 \text{ deg})$  and  $R_s(0 \text{ deg})$  would be included, as would the comparison between images  $R_s(+5 \text{ deg})$  and  $R_s(+15 \text{ deg})$ . If we define  $\mathbf{Y}$  for the match scores from the non-mated pairs with length  $Q$ ,  $G_{\Delta}^A$  is expressed as:

$$G_{\Delta}^A(t) = \frac{1}{Q} \sum_{q=0}^Q 1(Y_q \leq t) \quad (2)$$



**Figure 1** (a) Genuine and Impostor distribution of simulated data. (b) DET (FMR vs. FNMR) for simulated data (circle) and bootstrapped data (square). Green line shows the EER line. Thin lines (parallel to the EER) along which the bootstrapped distribution is calculated. Note that the sampling does not interpolate between data points. The samples connected with quantization steps.

### 3 Statistical Analysis

In order to statistically distinguish variability in ROC curves between different pose variations, it is necessary to have a measure of the distribution an ROC curve.

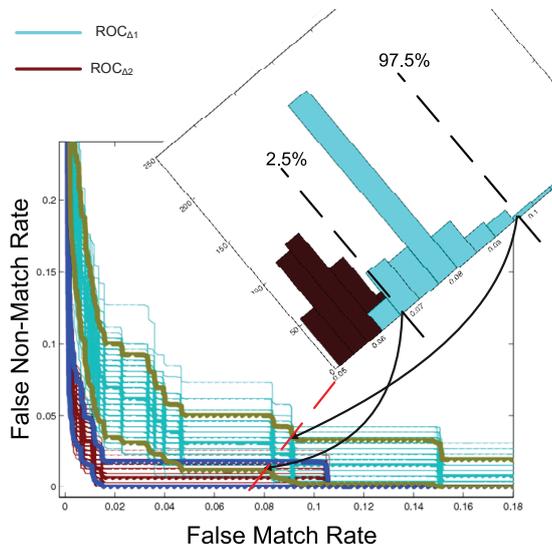
Hence for each algorithm  $A$  and pose difference  $\Delta$ , a curve,  $\text{ROC}_{\Delta}^A$ , was calculated from distributions  $F^A$  and  $G_{\Delta}^A$ . For similarity score value ( $\tau$ ), a FMR (false match rate) and FNMR (false non match rate) was calculated as:

$$\begin{aligned} \text{FMR}(\tau) &= \int_{\tau}^{\infty} F^A(t) dt \\ \text{FNMR}(\tau) &= \int_{-\infty}^{\tau^-} G_{\Delta}^A(t) dt \end{aligned} \quad (3)$$

Where a match score equal or greater than threshold is accepted as true match. Similar to all other biometric studies, the measure of FMR is more accurate than the FNMR since there are more negative samples than positive samples. Each ROC curve was used to study the impact of 5, 10, 15, 20, and 25 degrees roll, pitch and yaw angle variability on the biometric performance. Additionally, the change in performance for various algorithms over the years can be quantified. Several ways to measure this distribution have been proposed(1; 15; 14); we use an analysis of the pointwise ROC using a linear sweep methodology, based on the radial sweep method of (14). Given a single pose variation, and distributions,  $F^A$  and  $G_{\Delta}^A$ , we calculate re-sampled and bootstrapped ROC curves as illustrated in Fig. 1. The figure shows a variant of the ROC known as the DET curve which is better suited to describe the radial sweep method and sampling. The bootstrapped curve is created by defining new distributions  $F^{A*}$  and  $G_{\Delta}^{A*}$ , using nonparametric bootstrapping by sampling each original distribution with replacement. For clarity, the figure shows single bootstrapped ROC curve and the original ROC curve. The figure only shows a single bootstrapped ROC curve for illustrations. In real experiments, we use 1000 bootstrapped curves to determine the confidence interval for each ROC. The radial sweep method transforms each curve from the (FMR, FNMR) space to polar coordinates. For any point,  $(x,y)$ , on a ROC curve, we calculate an angle,  $\theta$  and distance  $r$  from a center point  $(c_x, c_y)$ .

