NIST FACE IN VIDEO EVALUATION (FIVE)

AUSTIN HOM, PATRICK GROTHER, MEI NGAN
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

IFPC
April 1st, 2025

NIST'S Open Benchmarks

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

FRTE

FACE **RECOGNITION** TECHNOLOGY EVALUATION

RECOGNITION: **WHO** IS IN AN IMAGE

1:1 VERIFICATION	SAME PERSON OR NOT?
1:N SEARCH	WHO? WHERE? WHEN?
FACE IN VIDEO 2024	**1:N ON NON-COOP PEOPLE**
TWINS DISAMBIGUATION	SAME PERSON, OR TWIN?

FATE

FACE **ANALYSIS** TECHNOLOGY EVALUATION

ANALYSIS: **ABOUT** IN AN IMAGE

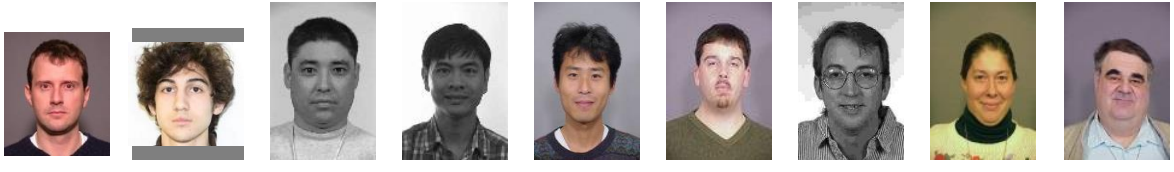
MORPH DETECTION	TWO PEOPLE IN ONE FACE?
QUALITY + DIAGNOSTICS	HOW BAD IS THIS PHOTO
PAD	SUBVERSIVE PHOTO?
AGE ESTIMATION	HOW OLD? OLD ENOUGH?

Benchmarks are:

- Independent
- Free
- Regular
- Fast
- Repeatable
- Fair
- Black box
- IP-protecting
- Open globally
- Large-scale
- Sequestered datasets
- Statistically robust
- Public
- Transparent
- Extensible
- **ABSOLUTE ACCU**
- **RELATIVE ACCU**

Challenges of FR in video

Surveillance Video Related to Boston Bombings



Challenges for FR

>> Pose

- Compound rotation of head to optical axis

>> Resolution

- Range to subject
- Legacy camera
- Adverse compression for storage or transmission
- Motion blur

>> But

- Multiple frames …

NIST Face in Video Evaluation (FIVE) 2024

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

Goals

- Assessment of the state-of-the-art of **1:N face recognition (FR)** on video sequences (and relative improvements since FIVE 2015)
- Assessment of FR on videos with low quality
  - Low resolution including compressed video and long range imaging affected by atmospheric turbulence
  - Elevated cameras resulting in high look-down angle
  - Passively observed subjects who at no point face the camera directly
  - Face detection in wide field-of-view imagery
- Absolute accuracy
- Comparative accuracy of algorithms
- Comparative computational cost
- Threshold calibration - ability to target specific false positive identification rates

Face In Video Evaluation

Out of Scope

- Re-identification
- Anomaly detection
- Detection of un-cooperative actions, evasion
- **Other modalities (e.g., body and GAIT recognition)**
- **Clothing and other non-facial metrics**
- 1:1 verification

FIVE 2024 - Timeline

Date	Activity
2024-01-23	First draft of API published
2024-02-29	API comments due
2024-03-07	Final API published
2024-03-11	Phase 1 submission window opens
2024-05-18	Phase 1 submission window closes
2024-06-28	Phase 1 report cards distributed
2024-07-01	Phase 2 submission window opens
2024-08-30	Phase 2 submission window closes
2024-10-18	Phase 2 initial report cards distributed
2024-12-06	Phase 2 final report cards distributed
2025 – Q2/Q3	Public report published

FIVE 2024 - Who Participated

» 11 developers from around the world submitted 31 algorithms total

- Cognitec
- Corsight
- Dermalog
- Gpstechvn
- Idemia
- Innovatrics
- NEC
- Neurotechnology
- ROC
- Viante
- Videmo

