NISTIR 8173

Face In Video Evaluation (FIVE) Face Recognition of Non-Cooperative Subjects

Patrick Grother
George Quinn
Mei Ngan

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8173

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8173

Face In Video Evaluation (FIVE) Face Recognition of Non-Cooperative Subjects

Patrick Grother
George Quinn
Mei Ngan
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8173

March 2017

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Kent Rochford, Acting NIST Director and Under Secretary of Commerce for Standards and Technology

ACKNOWLEDGEMENTS

The authors would like to thank the Department of Homeland Security (DHS) Science and Technology (S&T) Directorate First Responders Group Office for Public Safety Research for support of this effort.

Additionally, we thank the DHS S&T Homeland Security Advanced Research Agency Air Entry/Exit Re-engineering (AEER) Directorate for their support.

We are indebted to the staff of the Pacific Northwest National Laboratory for their development and collection of imagery used in this study. Similarly, we are grateful to SAIC and the staff at the Maryland Test Facility for their development and collection of imagery used in this study.

Thanks also go to the Richard Lazarick, Michael Garris, Marek Rejman-Greene, Jason Prince, and Catherine Tilton for discussion and reviews of this material.

DISCLAIMER

Specific hardware and software products identified in this report were used in order to perform the evaluations described in this document. In no case does identification of any commercial product, trade name, or vendor, imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

STRUCTURE OF THIS REPORT

The report is organized with an executive summary, a high-level background, and a technical summary preceeding the main body of the report which gives more detailed information on participation, test design, performance metrics, datasets, and the results.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

EXECUTIVE SUMMARY

▷ *Overview:* This report documents the Face in Video Evaluation (FIVE), an independent, public test of face recognition of non-cooperating subjects who are recorded passively and are mostly oblivious to the presence of cameras. The report enumerates accuracy and speed of face recognition algorithms applied to the identification of persons appearing in video sequences drawn from six different video datasets mostly sequestered at NIST. These datasets represent video surveillance, chokepoint, and social media applications of the technology. In six cases, videos from fixed cameras are searched against portrait-style photographs of up to 48 000 enrolled identities. In one case, videos are searched against faces enrolled from other videos. Additionally, the effect of supplementing enrollment with non-frontal images is examined.

▷ *Participation:* FIVE was open to any organization worldwide, at no charge. The research arms of sixteen major commercial suppliers of face recognition technologies submitted thirty six algorithms, allowing FIVE to document a robust comparative evaluation. The algorithms were submitted to NIST in December 2015 so this report does not capture research and development gains since then. The algorithms are research prototypes, evaluated as black boxes without developer training or tuning. They implement a NIST-specified interface and so may not be immediately commercially available. They run without graphical processing units (GPU).

▷ *Difficulty:* Face recognition is much more difficult in non-cooperative video than with traditional portrait-style photos. The initial face detection task is non-trivial because a scene may contain no faces or may contain many, and these can appear over a range of resolutions (scales), orientations (poses), and illumination conditions. Second, subjects move, so their faces must be tracked through time and this is harder when motion blur occurs or when a face is occluded by closer persons or objects. Third, resolution in video is compromised by optical tradeoffs (magnification, field of view, depth of field, cost) and then by compression used to satisfy data rate or storage limitations. Finally, other adverse aspects of image quality and face presentation degrade recognition scores so that scores from unrelated individuals can be similarly high, making discrimination between known and unknown individuals error prone. This leads to the possibility that a member of the public can be falsely matched to someone on a watchlist; the occurence of such hazards is mitigated by elevating a recognition threshold.

▷ *Key conclusions:* This study was conducted to support new and existing applications of face recognition, particularly to assess viability and technology readiness. These range from surveillance, to expedited single-factor access control, and to the tagging of faces uploaded to social media. Whether face recognition can be used successfully depends on an operations-research analysis of the intended use given the most important design parameters in Table 1. This report provides data to inform that. For example, timing estimates imply that real-time processing on a single server is achievable only with certain algorithms.

This report documents situations where face recognition of non-cooperative persons is accurate enough to satisfy some operational requirements. It also demonstrates cases where the core recognition technology fails to compensate for deficient imaging. That notwithstanding, subjects appearing in video can be identified with error rates approaching those measured in cooperative face recognition, but only if image quality reaches the level attained in engineered systems such as e-Passport gates. That is a very difficult goal.

High accuracy recognition of passively-imaged subjects is only achievable with: a) a small minority of the algorithms tested here; b) a dedicated and deliberate design effort that must embed optical, architectural, human factors, operations-research, and face recognition expertise; c) galleries limited to small numbers of actively curated images; and d) field tests with empirical quantitative calibration and optimization.

None of the datasets used in this report represent peak attainable performance. Given better cameras, better design, and the latest algorithm developments, recognition accuracy can advance even further. However, even with perfect

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

design, some proportion of a non-cooperative population will not be recognized simply because their faces were not availble to the receiving system. Such *failure to acquire* cases occur when subjects never look toward the camera (e.g. at a mobile phone instead) or because their faces were occluded (e.g. by sunglasses, hats, or by taller people standing in front of them). While such failures can be mitigated by careful placement of, in principle, multiple cameras, recognition essentially only succeeds if a clear line of sight to a sufficiently frontal face can be engineered (assuming the reference image is frontal).

Deployment should proceed only after quantitative assessment of objectives, alternatives, ease of evasion or circumvention, enrolled population sizes, search volumes, the proportion of searches expected to have an enrolled mate, accuracy requirements, consequences and procedures for resolution of errors, and speed and hardware cost constraints. In particular, deployers must weight their tolerance for misses and their risk appetite. In addition, when non-cooperative face recognition is used to identify individuals nominated to a watchlist, human reviewers must be employed to adjudicate whether candidate matches are true or false positives. This has two consequences. First, the recognition system must be dynamically configured to produce (true or false) candidate alerts at a rate matched to available labor resources. Second, because humans also commit face recognition errors, the overall error rates of the hybrid machine-human system must be understood and planned for.

OUT OF SCOPE

The following are out of the FIVE scope: One-to-one verification of identity claims; identification from body worn cameras, license plate cameras, drones and other aerial vehicles; video analytics, scene understanding, anomaly detection, and spatial boundary violation; suspicious behavior and intent detection; estimation of emotional state; full body recognition (static and kinematic) including gait recognition; use of imagery from exotic sensors such as infrared and thermal emission.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

#	Application name, description	Criminal, civil, or private	Positive or negative identification[1] "Whitelist" vs, "Blacklist"	Usual Mode	Nominal enrolled pop. sizes, N	Volume of searches: a) Enrollees b) Non-enrollees c) Prior probability, P, that subject appears in search imagery	Needed FPIR. Low FPIR requires high decision threshold	Human labor requirements	Timescale on which decision/action needed	Environment and imaging under owner control	Relevant datasets and results in this report
1	Tokenless access control (e.g. gymnasium) Single factor authentication	Private sector	Positive, user implicitly asserts gym membership (some cooperation)	V2S	10^1 - 10^4	a) 10^3 daily b) 10^1 variable fraud rate c) P very high, close to 1, with small amount of fraud	Moderate to high, which may be tolerable	High as false rejections need quick remediation. Depends on cost and hazard of false positives	Seconds	Yes	DATASET U: PASSENGER GATE (Section 5.2)
2	Video surveillance in transport terminus	Typically criminal	Negative	V2S	10^2 - 10^4	a) Very few b) 10^4 daily per camera c) P is very low	Very low	Very low, given high threshold, low FPIR and low P	Minutes	Yes. Tightly specified installation of cameras, lighting, network, and online recognition.	DATASET H: TRAVEL WALKWAY (Section 5.6) — DATASET T: TRAVEL WALKWAY (Section 5.7) — DATASET P: SPORTS ARENA (Section 5.5)
3	Video surveillance in casino (humans with PTZ cameras)	Private sector	Negative	V2S	10^2	a) Low b) 10^4 daily c) P is low	Low	Low	Minutes	Yes	NA. PTZ images can be very high resolution, are not represented in FIVE
4	Forensic investigation of bank heist	Criminal	Positive, officer posits suspect is present in criminal files	V2S	10^3 - 10^6	a) Few robbers b) 0 c) P is moderate, perpetrator may be recidivist ex-con	High is acceptable as high human adjudication effort is assumed	One or few investigators, since the volume of video is very low	Hours	No, typically legacy CCTV installations are imperfect for face recognition	DATASET L: PASSENGER LUGGAGE (Section 5.4)
5	Asylum claim re-identification of inbound aircraft (same day)	Civil	Positive, authorities assert traveler is present in airport video	S2V	$< 10^4$	a) Few stills of a person b) 0 c) P is high, exception passenger may actively evade cameras	High is acceptable as high human adjudication effort is assumed	One or few investigators, since the volume of video is very low	Hours	Yes, if airport has cameras and maintains video buffer	DATASET H: TRAVEL WALKWAY (Section 5.6)
6	Confirming immigration exit on or near air loading bridge	Civil	Positive	V2S	< 500	a) Size of aircraft b) Low c) P is 1, if everyone enrolled, otherwise 0.1 - 0.9 visa in-scope.	Moderate. Low if biographics from passenger manifest used as fallback	Depends on goals	Hours, unless enforcement component is added	Yes, deliberate design is possible	DATASET J: PASSENGER LOADING BRIDGE (Section 5.3) — DATASET L: PASSENGER LUGGAGE (Section 5.4)
7	Finding, grouping individuals in TV archives	Private	Positive	V2V	10^2	a) and b) vary with size of collection c) P is high since curator may know a priori which videos are relevant	Low	High, as part of a curation process	Hours	TV has good light and cameras typically, but arranged for aesthetics, not for FR	DATASET C: PHOTOJOURNALISM (Section 5.8)
8	Enforcement of restraining order, ASBO[2]	Criminal	Negative	V2S	1	a) very low, b) low, depending on perimeter, c) P is low, cf. recidivism rate in such cases	10^{-3}	Low, policing in case of alarm,	Minutes	Yes, but outdoors or in entrance ways.	DATASET P: SPORTS ARENA (Section 5.5)

[1]Positive identification implies a claim that the search subject is enrolled in the gallery, while negative identification implies that an operator, for example, is making a claim that the search subject is not enrolled.
[2]Antisocial behaviour ordinance (UK).

