Performance comparison of human and automatic face recognition

Andy Adler
School of Information Technology and Engineering, University of Ottawa, Ontario, Canada
adler@site.uottawa.ca
James Maclean
3M Security Systems Division, 1545 Carling Avenue, Suite 700, Ottawa, Ontario, Canada
jemaclean@mmm.com

1. Introduction

Of all biometric modalities, the capabilities of automatic face recognition software are most often compared to the human ability to do the same task. Unlike other biometric modalities, such as fingerprints, where a small group of trained people are experts, most people use their face recognition abilities every day. There is a significant body of literature analysing the ability of humans to recognize faces, including a number of recent works which analyse the ability of people to perform security related face recognition tasks. For example, Kemp et al. [5] analysed the ability of supermarket cashiers to identify shoppers from photos on credit cards, and discovered overall poor performance. Chang Hong et al. [2] analyzed the ability of people to match poor-quality video footage against high-quality photographs, and showed a high level of performance. Such discrepancy in results illustrates the importance of the task context for human face recognition. People focussing on face recognition as their primary task will outperform those, such as supermarket cashiers, whose primary task is customer service.

There are few reports in the scientific literature of comparisons of human face recognition performance to that of automatic systems. Burton and collaborators [1,3] compared PCA based and graph-matching algorithms against human ratings of similarity and distinctiveness, and human memory performance. These studies were focussed on the extent to which automatic algorithms explain features of human performance, rather than as a comparison of recognition performance levels. Additionally, these studies did not use recent face recognition software.

In order to investigate this issue, we designed a test to measure human face recognition performance against that of automatic software.

2. Methods

Our study investigated the ability of interested and motivated non-specialist volunteers to perform face identification tasks in comparison to that of several commercial face recognition software packages. Images were obtained from the NIST Mugshot database [6]. Pairs of frontal pose face images were randomly created from this database. Two-thirds of the pairs were *impostors* (images of different persons), and one third were *genuines* (different images of the same person). No special effort was made to select images of the same gender or ethnicity for the impostor pairs.

Twenty one people (16 male, 5 female) participated in the experiments. Participants were predominantly causasian and in the age range 20-40. Participants were asked to log onto a web site, where an application server would present pairs of face images, and the participant was asked whether they were from the same person. The participant then selected from the following options: *same*, *probably same*, *don't know*, *probably different*, *different*. The results from each participant were then post-processed to choose a threshold which would give a minimum total error score. Participants were not given any information about the distribution of genuines and impostors in the tests, or any feedback about their success. The same image pairs were then presented to each commercial face recogntion software package to which we had access. Approximately two percent of images could not be enrolled; any face image pair with an image that could not be enrolled was assigned a match score of zero.

3. Results

Results were calculated in terms of false match rate (FMR) and false non-match rate (FNMR) for both participants and face recognition algorithms. Fig. 1 shows results for automatic software and human performance. Since participants did not provide a match score, it was not possible calculate an error curve, but rather a single point. Software results are shown for the highest performing software available to us in 1999, 2001 and 2003. The results for 2001 had corrected eye locations provided to the software.