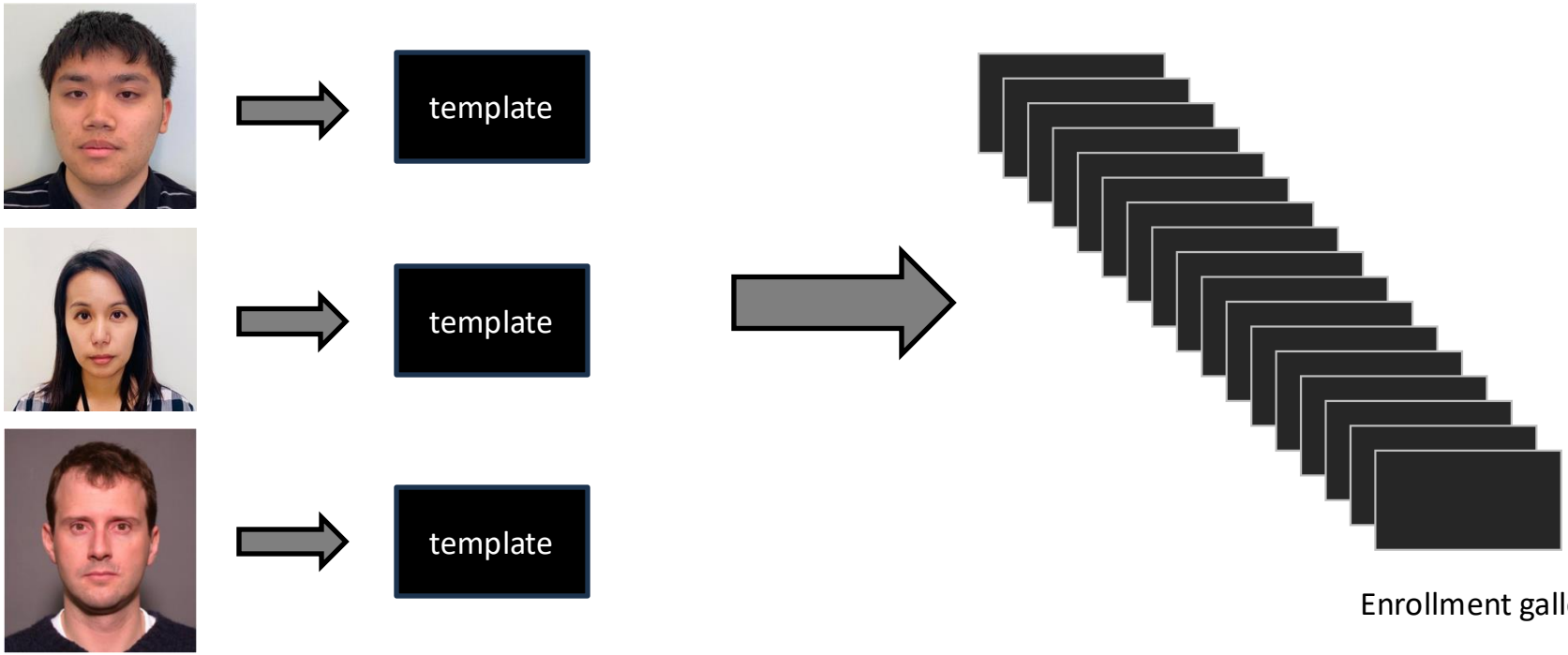
FIVE 2024 – How Algorithms Are Run

- Dynamically-linked C++ library (.so file)
- Run “bare metal” on Linux (Ubuntu 20.04.3)
- Hardware: Intel server-class CPUs (no GPUs)

- Hard duration limits - measured on a single CPU core
  - Still Enrollment (face detection + feature extraction + and encoding): 1.5 sec / image
  - Video Enrollment (face detection + tracking + feature extraction + encoding): 1.5 sec / frame / person
  - Finalize Enrollment (gallery size=10,000): *4000 seconds*
  - Search (gallery size=10,000): *1 second*
- Code that crashes is rejected.

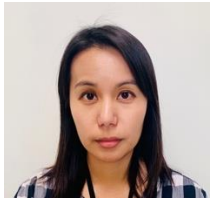
What's In The Gallery – Typical

» Galleries were typically composed of templates generated from still imagery



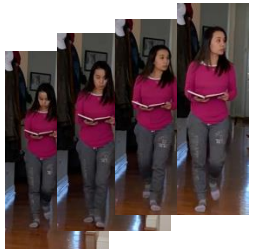
What's In The Gallery – Occasionally

» Galleries will occasionally be composed of templates generated from a combination of multiple stills and/or video sequences of the same subject



