Human versus Machine Performance

Alice J. O’Toole
The University of Texas at Dallas
Acknowledgements

• In collaboration with
  – P. J. Phillips - NIST
  – Fang Jiang - Univ. of Texas at Dallas (UTD)
  – Nils Pénard - UTD
  – Janet Ayyad - UTD
  – Hervé Abdi - UTD

• Work supported by TSWG
Overview

- rationale
- Face Recognition Grand Challenge
- human-machine comparison
Problem

• Are face recognition algorithms *ready* for security applications?
  – enormous improvements over last decade
  – accuracy of algorithms tested intensively

• *How accurate do they have to be to be useful?*
  – meet or exceed human performance
Why?

• *humans are the competition!*
  – human-machine comparisons *virtually* never done

• putting algorithms in the field
  – security improved or put at greater risk?
How accurate are algorithms?
U.S. Government-sponsored Competitions

• standardize comparisons
  – test multiple algorithms
  – identical, LARGE sets of face image data

  – Face Recognition Grand Challenge
    • (2004-ongoing)
Present work

• purpose
  – extend standardization of FRGC to compare humans and algorithms on a challenging face recognition task
  – matching face identity across changes in illumination (FRGC Exp. 4)
Why Illumination Change?

• recognized to be difficult for:
  – **humans** (e.g., Braje et al., 2000; Troje & Bülthoff, 1998)
  – **algorithms** (e.g., Phillips et al. 2005; Gross et al. 2005)
Most Challenging FRGC Experiment

- controlled illumination experiment (Exp. 1)
  - match images with controlled illumination
  - 20 participating algorithms
  - median performance of
    - .91 verification rate
    - .001 false acceptance rate
• uncontrolled illumination (Exp. 4)
  – match images with controlled and uncontrolled illumination
  – 7 participating algorithms
  – median performance
    • .42 verification rate
    • .001 false acceptance rate
FRGC Uncontrolled Illumination Test

• Match identity in target and probe faces
  – *target* - controlled illumination
  – *probes* - uncontrolled illumination
Specifics

• similarity matrix
  – target \( (n = 8014) \)
  – probe \( (n = 16028) \)

  – \( s(i,j) = \text{similarity between the } i^{th} \text{ and } j^{th} \text{ faces} \)
    • 128,041,040 similarity scores
    • 407,352 of same people
    • remainder of different people
Results

• ROC
  – verification rate
  – false acceptance rate
Algorithm Performance

![Algorithm Performance Graph]

March 22, 2006
Evaluating Algorithms with Human Benchmarks - supported by TSWG
Comparing Humans and Algorithms

- **problem**
  - 128 million face pairs?

- **solution**
  - sample face pairs
    - most difficult
    - easiest
Sampling

- homogeneous
  - caucasian males/females 20-30 yrs
  - comparisons made on identity not
    - age, race, sex

- caution on the FRGC results
Easy and Difficult

- PCA Baseline Algorithm
  - scaled and aligned images (SAIC)
  - available and widely used since the 90’s
  - but not state-of-the-art
Match Pairs

- **“easy” match pairs**
  - 2 “similar” images of same person
    - Similarity scores $> 2$ sd above mean similarity of match pairs

- **“difficult” match pairs**
  - 2 “dissimilar” images of same person
    - Similarity scores $< 2$ sd below mean similarity of match pairs
No-Match Pairs

- “easy” no-match pairs
  - 2 “dissimilar” images of different people
    - similarity scores < 2 sd below mean similarity of no-match pairs

- “difficult” no-match pairs
  - 2 “similar” images of different person
    - similarity scores < 2 sd above mean similarity of no-match pairs
• Experiment 1
  – unlimited exposure time
  – male face pairs

• Experiment 2
  – 2 sec. exposure time
  – male and female face pairs

• Experiment 3
  – 500 msecs. exposure time
  – male and female face pairs
Methods

• Stimuli
  – 240 pairs of faces
    • 120 male pairs
      – 60 easy
      – 60 difficult
    • 120 female pairs
      – 60 easy
      – 60 difficult
Subjects