### 3.1 ROC confidence intervals

The pointwise confidence for each ROC is determined as follows. For each curve in Fig. 1, we draw diagonal lines parallel to the EER (nine linear are shown). Along each line, the intersection with each bootstrapped ROC is calculated. Using these points, we are able to calculate a distribution of intersection points to the ROC. From this distribution, the 95% confidence interval is determined by finding the 2.5% and 97.5% percentile distribution point as shown in Fig. 2. These two points indicate the pointwise confidence interval of the ROC along the current diagonal line. In order to determine the complete confidence interval, we repeat this procedure for each diagonal line. The use of diagonal lines is equivalent to a choice of  $c = \infty$  in the radial sweep approach of (1). Our choice is motivated by trying to keep the confidence interval detection perpendicular to the ROC itself. Using this approach, we are able to calculate a 95% confidence interval for each ROC. These confidence intervals are then plotted in our results to graphically represent the change in algorithm performance with changes in pose.



**Figure 2** Determining Statistical Significance: The distribution along the sampling line (Red) is used to calculate the confidence interval of  $ROC_{\Delta_1}$  and  $ROC_{\Delta_2}$ . From this distribution (insert figure), the 95% confidence interval is determined by finding the 2.5% and 97.5% percentile distribution point for the two ROC's.

### 3.2 Significance of ROC differences

Finally, we wish to be able to detect, given two ROC curves,  $ROC_{\Delta_1}^A$ , and  $ROC_{\Delta_2}^A$ , for the same AFR algorithm,  $A$ , and different pose angles  $\Delta_1$  and  $\Delta_2$ , whether: 1) they are different, and 2) the statistical significance ( $p$ ) of the difference. We perform this test on the confidence intervals calculated above as follows. ROC differences are determined by iteratively sampling a random point from the distributions of  $\Delta_1 = ROC_{\Delta_1}^A$ , and  $\Delta_2 = ROC_{\Delta_2}^A$ , on a random choice of diagonal line, and comparing their magnitude, see Fig. 2. From these values  $p(\Delta_1 < \Delta_2)$  is calculated from the fraction of sampled of points which meet the criteria. This  $p$  value may be interpreted as follows: if  $p(\Delta_1 < \Delta_2) = 0.5$ , the mean of the ROC curves cannot be distinguished statistically. We consider the curves to be statistically distinct at the 5% level, which corresponds to  $p$  values of  $p < 0.025$  or  $p > 0.975$ .

## 4 Photo Capture

Photographic image acquisition was performed by Citizenship and Immigration Canada (CIC); the photo capture and privacy protocol was reviewed and approved by a departmental review board, and all participants provided written consent. The photo acquisition protocol was designed to acquire images of each subject at accurate pose angle measurements. For each subject, images at different roll, pitch and yaw angles were captured to provide a total of 38 images. Additional images were captured with/without glasses, with/without hair partially covering eyes, and



**Figure 3** Photo capture configuration: A subject in front of the photo capture and pose calibration apparatus. A camera tripod (shown) is placed at 2.0 m from the subject. For pose measurement, the subject wears a custom cap calibrated to point vertically. Each pose variation image is based on target lines as described.

with varied facial hair. However, these additional images were not analyzed as part of the pose results of this paper. The image capture apparatus is shown in Fig. 3. A subject sits in front of a calibrated board wearing a cap with a vertical wire indicator (not shown in the figure). The cap was designed from felt pipe-cleaners to obscure only a small fraction of the head while providing a tight fit. With the subject sitting in the baseline pose (full frontal with 0 deg roll, yaw and pitch) the wire indicator on the cap was fixed to be vertical. Since we were not studying the effect of image quality in this protocol, all images were designed to be captured at high resolution with good quality lighting. The choice to use only the data captured in this study, in spite of the small size, stems from the need to keep all other factors that contribute to data variability to a minimum. This approach guaranties changes observed in the results will be due to change in pose angle and algorithm. Images were captured with an HP Photosmart 707 camera with an HP 8mm–24mm, 24× zoom lens set to capture 5.1 megapixel images of size  $1108 \times 1034$  pixels (Hewlett-Packard Company, CA) The camera was positioned at 2.0 m from the subject at which position it captured a field of view of 0.65m (vertical)  $\times$  0.6m (horizontal). For the average subject, this gave 190 pixels between the eyes, which easily meets the ISO recommended best practice of 120 pixels between the eyes(9). Image acquisition was performed in a single session on the same day. Angles of *roll*, *yaw* and *pitch* are defined in correspondence with the aeronautical usage, as rotation about, respectively, the longitudinal axis, vertical axis, and an axis perpendicular to the longitudinal plane of symmetry. Neutral angles (0 deg) were defined with the face vertical pointed flat toward the camera.