still image

+



video sequence

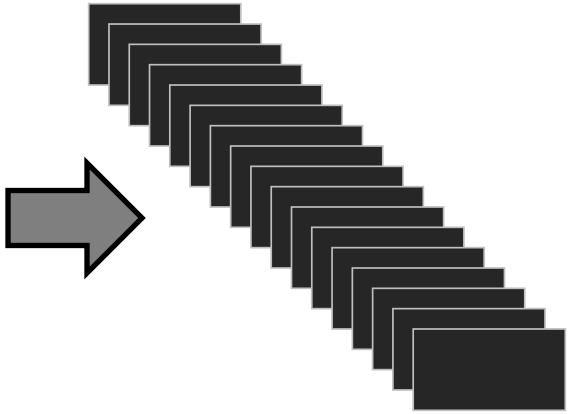
template

video sequence

+

video sequence

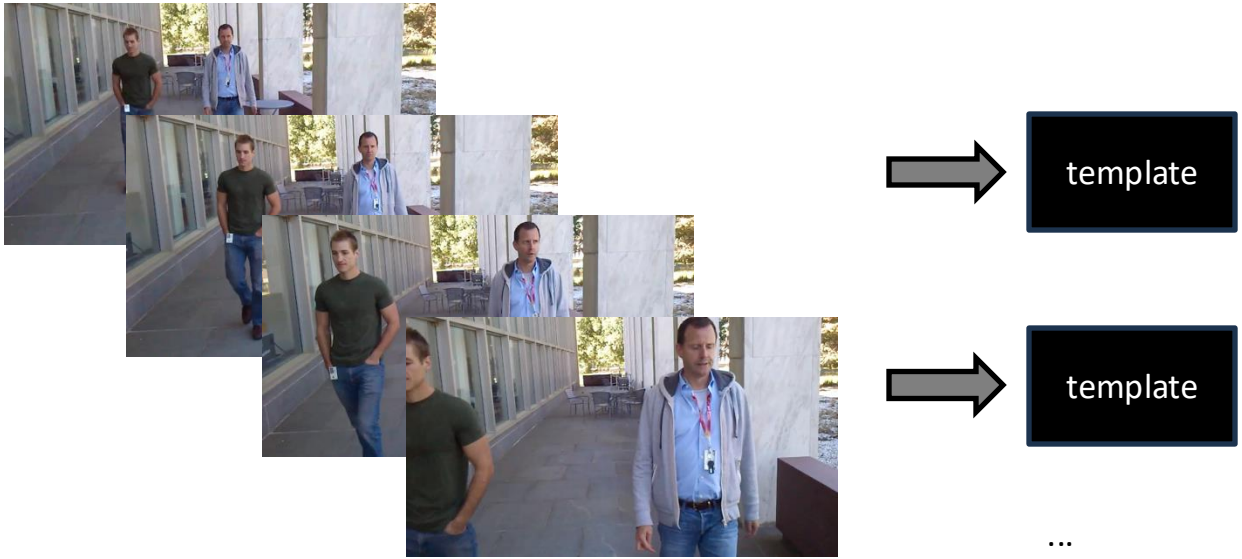
template



Enrollment gallery

Probes

» Probes were single video clips (sequence of frames) with one or more people in the scene



RECOGNITION IS THE GOAL: NOT DETECTION, NOT TRACKING

Software should maximize recognition performance by
- detecting person,
- tracking that person through time,
- determining which is most recognizable imagery, and
- extracting features / embeddings

We do **not** report metrics for
- false detection
- missed detection
- spatial (bounding box) location accuracy
- track integrity
- restoration

One video, one person, one track → K = 1 search

Non-detection is immaterial if subject is (later) found correctly and identified

Score	ID	
4.498	Marcia	TP
1.616	Mei	
0.750	Mae	
0.300	Maria	
0.128	Melissa	
0.072	Marissa	
0.012	Melani	
0.007	James	

ALGORITHM SOFTWARE

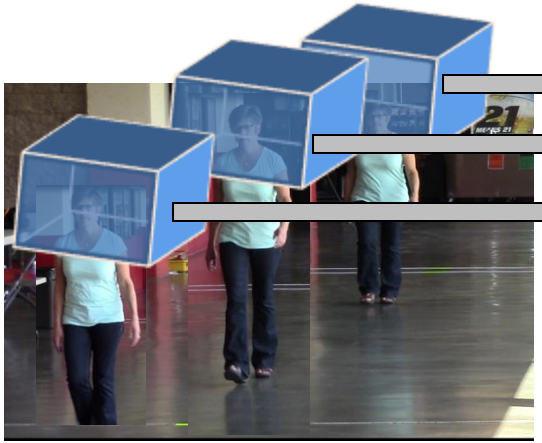
1. Detects K = 1 faces
2. Software extracts a set of features
3. Software searches gallery producing a candidate list of fixed length L ≤ N. The value of L is an input specified by NIST via the API

NIST EVALUATION

1. Chooses a threshold T e.g. 4.0
2. Records a false negative error unless the candidate list includes ID=123 at or above T

Repeats this for many probes and many threshold values to produce FNIR vs. T.

One Video, One Person, Multiple Tracks ⟶ K ≥ 0 Searches



template

template

template

Score	ID
3.142	Mary
2.998	Maria
1.626	Marcia
0.707	Mae
0.330	Mei
0.198	Melissa
0.074	Marissa
0.016	Melani

FN

Score	ID
2.901	Mary
2.798	Marcia
1.616	Mei
0.750	Mae
0.300	Maria
0.128	Melissa
0.072	Marissa
0.012	Melani

FN

Score	ID	
4.498	Marcia	TP
1.616	Mei	
0.750	Mae	
0.300	Mary	
0.128	Melissa	
0.072	Marissa	
0.012	Melani	
0.007	James	

NIST

1. Chooses a threshold T, e.g. 4.0
2. Records a false negative error unless ANY of the K candidate lists includes ID=Marcia at or above threshold T

Repeats this for many probes and many threshold values to produce FNIR vs. T.

