Facial Recognition Performance and Its Measurement

PRESENTED BY:
Patrick Grother
National Institute of Standards and Technology

MODERATED BY:
Stephen Redifer

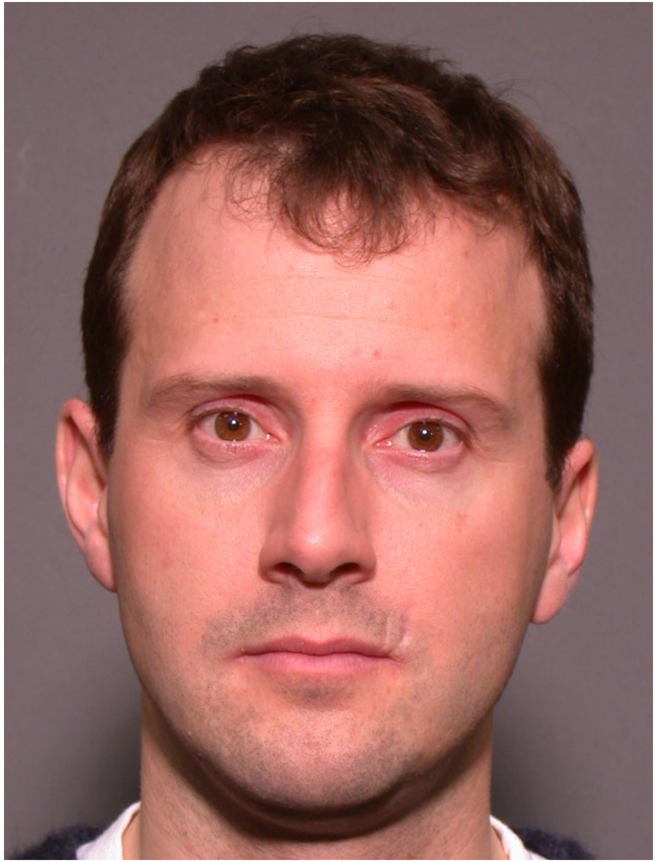
2020-09-24

Face Recognition Performance and its Measurement

Patrick Grother
NIST
U. S. Department of Commerce

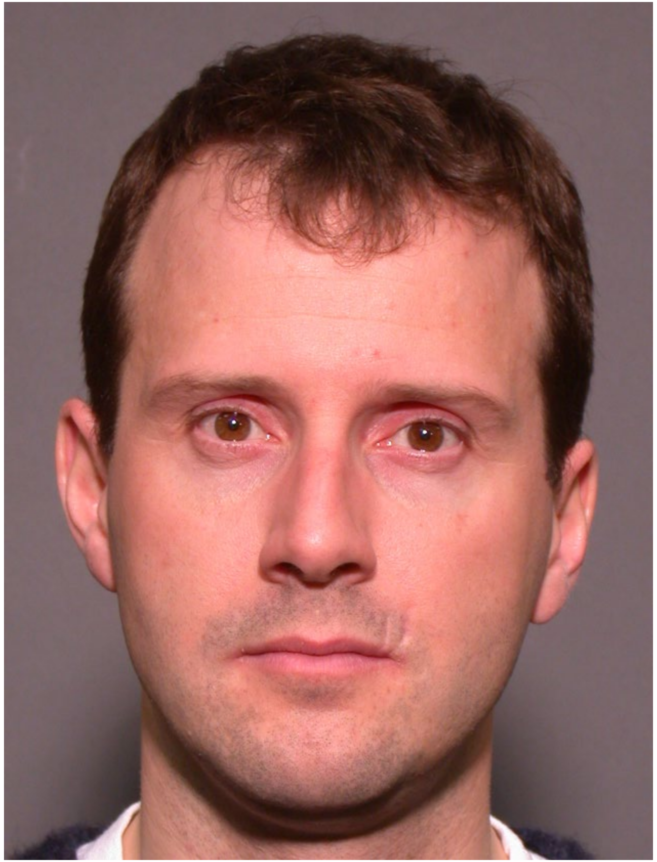
Homeland Defense and Security Information Analysis Center
September 24, 2020

What’s in a face?



How many biometrics here?

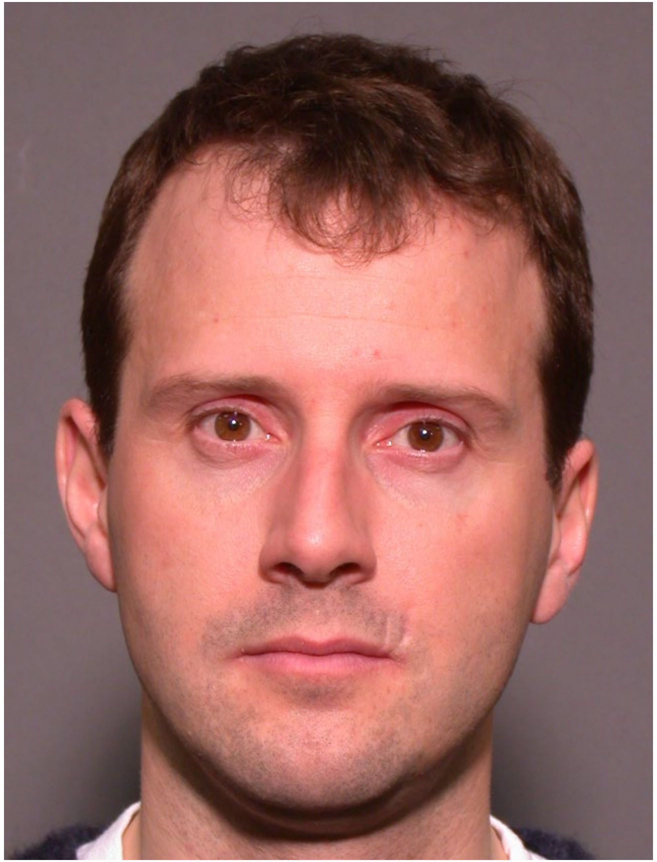
What’s in a face?



How many biometrics here?

1 **Face**

What's in a face?

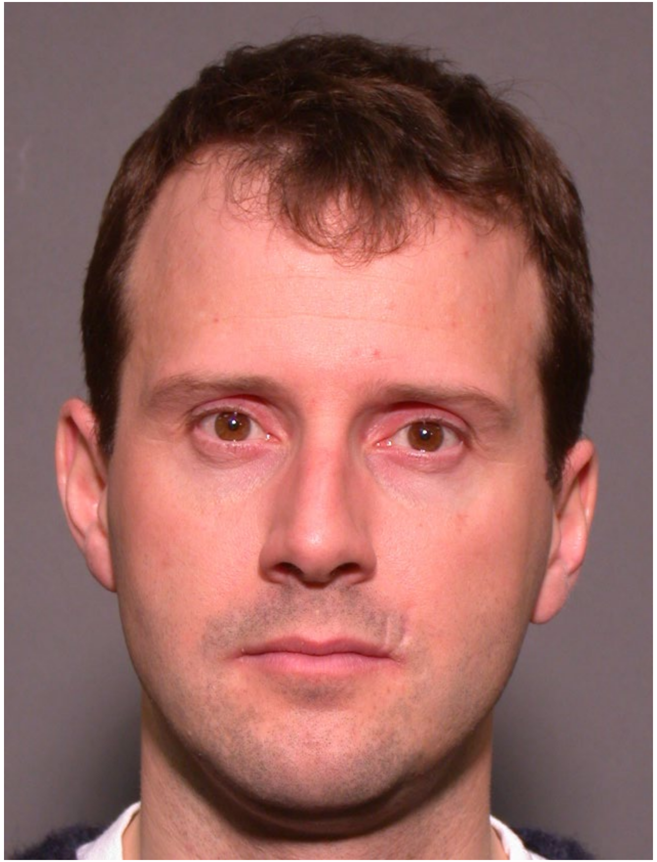


How many biometrics here?

1 Face

2 Irides + periocular

What's in a face?



How many biometrics here?

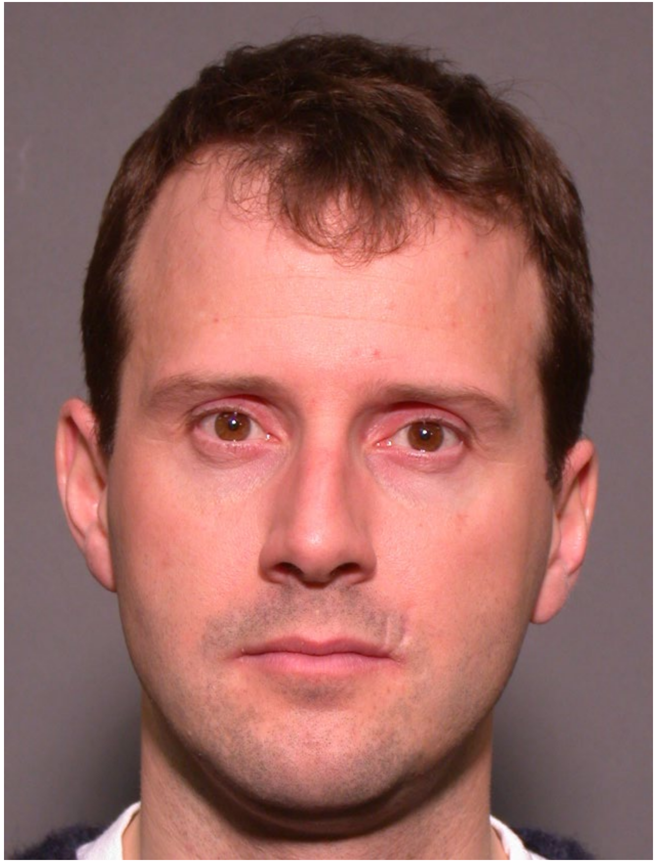
1 **Face**

2 **Irides + periocular**

3 **Skin texture**

https://patents.google.com/patent/US7369685B2/

What’s in a face?



How many biometrics here?

1 **Face**

2 **Irides + periocular**

3 **Skin texture**

https://patents.google.com/patent/US7369685B2/

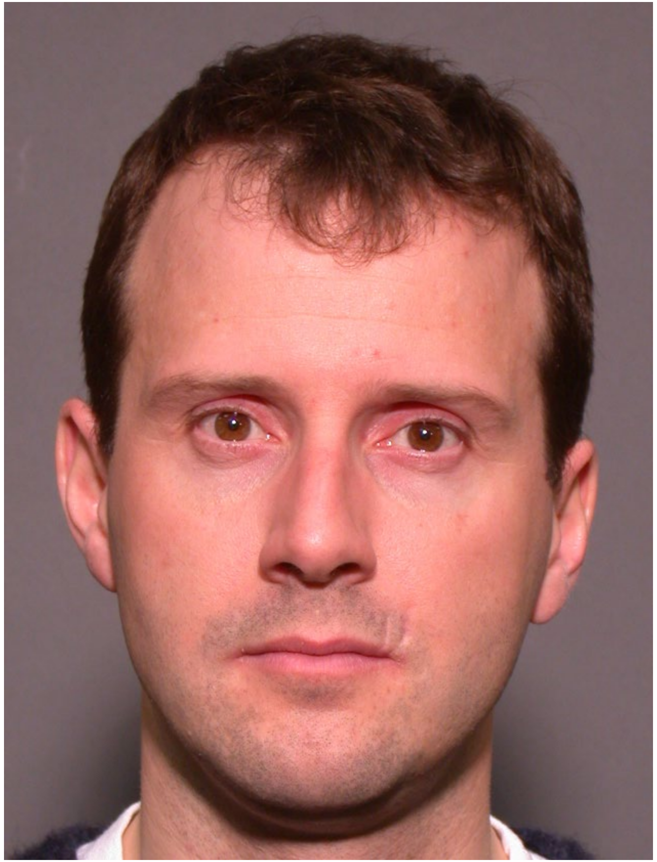
4 **Head shape**

5 **Ears**

6 **Scars**

Human review: See ASTM E3149
Standard Guide for Facial Image Comparison Feature List for Morphological Analysis

What's in a face?



How many biometrics here?