#### 4.1 Roll Tests

Images were acquired for *roll* angles of  $-20$  deg to  $+20$  deg in increments of 5 deg with neutral *yaw* and *pitch*. The 0 deg image corresponds to the baseline (neutral pose) acquisition. For an accurate capture of pose with each roll angle, a white flip chart paper with a center perpendicular line identifying the 90 deg mark and lines at 5 deg interval was set as background (Fig. 3). Each subject was required to wear a cap constructed of wire pipe cleaners pointing straight up. The crown was centered to make sure it was parallel to the 90 deg angle. Subjects were then asked to roll their head at a five degrees increment. The crown was then removed and the image taken. Images were captured at 5 deg, 10 deg, 15 deg and 20 deg degrees head roll both on the right and left side. A white flip chart paper covered this chart when the photos were taken.

#### 4.2 Yaw Tests

Images were acquired for *yaw* angles of  $-20$  deg to  $+20$  deg in increments of 5 deg with neutral *roll* and *pitch*. The 0 deg image corresponds to the baseline (neutral pose) acquisition. A wood dowel was used to measure the yaw angle. Pipe cleaner indicators were placed at each 5 deg increment on both sides of the center on a horizontal wood dowel that was attached to the camera tripod. In order to accurately measure the 5 deg increment a protractor and a laser diode pen were used.

#### 4.3 Pitch Tests

Images were acquired for *pitch* angles of  $-20$  deg to  $+20$  deg in increments of 5 deg with neutral *roll* and *yaw*. The 0 deg image corresponds to the baseline (neutral pose) acquisition. Another wood dowel was used to measure the head pitch of the models. Again, pipe cleaners were placed at each 5 deg increment on both side of the center of the vertical wooden dowel that was attached to the camera tripod. Again, in order to accurately measure increment a protractor and a laser diode pen were used.

## 5 Results

Using the images captured (Sec. 4) biometric comparisons were performed with three commercially available AFR software algorithms. These algorithms are widely considered to be amongst the top performers in recent technology evaluation tests(12). The highest performing AFR algorithms available to us in each of the years 2006, 2007, and 2008 were used, in order to measure the progress of AFR technology. For each algorithm, software parameters were set to maximize comparison accuracy at the expense of template size and processing speed. In total, 22 subjects (11 male and 11 female) agreed to participate the facial recognition experiment; the average age was 37 years with a range of 25–56. The number of comparisons used in the the impostor distribution was  $6.8 \times 10^5$ . The number of comparisons used in each genuine distribution varied, with an average of  $121 \pm 41$ . For each algorithm, we calculated ROC curves and their pointwise 95% confidence

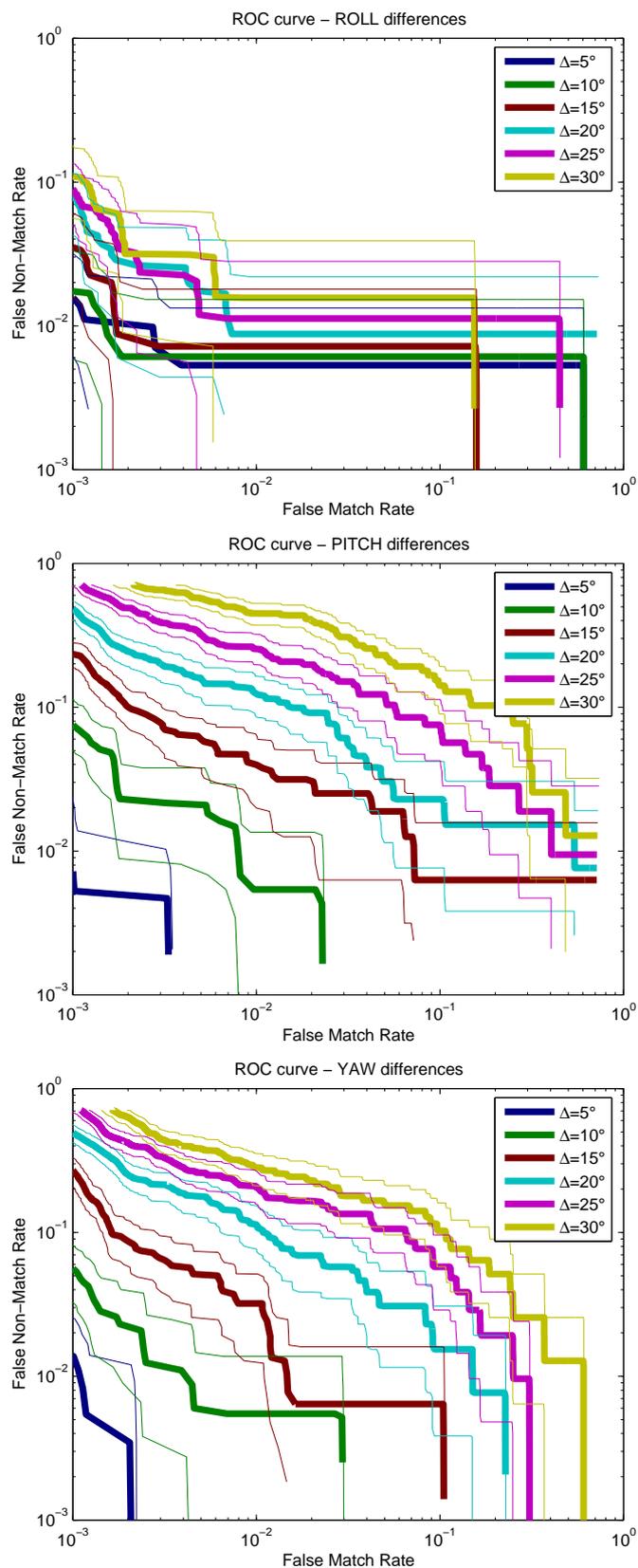
**Table 1** Significance  $p$  of ROC differences