DISCUSSION

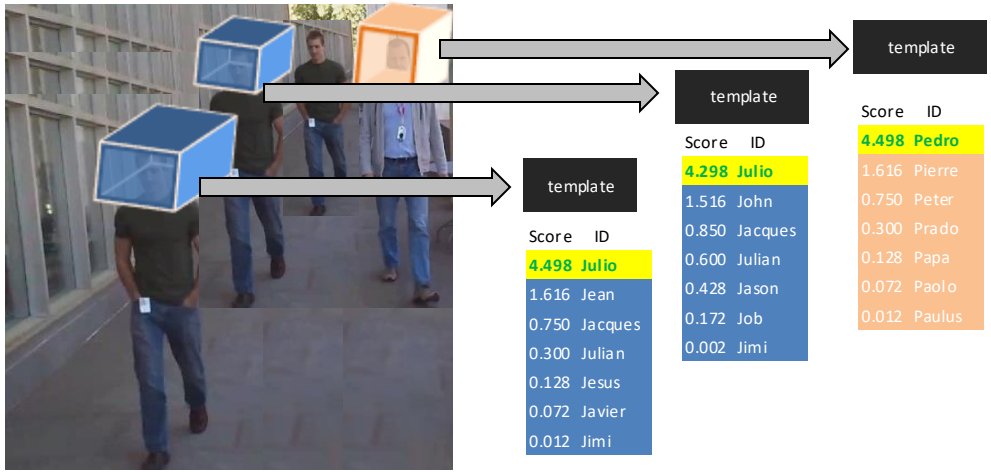
This method allows tracks to be broken. NIST doesn't care about track integrity per se, only that recognition succeeds.

The algorithm implementation is free to select best quality frames, to perform restoration, to perform feature level fusion, to produce a template that internally contains multiple embeddings to allow score-level fusion.

ALGORITHM SOFTWARE

1. Even if the person is present in entire clip, as she is here, an algorithm might find the person say K = 3 times (broken tracks)
2. Software extracts K sets of features (aka templates)
3. Software searches gallery producing K candidate lists each of fixed length L ≤ N.

One Video, Two Persons ⟶ K ≥ 0 Searches



DISCUSSION

If say only Julio is in the gallery, then the algorithm is rewarded for correctly finding him at some point. The scoring software does not reward twice for finding the person twice.

If say the person on the right is not in the gallery, then the high score against gallery person Pedro could be accumulated into a count of false positives. **That said, false positives are usually measured over sets where the gallery and probes are subject-disjoint.**

If the gallery is unconsolidated, and Julio is enrolled multiple times, the algorithm is rewarded for finding **any** occurrence of Julio in the gallery.

1:N SEARCH FALSE POSITIVES IN OPERATIONS

Rite Aid's A.I. Facial Recognition Wrongly Tagged People of Color as Shoplifters

Under the terms of a settlement with the Federal Trade Commission, the pharmacy chain will be barred from using the technology as a surveillance tool for five years.

https://www.nytimes.com/2023/12/21/business/rite-aid-ai-facial-recognition.html Eduardo Medina, 2023-12-21

FTC REPORTS THAT "*THE SYSTEM GENERATED THOUSANDS OF FALSE-POSITIVE MATCHES*"

https://www.ftc.gov/news-events/news/press-releases/2023/12/rite-aid-banned-using-ai-facial-recognition-after-ftc-says-retailer-deployed-technology-without

https://www.engadget.com/ftc-bans-rite-aid-from-using-facial-surveillance-systems-for-five-years-053134856.html

Measuring False Positives

FALSE POSITIVE IDENTIFICATION RATE

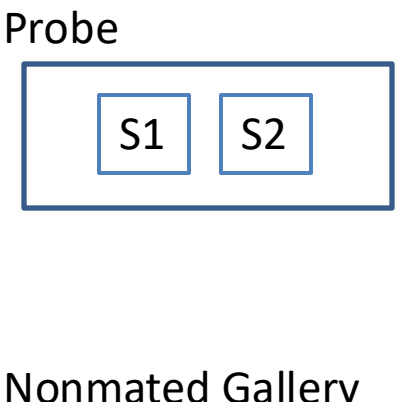
- Conventional to measure the **false positive identification rate** (FPIR)
  - Run searches of individuals who are known to be absent from the enrolled gallery
  - FPIR computed as the proportion of searches that produces one or more false positives above a threshold, T
- In FIVE 2024, false positive error is reported as FPIR given the availability of videos where we know the exact number of people present

NUMBER OF FALSE POSITIVES

- In FIVE 2015, calculating FPIR was not possible, because the number of individuals in the search imagery was not known
- Instead, false positive errors were stated as the **number of false positives** from running searches of individuals who are known to be absent from the gallery

Calculating False Positive Identification Rate

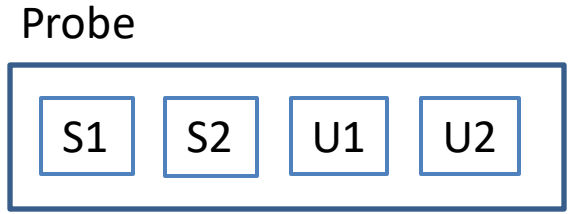
- Gallery containing only **nonmated** subjects
- Probe videos containing a **known number of people** known not to be in the gallery
- Any subjects who come up above threshold contribute to FPIR
- FPIR = # of subjects with any track with hit above threshold T (max 1 per subject in probe) / total # of subjects in probes