1 **Face**

2 **Irides + periocular**

3 **Skin texture** | https://patents.google.com/patent/US7369685B2/

4 **Head shape**

5 **Ears**

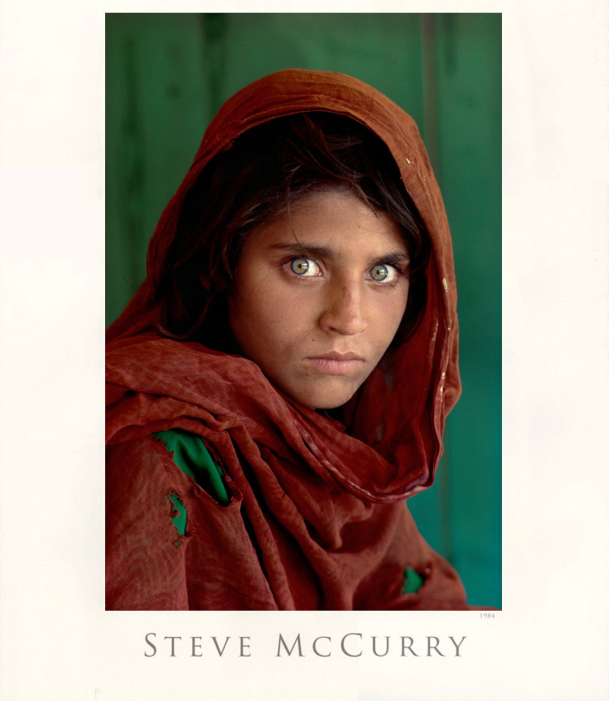
6 **Scars**

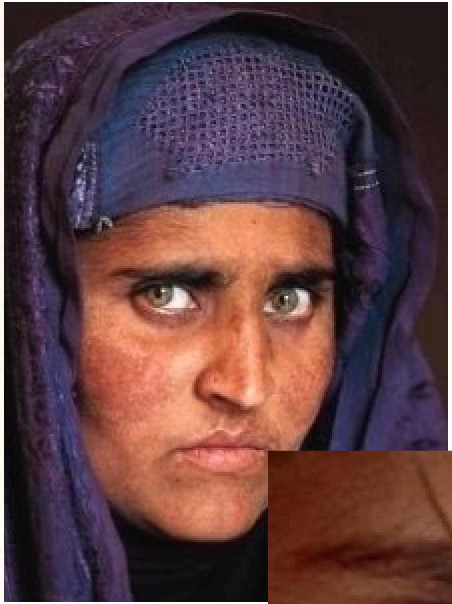
Human review: See ASTM E3149 *Standard Guide for Facial Image Comparison Feature List for Morphological Analysis*

7 **Anything else unique**

- Short + long wave infrared
- Hyperspectral
- 3D

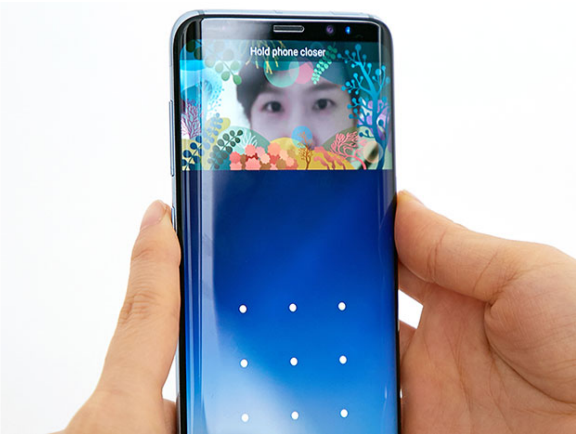
The Afghan Girl





https://www.nationalgeographic.com/magazine/2002/04/afghan-girl-revealed/
c. National Geographic, photographic portrait by journalist Steve McCurry, 1984

Face authentication: Closed system



https://spectrum.ieee.org/tech-talk/consumer-electronics/gadgets/new-samsung-galaxy-s8-unlocks-with-facial-recognition-iris-scanning

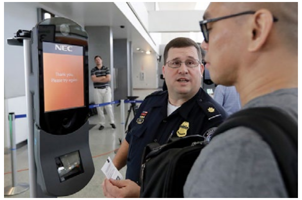
https://www.macrumors.com/2017/10/25/apple-reduced-face-id-accuracy-iphone-x/

https://support.apple.com/en-us/HT208109

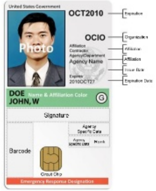
Face Recognition: How? By comparing faces

https://securitytoday.com/articles/2018/02/27/us-border-patrol-unable-to-validate-epassport-data.aspx

https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/eborder/eborder-abc



Georgetown Law. Center on Privacy + Technology
https://www.airportfacescans.com/
Figure 2: A traveler has his face scanned as a Customs and Border Protection agent provides instruction. (Photo: Associated Press, all rights reserved)



https://en.wikipedia.org/wiki/FIPS_201

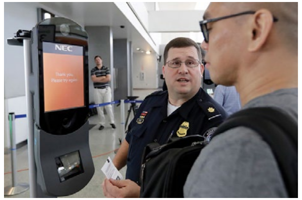
- Same identity?
- Different identity?

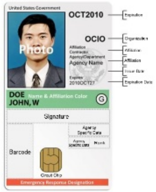
Face Recognition: How? By comparing faces

https://securitytoday.com/articles/2018/02/27/us-border-patrol-unable-to-validate-epassport-data.aspx

- Same identity?
- Different identity?

https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/eborder/eborder-abc





https://en.wikipedia.org/wiki/FIPS_201

Georgetown Law. Center on Privacy + Technology
https://www.airportfacescans.com/
Figure 2: A traveler has his face scanned as a Customs and Border Protection agent provides instruction. (Photo: Associated Press, all rights reserved)

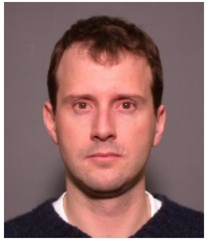
Source: FRVT staff and sister, with permission

Inbound border crossing using passport verification

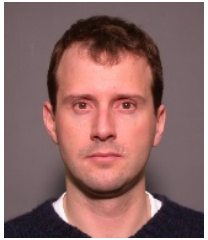
LIVE

https://en.wikipedia.org/wiki/EPassport_gates
CC BY 2.0. File:Heathrow Terminal 5 ePassport gates.jpg
Created: 16 July 2010

CHIP



Wait for luggage

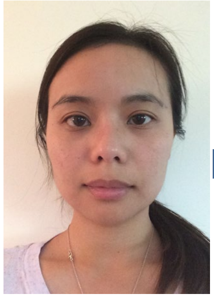


See an officer!

Two factor authentication:

1. Something you have: Possession of passport
2. Something you are: Successful recognition of a biometric

Face recognition: How?



1. Face Recognition (Identity Extractor)

TWO IMAGES

1. Face Recognition (Identity Extractor)

Feature vector

2. Face Recognition Comparator

Feature vector

Match Score

No, different person! Is Score above threshold? **Yes, same person!**

- DCNNs
- ML / AI
- Not commoditized
- Trade secrets

Templates aka feature vectors

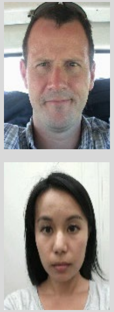
- 0.2 - 4KB, 2KB is most common
- 0.1 to 1 second on CPU

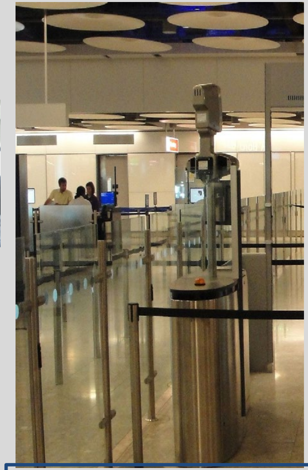
Templates

- Templates are reversible
- Images retained

FR in operations: Passport verification at a border

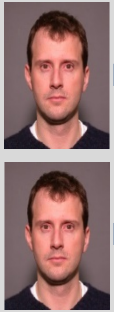
LIVE





https://en.wikipedia.org/wiki/EPassport_gates
CC BY 2.0. File:Heathrow Terminal 5 ePassport gates.jpg
Created: 16 July 2010

CHIP



Wait for luggage

See an officer!

Two factor authentication:
1. Something you have:
2. Something you are: biometric

Possession of passport
Successful recognition of a

1. No central database

2. Two images involved: live capture and chip image

3. Trusted passport?
   - Digital signature
   - Morphed image

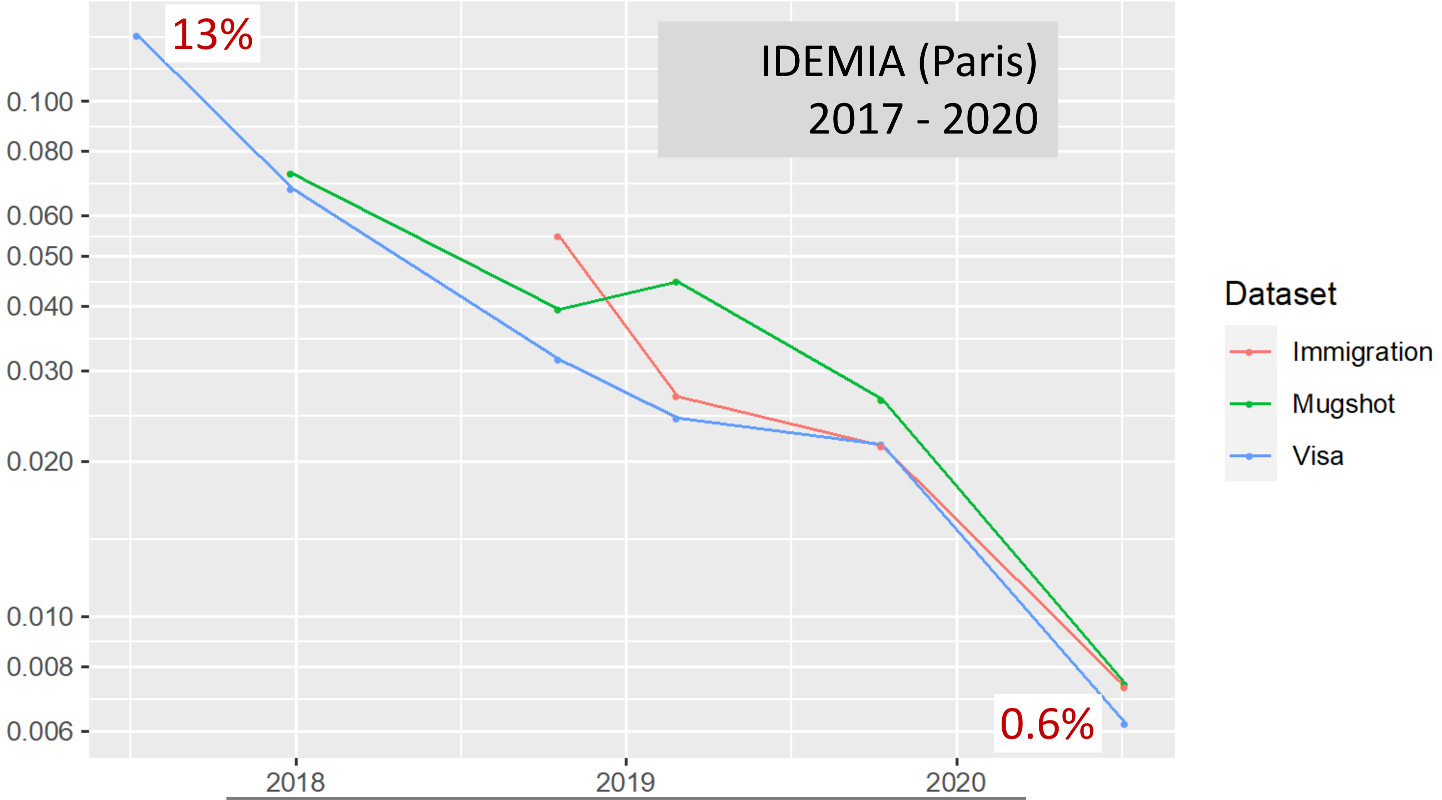
4. Error and consequences
   - False Accept ⟶ Border security
   - False Negative ⟶ Inconvenience