- 91 volunteers from UTD
  - Expt. 1
    - $n = 22$ (12 males; 10 females)
  - Expt. 2
    - $n = 49$ (24 males; 25 females)
  - Expt. 3
    - $n = 20$ (10 males; 10 females)
Procedure

• 2 faces appear side by side

• Human subject raters respond…
  – 1. sure they are the same person
  – 2. think they are the same person
  – 3. not sure
  – 4. think they are not the same person
  – 5. sure they are not the same person
Results

• **PCA predicts difficulty (d’ analysis)**
  – Experiment 1
    • $F(1,20) = 19.78, p < .002$
  – Experiment 2
    • $F(1,48) = 96.53, p < .0001$
  – Experiment 3
    • $F(1,18) = 62.65, p < .0001$
Experiment 2

![Graph of Human Face Matching]

- Blue line: difficult pairs
- Red line: easy face pairs

false accept rate

March 22, 2006
Evaluating Algorithms with Human Benchmarks - supported by TSWG
Experiment comparison

• Humans no more accurate with unlimited time than with 2 secs. presentations
  – $F(1,176) = 2.01, \text{ ns.}$

• Human accuracy declined with exposure times of 500 msecs.
  – $F(1,176) = 26.97, p < .0001$
Stability of human performance

• supports use of these data for benchmark comparisons with machines
Human-Machine Comparisons

• Seven state-of-the-art algorithms
  – 4 from industry
  – 3 from academic institutions

• Comparisons
  – 120 difficult face pairs
  – 120 easy face pairs
Results
Identity Matching for Difficult Face Pairs

False Accept Rate

NJIT
CMU
Viisage
Human Performance
Algorithm A
Algorithm B
Algorithm C
Algorithm D
Chance Performance
Results Summary

Difficult Face Pairs

• 3 algorithms surpass humans
  – NJIT (Liu, \textit{IEEE: PAMI, in press})
  – CMU (Xie et al., 2005)
  – Viisage (Husken et al., 2005)

• 4 less accurate than humans
Identity Matching for Easy Face Pairs

- CMU
- NJIT
- Algorithm B
- Algorithm D
- Viisage
- Algorithm C
- Algorithm A
- Human Performance
- Chance Performance
Results Summary
Easy Face Pairs

• 6 algorithms surpass humans!

• 7th less accurate than humans at high false acceptance rates
Human Attention

• Did attention waver during experiment?
  – no correlation between accuracy and trial
    • verification ($r = .07$, $ns$)
    • false acceptance rate ($r = -.04$, $ns$.)
Are human skills overrated?

• “familiar” versus “unfamiliar”

• unfamiliar matching
  – correct task for comparing “human” and machine security systems

• evidence that human expertise for faces may be limited to recognizing “familiar faces” (Hancock et al., 2001; O’Toole et al., 2003)
Familiarization

• Can we improve human performance?
• Experiment 4
  • select face pairs that generated errors in Exp. 2
  • familiarize subjects with people in pairs
    – 5 exposures to one face in pair
  • $n = 77$ subjects
• results
  – improvement, but not significant ($F(1,76)=1.3, p < .25$)
Human Performance with Familiarization

- Blue line: familiarized
- Red line: unfamiliarized

$p(false \text{ alarm})$

0 0.2 0.4 0.6 0.8 1
Conclusions

• Algorithms compete favorably with humans on the difficult task of matching faces across changes in illumination

  – some algorithms are *better* than humans on “difficult” face pairs
  – nearly all are *better* than humans on “easy” face pairs
Implications

• Algorithms may improve security in some situations
  – even if they perform poorly in absolute terms
Implications

• We accept on “face” value the need to test any algorithm that we put in the field for an important security application

• Tools available for testing humans
  – We rarely do!?
What next?

• Why?
  – Analysis of the variability of algorithms
  – Which face pairs separate algorithms?
    • Hybrid strengths & weaknesses