Table 1: Applications of video-based face recognition. The rows give applications and domains of use. The columns give the defining and salient properties. The values are nominal and the technology may be viable outside the stated values and ranges.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

BACKGROUND

Non-cooperative face recognition: Prior NIST reports on the evaluation of face recognition prototypes [19, 21] have quantified the accuracy available from still photographs that were collected with cooperation of the subject and for which the imaging apparatus, environment and configuration were controlled according to formal face recognition standards [15,18]. In contrast, this report details recognition performance available from video sequences of human subjects who do not make explicitly cooperative attempts to present their faces to a camera; instead they are observed passively and are usually oblivious of the presence of the cameras. Non-cooperative imaging of subjects is generally adverse to recognition accuracy, because face recognition algorithms are sensitive to head orientation relative to the optical axis of the camera, accompanying shadows, and to variations in facial expression. Note also that cooperation is not all-or-nothing: It can be induced by using a video display attractor. We investigate this with Dataset J.

Challenges of face recognition in video imagery: Video sequences present additional challenges to face recognition systems. First, subjects may be entirely absent for extended periods, and with a cluttered background, algorithms will yield false detections. There is generally a tradeoff between missed detections and adequate suppression of false detections. Second, the subjects are generally moving, and this motion is not necessarily parallel to the optical axis. Motion, which can cause blur, requires algorithms to track faces through time, a task that is not trivial in cluttered scenes with variable lighting and the possibility of occlusion by other subjects. Third, there may be multiple subjects present, each of which must be localized. Moreover, this must occur over varying resolutions (scales) and head orientations (poses). Fourth, facial detail may be lost to motion blur and to the lossy compression that remains a ubiquitous requirement in video transmission and storage. Finally, optical distortion can occur when subjects are displaced off the optical axis, or, more rarely, are too close to the camera.

The FIVE study leverages seven video datasets whose characteristics differ, with variations in duration, frame rate, camera quality, camera configuration, illumination environment and geometric placement - see Table 3. These factors are known to influence accuracy, and sometimes computational cost. Note however, that all of the cameras used in the collection had fixed position, orientation and optics, and none were under active human control. Even without variation of those parameters, face recognition in video presents considerable challanges beyond those faced in traditional still-image tasks.

Applications: Non-cooperative face recognition can be applied in a number of applications, beyond the video surveillance task popularized in television and films. These can be categorized in several ways: First, many applications search video against enrolled still images (V2S), others could search still to video (S2V), and yet more might search video to video (V2V). This will have architectural and computational implications. Second, they can be categorized by whether the technical specifications (camera, optics, illumination) are under the control of the operator vs. provided to a receiving agency, as is. Third, applications vary by the frequency with which enrolled subjects are encountered in a search. Fourth is whether they embed positive or negative identity claims - would a typical user claim to be enrolled, or claim *not* to be enrolled. The canonical application, video surveillance, involves searching live video feeds against enrollment imagery from known persons of interest. This "blacklist" search is a negative identification application in that there is an implicit claim that most searches are not expected to have an enrolled mate. This prior probability that an enrolled subject is actually encountered can be very low. If it is too low, then this contraindicates deployment of the system, unless the risk of not doing so is to miss a person of such interest that it has large adverse effects. An additional trouble with "rare hit" applications is boredom and lack of reward for the human reviewers who adjudicate candidate matches.

The report also seeks to support positive "whitelist" identification applications such as expedited access control in which non-cooperative subjects are not required to make a deliberate interaction with a biometric sensor, such that the duration of the transaction is by-definition zero. This is contrary to the more typical access control application in which a subject presents a credential to a reader (swipe or chip read, for example), makes one or more cooperative presentations to

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

a sensor (finger placement, for example), and then proceeds. In contrast, passive non-cooperative imaging elides the possession and presentation of a credential. The downside to this approach is the likely reduction in true acceptance rates associated with configuring the system to limit false positives inherent in comparing imagery with that of N enrolled identities.

Note that in some applications, the enrolled subjects may be entirely unknown i.e. there may be no biographic metadata - and the goal of biometric search is simply to link images across two or more events, without any actual identification.

Table 1 gives a matrix of applications and their defining properties. This is intended to support prospective users in analysis of their application and mapping that onto results available later in this report.

Evasion of face recognition: This report does not cover a further operational challenge for recognition: subjects who actively impede face processing. Deliberately uncooperative subjects can expend arbitrary effort to evade detection and recognition by avoiding the camera, either completely or by keeping their head down, or by altering the appearance of their face, for example by wearing (sun)glasses [32], hats, cosmetics, and masks. The likelihood that such techniques will be used - they have been ubiquitous in bank robberies since the advent of the CCTV camera - may well contraindicate the use of face recognition. Such techniques can be 100% effective. Effective evasion is however predicated on knowledge of when cameras are present and face recognition may be in use. Without that knowledge, successful evasion would require continuous persistent effort.

State of face recognition research: While face recognition has been a perenial popular research challenge since its inception, there has been a marked escalation in this research in recent years due to the confluence of several supporting factors. These are the development of convolutional neural networks and deep learning [27] schemes, particularly for face recognition [30], which invariably leverage [31] the ever increasing amount of web-downloadable face images. The face imagery is taken from identity-labeled photographs downloaded automatically from social media. These are available in large part due to the advent of the digital camera, particularly on the smart phone, and critically, the internet as a distribution mechanism.

Open-universe identification: Non-cooperative face recognition is implicitly a one-to-many identification task. It is not a one-to-one verification task because it does not involve the subject in making an identity claim. As such, it requires unknown video imagery to be compared with that previously collected from multiple individuals. Given N individuals, the task is called 1-to-N identification, the goal being to determine the identity of the query subject. However, a key point is that most applications are *open universe* where some proportion of query subjects will not have a corresponding mate entry among the previously enrolled identities. The face identification algorithm must then minimize both false positive errors where an unenrolled person is mistakenly returned, and a false negative error where an enrolled person is missed. This is critical whenever the proportion of non-mated searches is naturally large particularly in the canonical "watch-list" surveillance application where a large majority of individuals in the field of view are not enrolled and should not be returned by the system. So a face recognition system looking for terrorism suspects in a crowded railway station must be configured to produce few false positives, no more than what can be sustainably adjudicated by trained reviewers who would determine the veracity of the candidate match, and then initiate action. If, on the other hand, the proportion of unmated searches is low, as is the case, for example, with patrons entering their gymnasium, the system must be configured to tolerate a few false positives, i.e. to admit the infrequent non-customer who attempts access.

Relevant accuracy measures: It is necessary to report accuracy in terms of *both* false negative identification rate quantifying how often enrollees are not recognized, and the false positive identification rate stating how often algorithms incorrectly issue false alarms. Accordingly, FIVE supported measurement of *open-set* accuracy, including both actor false negatives ("misses") and non-actor false positives ("false alarms"). In this respect, this benchmark differs from recent face recognition evaluations [25] that assume all searches have an enrolled mate.

Academic research focuses on closed-universe applications: Most research papers on face identification have two

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

related limitations. First, their experiments do not include searches for which a mate does not exist in the enrolled gallery. This is a closed universe application of biometrics; it derives from the approach used in the computer vision and pattern recognition community in which benchmarks involve the classification of the image of an object to a closed set of exemplar objects. This almost always does not apply in biometrics, because some subjects have usually not been encountered before. For example, in a criminal justice application, the closed universe assumption is tantamount to assuming everyone has been previously arrested and photographed. By failing to execute non-mate searches, academic experimentation does not measure false positive identification rates i.e. proportion of non-mate searches that incorrectly return an enrolled identity. Instead, most research reports accuracy using rank based recognition rates drawn from the cumulative match characteristic. While some publications [31] do execute 1:1 verification, with many impostor comparisons, they do not explicitly consider the largest of N impostor scores generated in a 1:N search. It is this extreme value which, if high enough, leads to false alarms in identification systems[1].