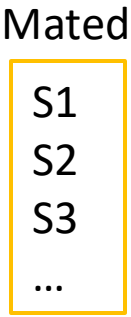


Calculating False Negative Identification Rate

- Gallery containing **mated** subjects
- Mated probe videos containing people known to be in the gallery, can also include unknown subjects
- Any subjects who does not come up above threshold contribute to FNIR
- FNIR = # of mated subjects with no tracks with hit above threshold T /

total # of mated subjects in probes





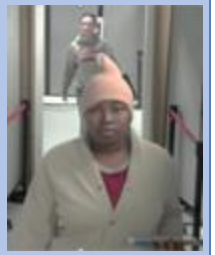
FIVE RESULTS

FIVE 2015 Datasets: People On The Move

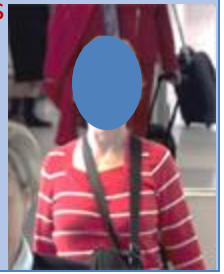
NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

1) Sports Arena
11 consumer cameras
50 actors

2) Passenger Loading Bridge
10 pro cameras
50 actors



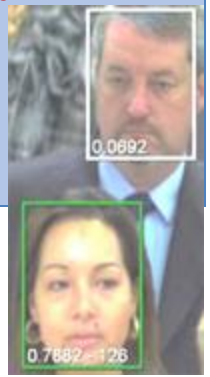
3) Concourse
10 pro cameras
50 actors



4) Self Boarding Gate
Classic chokepoint
1 webcam
250 actors

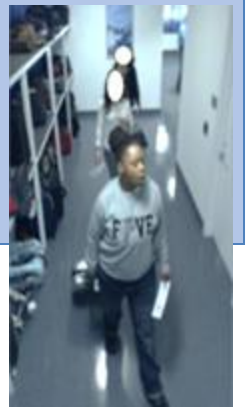
5) In the Wild Photojournalism
Not social media
TV cameras
500 actors

6) Public Space
Multiple professional
+ legacy cameras
80 actors



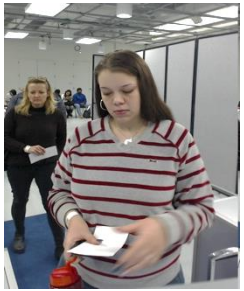
7) Luggage
2 webcams
375 actors

Adverse res, pose



See details on some datasets in the FIVE 2015 report

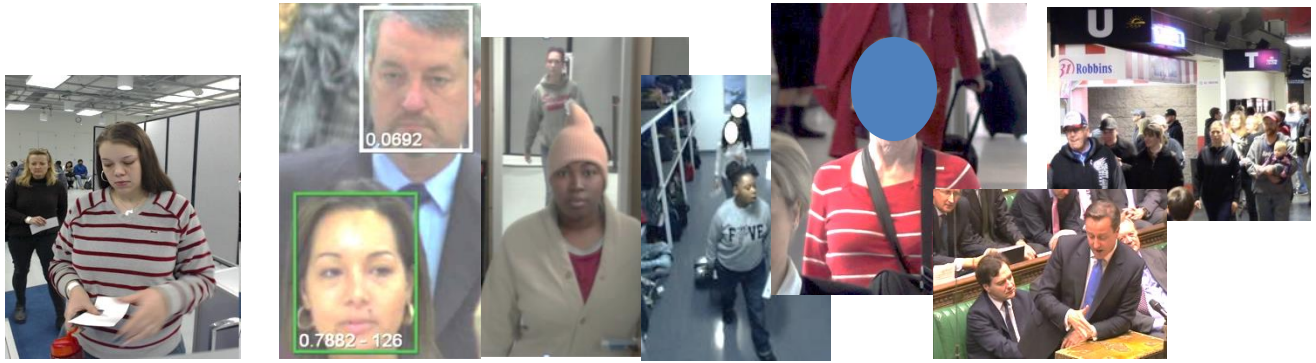
Accuracy gains since 2015




Dataset	Self Boarding Gate	Luggage	Sports Arena	Concourse	Public Space	Video Journalism
Gallery Size	48000	4800	480	48000	4800	935
Total video footage duration (minutes)	18	48	7995	2883	600	699
# False Positives	1	1	200	10	10	100
Best 2015	0.09	0.45	0.24	0.35	0.26	0.62
Best 2024	0.00	0.00	0.03	0.05	0.01	0.20

Miss rates across different datasets (best 2015 vs. best 2024)

FIVE 2024 Results



	Self Boarding Gate N=48000	Public Space N=48000	Jetway N=48000	Luggage N=48000	Concourse N=48000	Video Journalism N=935	Sports Arena N=48000
azumane_2	0.0040	0.0151	0.0117	0.0122	0.0667	0.2067	0.0439
hinata_0	0.0000	0.0120	0.0064	0.0152	0.0548	0.1985	0.0474
kageyama_0	0.0000	0.0482	0.0959	0.0366	0.0800	0.3797	0.1746
nishinoya_0	0.0000	0.0753	0.0222	0.0457	0.0889	0.2851	0.0992
sawamura_0	0.2460	0.7319	0.8431	0.9116	0.6637	0.8552	0.9538
shimizu_1	0.0000	0.0211	0.0117	0.0213	0.1244	0.2707	0.0916
sugawara_2	0.0000	0.0120	0.0108	0.0061	0.0474	0.1981	0.0609
tanaka_1	0.0000	0.0120	0.0598	0.0396	0.0593	0.2757	0.0881
tsukishima_1	0.0000	0.0211	0.0190	0.0183	0.0859	0.3956	0.1554
yachi_2	0.0081	0.0512	0.0563	0.1189	0.1007	0.4256	0.2707
yamaguchi_2	0.0040	0.0271	0.0370	0.0701	0.0859	0.3711	0.1684