2006 Algorithm	Pitch	Roll	Yaw
$p(\Delta_{5 \text{ deg}} < \Delta_{10 \text{ deg}})$	.0000	.1838	.0000
$p(\Delta_{10 \text{ deg}} < \Delta_{15 \text{ deg}})$	.0002	.2995	.0024
$p(\Delta_{15 \text{ deg}} < \Delta_{20 \text{ deg}})$	.1118	.0797	.0008
$p(\Delta_{20 \text{ deg}} < \Delta_{25 \text{ deg}})$	.1740	.3976	.0116
$p(\Delta_{25 \text{ deg}} < \Delta_{30 \text{ deg}})$	.1012	.2705	.0256
2007 Algorithm	Pitch	Roll	Yaw
$p(\Delta_{5 \text{ deg}} < \Delta_{10 \text{ deg}})$	.1975	.2832	.1847
$p(\Delta_{10 \text{ deg}} < \Delta_{15 \text{ deg}})$	.1176	.2164	.0638
$p(\Delta_{15 \text{ deg}} < \Delta_{20 \text{ deg}})$	.0394	.2168	.0000
$p(\Delta_{20 \text{ deg}} < \Delta_{25 \text{ deg}})$	.0560	.2049	.0128
$p(\Delta_{25 \text{ deg}} < \Delta_{30 \text{ deg}})$	.0771	.2280	.1017
2008 Algorithm	Pitch	Roll	Yaw
$p(\Delta_{5 \text{ deg}} < \Delta_{10 \text{ deg}})$	.2205	.2797	.2352
$p(\Delta_{10 \text{ deg}} < \Delta_{15 \text{ deg}})$	.1633	.2349	.1494
$p(\Delta_{15 \text{ deg}} < \Delta_{20 \text{ deg}})$	.1251	.2339	.0025
$p(\Delta_{20 \text{ deg}} < \Delta_{25 \text{ deg}})$	.0179	.2210	.0887
$p(\Delta_{25 \text{ deg}} < \Delta_{30 \text{ deg}})$	.0422	.2158	.0653