The underlying point here is that much research is directed at optimizing a metric that is often inappropriate to typical use cases. Specifically, the research evalutions assume the prior probability that a search photograph has a mate is 1. This departs from operational reality in that the prior probability can span many orders of magnitude below 1. At the high end, for example, a majority of faces uploaded to a face recognition tagging service on a social networking website will belong to the universe of friends of the uploader. In law enforcement applications, the prior probability that a photograph searched against a criminal database would have an enrolled mate will be related to the recidivism rate which has been measured around 0.6 in the United States [10]. In a casino, the proportion of card-sharps among the gambling population might never exceed 1:1000, while in a border crossing drug interdiction setting the prior probability could be below 1:10000. In a counter-terrorism application, the probability could be much lower still. The point is that for all the unenrolled subjects, the face recognition system must correctly not issue a false alarm, or, least do so rarely. This necessitates good algorithmic discrimination and high, stringent, similarity thresholds. This is at odds with much of the academic research which focuses on employing deep convolutional neural networks (DCNNs) designed for their invariance to pose, translation, illumination. That is appropriate in a social media setting perhaps. It is less so for governmental applications where, often, the camera and illumination can be controlled even with a non-cooperating population. There remain open questions with DCNNs: Can they surpass traditional approaches at very low false positive rates, at least in cooperative portrait-style imagery? Can they be used "off-the-shelf", i.e. without either training from scratch, or without some kind of adaptation e.g. transfer learning.

Human adjudication of true and false matches: Most applications of automated face recogition are monitored by human operators who adjudicate recognition decisions. So, for example, it is typical for sets of five e-Passport verification gates to be observed by an immigration officer who can open the gate if a traveler (who may be a legimate passport holder or an impostor) is not authenticated automatically, or refer him to a traditional process. In one-to-many situations, for example, the detection of duplicate visa or driving license submissions, the examining official must determine whether the identity of the search image and the hypothesized gallery entry are indeed the same, or if a false positive has occured. The system managers must ensure sufficient labor is available to adjudicate such outcomes. Practically, the recognition threshold is set to limit the number to manageable levels.

[1]For quick background see http://www.wjscheirer.com/misc/openset/cvpr2016-open-set-part2.pdf, or [22] for the effect of extreme values on 1:N accuracy.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

TECHNICAL SUMMARY

Face detection: The first step in automated identification in video sequences is face detection. This is challenging due to variation in resolution and head orientation (pose). Given the same input video clips, algorithms vary by an order of magnitude in the number of faces they detect. Variation arises due to false detections (of non-faces), missed detections (of actual faces), and fragmented tracking through time. The result is that in the highest quality dataset we ran (H), using over 20 hours of video recorded in a transport terminus, the number of detections varied from 8 to 150 per minute, with a consensus (from the more accurate algorithms) of between 10 and 15 per minute - the actual number is not known, nor well defined. Similarly in 43 hours of video shot using 11 cameras mounted in the access areas of a sports arena (Dataset P), the number of detections also varies by a factor of 20. See Tables 19 and 24

Verbose detection algorithms - those that report more faces in video - cause face recognition implementations to incur computational expenses associated with searching their outputs. Their accuracy is not necessarily worse, becoming so only if the higher volume of detections yields more false positives, and that occurs only if non-faces, small faces or otherwise poor quality faces yield high impostor scores. Terse algorithms, those that produce so few detections that they miss some actual faces can clearly give increased false negative identification rates. We note a few instances of algorithms reporting fewer detections than there are actors. See Figure 22

Variation across algorithms: As with most other tests of biometric recognition algorithms, there is a massive variation in accuracy between algorithms: some give many fewer errors than others. This is particularly pronounced here as non-cooperative video stresses algorithms. For even the easiest dataset, with small ($N = 480$) galleries, the proportion of searches that do not yield the correct identity at rank 1 [2] ranges from below 1% to above 40%. See Table 6.

Similarly, using professional-grade ceiling-mounted cameras, miss rates can range from 20% to over 90% even for algorithms provided by experienced developers. While accuracy is just one factor necessary to successful deployment - others are speed, scalability, stability, ease of integration and configuration, and cost - such results imply a "procurement risk" associated with fielding the wrong algorithm. See Table 24.

Variation across datasets: The FIVE activity employed seven datasets representative of applications given at the foot of this page. These range from high quality datasets typical of what would be expected if a dedicated design and installation effort had been made, to a dataset more representative of many legacy installations where CCTV is available at a crime scene but is poorly-suited to face recognition. Dataset C is difficult because it is comprised of celebrity videos, where high yaw angles are typical, and the enrollment images are also unconstrained. All other datasets used controlled portraits. The inset table shows rank one miss rates from five of the more accurate providers. The enrolled population size is small, $N = 480$. These numbers are applicable to investigational applications where ranked candidate lists are examined by reviewers, without any thresholding. Miss rates range widely, and certain algorithms are ill-suited to certain datasets. For example, while several algorithms perform well on the professional surveillance Datasets H and T, only one handles the lower resolution Datasets L and P, yielding about half as many errors as algorithms from the next most accurate developer.

Alg \Dataset	U	J	L	P	H	T	C
M30V	1%	6%	12%	14%	15%	13%	42%
I30V	8%	10%	24%	23%	15%	10%	49%
H30V	4%	12%	25%	39%	16%	16%	60%
N33V	4%	14%	32%	44%	17%	28%	53%
G32V/G30V$^+$	6%	$^+$30%	40%	35%	$^+$21%	23%	36%

We summarize accuracy by stating the best single camera results here. Later we report results for fusion across cameras.

Surveillance: When using ceiling-mounted professional-grade cameras to identify previously enrolled actors appearing in about 20 hours of transport-terminus surveillance video, the most accurate algorithm detects and reports more than 17000 faces, some of those actors, most of them un-enrolled passers-by. Of the actors, the algorithm incorrectly

[2] This report universally uses "miss rates" i.e. false negative identification error rates, x, where smaller values are better. It is more common to see "hit rate" values, $1 - x$. We do not quote these because they can lead to a framing bias when success rates are high. For example, 97% and 99% correspond to three times as many errors.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

misses about 20% of them in a gallery constructed from N = 480 individuals' frontal photographs. This degrades with population size as shown in the inset table. This applies to identification mode where a high threshold is set to limit false positives to just 10; this is a very low count given the 20 hour, 2 million frame, duration of video and the presence of thousands individuals. If, instead, forensic investigators had isolated a clip of interest from the same video collection, say resulting from a terrorism incident at a particular time, then they could review candidate identities reported by face algorithms. In such cases, the recognition threshold is set to zero, and trained human reviewers adjudicate photos on candidate lists. As illustrated in the second row, rank-based identification miss rates are better because they allow weak-scoring hits to count. Elevating a threshold, on the other hand, causes some hits to become misses, but is necessary in most practical cases to limit false positives. Further gains are possible considering longer candidate lists. Here for a dataset of N = 48 000 persons, the two most accurate algorithms would miss about 24% of suspects within the top-20 ranks vs. 26% at rank 1. See Tables 23 - 25.

Dataset H, Algorithm M30V	N = 480	N = 4800	N = 48000
Identification miss rate	20%	21%	35%
Investigation miss rate	14%	16%	26%

For this reason, it is imperative that enrolled population sizes be kept as small as possible as part of an active gallery curation process. The high accuracy values reported later are achieved with gallery sizes three order of magnitude below those used in contemporary cooperative passport and driving license de-duplication systems.

Boarding gate: In the case where a turnstile or gate is equipped with a camera that passively records subjects presenting paper barcode tokens, the most accurate algorithm will miss just 6% of subjects present in an enrolled database of $N = 480$ individuals. This degrades as shown in the inset table. This benign approximately logarithmic increase in miss rates with database size has been reported in prior tests of face recognition with cooperative portrait-style images [23], and is the underlying reason why face recognition has utility for large scale face identification. Accuracy may be improved with refined camera placement or additional cameras. See Tables 6 - 8.