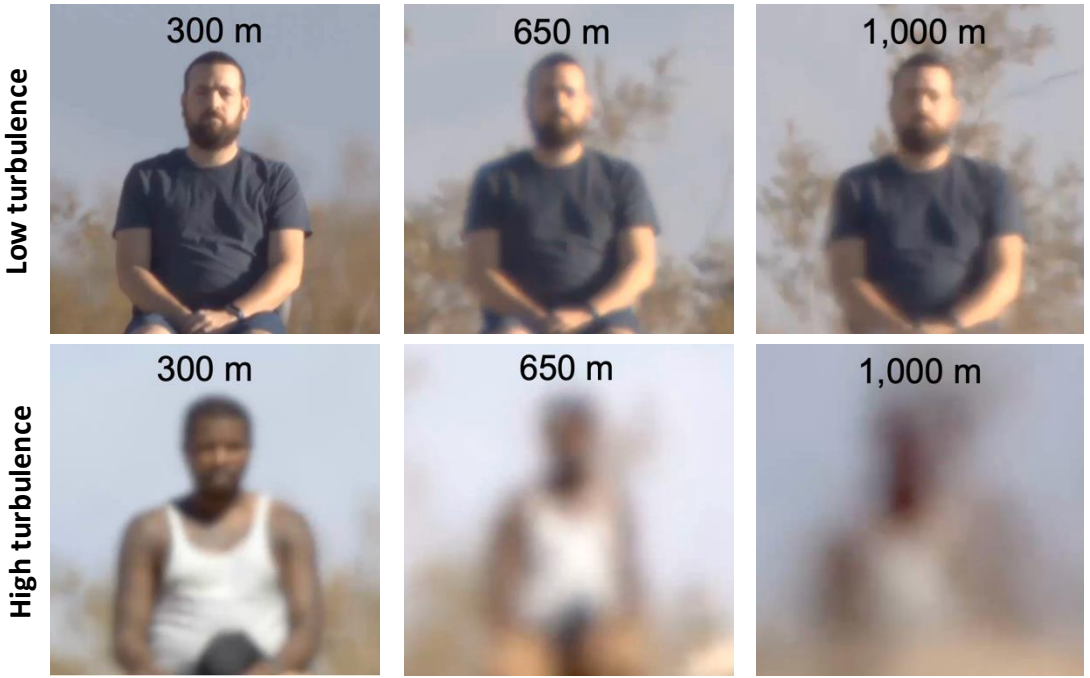
SUMMARY RESULTS | FNIR @ FPIR = 0.01

- Algorithm matters!

Video Journalism is difficult, because it is comprised of celebrity videos, where high yaw angles are typical, and the enrollment images are also unconstrained

New in FIVE 2024: long range with turbulence

Imagery collected at long range can potentially be distorted by atmospheric turbulence. Turbulence here refers to the distortions in an image caused by the movement of air due to temperature differentials. Here are examples of imagery collected at 300 meter, 650 meter, and 1000 meter ranges, at low and high turbulence levels. Note that some videos collected at long range will have turbulence, some will not.



Phase 2 Results:

- Long range
- Cooperative subjects
- Long lens

Low Turbulence $C_n^2 < 1.93 \times 10^{-13}$ | High Turbulence $C_n^2 > 1.93 \times 10^{-13}$

	300m	650m	1000m	300m	650m	1000m
hinata_0	0.0000	0.6027	0.7534	0.0133	1.0000	1.0000
sugawara_2	0.0000	0.6027	0.7534	0.0400	1.0000	1.0000
sugawara_3	0.0000	0.6164	0.7534	0.0267	1.0000	1.0000
sugawara_0	0.0000	0.6301	0.7534	0.0267	1.0000	1.0000
azumane_2	0.0000	0.6438	0.7534	0.0800	1.0000	1.0000
tanaka_0	0.0000	0.6575	0.7945	0.0800	1.0000	1.0000
tanaka_2	0.0000	0.6575	0.7671	0.1467	1.0000	1.0000
shimizu_0	0.0000	0.6712	0.8082	0.2000	1.0000	1.0000
shimizu_1	0.0000	0.6712	0.7397	0.0667	1.0000	1.0000
tanaka_1	0.0000	0.6712	0.7671	0.1467	1.0000	1.0000
azumane_1	0.0000	0.6849	0.7534	0.0933	1.0000	1.0000
shimizu_2	0.0000	0.6849	0.8356	0.2000	1.0000	1.0000
hinata_2	0.0000	0.6986	0.7671	0.0400	1.0000	1.0000
hinata_1	0.0000	0.7123	0.8082	0.0667	1.0000	1.0000
tsukishima_0	0.0000	0.7123	0.8493	0.2800	1.0000	1.0000
...						