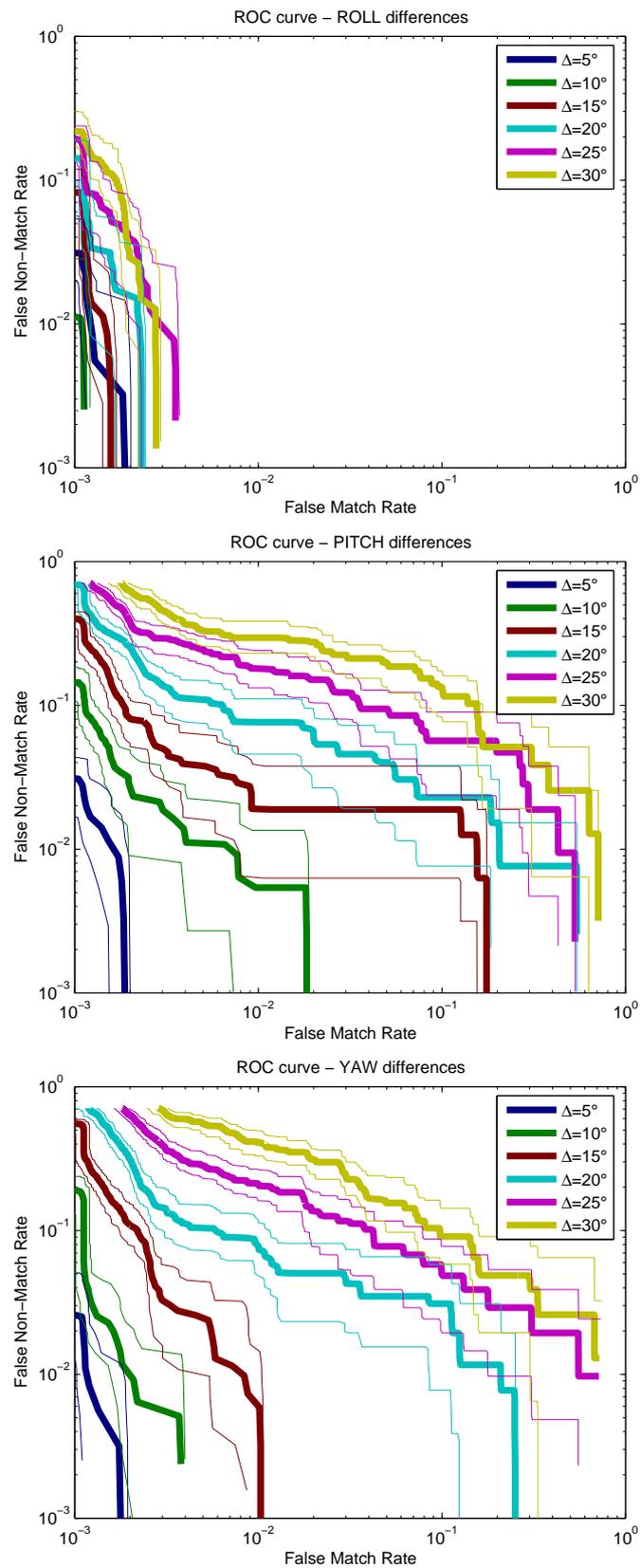
intervals. ROC curves for roll, pitch and yaw are shown for the 2006 (Fig. 4), 2007 (Fig. 5) and 2008 (Fig. 6). AFR algorithm. For the *roll* performance, the effect of variability is much lower than for the other pose changes. Both *pitch* and *yaw* changes result in significant and much more dramatic decreases in ROC performance and error rates. The statistical significance of the ROC curves is shown in Table 1. In each case the significance of the difference between each ROC curve with the next angle increment is calculated. For example, in the 2006 algorithm the difference in the ROC curves of  $\Delta_{5 \text{ deg}}$  and  $\Delta_{10 \text{ deg}}$  shows the change in Roll angle has an 18.38% overlap. Therefore, the null hypothesis that these two curves are the same can not be ignored. And, the difference in the ROC curves of  $\Delta_{5 \text{ deg}}$  and  $\Delta_{10 \text{ deg}}$  for Pitch has 0% overlap. Hence, the hypothesis that these ROC curves are the same does not hold true. Overall, for pose changes of 10deg or larger, the biometric error rates become dramatically larger. The 2008 algorithm does perform better or equal to the 2006 algorithm in all cases, although the improvement is not dramatic except for improvement in the *yaw* response below 10deg. This improvement with time is also present in the 2007 algorithm, which shows some of the improvements seen in the 2008 algorithm.