Dataset U: Gate	N = 480	N = 4800	N = 48000
Identification miss rate	6%	9%	18%
Investigation miss rate	1%	2%	4%

Aircraft loading bridge: Consider the use of ceiling mounted cameras deployed on the passenger loading bridge. The video imagery could be searched simultaneously against "blacklist" and "whitelist" galleries. In the former case the application would be to prevent certain individuals from boarding, essentially as a public safety or law enforcement (LE) matter. In the latter, there are at least two possible applications. One would be simple access control - to allow only ticketholders to board - where some gate mechanism would deny boarding to a traveler whose video imagery did not match the enrollment photo. A second whitelist application would be immigration exit - a facilitation application - in which the goal is for visa holders to confirm exit from a country by achieving a positive match against their enrollment record. This differs from the access control situation in terms of security goals and error consequences, and relevant performance metrics as shown in the inset table. Using a purpose-built simulated aircraft passenger loading bridge, the most accurate algorithm correctly identifies all but 6% of subjects when searching a gallery of N = 480 individuals. This is achieved with the aid of an attractor to induce a frontal view and a bottleneck to force a delay, and when the threshold is set to allow 10 false positives. In a video surveillance application intended to detect rare "blacklisted" individuals this number of false positives may be tenable during the boarding of individuals onto an aircraft, as officers could review those ten candidate hits. However, for the "whitelist" access control application, where positive match above threshold opens a gate, the threshold is probably too low because, from a security perspective, 10 false positives during the boarding of a 480 person aircraft is poor: the implied false positive rate (~ 0.02) is much higher than is typically configured in an access control system (≤ 0.001). Absent fraud, all persons should legitimately be able to board an aircraft (by matching any enrolled identity), any misses would

Application	Blacklist / Whitelist	N	Prob Mate	FPIR Required	False positive Consequence	False negative Consequence
Public safety LE	Black	O(10^4)	$\rightarrow 0$	Very low	Detain innocent	Safety, crim. evasion
Access control	White	< 500	$\rightarrow 1$	0.001 cf. eGate	Wrong boarding	Deny boarding
Immigration exit	White	< 500	P(visa)	Low	Overstay fraud or wrong status	Wrong immi. status rest on biographic

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

require human assistance and that would delay boarding. Now consider the immigration exit alternative where no gate is present, everyone proceeds regardless of the recognition results. A "miss" will later be met with an assertion from the traveller that he did indeed leave, which will be supported by biographic logs from passenger manifests. However, any visa overstayer could send a confederate, ensure that their face was missed (by looking down at a mobile phone). The overstayer could later leave and re-enter relying on that biographic claim. See Figures 17 and 18.

Legacy activity monitoring cameras: We regard the webcam imagery of Dataset L as more representative of the sub-optimal imagery acquired in, for example, banks and convenience stores where the goal was recording but not specifically face recognition. The imagery is characterized by adverse elevation angles and lower resolution. In this case the M algorithms give fewer than half the errors of any other algorithm.

Effect of camera height: The algorithm effect is prominent for recognition when cameras are mounted above head height. Such installation is typical. For subjects in a concession stand queue of a sports arena, recognition accuracy in a gallery $N = 480$ frontal stills is superior for cameras mounted at a height of 1.83 vs. 2.44 meters (6 vs 8 feet). However, the magnitude of this depends strongly on the algorithm - see the miss rates in the inset table for algorithms from the more accurate developers on this dataset. See Table 19.

Miss rate	G	M	H	I	R
High mount	47%	15%	63%	39%	55%
Low mount	42%	13%	43%	32%	43%

Error rate tradeoff: As in all biometrics systems, false negative outcomes can be traded off against false positives. When a stringent high threshold is imposed to suppress false positives, the false negative rate rises. When false positives are tolerable, lower false negative rates are achievable. In conventional biometrics, it is ubiquitous to compute an error tradeoff characteristic and plot false negative vs. false positive identification rates. In surveillance, the actual number of subjects appearing within the field of view is unknown (and poorly defined) such that the false positive identification *rate* is replaced by the false positive *count*. In operations, the relevant metric becomes the false positive count per unit time given some canonical population flow rate. See Figure 10.

In any case, both kinds of identity error are influential. Prospective deployers should consider the consequences of both types of error, and ensure (including contractually) that they have an ability to alter the threshold.

Algorithms for investigation vs. identification: It appears that some FIVE developers have not targeted recognition accuracy in cases where high thresholds are necessary to suppress false positives. Instead, as with much academic research, the focus has been on rank-based reporting. This manifests itself with high-threshold accuracy being far inferior to rank-based zero-threshold accuracy. Thus algorithms developed for investigational applications with low search volume (e.g. bank robberies) are not automatically appropriate to identification applications such as surveillance with crowds or sustained volumes.

Effect of enrolling multiple images per person: As face recognition has a well known dependence on head orientation, accuracy can be improved if a gallery is populated with additional views of a subject. Here, by enrolling three views together - full frontal and two images for which head yaw is $\pm 22^o$ - we see accuracy benefits with many algorithms. Modest gains are available in surveillance like applications, but more substantial gains appear for investigations where humans will review candidate lists. The inset table shows the proportion of actors not returned among the top 10 candidates. The technique is effective at promoting weak-scoring identity hypotheses up the candidate list. The technique has implications for standardized photo capture. While the face recognition industry has grown around the collection and enrollment of frontal portrait photographs, supported by formal standards, the forensics community, motivated by their involvement in the adjudication of outputs of automated face recognition systems, has called for a) higher resolution photographs, and b) additional views. Greater resolution is available with virtually all cameras today. Collection of photographs from different viewpoints will likely require additional labor and possibly equipment, so further standardization should be predicated on an understanding of which views are most beneficial, the consistency with which they can be achieved,

Enrollment	M	I	G	H
One view	4%	5%	19%	18%
Three views	3%	4%	10%	9%

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

and on the implementation cost. See Table 20 and Figures 34 - 35.

Effect of stalling a queue: When a ceiling mounted camera is used to identify subjects traversing a loading bridge, identification of subjects walking is much less successful than if an artificial bottleneck is used to cause the subjects to wait in a queue. The inset table shows miss rates for $N = 480$ when a modest threshold is used to limit false positives to ten during the 248 person experiment (without an attractor). These accuracy gains are available but with the cost that processing durations that are up to three times longer with queued subjects. See Figure 17 and Table 33.

Accuracy	M	I	J	H
Freely walking	17%	46%	54%	53%
Queue stalled	9%	27%	36%	23%

Effect of using an attractor: As face recognition algorithms have well documented sensitivity to head pose elevation, identification error rates increase with both the angle of the head relative to the optical axis of the camera, and with the difference of that angle to whatever it is in the enrollment photograph. It is common to install some device that attracts subjects' attention, and thereby improve pose. The audio-video attractor reduced identification misses substantially. An alternative attractor, a live display of the subjects walking was less effective as it induced adverse behaviors (hand waving, exaggerated expressions). Subjects walked freely. The threshold is set to give 10 false positives in each condition. See Figure 17.

Attractor	M	I	N	H
Off	26%	46%	56%	53%
Digital mirror	20%	40%	54%	45%
Agent audio video	10%	26%	33%	20%

Effect of multiple co-located cameras: Most results in the report relate to the failure to recognize an individual appearing before a single camera. However, when subjects walk freely through a volume that is concurrently imaged by three cameras, identification error rates vary between the cameras, and can be improved by fusion of scores produced during search. This was achieved by NIST implementing a max-score fusion across the three cameras. This step should more appropriately be done by passing multiple video streams to the algorithms, but this was not supported by the FIVE API. The threshold is set to give 10 false positives from each camera, and in the case of fusion, 10 from all cameras. The attractor was off. See Figure 17.

Position	M	I	N	H
Ceiling	26%	46%	63%	53%
Right	35%	64%	70%	62%
Left	26%	50%	60%	53%
Fusion	16%	38%	58%	41%

Capability to identify over multiple cameras and multiple locations: The prior paragraph addressed fusion of recognition hypotheses from multiple cameras observing a subject simultaneously. We ask two similar questions here. First, what proportion of subjects are missed if we fuse over all appearances before a single camera. This fusion over time (in the case of Dataset P, a few hours), is effective if subjects are *ever* recognized. Second, we ask what proportion of subjects are missed if we fuse over all appearances before *all* cameras. This is fusion over space and time. The inset table shows fusion is very effective, suggesting installation of more cameras will improve accurary. Some obvious caveats, however: First, fusion over time will be ineffective if subjects don't appear on more than one occasion. Second, fusion over space will be ineffective if subjects only appear in one location. Third, fusion over time changes the operational relevance in that it is largely useless to know that a kleptomaniac has entered a shop only at the end of a day. The fused error rates are germane only to retroactive applications, like asylum claim backtracking in an airport, or non-repudiation. See section 5.5.1.

Fusion	Cameras	M	I	H	G
None	low-near	14%	32%	43%	42%
Over time	low near	2%	5%	8%	6%
None	All	31%	54%	67%	61%
Over time and space	All	1%	4%	6%	1%

Applicability in photojournalism: Imagery captured and published by professional photographs and television camera operators is distinguished by good illumination, focus, and composition. However it remains challenging because of diverse head poses and facial expressions. In a video-to-still mode, the inset table shows rank-1 miss rates are much higher than a ceiling-mounted surveillance setting, despite use of a gallery five times smaller. The reasons for this is wider variation in pose in the photojournalism data, and that *both* the reference photograph and the search video are

Set	Use	Mode	Gallery	G	M	I	N
C	Photojournalism	V2S	N = 935	36%	42%	50%	52%
H	Surveillance	V2S	N = 4 800	28%	16%	17%	19%
H	Surveillance	V2S	N = 48 000	48%	26%	50%	37%

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

unconstrained. For the surveillance dataset (H), the approximately frontal videos are searched against pristine frontal photographs. See Table 31.