- Recognition is possible at 300m, even with high turbulence
- Algorithm matters!
- Recognition accuracy decreases significantly at 650m and above

FNIR @ FPIR = 0.01

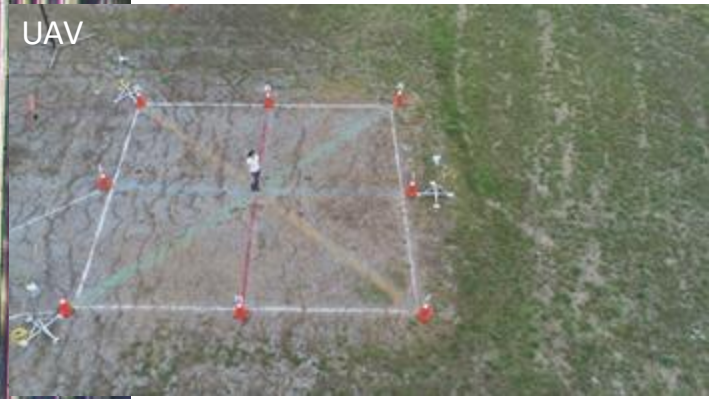
aka

miss rate with T set to have a 1% false alarms rate

N = 48000

New in FIVE 2024: long range and high altitude

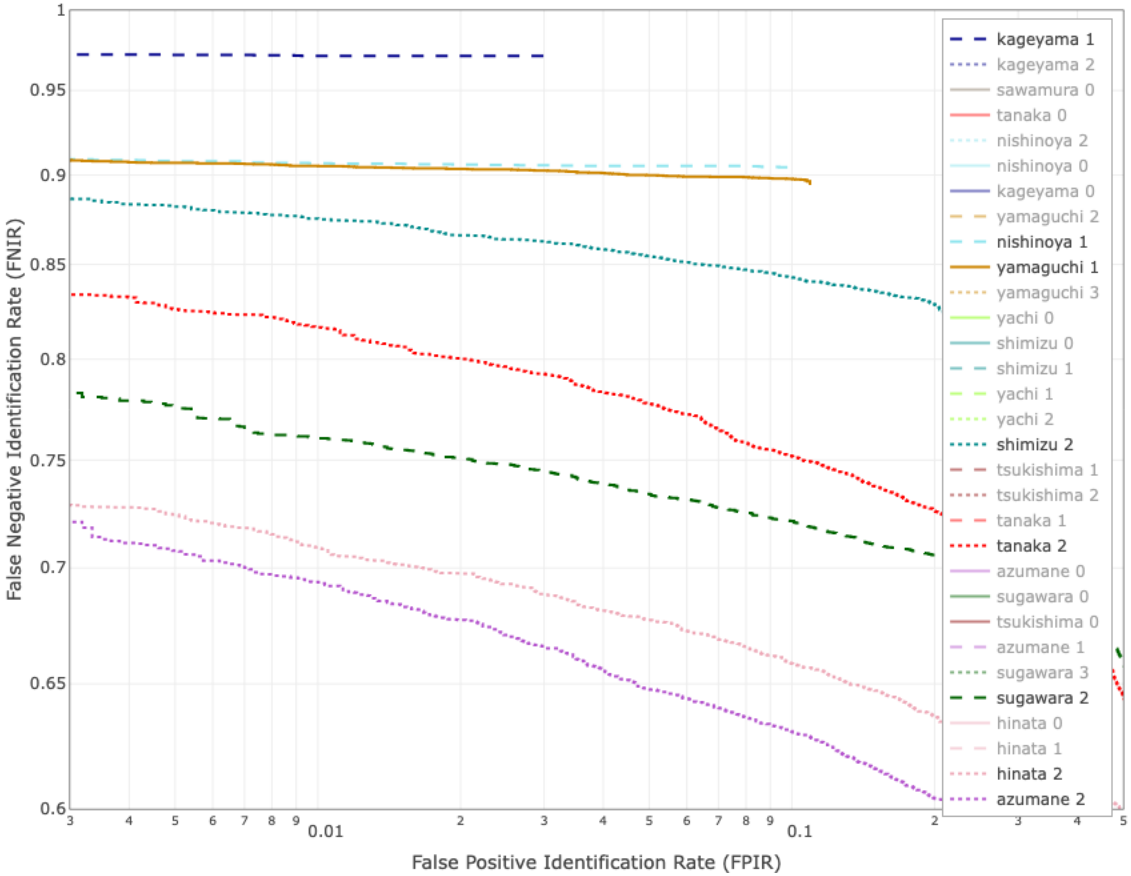
Elevation angle: 18 degrees



Dataset Description

- Wide range of optical setups for probes
  - Various ranges (close range to 1km+)
  - Various pitch angles
  - Specialized sensors, non specialized sensors, UAVs
- Detailed enrollment data
  - High quality stills at various pitch and yaw angles
  - High quality, close range enrollment videos
    - Random walk
    - Structured walk
- Multiple different collection locations and scenarios