FRVT 1:1 Leaderboard 2020-07-27

Developer	VISA Photos FNMR @ FMR ≤ 0.000001	MUGSHOT Photos FNMR @ FMR ≤ 0.00001	MUGSHOT Photos FNMR @ FMR ≤ 0.00001 DT>=12 YRS	VISABORDER Photos FNMR@ FMR ≤ 0.000001	BORDER Photos FNMR @ FMR = 0.000001	WILD Photos FNMR@ FMR ≤ 0.00001
sensetime-003	0.0027 (3)	0.0027 (1)	0.0027 (1)	0.0051 (6)	0.0100 (7)	0.0355 (45)
deepglint-002	0.0027 (2)	0.0032 (7)	0.0033 (2)	0.0043 (2)	0.0084 (3)	0.0301 (1)
paravision-004	0.0046 (7)	0.0030 (4)	0.0036 (3)	0.0091 (18)	0.0188 (27)	0.0311 (16)
visionlabs-008	0.0036 (4)	0.0031 (6)	0.0040 (4)	0.0045 (3)	0.0079 (1)	0.0308 (10)
...						
toshiba-003	0.0214 (64)	0.0085 (41)	0.0131 (40)	_	0.0241 (37)	0.0321 (26)
fujitsulab-000	0.0212 (63)	0.0091 (45)	0.0133 (41)	0.0251 (71)	0.4200 (105)	0.0481 (73)
asusaics-000	0.0209 (62)	0.0085 (39)	0.0134 (42)	0.0143 (38)	0.7189 (112)	0.0332 (35)
cogent-004	0.0116 (33)	0.0096 (49)	0.0134 (43)	0.0157 (41)	0.0325 (54)	0.0436 (66)

https://pages.nist.gov/frvt/html/frvt11.html

Accuracy Gains: Typical Example

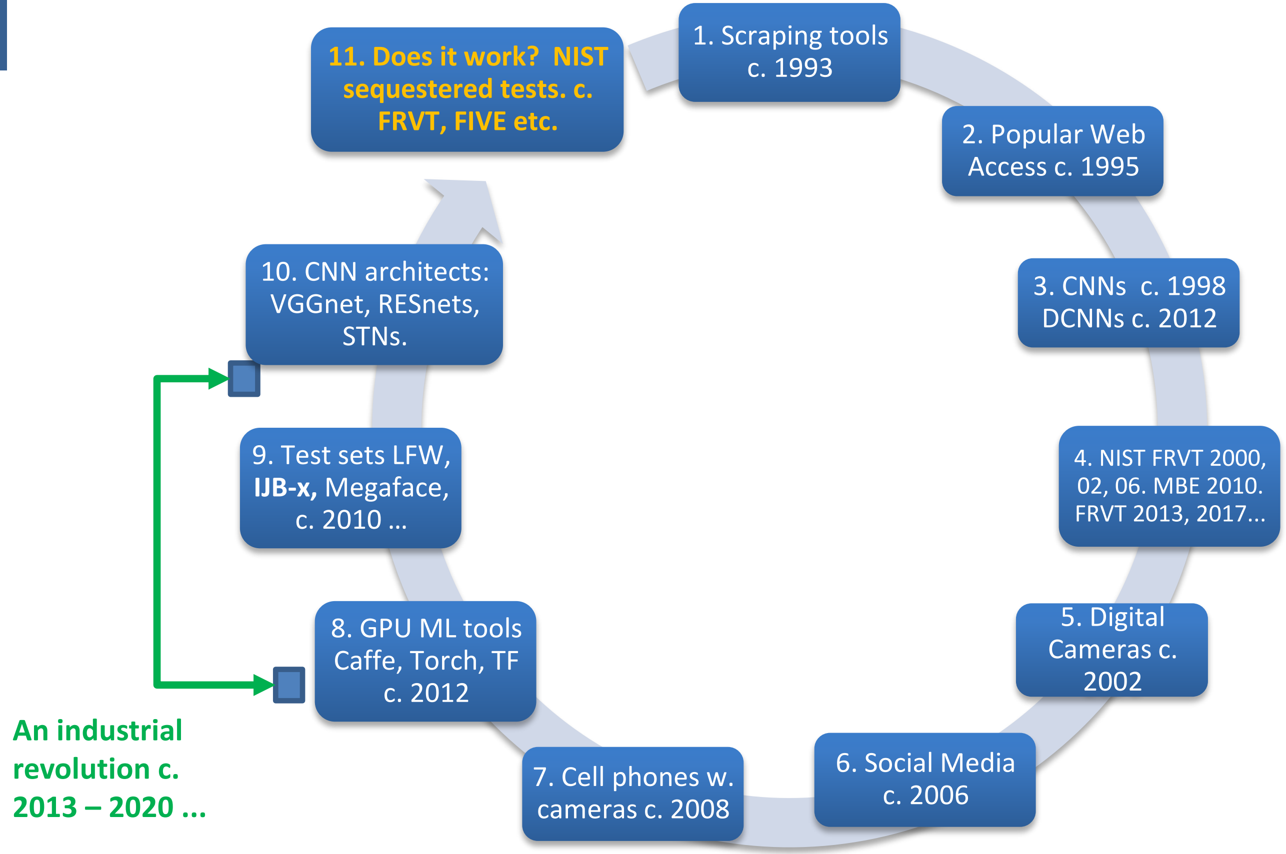


False Rejection

FNMR at FMR = 10^{-6}

Date Algorithm Delivered to NIST

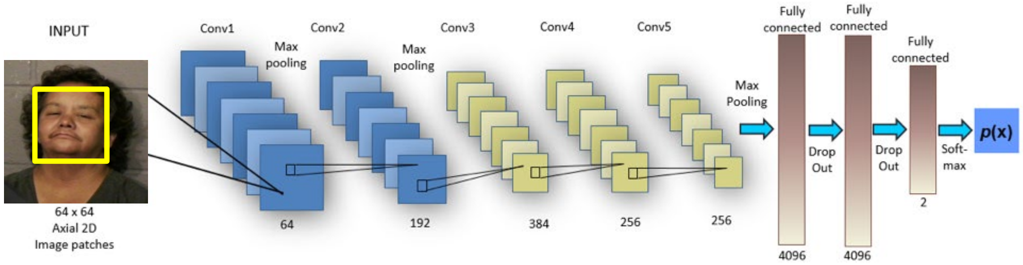
Enablers of Better Face Recognition



Black box: What is a DCNN?

$$F(\mathbf{x}) = F_N(F_{N-1}(\ldots F_2(F_1(\mathbf{x}, \mathbf{w}_1), \mathbf{w}_2) \ldots, \mathbf{w}_{N-1}), \mathbf{w}_N)$$

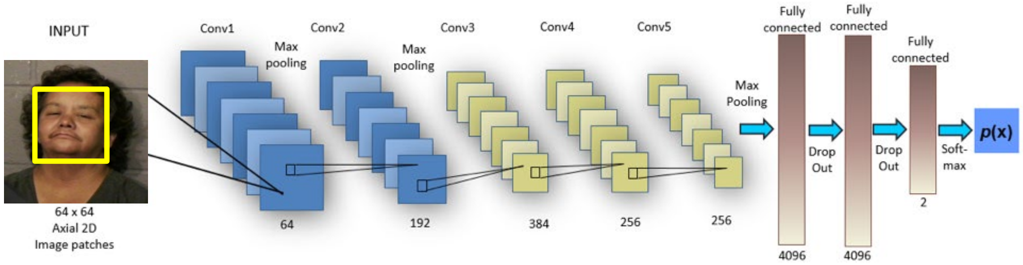
- CNN is a composed function, F, implementing local (image) filters
- Operating on an image, $\mathbf{x}_1$, input to the first layer
  - Dimensions are W x H x K
- Producing intermediate feature maps, $\mathbf{x}_n$, $1 < n \leq N$
- Each layer has a function, F_n, which perform various operations and are handcrafted
- Each layer has parameters, $\mathbf{w}_n$, which are **learned from some training data**



Black box: What is a DCNN?

$$F(\mathbf{x}) = F_N(F_{N-1}(\dots F_2(F_1(\mathbf{x}, \mathbf{w}_1), \mathbf{w}_2) \dots, \mathbf{w}_{N-1}), \mathbf{w}_N)$$

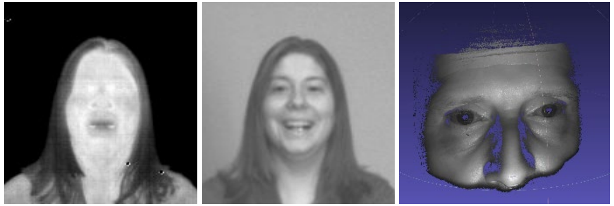
- CNN is a composed function, F, implementing local (image) filters
- Operating on an image, $\mathbf{x}_1$, input to the first layer
  - Dimensions are W x H x K
- Producing intermediate feature maps, $\mathbf{x}_n$, $1 < n \leq N$
- Each layer has a function, F_n, which perform various operations and are handcrafted
- Each layer has parameters, $\mathbf{w}_n$, which are **learned from some training data**



Multi-biometrics

Multimodal

Multisensor



Multi-instance
(contemporaneous)

Repeated-instance
(longitudinal)

Multiple algorithm

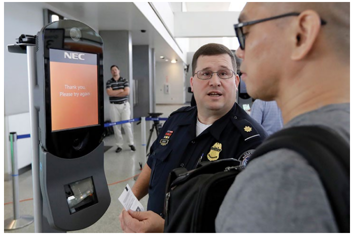
$\text{Score} = \text{Fusion}\,[\,\text{Algorithm}_B(X,Y),\ \text{Algorithm}_A(X,Y)\,]$

1:1 Authentication: Live-to-document

https://en.wikipedia.org/wiki/FIPS_201

https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/eborder/eborder-abc

https://securitytoday.com/articles/2018/02/27/us-border-patrol-unable-to-validate-epassport-data.aspx



Georgetown Law. Center on Privacy + Technology
https://www.airportfacescans.com/
Figure 2: A traveler has his face scanned as a Customs and Border Protection agent provides instruction. (Photo: Associated Press, all rights reserved)

1:N Identification

Scalability to Large Populations

High volume applications

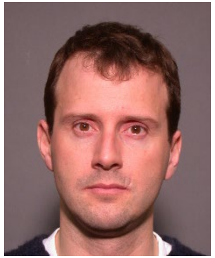
- Duplicate detection (passports, visa fraud, National ID)
- Casino persons of interest
- Aircraft boarding
- Surveillance

Human review usually infrequent

Low volume applications, with human review:

- Criminal investigation
- Clustering media

Face recognition: How?



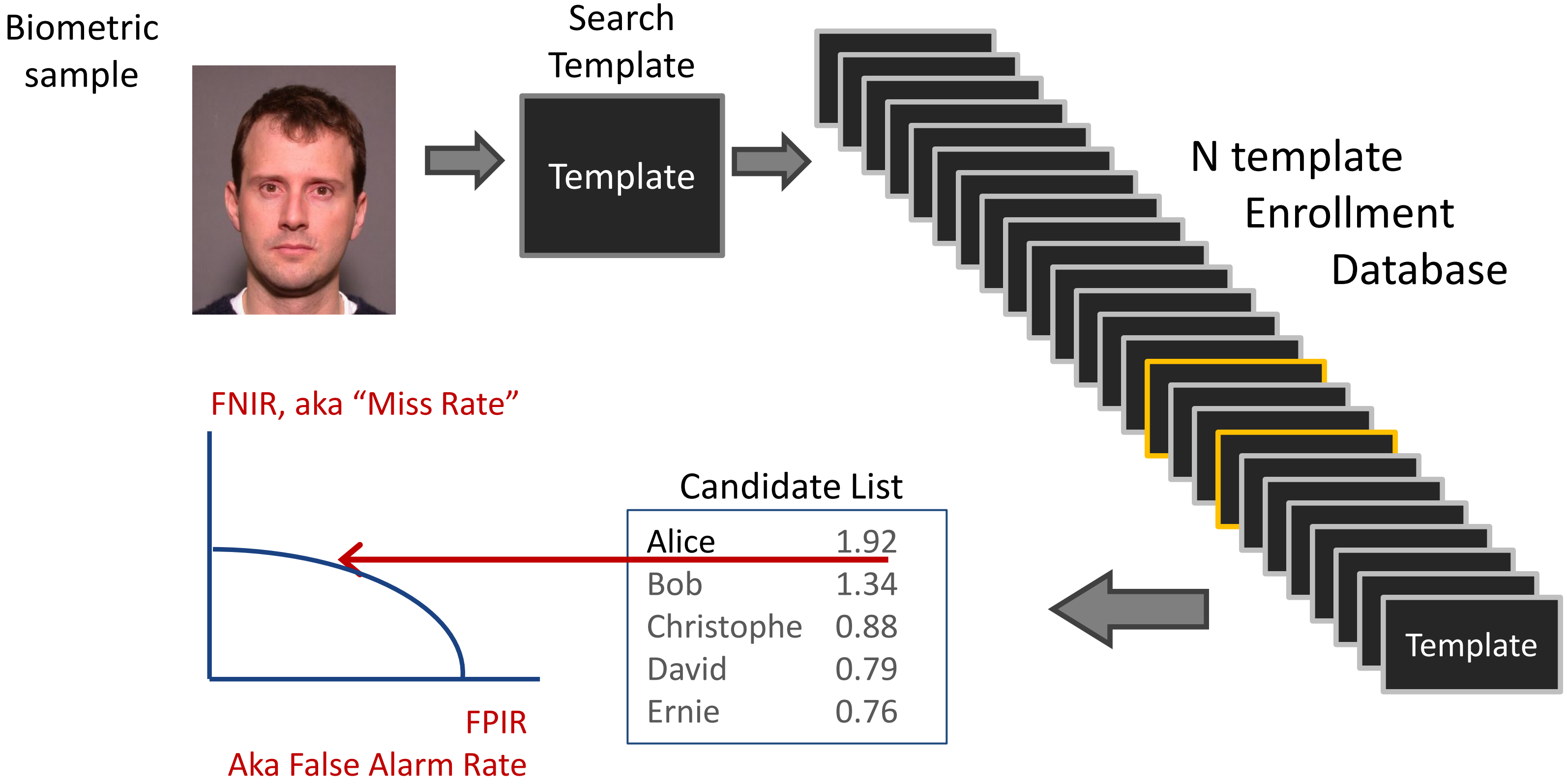
Face Recognition Engine

Person is Patrick

THIS IS NOT HOW FR WORKS. INSTEAD:

- An FR engine only knows people who are enrolled into it
- FR implements comparisons of new photos

1:N Search = N 1:1 comparisons (sometimes)

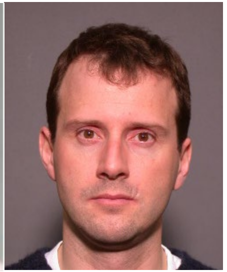


A demonstration of 1:N face recognition

» Enroll border crossing images
  - 104.1 million
  - 32.6 million people

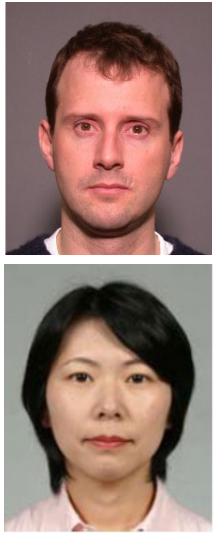

» Mated searches
  - 2.3 million “visa” APPLICATION images → FNIR, aka “miss rate”

» Non-mated searches
  - 1.8 million “visa” APPLICATION images → FPIR, aka “false alarm rate”

A demonstration of 1:N face recognition

Step 1:
- Enrol N = 104 million photos, of 32.6 million people
- Images are examples, from NIST Special Database 32, representative of pose, illumination, compression



Step 2:
- Search with almost ISO compliant "visa" portraits

104 Million: “visa” to “border crossing” search accuracy

NIST

INVESTIGATION FALSE POS ID RATE = 100%	NEC-3 (2018-11) (0.7 + 1.1 seconds)	RankOne-006 (2019-06) (0.1 + 18 seconds)
Searches not returning ANY image of the correct person at rank 1	**0.4%**	**2%**
Searches not returning ALL images of the correct person in the top ranks	1.6%	11%

HIGH VOLUME, HIGH THRESHOLD IDENTIFICATION, FALSE POS ID RATE = 1%	NEC	Rank One
Searches not returning ANY image of the correct person above threshold	**0.6%**	**8.3%**
Searches not returning ALL images of the correct person above threshold	4.5%	41.0%

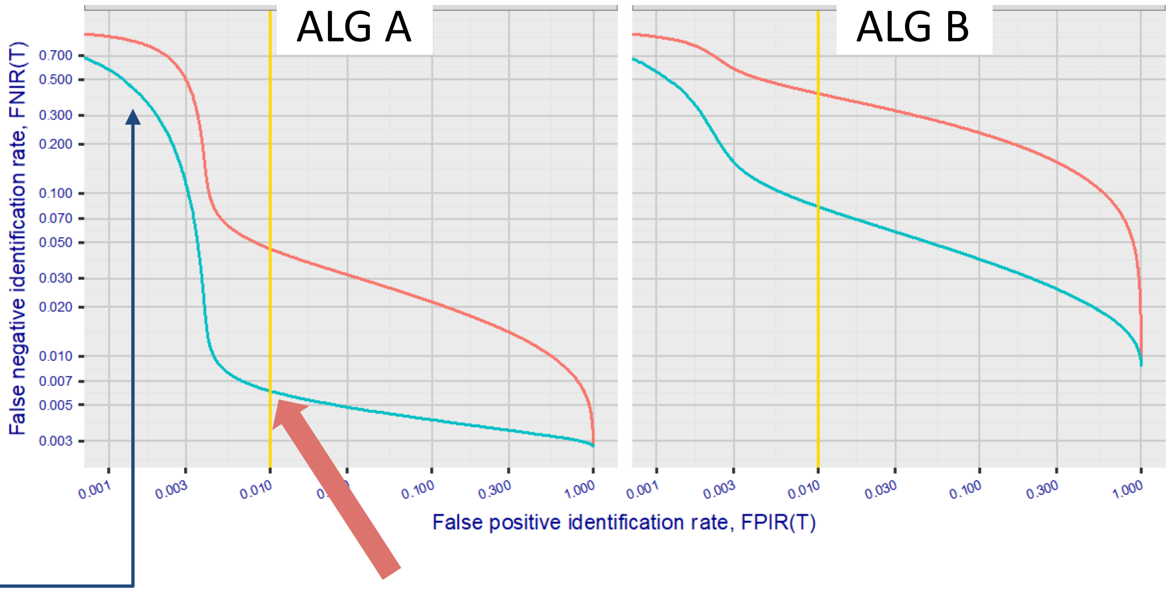
1:N search accuracy
Enroll N = 104 million ENTRY images; Search CIS Portraits

NIST

1. But version control matters:
   - ⇨ NIST eval vs. Productized

2. **Investigative search with N > 100M is possible, defensible**

3. Low FPIR is not attainable, limited by
   - ⇨ Unconsolidated IDs
   - ⇨ So do presence of twins > siblings > families



Miss rate: 0.6% ⇨ Hit rate: 99.4%

With threshold set so that only 1 in 100 non-mate search produces a false positive

FRVT 1:N Leaderboard 2020-08-12

NIST

Algorithm	Mugshot Mugshot N = 12000000 FBI FBI	Mugshot Mugshot N = 1600000 FBI FBI	Mugshot Webcam N = 1600000 FBI CBP	Mugshot Profile N = 1600000 FBI FBI	Visa Border N = 1600000 VISA AIRPORT
deepglint_001	_	0.0025 (4)	0.0116 (2)	0.7914 (24)	0.0051 (1)
sensetime_003	0.0024 (1)	0.0015 (1)	0.0105 (1)	0.1953 (4)	0.0067 (2)
nec_3	0.0031 (2)	0.0021 (3)	0.0149 (3)	0.5136 (13)	0.0070 (3)
paravision_005	0.0065 (4)	0.0030 (5)	0.0199 (5)	0.2335 (6)	0.0098 (4)
pixelall_004	0.0230 (14)	0.0109 (13)	0.0497 (17)	0.9992 (136)	0.0227 (5)
microsoft_6	0.0184 (9)	0.0086 (10)	0.0298 (9)	0.1174 (1)	0.0234 (6)
ntechlab_008	0.0218 (11)	0.0099 (12)	0.0364 (11)	0.1998 (5)	0.0284 (7)
idemia_007	0.0242 (15)	0.0123 (17)	0.0419 (16)	1.0000 (168)	0.0350 (8)
rankone_009	0.0258 (18)	0.0124 (18)	0.0597 (24)	0.8180 (25)	0.0427 (9)
dermalog_007	0.1097 (81)	0.0594 (95)	0.1202 (84)	0.9341 (39)	0.1027 (10)
gorilla_004	0.1109 (82)	0.0645 (107)	0.1317 (97)	0.8521 (26)	0.1059 (11)

- Values are threshold-based FNIR at FPIR = 0.003
- https://pages.nist.gov/frvt/html/frvt1N.html

State of the Industry

Performance

- » Massive expansion of industry
  - International markets + adoption
- » Massive gains in accuracy
  - Very accurate on high quality images
  - Better tolerance of poor image quality
  - Better tolerance of ageing (time lapse < 20 years)
  - Operate with larger databases
- » Accuracy varies greatly across the industry
  - China – EU – Japan – Russia – US
  - Buyer beware!
- » Some high volume applications (e.g. duplicate detection) require a high threshold for low false positives
  - Leads to higher false negatives
  - Image quality remains critical
- » Face-aware cameras
  - ISO/IEC 24358 camera capabilities

Limitations

- » Demographic differentials ”bias”
  - False positive >> False negative
    - False negatives from poor quality photos
    - Large false positive variations by race
    - Higher false positives among women, elderly, young
  - Algorithm matters
    - Better accuracy ⟶ smaller inequities
    - Only some Chinese algorithms give false positive rates on Chinese faces similar to those in Caucasian
    - Some one-to-many algorithms mitigate differentials
    - “Know-your-algorithm”
- » Twins not separable (false positives)
- » Attacks
  - Easy to “steal” a face for impersonation
  - Systems may be deployed without attack detection
  - Morphing
  - Adversarial
- » Human review capability is poor