## 6 Discussion

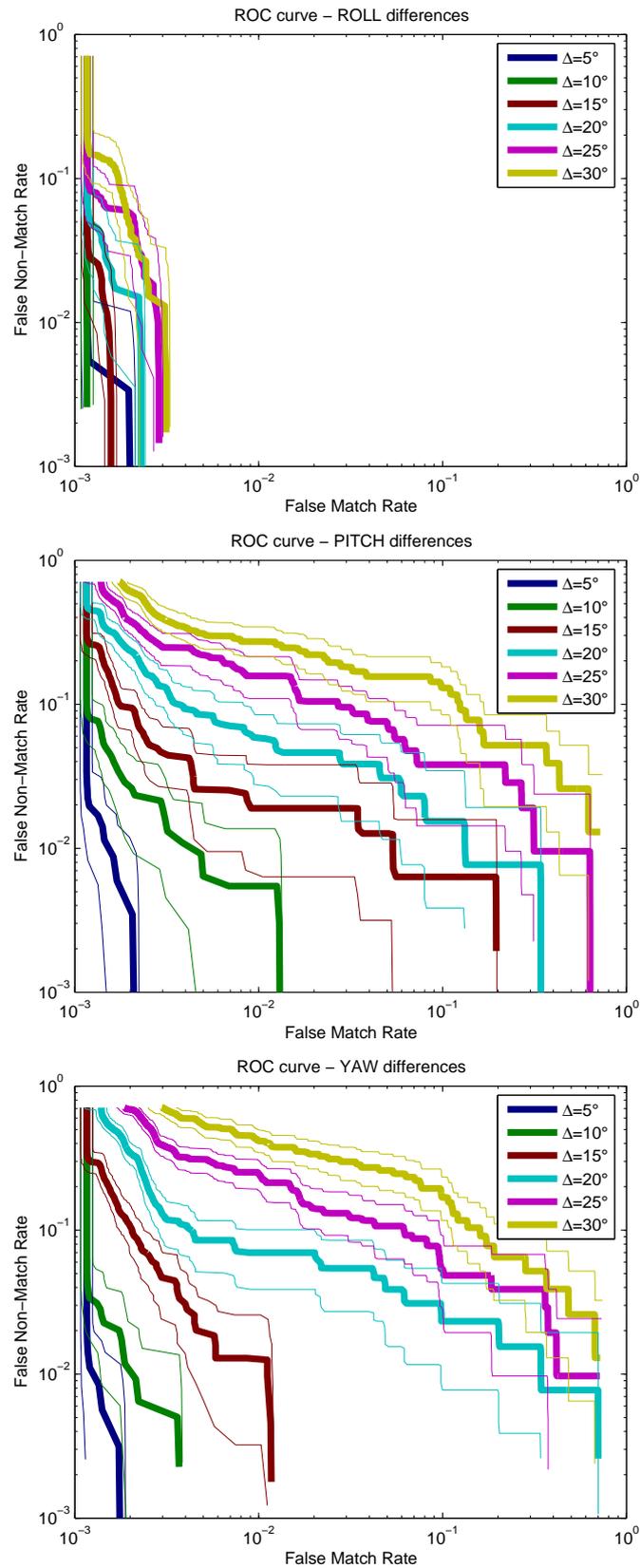
In this paper, we have presented a method and an analysis of the consequences of variability in face pose on error rates in face recognition performance. Experiments were conducted with volunteers who were asked to assume specific poses from neutral pose to  $\pm 20$ deg in each of the roll, pitch, and yaw directions. Data



**Figure 4** Receiver operating characteristics (ROC) (thick lines) and 5% confidence intervals (thin lines) for the 2006 algorithm for pose variation in *roll* (top), *pitch* (middle), and *yaw* (bottom). For each curve, pose variations of 5,10,15,20,25 and 30 deg are shown.



**Figure 5** (thick lines) and 5% confidence intervals (thin lines) for the 2007 algorithm for pose variation in *roll* (top), *pitch* (middle), and *yaw* (bottom). For each curve, pose variations of 5,10,15,20,25 and 30 deg are shown.



**Figure 6** (thick lines) and 5% confidence intervals (thin lines) for the 2008 algorithm for pose variation in *roll* (top), *pitch* (middle), and *yaw* (bottom). For each curve, pose variations of 5,10,15,20,25 and 30 deg are shown.

were analyzed to determine changes in recognition performance, using three leading commercial face recognition algorithms from 2006, 2007, and 2008. For each pair of images of a given pose difference, a genuine distribution was calculated, while the impostor distribution was calculated from all non-matching images. Confidence intervals were determined using a pointwise non-parametric bootstrapping technique along lines parallel to the EER. Results indicate that roll variations have a relatively small effect on performance, while pitch and yaw variations have a large and significant impact on error rates. In comparing the changes in performance of the algorithms over the period considered (2006–2008), we observe a general improvement in error rates, yielding an approximately 10 fold reduction in EER for low pose variations (below 10 deg in pitch and yaw). For larger pose variations ( $\geq 15$  deg), there appears to be no significant change in algorithm performance over the period. This would appear to suggest that the relative consequence of pose variability has become more severe as AFR algorithms have improved in the from 2006 to 2008. Automatic face recognition technology has important applications for verification of identity documents, and is part of the requirements for ICAO compliant biometrically enabled travel documents(8). For such travel documents, it is important to understand the biometric performance implication of any variability in photo quality. More strict standards for acceptability of photos can dramatically improve the utility of the biometric technology; however, such strict standards mean that a larger fraction of submitted photos must be rejected. This imposes an inconvenience on the population, which may, in turn, result in political pressure the governments issuing these documents.