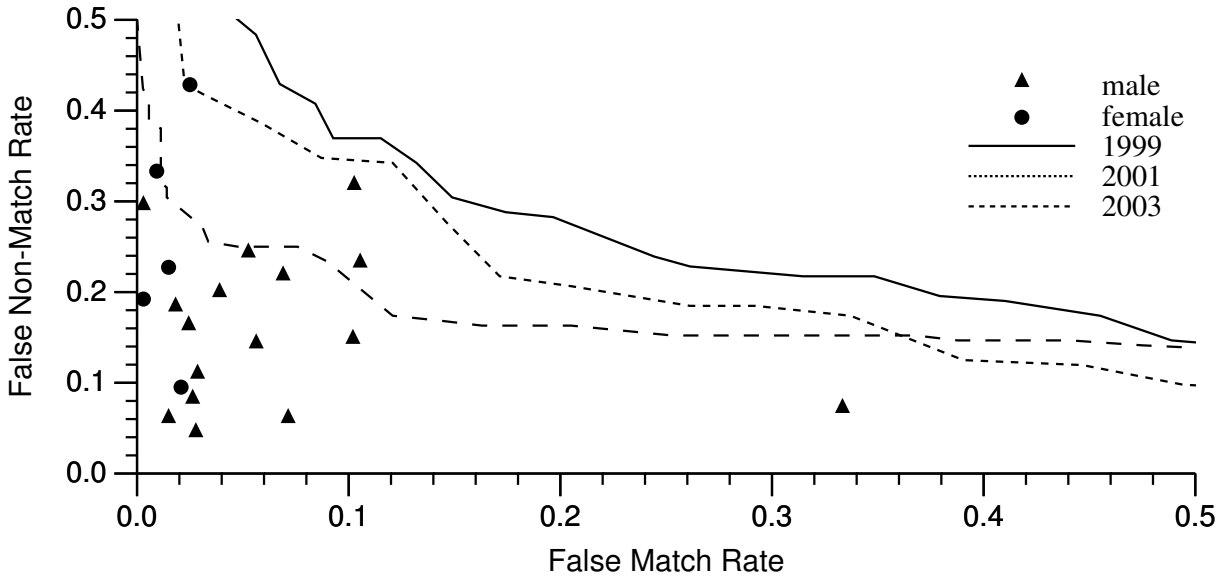

Fig. 1: Detection error tradeoff curve for Face recognition results. Curves indicate results for the highest performing software available to us in the years 1999, 2001 and 2003, respectively. Human results are shown by symbols at the measured FMR and FNMR point.

4. Discussion

These results suggest that:

- Most motivated, attentive humans are currently able to significantly outperform face recognition software. However, recent advances mean that face recognition software now outperforms approximately 20% of human participants.
- There is a wide variability in the face recognition ability of humans. Differences in error rates of an order of magnitude were observed.
- There does not appear to be a significant difference in error rates between male and female participants, although results from females show a preference for false non-matches to false matches in comparison to that from males. Since the mugshot database [6] is primarily male faces, the increased capability of females to recognize female faces [5] is not evident in these data.

The choice of database was based on the "Goldilocks" criterion: it was necessary to choose a sufficiently difficult database in order for error levels to be sufficiently large to make meaningful comparisons. Thus the error rates reported here should not be taken as reflective of the general abilities of human or software face recognition.

This work has studied the abilities of untrained, motivated, human volunteers. Future work should address these questions: 1) How do humans perform as familiarity increases? 2) What is the effect of motivation? 3) What is the effect of routine and boredom? 4) Do experts outperform untrained recognizers? and 5) What characterizes good recognizers from poor ones?

5. References

1. Burton A M, Miller P, Bruce V, Hancock P J B, Henderson Z (2001) Human and automatic face recognition: a comparison across image formats *Vision Research* **41** 3185-3195
2. Chang Hong L, Seetzen H, Burton A M, Chaudhuri A (2003) Face recognition is robust with incongruent image resolution: Relationship to security video images. *Journal of Experimental Psychology: Applied.* **9** 33-41.
3. Hancock P J B, Bruce V, Burton M A (1998) A comparison of two computer-based face identification systems with human perceptions of faces *Vision Research* **38** 2277-2288
4. Herlitz C L (2002) Sex differences in face recognition−women's faces make the difference. *Brain &Cognition.* **50** 121-128.
5. Kemp R, Towell N, Pike G (1997) When seeing should not be believing: Photographs, credit cards and fraud. Applied Cognitive Psychology. **11** 211-222
6. NIST, *NIST Special Database 18: Mugshot Identification Database (MID),* http://www.nist.gov/srd/nistsd18.htm (current May 2004)