AGEING


Images from presenter

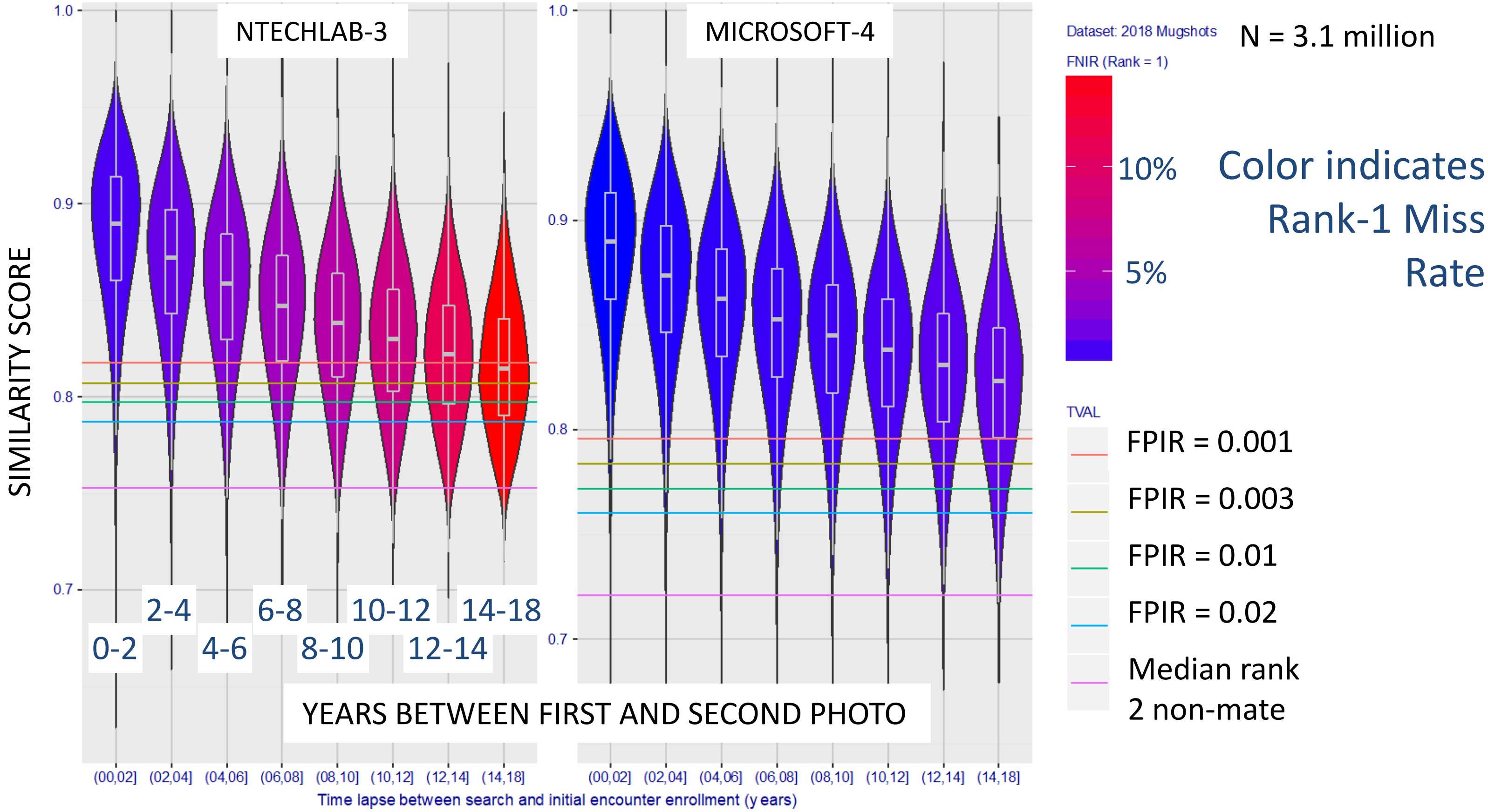
Ageing

2002-08 2004-10 2010-05 2012 2013-08 2018-06

Images from presenter

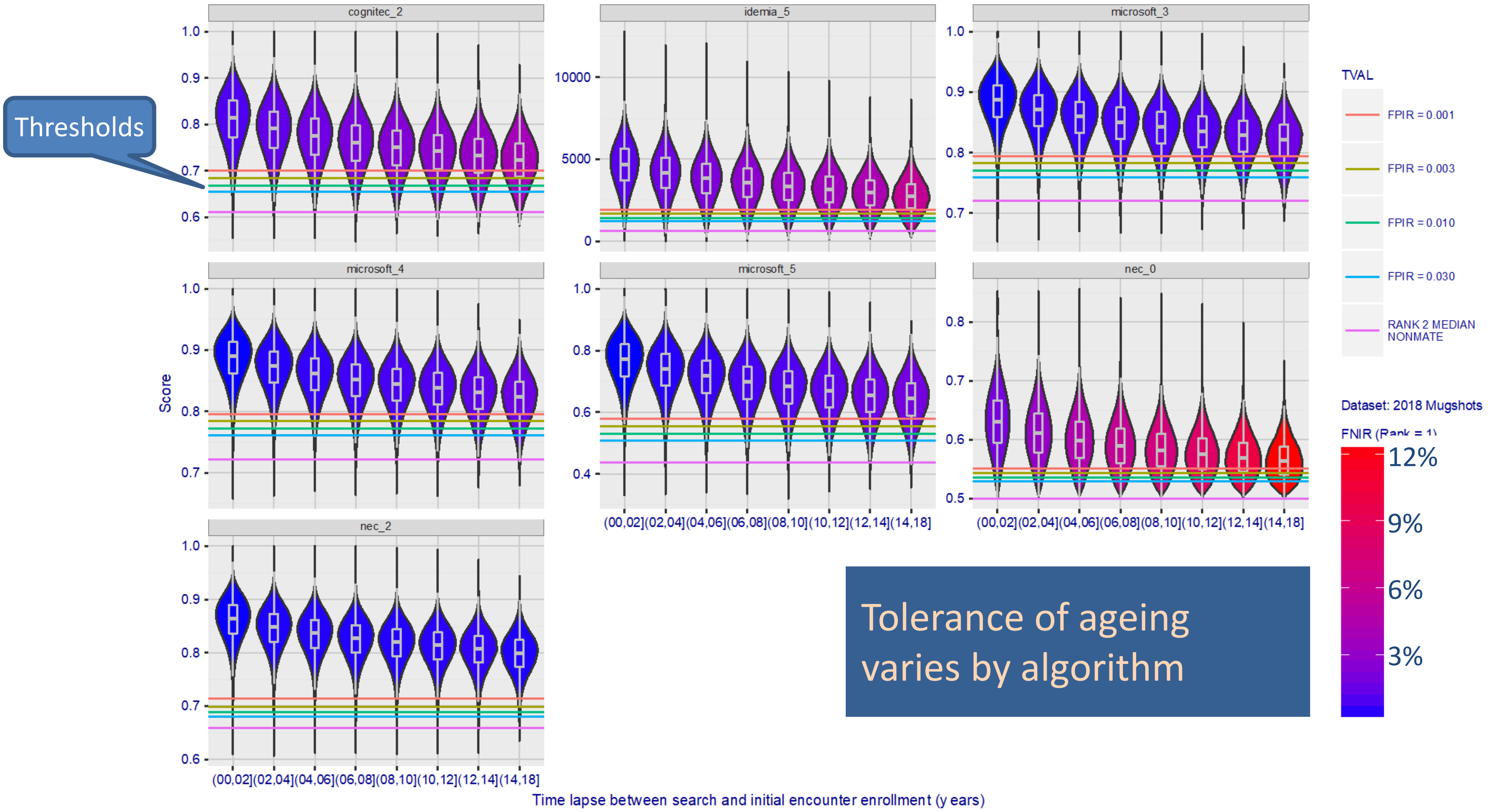
https://www.bellingcat.com/news/uk-and-europe/2018/09/26/skripal-suspect-boshirov-identified-gru-colonel-anatoliy-chepiga/

Mate score distributions under ageing



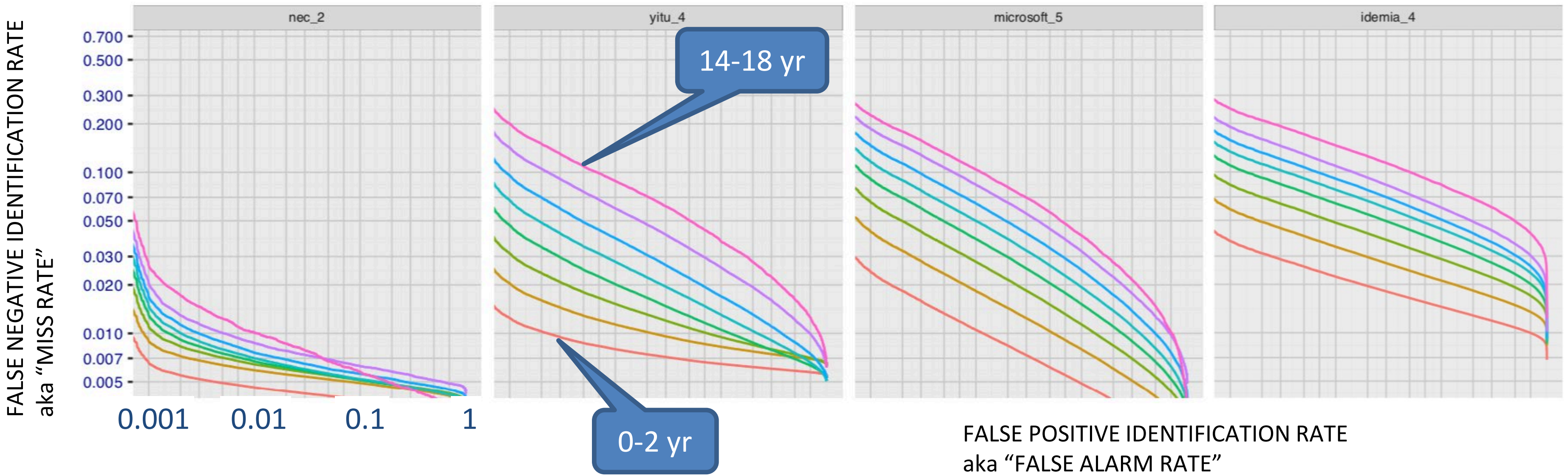
N = 3.1 million

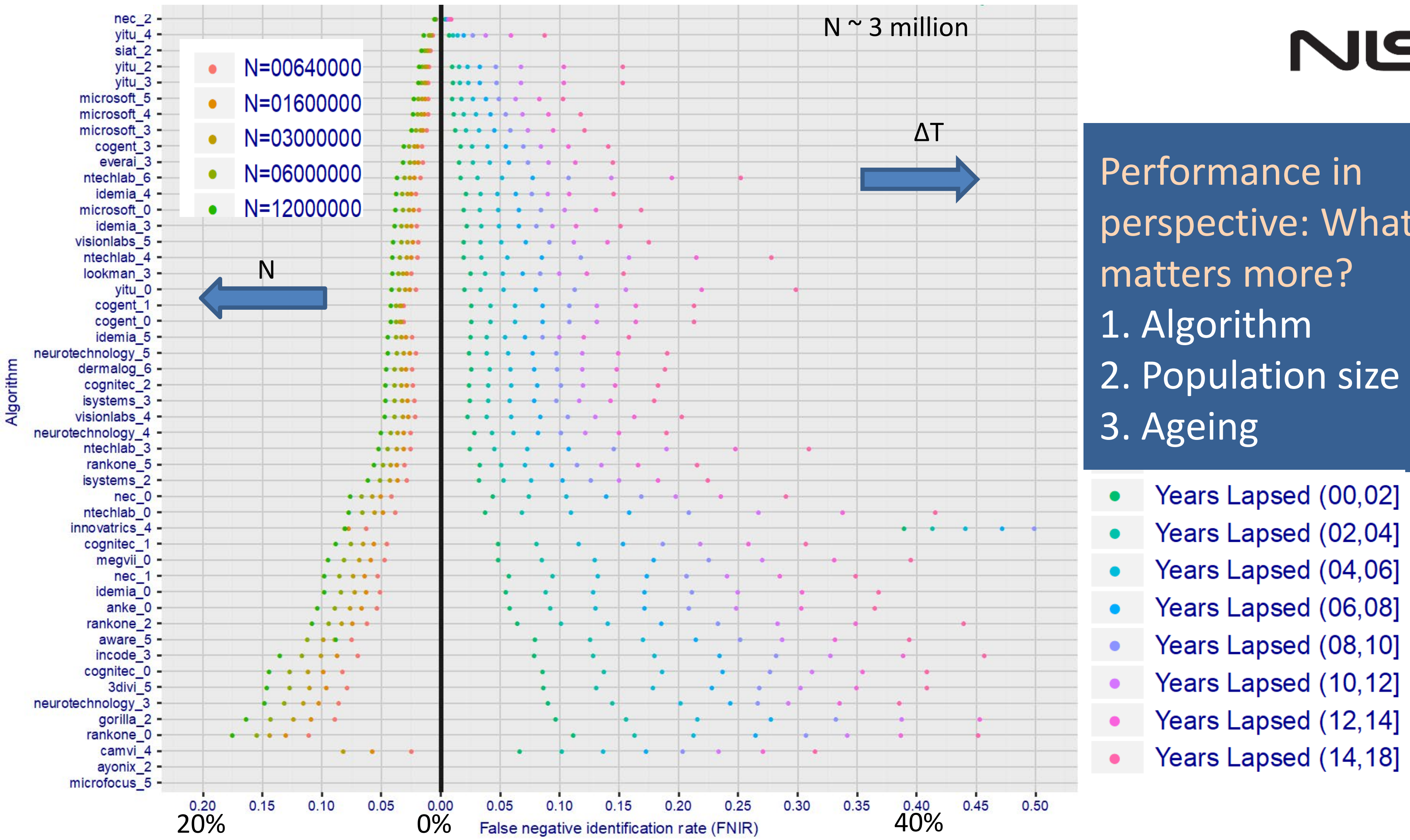
Color indicates Rank-1 Miss Rate



Tolerance of ageing varies by algorithm

Ageing: N = 3.1 million





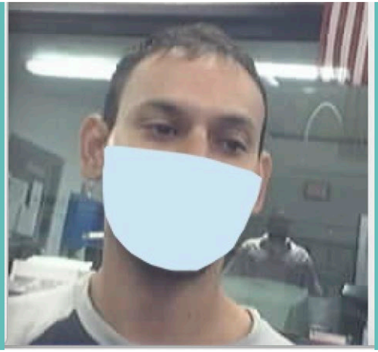
NIST

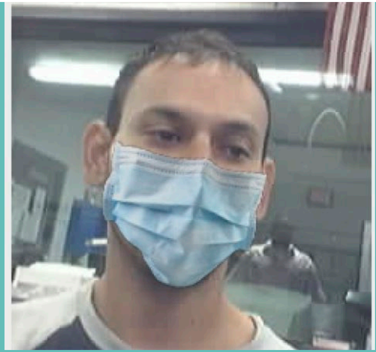
Performance in perspective: What matters more?

1. Algorithm
2. Population size
3. Ageing

Masks

What happens when you hide 40-70% of the face?