This work was motivated by the Canadian work on photo capture requirements as part of the ISO standard(9). The key concern was to establish a tolerance limits for face pose variation for identity documents. These results support the photo capture best practice recommendations of the standard, which indicate a maximum of  $\pm 5$  deg variability in pitch and yaw. Such variability potentially results in a 10 deg difference between two images under comparison; our results show that differences in pitch and yaw greater than these limits result in a dramatic and statistically significant decrease of face recognition performance. Finally, the proposed method can be used to study any quantifiable intra-class variability such as change in number of minutiae points for fingerprint and percentage of iris occlusion for iris recognition.

## References

- [1] A. Adler, M. Schuckers, “Comparison of Human versus Automatic Face Recognition Performance” *IEEE Trans. Systems, Man and Cybernetics: Part B* 37:1248-1255, 2007.
- [2] K. W. Bowyer, K. Chang, P. Flynn “A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition” *Computer Vision and Image Understanding* 101:1–15, 2006
- [3] Burton, A. M., Miller, P., Bruce, V., Hancock, P. J. B., Henderson, Z.: “Human and automatic face recognition: a comparison across image formats” *Vision Research*, 41:3185–3195, 2001.

- [4] P. Grother, E. Tabassi, "Performance of Biometric Quality Measures" *IEEE T. Pat. Anal. Mach. Intel.* 29:531-543, 2007.
- [5] Huang, F.J., Zhou, Z., Zhang, H. J., Chen, T.: "Pose Invariant Face Recognition," *Proc. 4th IEEE Inter. Conf. on Automatic Face and Gesture Recognition* 245-250, 2000
- [6] Gross, R., Baker, S., Matthews, I., Kanade, T., "Face Recognition Across Pose and Illumination," *Handbook of Face Recognition* Springer-Verlag, 2004
- [7] Beymer, D., "Face recognition under varying pose" *Technical Report* MIT AI Laboratory, 1993.
- [8] ICAO, "Machine Readable Travel Documents" Document 9303, October 2004.
- [9] ISO/IEC, "Information technology – Biometric data interchange formats – Part 5: Face image data." Document 19794-5:2005.
- [10] ISO/IEC, "Information technology – Biometric Sample Quality – Part 1: Framework." Document 29794-1:2009.
- [11] Passports Canada, "Photo Specifications", PPTC-320 (08-03) Cat. No. FR5-11/2-2008 ISBN 978-0-662-05446-7 [www.ppt.gc.ca/form/pdfs/pptc320\\_eng.pdf](http://www.ppt.gc.ca/form/pdfs/pptc320_eng.pdf)
- [12] P.J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, M. Sharpe, "FRVT 2006 and ICE 2006 Large-Scale Experimental Results" In press *IEEE T. Pat. Anal. Mach. Intel.*, 2009.
- [13] P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, "The FERET evaluation methodology for face-recognition algorithms", *IEEE T. Pat. Anal. Mach. Intel.*, 22:1090–1104, 2009.
- [14] M.E. Schuckers, Y. Mineev, A. Adler, "Curvewise DET confidence regions and pointwise EER confidence intervals using radial sweep methodology" *2nd Int. Conf. Biometrics* Seoul Korea, 27-29 August 2007
- [15] Bradley, A. P.: "The use of the area under the ROC curve in the evaluation of machine learning algorithms." *Pattern Recognition* 7, 1145–1159, 1997.
- [16] Zhao W, Chellappa R, Phillips P J, Rosenfeld A "Face recognition: A literature survey" *ACM Computing Surveys* 35(4):399–458, 2003.
- [17] Hube, J.P "Using Biometric Verification to estimate identification performance" *2006 Biometric Symposium: Special Session on Research at the Biometric Consortium* 1–6, Sept. 19 2006-Aug. 21 2006.