When photojournalism imagery is used in a *video-to-video* search application, identification miss rates are improved, as shown in the inset table. Here the task is to detect and enroll faces from 393 multi-person videos. This results in galleries containing from 1 000 to 3 000 templates, varying by algorithm. Then 1356 new videos each featuring one of the 393 main actors are then searched against the gallery. Accuracy is improved because any enrollment video will have more pose variation and a better likelihood to have some head pose in common with the search imagery. See Figure 39.

Set	Use	Mode	Gallery	G	M	I	N	H
C	Photojournalism	V2S	N = 935	36%	42%	50%	52%	60%
C	Photojournalism	V2V	N = Variable	21%	6%	22%	25%	27%

Recently enormous development has been done using images of celebrities and politicians and more general amateur photos due to their internet availability. The important result here is that such imagery is more challenging than with video acquired from cameras whose installation is favorable. It suggests that cross-pose recognition of the human head (which often cannot be controlled) remains the primary impediment to better accuracy.

Computational resources for enrollment of images: For enrollment of single faces in still photographs, algorithms vary in their their execution duration and in the size of the template they produce. Template size can be important because it will drive memory requirements, and speed of communication (e.g. across a computer bus). It will be more important for embedded processors, and low power devices, and less important on server-class hardware. Template sizes range over three orders of magnitude, from 128 to 109 150 bytes. The most accurate algorithm has size 2 585 bytes. This range is indicative that the industry is far from consensus on the mathematical encoding of features from faces. Using a single server-class core, template generation durations range from 0.06 seconds to 1.49 seconds. See Table 34.

Computational resources for processing of videos: Recognition in video is computationally much more expensive than in still photographs. The duration of the entire video processing function scales with (at least) the image dimensions, the length of the video, and the number of faces. If an implementation extracted features from video frames using the same still-image technology - they don't - none of these algorithms would be capable of extracting features from 24 frames per second video in real-time on a single core. Instead, the implementations process entire video clips using dedicated algorithms implemented behind a FIVE API function call. This function is passed $k \gg 1$ frames (i.e. *not* single frames) thereby devolving responsibility of video processing to the algorithm. Internally the algorithm finds zero or more faces, tracks them through space and time, and extracts features, producing a template for each track. An algorithm could extract features from a best-frame, or it might integrate features over time. As a black box test, the particular techniques remain proprietary trade secrets. Only a small minority of the algorithms have speed sufficient to execute face recognition in video in real-time using a single server-class processing core. While the duration depends on how crowded a scene is, many of the more accurate algorithms can sustain processing of a single video stream in real-time on a single multi-core server. Deployers will need to weigh network and hardware costs, given expected search volumes. In addition, the offline nature of the FIVE test methodology does not simply translate to a mode of operation in which video is continuously streamed to a recognition engine, so software (and hardware) architectural effort will be necessary. See Table 33

Privacy implications: The availability of effective means to biometrically identify persons in public spaces without their knowledge has been discussed in everything from Hollywood movies and dystopian novels to policy documents, government-led forums [4], public procurements, reports on regulation [11, 29], and national strategy drafts [6]. In the the United States, legal issues arise primarily from the Fourth Amendment to the Constitution which prohibits "unreasonable searches and seizures" without "probable cause, supported by oath or affirmation" i.e. a warrant or "consent" which retires the requirement for a warrant. The constitutionality of acquiring faces (or other biometric data) in a public space without a warrant in order to identify or track an individual has not been considered: *"The Supreme Court has*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

considered the Fourth Amendment implications of the police use of tracking beepers, electronic eavesdropping devices, photographic cameras with zoom lenses, and thermal-imaging devices, but not the use of video surveillance systems [9]. At issue will be whether any consent to be photographed additionally implies consent to storage (recording) of the result, enrollment by a feature extraction algorithm, and automated search. Would such consent be tied to a specific purpose, and would it last in perpetuity?

FIVE is germane to these issues because imaging resolution and duration have been deemed relevant in the courts. In 2016, in U.S. v. Houston [7], a warrantless search was permissible because *"the use of the pole camera did not violate Houston's reasonable expectations of privacy because the camera recorded the same view of the farm as that enjoyed by passersby"*. In 2014, a lower court had reached the opposite conclusion in a very similar case, U.S. v. Vargas [5], which reflects why the opinion in Dow Chemical v. United States [1] noted *"Fourth Amendment cases must be decided on the facts of each case, not by extravagant generalizations"*. In that case, the government prevailed because the aerial surveillance (of a Dow facility) was of low enough resolution that *"No objects as small as 1/2-inch in diameter such as a class ring, for example, are recognizable, nor are there any identifiable human faces or secret documents captured in such a fashion as to implicate more serious privacy concerns."* The implication, albeit in 1986, that face capture would trigger the Fourth Amendment, leads us to report image resolution needed for face recognition. Among the datasets evaluated here, the best accuracy is obtained from faces appearing in turnstile video clips with mean minimum and maximum interocular distances of 20 and 55 pixels respectively (See Table 9). These distances, which are lower than the 120 pixels mandated by the ISO/IEC 19794-5 passport specification, can readily be achieved with modern sensors and cameras placed tens of meters from the subject[3], given adequate light. For resolution, the argument is *"magnification may be functionally equivalent to a physical search of persons, papers, or effects"* [9] in the same way that training specialized optics on a home can be tantamount to an unconstitutional search viz. Kyllo v. United States [2] which held that use of a *"device that is not in general public use, to explore details of a private home that would previously have been unknowable without physical intrusion, the surveillance is a Fourth Amendment 'search,' and is presumptively unreasonable without a warrant"*.

Regarding duration (of surveillance), algorithms are capable of identifying faces that appear for very short durations - below one second, and at frame rates well below that of broadcast video - and this means that a public figure might be identified at the entrance to a restaurant, or a protestor could be identified at a point on a march. This presumes the existence of a database of prior reference imagery, typically an authoritative database including portrait photographs and biographical data, but could also be any curated collection of photographs. This requires that subjects approximately face the camera (conservatively, to within 15 degrees) for a short interval (conservatively, for one second). See Figures 22 and 19.

Human involvement and human fallibility: False positive and negative error rates are high enough that all consequential deployments of non-cooperative face recognition will require human adjudication of candidate matches. Prospective deployers must integrate human labor requirements into the workflow. In negative blacklist applications, where hits are expected to be rare, the system should be configured to yield (false) positives at a tractable rate. In positive whitelist applications, where all users are expected to yield a system response, and human review is not the default, the system must be configured to yield few enough false positives to satisfy security goals. Given that human review is implied and assumed in the operation of non-cooperative face recognition systems, and that humans commit recognition errors [14,33], readers should consider how the various errors from the human and automated parts of the hybrid system interact and how they can be mitigated. Note humans may be able to exploit non-face cues in video to video comparison [35].

Standards: There are standardization gaps associated with the use of automated face recognition. These are listed below. The first two are related to the necessity of human review of candidates from automated identification systems.

[3]For example, it is possible to sample 80 pixels across an interocular distance of 8cm on a subject standing at 60 meters from a modern 36 megapixel SLR camera by using a 300mm lens set with an aperture of f8, achieving a depth of field of 3.4 meters. This configuration is likely expensive c. $10 000.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

The remaining items support deployment of robust face recognition infrastructure.

- While a vibrant industry has developed around the frontal image types of the the ISO/IEC 19794-5 standard, this has had two unintended consequences. One is that because the standard recommends only 300 kilopixel images (640x480 "Token" geometry photographs), high resolution (multi-megapixel) images that are collected and used in their preparation are often not retained to support downstream review purposes. The standardized practice should be to collect high resolution, retain it for human review, and prepare the Token imagery from it for ingest into automated face recognition systems. Secondly, while the collection of frontal and profile views has existed in law enforcement since at least 1900[4], it is not universally done, and collection of non-frontal views has been deprecated in most civil applications. The standardization gap is to formalize which non-frontal views best support human adjudication of potential false positives in one-to-many recognition.

- There is no standard for the display of photographs to human reviewers. The standard should specify: the optical display properties of the display device (resolution, color depth, gamma correction etc.); rotation and interpolation (magnification) algorithms; cropping; graphical user interface functionality, including with an integrated face recognition engine; procedures for the display of single photographs (e.g. for markup), and comparison of two hypothesized mated photographs. The standard should prohibit certain steps too, for example, certain image manipulation techniques such as alterations to parts of the face.

- Testing of biometric performance of video-based systems, e.g. surveillance systems, is non-trivial. Measurement techniques used in the FIVE study have been contributed toward development of a new standard: *ISO/IEC 30137 - Use of biometrics in video surveillance systems - Part 2: Performance testing and reporting*. In early 2017, the standard is under development at the working draft stage.

- Installation of cameras for passive collection of faces for automated face recognition is a non-trivial process, particularly regarding geometry and optics. The FIVE results will be provided as quantitative support to the developement of the *ISO/IEC 30137 - Use of biometrics in video surveillance systems - Part 1: Design and specification*, now under development.