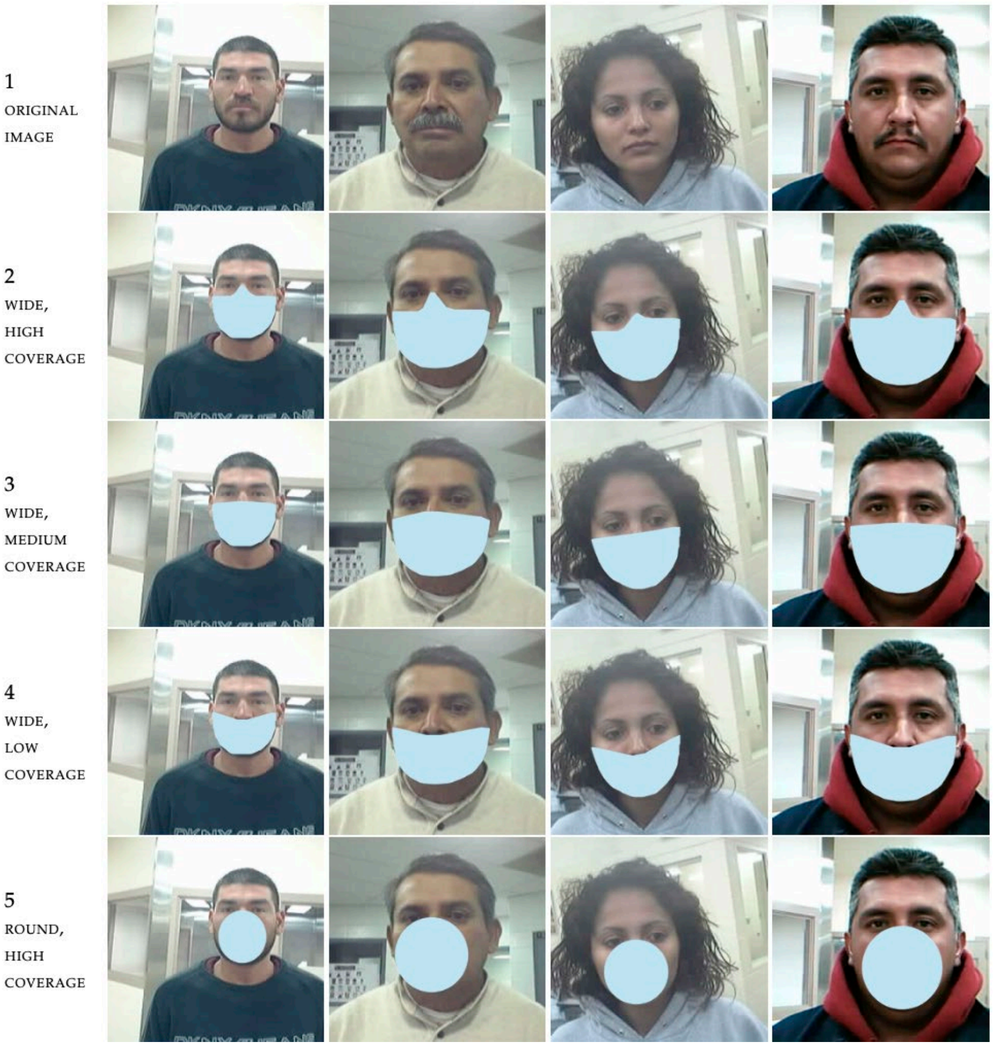




Synthetic masks

» NIST will vary
  • Shape, color, extent

» Positioning
  • Relative to landmarks reported by "dlib"
  • If "dlib" fails, then relative to detected eyes from good FRVT FR algorithms



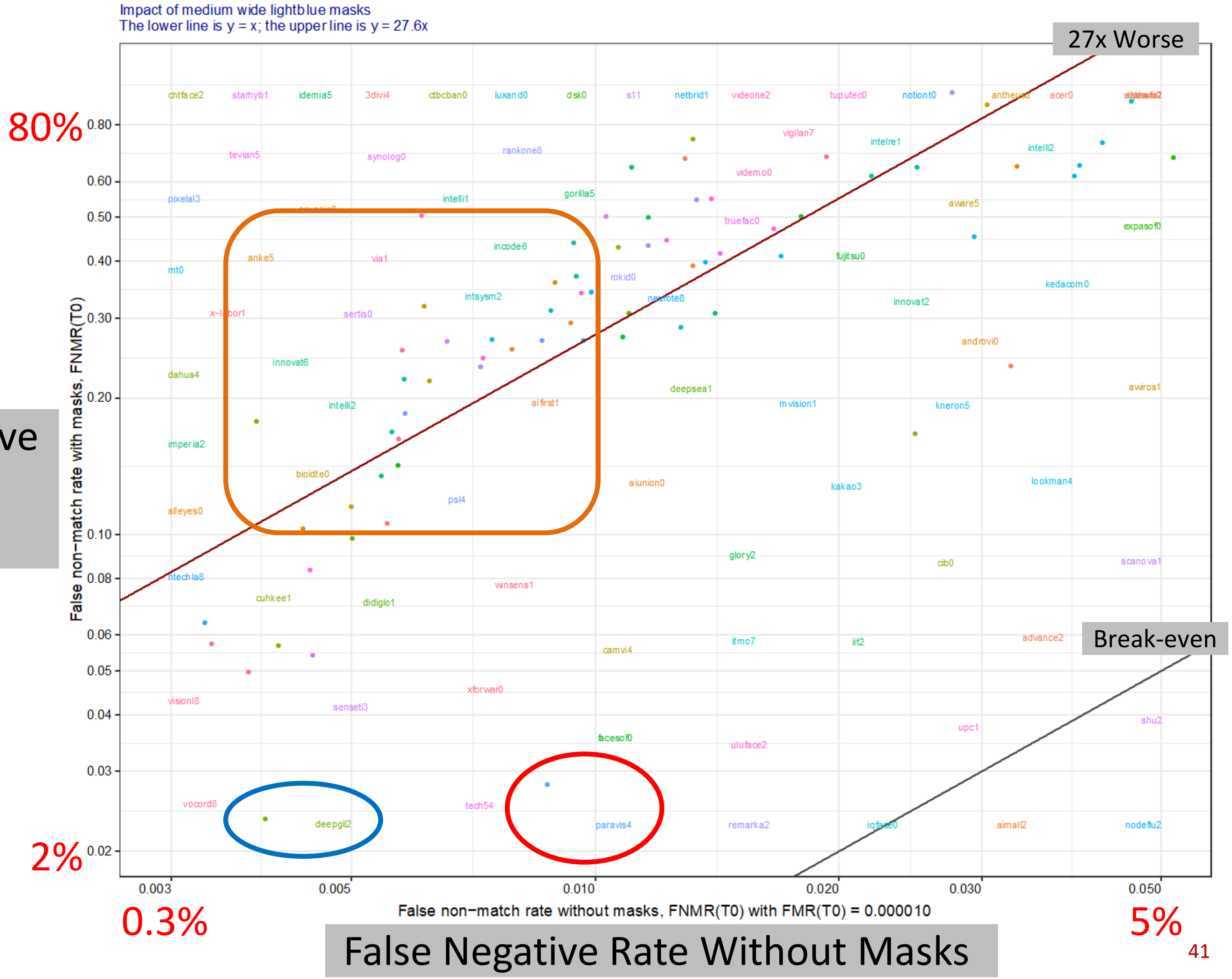
FRVT Leaderboard (all without masks)

https://pages.nist.gov/frvt/html/frvt11.html

Developer	VISA Photos FNMR @ FMR ≤ 0.000001	MUGSHOT Photos FNMR @ FMR ≤ 0.00001	MUGSHOT Photos FNMR @ FMR ≤ 0.00001 DT>=12 YRS	VISABORDER Photos FNMR@ FMR ≤ 0.000001	BORDER Photos FNMR @ FMR = 0.000001	WILD Photos FNMR@ FMR ≤ 0.00001	CHILD EXP Photos FNMR@ FMR ≤ 0.01
visionlabs-008	0.0036(4)	0.0031(6)	0.0040(4)	0.0045(3)	0.0079(1)	0.0308(10)	_
ntechlab-008	0.0061(10)	0.0056(17)	0.0108(29)	0.0042(1)	0.0080(2)	0.0312(20)	_
deepglint-002	0.0027(2)	0.0032(7)	0.0033(2)	0.0043(2)	0.0084(3)	0.0301(1)	0.3422(6)
dahua-004	0.0058(9)	0.0036(8)	0.0048(8)	0.0051(5)	0.0086(4)	0.0304(5)	_
vocord-008	0.0038(5)	0.0042(11)	0.0055(12)	0.0045(4)	0.0086(5)	0.0310(14)	_
cuhkee-001	0.0045(6)	0.0031(5)	0.0046(7)	0.0051(7)	0.0095(6)	0.1524(102)	_
sensetime-003	0.0027(3)	0.0027(1)	0.0027(1)	0.0051(6)	0.0100(7)	0.0355(45)	0.3683(7)
alleyes-000	0.0090(21)	0.0055(15)	0.0087(21)	0.0068(10)	0.0105(8)	0.0306(8)	_
tech5-004	0.0234(72)	0.0086(42)	0.0162(53)	0.0065(9)	0.0112(9)	0.0311(17)	_
yitu-003	0.0026(1)	0.0066(25)	0.0085(17)	0.0064(8)	0.0114(10)	0.0360(49)	_

Accuracy with and without masks

False Negative Rate With Masks



But... further challenges

Some algorithms may be usable

Algorithm	VISABORDER Photos FNMR @ FMR ≤ 0.00001 (NOT MASKED)	VISABORDER Photos FNMR@FMR ≤ 0.00001 (MASKED PROBE) lightblue, wide, medium coverage
deepglint-002	0.0039$^{(9)}$	0.0237$^{(1)}$
paravision-004	0.0088$^{(48)}$	0.0281$^{(2)}$
visionlabs-009	0.0028$^{(1)}$	0.0355$^{(3)}$
iqface-002	0.0086$^{(46)}$	0.0445$^{(4)}$
pensees-001	0.0106$^{(60)}$	0.0461$^{(5)}$
vocord-008	0.0038$^{(7)}$	0.0500$^{(6)}$
idemia-006	0.0048$^{(17)}$	0.0539$^{(7)}$
rankone-008	0.0134$^{(52)}$	0.5470$^{(58)}$
videmo-000	0.0140$^{(54)}$	0.5509$^{(59)}$
scanovate-001	0.2403$^{(80)}$	0.5973$^{(60)}$
intelresearch-001	0.0220$^{(61)}$	0.6184$^{(61)}$
kedacom-000	0.0391$^{(71)}$	0.6188$^{(62)}$
innovativetechnologyltd-002	0.0251$^{(64)}$	0.6454$^{(63)}$
idemia-005	0.0111$^{(44)}$	0.6469$^{(64)}$

Failure to verify rate rises from 0.4% to 2.4%

Most pre-pandemic algorithms do not tolerate masks

Failure to verify rate rises from 1% to 65%

NISTIR 8311 - Ongoing FRVT Part 6A: Face recognition accuracy with face masks using pre-COVID-19 algorithms
https://pages.nist.gov/frvt/html/frvt_facemask.html

Demographic Effects

FR accuracy varies by population

Landscape

- Race? Sex? Age? What else?
- Algorithms, cameras?
- 1:1 vs. 1:N
- False positives? Or Negatives?

NIST tests and results

- Criminal investigation
- Clustering media

Scope of NIST demographics work

NIST

» Algorithms
  - 187 algorithms, 99 developers
  - Mostly commercial, some universities
  - Prototypes from R&D labs

» Modes
  - One to one verification (DHS, DoS)
  - One to many identification (mugshots)

» Metrics:
  - False positives
  - False negatives
  - Failure to enroll

» Relevance to applications

» 18.3 million cooperative photos of 8.5 million people
  - DHS/CIS Application Photos
    - High quality
    - Race: 24 countries, 7 regions
    - Sex: M, F only
    - Age groups: [12-20], [20-35], [35-50], [50-65], [65-99].
  - DHS/CBP Entry Photos
    - Mediocre quality
    - Compare with CIS photos
  - DOS Visa photos
    - Age
  - FBI mugshots
    - Sex: M, F, only
    - Age groups: Adults above or below 45.
    - Race: Asian, Black, White, Native American