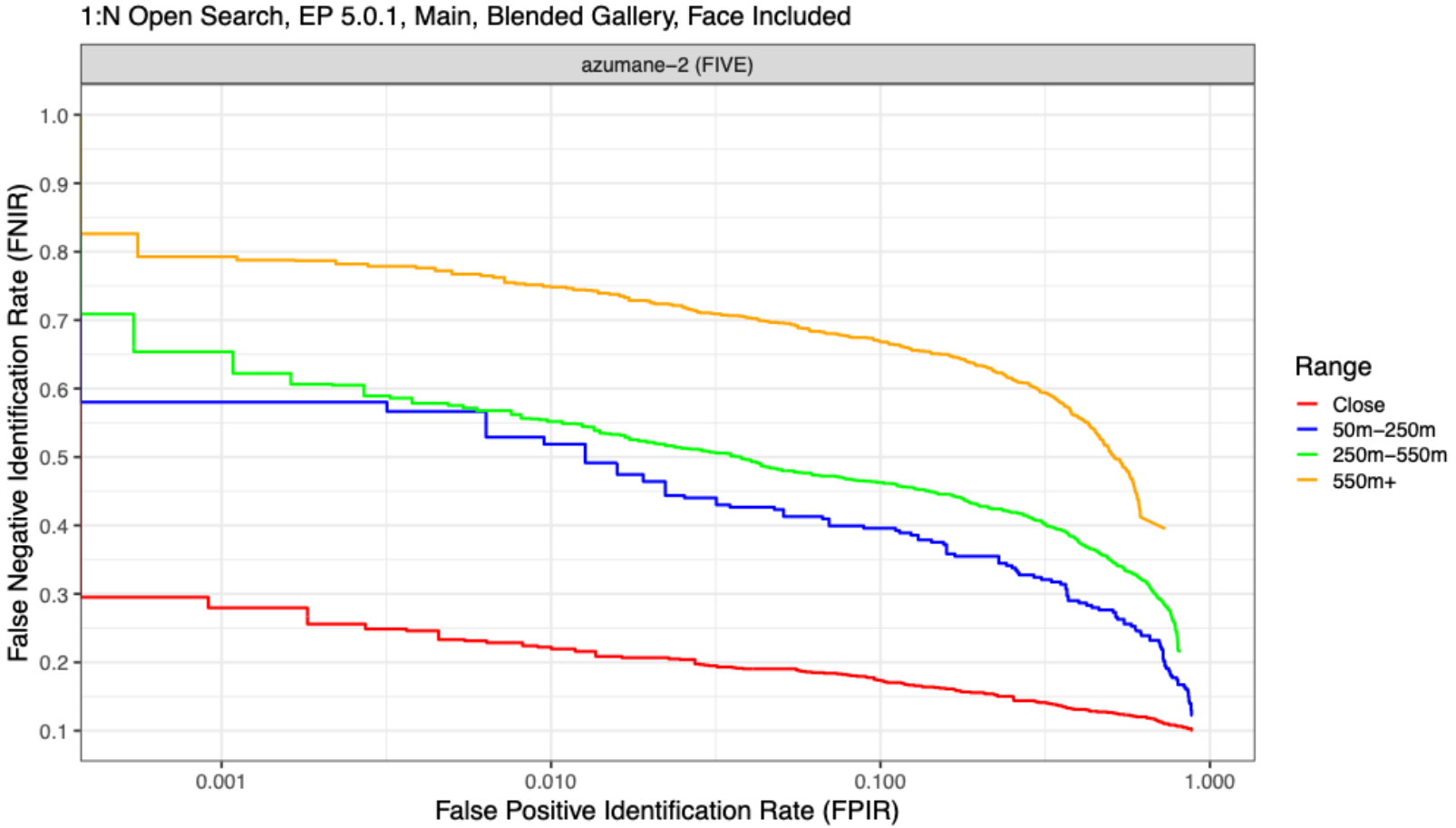
Longe Range Dataset, 1:N Open Search (Main, Blended Gallery, Face Included)



Probes: Various types: Some collected outdoors, **sometimes at close distance but often at longer range distances**, using pole-mounted cameras and unmanned aerial vehicles (UAVs), cameras facing down. Subjects were sometimes stationary, sometimes walked around.

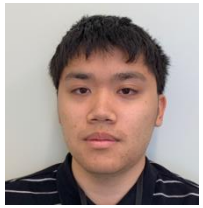
Gallery: Two galleries, average N = 559 people. Each gallery subject has a variable amount and type of imagery, which could include any combination of videos and still imagery.

Long Range Dataset – By Range



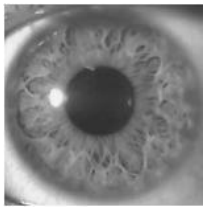

Patrick Grother


Kayee Hanaoka (+ Mei Ngan)


Austin Hom


(not) Mei Ngan


Joyce Yang


Jim Matey

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

THANK YOU

FRVT@NIST.GOV