Conclusion: The accuracy of face recognition algorithms applied to the identification of non-cooperating individuals can approach that for the case of recognition of cooperating individuals in still photographs. This, however, will only be achieved if it is possible to repeatably capture similarly high quality frontal photographs. This is unlikely to occur as some proportion of a non-cooperating, passively imaged population will inevitably not present a suitable face image to the camera - for example, that population who happen to be wearing peaked hats while looking at a mobile phone. Others will present non-frontal faces. Further, high accuracy can only be achieved by deliberate installation and configuration of cameras and the environment, and such control over the deployment may sometimes be impossible, for physical, economic or societal reasons.

On the basis of the results in this report, the largest drivers of recognition accuracy are, in decreasing order of influence: algorithm selection; camera placement; the tolerable false positive rate for the particular application; subject walking speed and duration; gallery composition (portraits vs. selfies, for example); enrolled population size; camera type and the number of co-located cameras. This report provides some of the complex narratives around these variables, including interdependency.

[4]See Bertillion's standard https://en.wikipedia.org/wiki/Alphonse_Bertillon

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

CAVEATS

Readers are cautioned that, as documentation of a laboratory test of core technology, this FIVE report is not a guide to the procurement of a face recognition enabled video based system, nor is it a complete mechanism to predict performance. The following specific cautions supplement those in the main text.

- **Streaming:** The FIVE study was conducted by passing single video clips to algorithms which output proprietary recognition templates and spatial face track information. This is done offline, in a batch process. This differs from online processing in which an algorithm receives a continuous video stream, and is asked to either detect and track faces in real-time or, further, to render identity hypotheses or decisions. Our approach is adopted to allow evaluation on large datasets and to achieve repeatability, but has two effects: First FIVE does not force algorithms to operate within some defined computation budget. Stated another way, it allows the algorithms to consume as many computer cycles as the developer deems necessary. To expose this, we measure and report computational expense. Secondly, it avoids the need to more use more complex accuracy measuring techniques. An online evaluation that allows algorithms to render real-time decisions from partial video sequences would measure the tradeoff between false negative and false positive identification rates *and* recognition time. The issues and proper conduct of such an evaluation has has been succintly described in the e-Passport gate context [16].

- **Recent Development**: The algorithms whose results appear in this report were submitted to NIST in Decemmber 2015. This report therefore does not document possible performance gains realized since that time. This is a consequence of mismatched evaluation vs. development schedules. It is not ideal, yet inevitable, that a report that attempts to *broadly* document the landscape of non-cooperative face recognition performance will take time longer than the development timescales of the core face recognition algorithms. This means that some suppliers will have progressed capability beyond that available in December 2015 when these algorithms were provided to the FIVE evaluation. This has motivated NIST to launch, in March 2017, an ongoing automated face recognition benchmark[5].

- **Same day recognition**: Many of the trials described in the FIVE report include recognition of video collected on the same day as the enrollment still photographs. This is generally deprecated in performance testing standards such as ISO/IEC 19795 because it is known that accuracy is better for most biometric modalities when same-day matching is involved. This aspect is necessary for the cost-efficient collection of video data. We suggest that the increase in error rates from using say 30-day old photographs is small compared to the overall error rates, and the uncertainty in error rates from other sources. These include: population variance; population age, race and sex sample bias[6]; difficulty in reproducing cameras, compression procedures, environments, and algorithm parameterizations; and actor behavior departing from that expected operationally.

- **Accuracy is not the entire story**: Implementers should consider the following when considering which algorithms to integrate: cost; software maturity; software documentation; ease of programming; extensibility across servers and databases; performance requirements; accuracy reported here; accuracy reported in other independent test reports; accuracy dependence on image properties, such as resolution; dependence of accuracy on enrolled population size; template generation duration; search duration, and its dependence on enrolled population size.

[5]See https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing.
[6]For example, younger people are harder to recognize

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

RELEASE NOTES

FIVE Reports: The results of the FIVE appear as a series of NIST Interagency Reports. All reports are linked from https://www.nist.gov/programs-projects/face-video-evaluation-five and its sub-pages.

Typesetting: Virtually all of the tabulated content in this report was produced automatically. This involved the use of scripting tools to generate directly type-settable LaTeX content. This improves timeliness, flexibility, maintainability, and reduces transcription errors.

Graphics: Many of the figures in this report were produced using Hadley Wickham's ggplot2 package running under R, the capabilities of which extend beyond those evident in this document.

Contact: Correspondence regarding this report should be directed to FIVE at NIST dot GOV.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Contents

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

List of Tables

List of Figures

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

1 Introduction

The Face In Video Evaluation (FIVE) was conducted to assess the capability of face recognition algorithms to correctly identify or ignore persons appearing in video sequences i.e. the open-set identification problem. This test was intended to support a plural marketplace of face recognition in video systems as there is considerable interest in the potential use of face recognition for identification of persons in videos.

NIST initiated FIVE in the second half of 2014. The evaluation was focused on one-to-many identification tests for video sequences. The test was intended to represent identification applications for face recognition in video, which included:

- *Video-to-Still*(V2S): This scenario supports identification of persons in video sequences against a gallery of enrolled stills, which has application in watch-list surveillance for example.
- *Still-to-Video*(S2V): This scenario supports identification of persons in stills against a gallery of identities enrolled from videos, which may have application in media search and asylum re-identification.
- *Video-to-Video*(V2V): This scenario supports identification of persons in video sequences against a gallery of identities enrolled from videos, which may have application in identity clustering and re-identification.

Out of scope: Areas that are out of scope for this evaluation and were not studied include: One-to-one verification of identity claims; identification from body worn cameras, license plate cameras, and aerial vehicles; video analytics, scene understanding, anomaly detection, and spatial boundary violation; suspicious behavior and intent detection; estimation of emotional state; gait recognition.

2 Participation

The FIVE program was open to participation worldwide. The participation window opened on November 17, 2014, and submission to the final phase closed on December 11, 2015. There was no charge to participate.

The process and format of algorithm submissions to NIST was described in the FIVE Concept, Evaluation Plan and Application Programming Interface (API) document [20]. Participants provided their submissions in the form of libraries compiled on a specified Linux kernel, which were linked against NIST's test harness to produce executables. NIST provided a validation package to participants to ensure that NIST's execution of submitted libraries produced the expected output on NIST's test machines. FIVE had three submission phases where participants could submit algorithms to NIST. Results from phase 1 and 2 were provided back to the participants and are not documented in this report. This report documents the results of all algorithms submitted in the final phase (phase 3). Table 2 lists the FIVE participants, the letter code associated with the submitting organization, and the number of submissions made in each phase. The letter codes assigned to the participants are also located at the bottom of each page for reference.

Note that neither social media companies nor academic institutions elected to submit algorithms, and this report therefore only captures their capabilities to the extent that those technologies have been adopted or licensed by FIVE participants.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Letter Code	Organization	Phase 1 (February 2015)	Phase 2 (June 2015)	Phase 3 (December 2015)	Total # Submissions
A	Digital Barriers	1	2	2	5
B	HBInno	1	1	1	3
C	Vigilant Solutions	1	1	2	4
D	Ayonix	1	1	2	4
E	Neurotechnology	1	2	2	5
F	Vapplica	1		1	2
G	Safran Morpho	1	1	3	5
H	3M Cogent	1	1	3	5
I	Eyedea Recognition	1	2	2	5
J	Beijing Hisign Technology		2	3	5
K	Cognitec Systems		1	4	5
L	CyberExtruder			1	1
M	NEC Corporation		2	3	5
N	Toshiba Corporation		1	4	5
Q	Imagus			2	2
R	Rank One Computing			1	1

Table 2: FIVE participants, and the number of algorithm submissions, by phase.

3 Test design

Offline evaluations: The evaluation was conducted by applying algorithms to video and still imagery that is sequestered on computers controlled by NIST. Such offline tests are attractive because they allow uniform, fair, repeatable, and large-scale statistically robust testing. Most of the imagery was collected in separate collection activities staged over the last few years, the one exception being the photojournalism imagery which was produced over many years, gathered from the internet, and assembled for use at NIST.

No algorithm bias: The collection activities were conducted without any recognition algorithm involvement i.e. there is no quality bias that would result if a face recognition algorithm had been involved in the selection or retention of any images.

Operational representativeness: The degree to which the results realized in this test can be replicated in operational settings depends on many factors, the foremost of which would be collection of imagery with the same properties as that used here. The germane properties are those that generically affect face recognition performance, i.e. orientation of the face relative to the optical axis (pose), optical resolution, illumination, video compression and frame rate, and human behavior.

This test is advantaged over many biometric evaluations in that most of the video imagery is collected in a single pass without explicit subject cooperation and system feedback. The absence of a live interactive component means that offline evaluation can readily repeat what would have happened if recognition had been attempted as soon as the imagery was streamed to a receiving system.