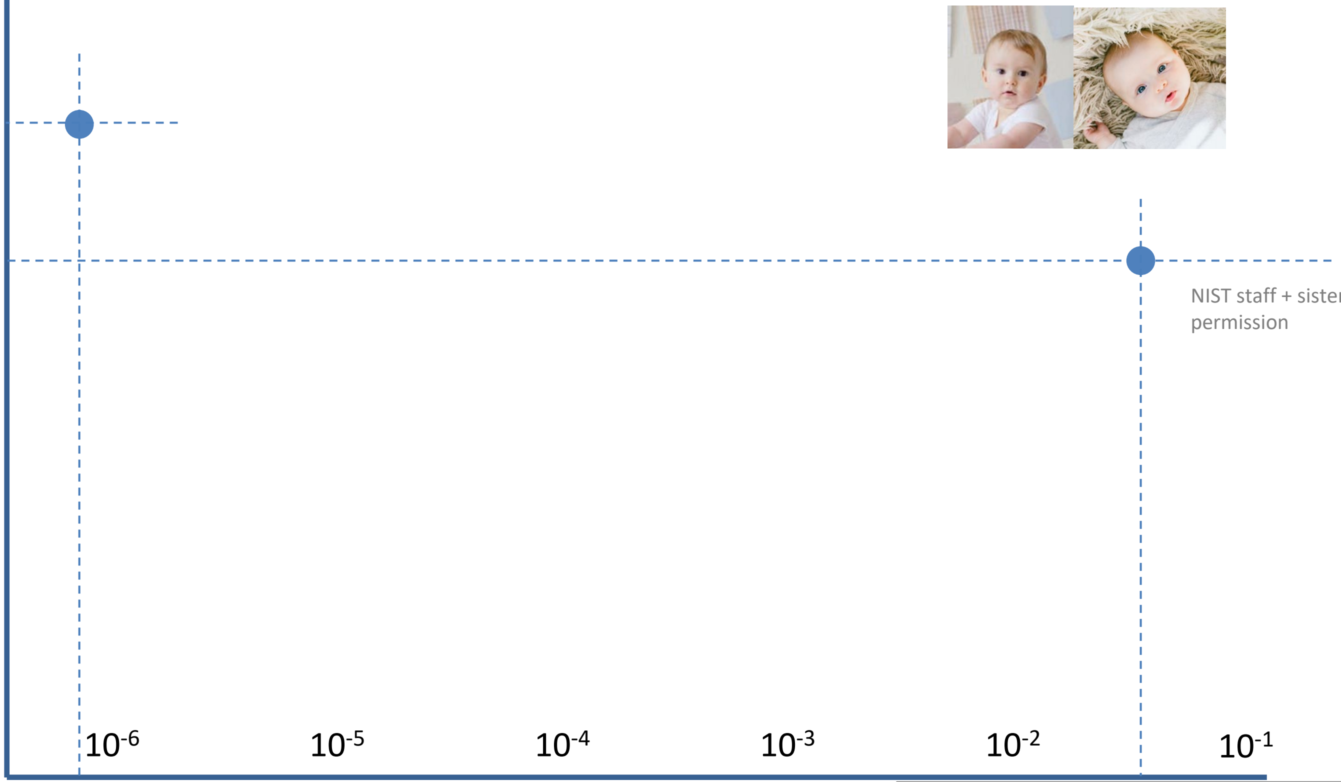
Declared instance of demographic differential

FNMR
False non-match rate

Proportion of genuine comparisons producing score below threshold, T.

See ISO/IEC 19795-1

Log-scale is typical to show small numbers.



NIST staff + sister, with permission

Log-scale is often required because low FMR values are operationally relevant.

FMR False match rate
Proportion of impostor comparisons searches yielding any candidates at or above threshold, T.

Declared instance of demographic differential

NIST

FNMR
False non-match rate

Proportion of genuine comparisons producing score below threshold, T.

See ISO/IEC 19795-1

Log-scale is typical to show small numbers.

NIST staff + sister, with permission

Apple Face. ID claims FMR ~ 1:1 000 000
https://support.apple.com/en-us/HT208108

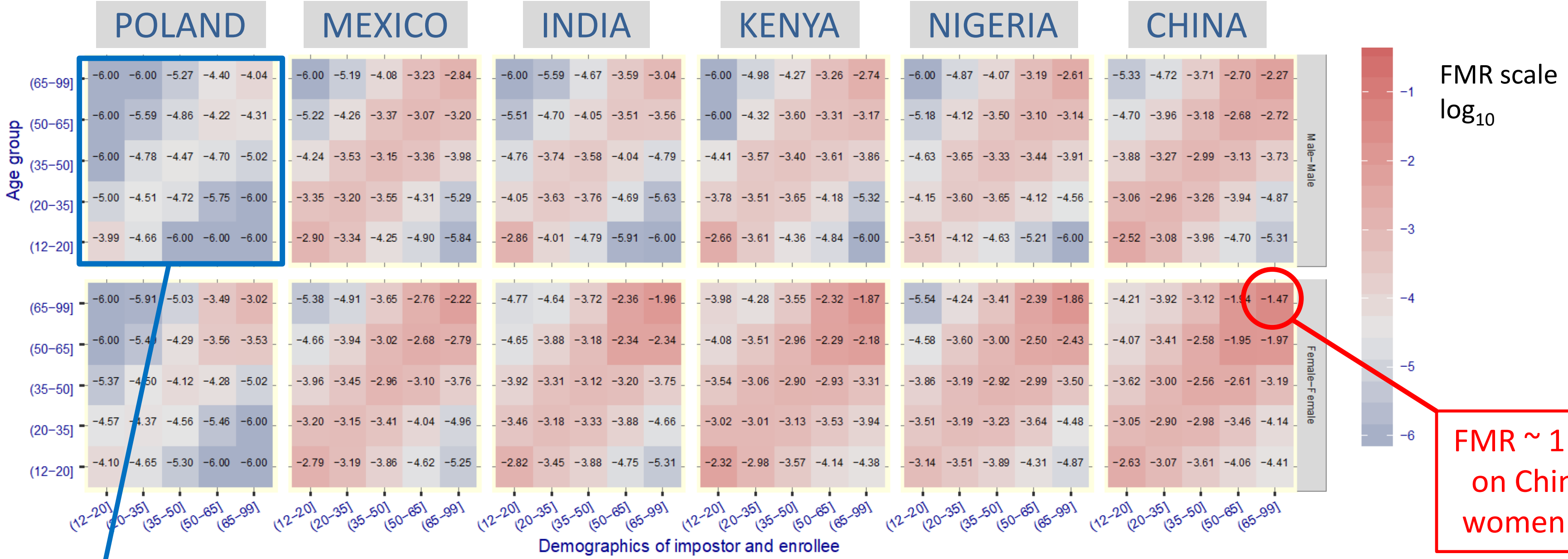
"The statistical probability is different for twins and siblings that look like you and among children under the age of 13, because their distinct facial features may not have fully developed."

10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1}

Log-scale is often required because low FMR values are operationally relevant.

FMR False match rate
Proportion of impostor comparisons searches yielding any candidates at or above threshold, T.

Cross-age false match rates in six countries, male x male, and female x female



FMR ~ 1 in 30 on Chinese women > 65

FMR on white males are below 1 in 30 000

Algorithm Imperial-002 with T = 1.381120
Nominal FMR = 0.00003
Dataset = Frontal cf. passports

Source: NIST IR 8280, 2019-12

Thinking through consequences: Three applications

1. Dispensing drugs

- Non-repudiation
- 1:1
- Volume: 100s per day
- Transactions are almost always mated
  - Prob(Impostor) is LOW
- False negative ⟶ Inconvenience
- False positive ⟶ Prescription drug fraud
- Who is harmed by demographic differential in FP?
  - Some pharmacists

2. Boarding a plane

- Facilitation of recording immigration exit vs. Access Control
- 1:N
- Volume: 100s per flight
- Transactions are almost always mated
  - Prob (Impostor) is LOW
- False negative ⟶ Paper boarding with airline staff
- False positive ⟶ Stowaway
  - but manifest exists, and legitimate customer may board also so “low” consequences
- Who is harmed by demographic differential in FP?
  - Airline.

3. Watchlist

- Soccer stadium. Counter-terrorism. Compulsive gamblers
- 1:N
- Volume: 10s of thousands per day
- Transactions are almost always non-mated
  - Prob (Genuine) is LOW
- False negative ⟶ Undetected “bad guy”
- False positive ⟶ Incorrect enforcement action ... civil liberties
- Who is harmed by demographic differentials in FP?
  - Bystanders

Summary

- Leading contemporary algorithms
  - Are very accurate
  - Increasingly tolerate poor image quality
  - Generally distribute errors inequitably across demographics

Summary

- Leading contemporary algorithms
  - Are very accurate
  - Increasingly tolerate poor image quality
  - Generally distribute errors inequitably across demographics

- False positive differentials much larger than false negative differentials
  - More false positives in Asian and African faces
  - More false positives in women
  - More false positives in the old and very young

Summary

- Leading contemporary algorithms
  - Are very accurate
  - Increasingly tolerate poor image quality
  - Generally distribute errors inequitably across demographics

- False positive differentials much larger than false negative differentials
  - More false positives in Asian and African faces
  - More false positives in women
  - More false positives in the old and very young

- One-to-many algorithms don’t necessarily behave like one-to-one
  - Some one-to-many effect a stabilization of the impostor distribution

Summary

- Leading contemporary algorithms
  - Are very accurate
  - Increasingly tolerate poor image quality
  - Generally distribute errors inequitably across demographics
- False positive differentials much larger than false negative differentials
  - More false positives in Asian and African faces
  - More false positives in women
  - More false positives in the old and very young
- One-to-many algorithms don’t necessarily behave like one-to-one
  - Some one-to-many effect a stabilization of the impostor distribution
- Algorithm matters
  - Accuracy
  - Demographic sensitivity
  - Know-your-algorithm
  - Traceability to (NIST) tests is not easy

Summary

- Leading contemporary algorithms
  - Are very accurate
  - Increasingly tolerate poor image quality
  - Generally distribute errors inequitably across demographics
- False positive differentials much larger than false negative differentials
  - More false positives in Asian and African faces
  - More false positives in women
  - More false positives in the old and very young
- One-to-many algorithms don't necessarily behave like one-to-one
  - Some one-to-many effect a stabilization of the impostor distribution
- Algorithm matters
  - Accuracy
  - Demographic sensitivity
  - Know-your-algorithm
  - Traceability to (NIST) tests is not easy

- Application matters
  - Error impact can be grave or inconsequential.

Summary

- Leading contemporary algorithms
  - Are very accurate
  - Increasingly tolerate poor image quality
  - Generally distribute errors inequitably across demographics
- False positive differentials much larger than false negative differentials
  - More false positives in Asian and African faces
  - More false positives in women
  - More false positives in the old and very young
- One-to-many algorithms don’t necessarily behave like one-to-one
  - Some one-to-many effect a stabilization of the impostor distribution
- Algorithm matters
  - Accuracy
  - Demographic sensitivity
  - Know-your-algorithm
  - Traceability to (NIST) tests is not easy
- Application matters
  - Error impact can be grave or inconsequential.
- Incomplete reporting in the press and academia
  - Confusion of face “analysis” with “recognition”
  - Don’t identify which component is at fault
  - Missing reports on false positives
  - Differentiate false positives from false negatives

Twins:
The Forgotten Demographic

Source: Twins Day Ohio collected by Notre Dame

Same person or not?

	Identical	Fraternal
How	Monozygotic	Dizygotic
Proportion of individuals that are a twin	0.9%	3.1%
Same-sex	100%	50% in theory 58% actually
TR gain since 1980	x1.5 since 1980	x1.9 since
Demographics	~ constant with age, geography	varies with mothers age, order, geography

Twins, triplets ... constituted 140,000 out of 4M births in 2015
https://www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_01.pdf

Scenario: Identical Twins

Probe is an identical twin

Gallery Size: 1.6 million

Algorithm	Rank of sibling	Score	FPIR
Microsoft	1	0.78	0.0007
NEC	1	0.77	0.0010
Idemia	1	3066	0.0007

Almost all algorithms give high scores

Candidate List

…

Rank 1

…

Scenario: Fraternal Twins

Probe is a fraternal twin

Gallery Size: 1.6 million

Algorithm	Rank of sibling	Score	FPIR
Microsoft	1	0.18	0.878
NEC	1	0.64	0.986
Idemia	11	670	0.909

Candidate List

…

Rank 1
(NEC/Microsoft)