4 Accuracy metrics

This section describes how accuracy is measured and reported. Biometric algorithms commit three kinds of errors: failures to acquire[7], and false negatives and false positives. In the context of this report, failures to acquire occur when

[7]The term failure to acquire is an overloaded phrase: It can mean failure of the face detection algorithm to find a face; it can mean the choice by the algorithm to not compute features from a face it deems to have poor utility for recognition; it might even refer to software failures.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

a face is not detected in an image or video, false negatives are outcomes where a search should yield a specific enrolled identity but does not, and false positives correspond to a search returning enrolled identities when it should not. Before defining these more formally, we note that it is erroneous yet common to state performance using a single number, a practice which is incomplete and very often misleading. As described below, correct coverage should note several aspects: the tradeoff between false positives and negatives, the enrolled population size, the prior probabilities of known and unknown persons, the degree and type of human involvement, and the dependence on image quality which can render results only poorly portable to a new installation.

4.1 Limits on ground truth annotation accuracy

Before discussing recognition accuracy metrics, we first introduce the notion that measurement of accuracy in video surveillance is limited by the appearance of multiple people in a scene, their motion in and out of the scene, and their non-cooperative incidental presence in front of the camera. This complicates the question of actually how many people are present, something that is needed for a hard estimate of false positive identification rate.

*Figure 1: This image is an example of the difficulty in counting the number of faces available for recognition. This question comes in two parts: Given an enrollment image it is clear the main subject of the photo, President Bill Clinton, and the agent with the red tie could be included in a measurement of FNIR. While the bald man at right, and the police officer next to the car, could possibly be recognized by members of his family or inner circle, should an algorithm be tested on such images? The remaining faces are detectable as human beings but are arguably not recognizable. The second part is how many faces could be counted towards a measurement of FPIR? There are 13 individuals present in the photograph, but facial skin is only visible from seven. [Photo Credit: Marc Tasman. License: https://commons.wikimedia.org/wiki/File:Bill_Clinton_Wisconsin.jpg] **This is image 913 in the IJB-A dataset. It is distributed under a creative commons license.*

When images are collected in non-staged settings and public spaces, for example from surveillance cameras or from photojournalists, and in other "in the wild" settings, the number of persons is not normally known. This is true in extended video sequences and still photographs. More importantly, the number of visible faces is not known, not least because "visibility" is poorly defined. Any human reviewer could sum the number of people present in Figure 1 but if a face only appears in two frames with 8 pixels between the eyes, does that count as a ground truth face? Does a person wearing sunglasses, a baseball hat and looking down at a mobile phone such that only his chin is visible count as a recognizable face? In an operational utopia, both of these edge cases would be recognition opportunities, but in a practical sense, they are not - recognition will forever be uncertain. One rigorous, though not scalable, way forward is two establish two quantities. First is to count humans by providing imagery to an analyst with a play-forward play-backward reviewing capability and summing where the analyst can ascertain that a person was present, whether their face was at all visible or not. Second, is to count the number of recognizable faces by asking a skilled analyst to identify the faces in a very small gallery. This method is laborious, and subject to inter-rater variance. Given the volume of imagery, this procedure is not undertaken in FIVE.

Our conclusion here is that the ground truth number of faces present in non-cooperative video is often, not always, unknown *and unknowable*. The implication of this, as detailed below, is that the denominator in the false positive identification rate estimates is not known. Instead the operational metric is the count of false positives, rather than a rate. This

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

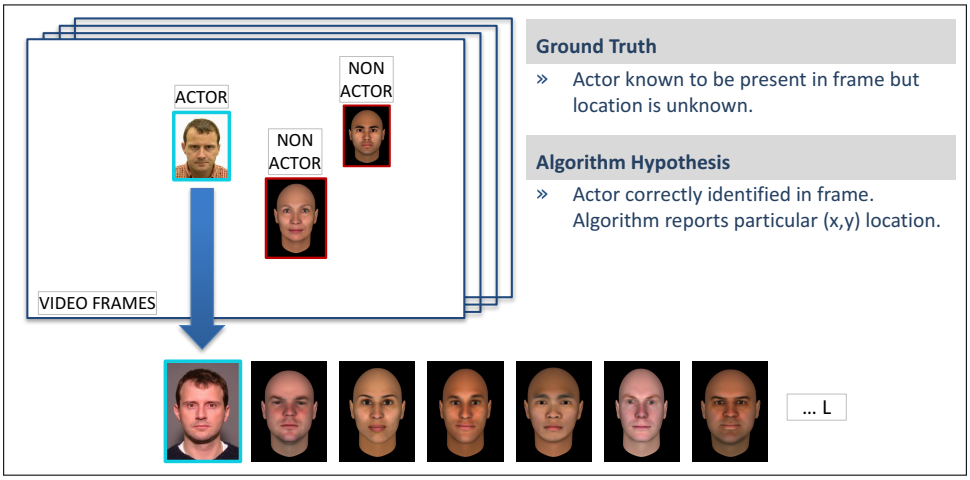


Figure 2: Illustration of a mated search. While three faces are present here, a video generally contains zero or more faces. The algorithm must detect and track faces across frames - it does not know a priori which, if any, faces are present in the enrolled gallery. The algorithm, which operates on the video sequence, produces a template for each track and this is searched against one or more enrollment galleries. The faces in this figure are of the first author or are synthetic.

does not have to be exact - an approximate estimate is sufficient as because other quality factors (resolution, pose) can alter false positive rates by larger amounts.

It could be argued that the number of known faces present is similarly unknown, such that the denominator in the false negative identification rate estimate is similarly unknown. However, in FIVE their presence *is* known, and their behaviour is expected to be that of the target population, they are all counted as recognition opportunities.

Annotation of faces in video imagery is now being standardized in ISO/IEC 30137 *Use of biometrics in video surveillance systems – Part 4: Ground truth and video annotation procedure*.

4.2 Quantifying false negative accuracy

4.2.1 Overview

The first error metric is the False Negative Identification Rate (FNIR) which can be regarded as a "miss rate". FNIR is the proportion of searches involving imagery of persons who are enrolled in the gallery which fail to produce the correct matching identity from the enrolled set of identities. FNIR is estimated by conducting mated searches (Figure 2) of people in videos or stills against an enrollment dataset where persons are known to be in both the search probe and the enrollment dataset.

4.2.2 Measurement in video

So far this definition is generic to biometrics; for faces in video clips we need more specificity. Particularly when the imagery is a video clip, two complications arise. First there can be many faces in the video. Second, a face detection and tracking algorithm may find one person on several occasions over the duration of the clip. Thus it may generate multiple templates for a single face track (see Figure 3), or incorrectly consolidate multiple face tracks into a single template. Such events must be appropriately reflected in the error metrics.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

*Figure 3: Example of how an algorithm may produce multiple templates for the same person/single face track in a video clip. **The face images in this figure are from a DHS S&T provided dataset. We obtained written consent from DHS / S&T to use these images in public reports.*

Thus, given an input video, a recognition algorithm detects zero or more face tracks and produces a template from each. Each is searched against the enrolled database. The algorithm is required to return the L closest (most similar) candidates. We quote accuracy by shortening the candidate list by either applying a threshold T, or by considering only the top $R \leq L$ candidates[8]. Each candidate is comprised of a hypothesized identity, and a similarity score. The algorithm sorts the resulting candidate list in descending order of similarity score. Even when a clip contains only one person, we need to handle the detection and production of $K \geq 0$ candidate lists. Thus, for a video known to contain imagery of individual i we take the maximum of K mate scores or 0 if there are none.

Formally, when searching N identities, let s_{ikr} be the comparison score associated with the $1 \leq r \leq L$-th ranked candidate from the $1 \leq k$-th candidate list from the ith video, and let p_{ikr} be the subject identifier for the r-th candidate from the same candidate list and video. Additionally, if J_i is the number of templates returned for the i-th video clip, where $0 < i \leq I$, and M_i is the set of identifiers for subjects actually present in the ith video then

$$\text{FNIR}(N, R, T) = \frac{\sum_{i=1}^{I} \sum_{j \in M_i} H\left[\max_{1 \leq r \leq R} \max_{1 \leq k \leq J_i} \left[s_{ikr} \mid p_{ikr} = j \right] - T \right]}{\sum_{i=1}^{I} |M_i|} \quad (1)$$

where $|M_i|$ is the size of M_i and the step function $H(x - T)$ is 1 if score x is at or above threshold T. The denominator counts the total number of times subjects in the enrollment dataset appear in the video clips (only counting a subject

[8] To clarify the relationship between L and R: L is the number of requested candidates that is communicated to the algorithm. Importantly setting L to a large value will generally cause search to go slower, so system administrators may not permit analysts to set L at all, or only to do with some bound. R, on the other hand, is the number of candidates that an analyst might look at in an application with a graphical user interface. The analyist is constrained to look at R ¡= L. The performance metrics in this report survey over R around, to see the effect on accuracy. Operationally, on a busy day, local workflow management software might limit analysts to look only at top R = 5 canidates even though the underlying system was requested L = 100.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

multiple times if they appear in different video clips). Note that for each video clip, this equation only considers the highest score for an enrolled subject regardless of how many candidate lists the subject appears on. Thus for a given search clip, a correct identification occurs if the subject appears on at least one candidate list for the video clip with a score at or above threshold T and rank no greater than R. This gives no additional credit for correctly identifying a particular subject in a video clip several times. Algorithms vary considerably in how many templates they produce given the same input clip. This arises due to varying face detection methods and imperfect tracking of individuals across frames. Algorithms that generate more templates (and thus more candidate lists) have more opportunities to find the correct person. A "gaming strategy" might involve submitting an algorithm that produces very large numbers of templates essentially guessing at identities. Although this could reduce FNIR, it might also substantially elevate the number of false positives, particularly because some galleries here contain no actors (for measurement of FPIR - see section 4.3 below).