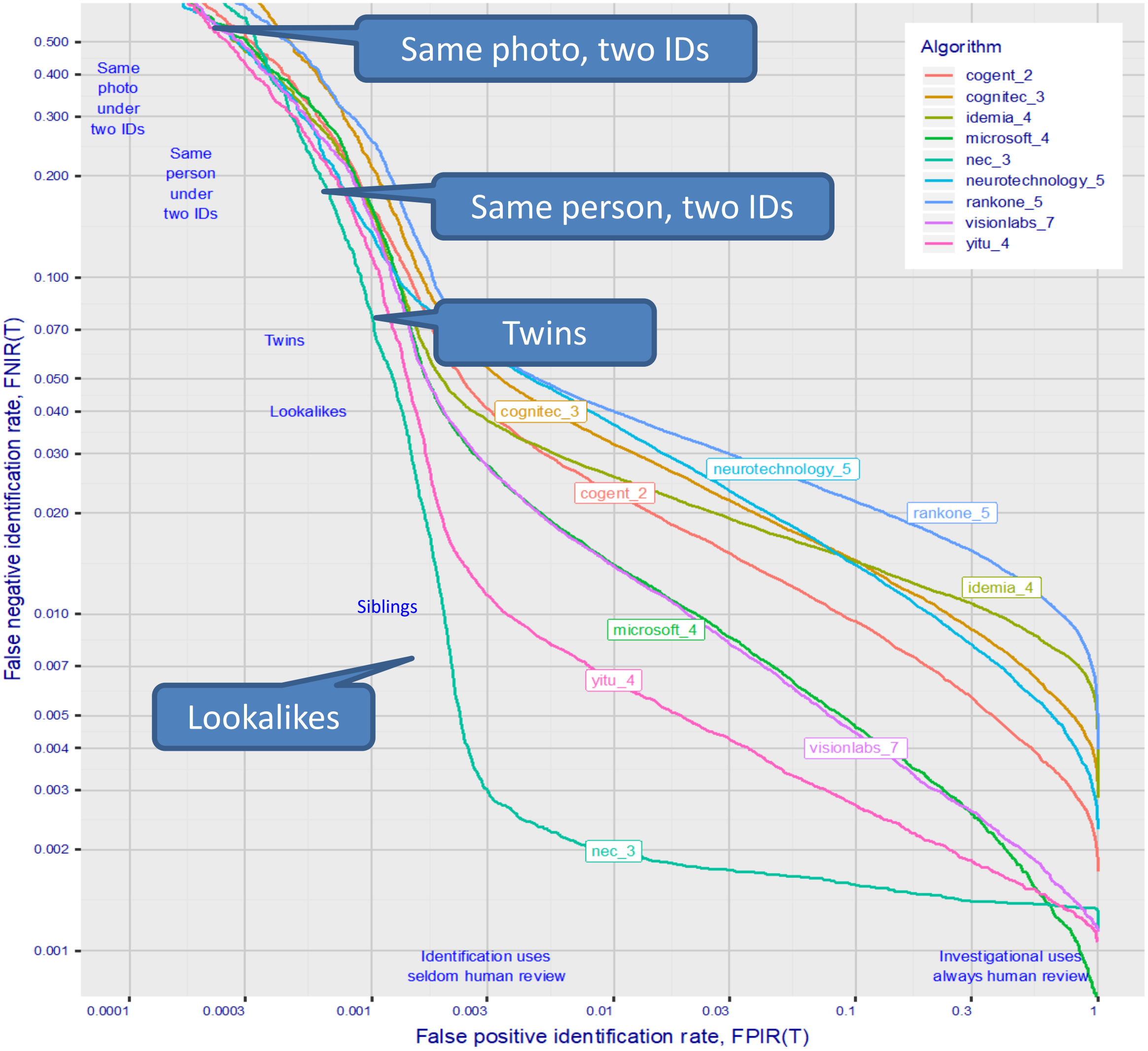
…

Rank 11
(Idemia)

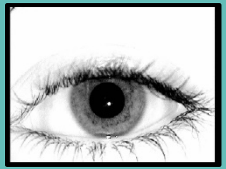
Face Recognition at National Scale

In a "closed" population (town, country):

- Low false positive rates cannot be achieved due to familial relationships
- Not expected with 10 fingerprints, and iris recognition



Why Face?
Versus Fingerprint, Iris.



Source: http://biometrics.it-sudparis.eu/english/index.php?menu=datasample

Modality selection

Modality	Image appearance standards	Availability (Ease of capture)	Permanence (ageing)	Uniqueness	Demographic problems	Twins	Retained reference images	Social acceptance
Face	Yes, compliance is difficult and not necessary	Fast Non-contact Socially accepted	Lower Low in children	Lower	Strong false positives in twins, families, same ethnicities, same sex, age	FMR ⟶ 1 identical twins FMR high also in fraternal	Social media, gov databases, (passport, drivers license)	Highest: Global ICAO passport
Finger Contact	Yes	Single fastest Four fast Ten slow (for gov use)	High Possibility of environmental damage	High Very high 10 fingers	No More false negatives in the elderly, very young, depends on sensor	FMR ⟶ 0	Legacy gov databases	Lower: Local cultural
Finger Contact-less	No: Interoperability problems with contact	Fast Four fingers for physical access control	High	High	No	FMR ⟶ 0	Yes, but only contact fingerprints	Higher: For PACS
Iris	Partial Guidance yes	Slower, optical tradeoffs. Capture both simultaneously	High, possibility of disease	High Very high two irides	No False negatives in elderly	FMR ⟶ 0	Few	Lower

- Nuanced discussion around many of these entries
- There are applications where property is not relevant

ONGOING BENCHMARKS

1. FRVT 1:1
Core Biometric Operation

2. FRVT 1:N
Search Performance

3. FRVT Morph
Morphed Photo Detection

4. FRVT Quality
Automated Quality Assessment

FRVT

Face Recognition Vendor Test

CURRENT PRODUCTS

Part 1:
Performance of 1:1 Verification Algorithms


NISTIR XXXX Draft

Ongoing Face Recognition Vendor Test (FRVT)
Part 1: Verification

Last: 2020-08-25
Next: 2020-07

Part 2:
Performance of 1:N Identification Algorithms


NISTIR 8271 DRAFT SUPPLEMENT

Face Recognition Vendor Test (FRVT)
Part 2: Identification

Last: 2020-03-27
Next: 2020-08

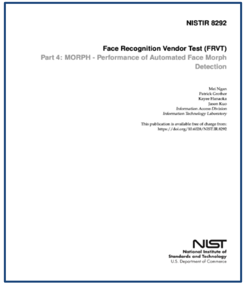
Part 3:
Demographic Effects in Face Recognition


NISTIR 8280

Face Recognition Vendor Test (FRVT)
Part 3: Demographic Effects

Last: 2019-12-19
Next: 2020-09

Part 4:
Performance of Morph Detection Algorithms


NISTIR 8292

Face Recognition Vendor Test (FRVT)
Part 4: MORPH - Performance of Automated Face Morph Detection

Last: 2020-07-24
Next: 2020-09

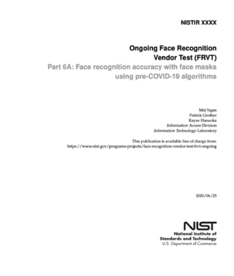
Part 5:
Performance of Image Quality Assessment Algorithms


Draft NISTIR XXXX

Ongoing Face Recognition Vendor Test (FRVT)
Part 5: Quality Assessment

Last: 2020-07-27
Next: 2020-09 est.

Part 6:
Performance of Face Recognition with Face Masks


NISTIR XXXX

Ongoing Face Recognition Vendor Test (FRVT)
Part 6A: Face recognition accuracy with face masks using pre-COVID-19 algorithms

Last: 2020-07-27
Next: 2020-08 est.

Part 7:
Performance of Face Recognition on Twins

Last:
Next: TBD

Current technical issues in face recognition

NIST

Impeding accuracy

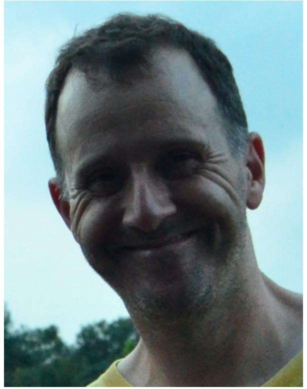
- Ageing
- Twins
- Demographics differentials
  - False positives WORSE THAN false negatives
- Poor quality images
  - Pose
  - Illumination
  - Resolution
  - Occlusion (face masks)
  - Cropping
  - Distortion
- Lack of capture standards

Impeding security

- Morph attack detection
- Presentation attack detection
- Tampering
- Fakes

ISO/IEC 24358

FASTER, BETTER, FACE-AWARE CAPTURE (QUALITY MATTERS!)



Images from presenter

Problems:

a) Non-frontal faces

b) No-faces, multiple-faces

c) Over-, under-exposure

d) Human review errors

e) Morphing

f) Inadequate presentation attack detection

Potential Solutions:

a) Face pose detector

b) Face detectors

c) 12 bits or closed-loop control

d) Higher resolution, better compression, 3D

e) Crypto for tamper-proofing

NIST IFPC Conference: October 27-29.

		IFPC 2020 - Tuesday Oct 27			IFPC 2020 - Wednesday Oct 28			IFPC 2020 - Thursday Oct 29
	07:20	Welcome		07:00	Welcome		07:00	Welcome
11	07:30	**Arun Vemury**, DHS Science + Technology Directorate (US): Welcome + DHS context	21	07:10	**Lars Ericson**, IARPA (US): Overview of the IARPA efforts on face recognition	31	07:10	**Rebecca Heyer**, DSTG (AU): Face recognition in Australia
12	07:40	**Istvan Szilard Racz**, EU-LISA: European Entry-Exit System	22	07:40	**Stergios Papadakis**, Johns Hopkins Applied Physics Lab (US): Results from the Odin program on presentation attack detection	32	07:40	**Martins Bruveris,** Onfido (UK): Reducing geographic performance differentials for face recognition
13	08:10	**Anna Stratmann**, BSI (DE): Biometric processes of the Entry Exit System	23	08:10	**Marta Gomez-Barrero,** Hochschule Ansbach (DE): Presentation attack detection and unknown attacks	33	08:10	**Mosalam Ebrahimi**, Trueface AI (US): A bias mitigation strategy: overcoming the problem of overly confident false matches
14	08:40	**Patrick Grother**, NIST (US): Measurement of face recognition performance for Entry-Exit	24	08:40	**Christian Rathgeb,** Hochschule Darmstadt (DE): Impact of facial beautification on face recognition: From plastic surgery to makeup presentation attacks	34	08:40	**Jacqueline Cavazos**, UT Dallas (US): Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?
	09:10	Break 15 mins		09:10	Break 15 mins		09:10	Break 15 mins
15	09:25	**Arun Ross,** Michigan State University (US): Look-alike disambiguation in face recognition	25	09:25	**Stéphane Gentric**, Idemia (FR): Synthetic faces: Are they new identities and can they be used in evaluation?	35	09:25	**John Howard & Yevgeniy Sirotin,** SAIC (US): Revisiting the Fitzpatrick Scale and Face Photo-based Estimates of Skin Phenotypes
16	09:55	**P. Jonathon Phillips,** NIST (US): Item response theory for designing calibrated face ability tests	26	09:55	**Mei Ngan,** NIST (US): Face morphing - threats, technology, what's next	36	09:55	**Michael Thieme,** Novetta (US): AI performance assessment standardization in SC 42 – implications for biometrics
17	10:25	**Laura Rabbitt & Yevgeniy Sirotin**, SAIC (US): Human-Algorithm Teaming in Face Recognition	27	10:25	**Christoph Busch**, NTNU/Hochschule Darmstadt (NO/DE): Face morphing attack detection in the iMARS project	37	10:25	**Johanna Morley**, Metropolitan Police (UK): Testing of demographic effects in an operational live facial recognition from video system
18	10:55	**Carina A. Hahn**, NIST (US): The effectiveness of fusion in face recognition	28	10:55	**Kiran Raja**, NTNU/MOBAI (NO): Morphing Attack Detection - obstacles for research to deployment	38	10:55	**Brendan Klare**, Rank One Computing (US): Efficiency considerations for face recognition algorithms
19	11:25	**Amy N. Yates**, NIST (US): Perceptual face abilities of face examiners for varying tasks	29	11:25	**Chen Liu**, **Zander Blasingame**, Clarkson U., **David Doermann**, U. at Buffalo, **Jeremy Dawson**, West Virginia U. (US): Center for Identification Technology Research (CITeR) Morph Attack Detection and Mitigation Projects	39	11:25	**Bhargav Avasarala**, Paravision (US): Challenges and considerations for masked face recognition
1a	11:55	**John Howard & Yevgeniy Sirotin,** SAIC (US): Quantifying Race and Gender Effects in Face versus Iris Algorithms	2a	11:55	**Pawel Drozdowski** Hochschule Darmstadt (DE): Workload reduction in FR identification with morphing	3a	11:55	**Tony Mansfield**, NPL (UK): The new ISO/IEC 19795-1 biometric performance testing and reporting standard
1b	12:25	**Patrick Grother**, NIST (US): Now under development: ISO/IEC 29794-5 face image quality standard ISO/IEC 24358 face-aware capture specifications	2b	12:25	**Mei Ngan**, NIST (US): Evaluation of face recognition accuracy for subjects potentially wearing face masks	3b	12:25	
		12:55 Close			12:55 Close			12:55 Close

https://www.nist.gov/news-events/events/2020/10/international-face-performance-conference-ifpc-2020

THANKS

PATRICK.GROTHER@NIST.GOV

FRVT@NIST.GOV

Facial Recognition Performance and Its Measurement

PRESENTED BY:
Patrick Grother
National Institute of Standards and Technology

MODERATED BY:
Stephen Redifer

2020-09-24