4.2.3 Application-specific FNIR metrics

The report includes extensive tabulation of FNIR reflecting two classes of use:

Forensic: In a high profile case, or in an application where only a few searches are ever conducted, a human analyst might examine say $1 \leq R \leq L$ candidates, where L is the maximum number available[9]selecting R according to the priority of the case, and labor availablity. The analyst might increase candidate list length, L, also to support a more laborious search for matching identities. In any case, the appropriate metric for this forensic use case, is a special case of equation 1, namely FNIR$(N, R, 0)$ where threshold is set to zero so that all candidates can be available for review. This is a "miss rate" and is related to the ubiquitous cumulative match characteristics which states the proportion of searches with "hit" at rank R or better:

$$\text{CMC}(N, R) = 1 - FNIR(N, R, 0) \tag{2}$$

By ignoring scores (T = 0), this metric allows "weak" hits to count as strongly as high-scoring "strong" hits. Note the CMC metric is relevant to operations in which (trained) human reviewers who will traverse candidate lists in pursuit of hits are required and assumed. Their presence, in conjunction with a face recognition engine, forms a hybrid automatic-human system. The system functions only when when the volume of searches is low enough, and when the CMC is favorable enough, to occupy the available labor (and no more).

Surveillance: On the other hand, in applications such as surveillance in a public-space, where the prior probability of a mate is low, or where search volumes are very high and where human labor has limited availability, it becomes impossible to review all candidate lists. To limit workload a threshold T is applied so that only candidates with score at or above threshold are provided for examination. The appropriate metric then is FNIR(N, L, T) where the rank criterion is relaxed by setting R equal to L, and a non-zero threshold is applied[10]. High thresholds suppress

[9]Some biometric search implementations return only high scoring candidates. It is more common, however, for systems to return a specified number, L, candidates, and this value is communicated to the algorithm. This can be set by system policy, or sometimes by the analyst. In general, the duration of the search depends on L, the number of nearest neighbors are being sought, because multi-stage templates might be used, and because some fast search algorithms depend on the data. R is the number of candidates that an analyst might look at in a GUI-enabled workstation. We distinguish the symbol R from the length L in order to that we may sweep it over its range $1 \leq R \leq L$ to see the effect on accuracy. Operationally, on a busy day, local workflow management software might limit analysts to look only at top R = 5 even though the underlying system was set to require $L = 100$ from the algorithm.

[10]This formulation allows a mate to be at any rank $R < L$ as long as it is above threshold. Practically FNIR$(N, L, T) \rightarrow$ FNIR$(N, 1, T)$ except when $T \rightarrow 0$. Once the threshold is elevated slightly sufficiently, mates are always found at rank 1.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

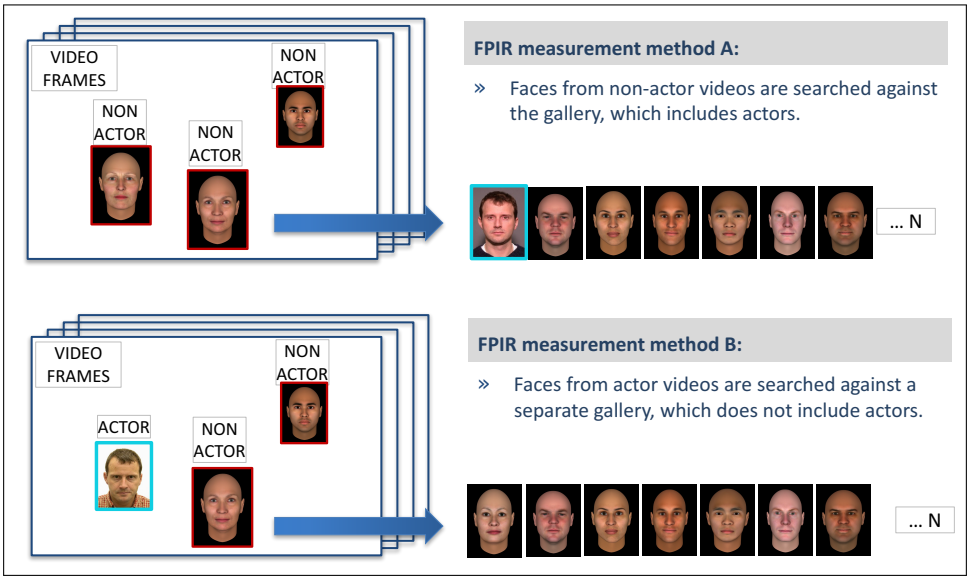


Figure 4: The figure shows two means of measuring FPIR. At top, videos containing faces not present in the gallery are used. Below the same set of videos used in measurement of FNIR are used but against a separate gallery not containing the actors. The faces in this figure are of the first author or are synthetic.

false positives, but elevate false negatives. For example, the 2007 German trial of a surveillance system in the Mainz train station [3] configured the threshold on each algorithms to target FPIR= 0.001.

The threshold is set to limit the number of false positives. This is discussed in the next section.

4.3 Quantifying false positive accuracy

It is conventional in testing of biometric identification systems to measure the false positive identification rate (FPIR). This is done by running searches of individuals who are known to be absent from the enrolled gallery. FPIR is then computed as the proportion of searches that produces one or more false positives above a threshold, T.

Here, this computation is not possible, because the number of individuals in the search imagery is not known, per the discussion of Figure 1. This leaves us to compute only a *number* of false positives from some searches. This still depends on the threshold, which we calibrate as follows. Given an input video, the recognition algorithm detects zero or more face tracks, producing a template from each. These are searched against a disjoint set of N individuals such that all reported candidates are, by definition, false positives. The threshold is set to the lowest value that results in a fixed number of observed false positives, denoted by NFP(T). This value is an integer rather than a proportion of the population.

As shown in Figure 4, NFP(T) is estimated in either of two ways: First by searching imagery of unrelated individuals against an enrolled actor gallery; and second by searching actor imagery against a gallery of unrelated indviduals. In a video-to-still experiment, the first of these methods necessitates collection of separate video, while the second method uses just the actor video but requires construction of a new, disjoint, gallery.

Most of the video clips used in this evaluation are fairly short ($\leq$ 20 or 30 seconds), and a particular subject only appears in a video clip once (although presentation of the face may be momentarily interrupted due to occlusion or pose changes).

To compute non-mated scores, in most cases the enrollment dataset was replaced with an equally sized set of frontal

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

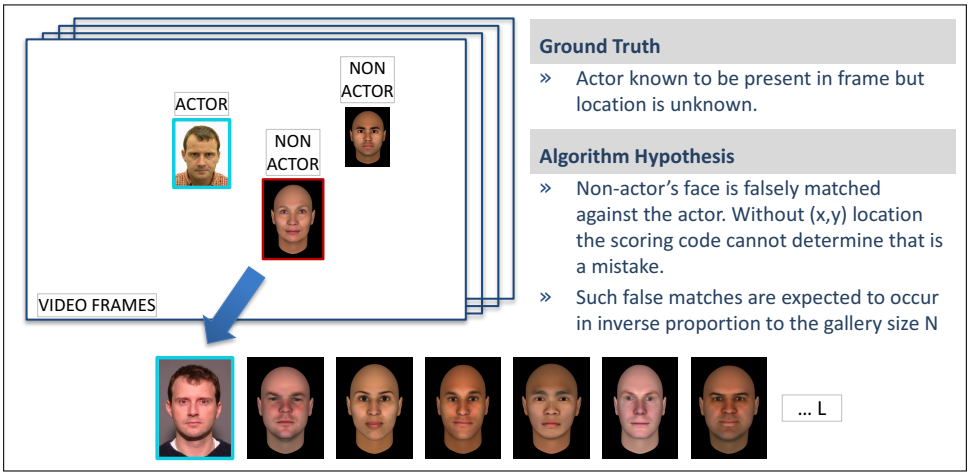


Figure 5: The figure shows that without location ground truth, it is possible for the algorithm to be credited with a correct identification of the actor, but for the wrong reason. The faces in this figure are of the first author or are synthetic.

stills of subjects that are not in any of the video clips. Let q_{ijk} be the comparison score associated with the kth candidate from the jth candidate list from the ith search. Since the enrollment dataset is populated with people not present in the video clips, all scores are non-mated. The number of false positives is computed as follows

$$\text{NFP}(T) = \sum_{i=1}^{I} \sum_{j=1}^{|J_i|} H(q_{ij1} - T) \tag{3}$$

NFP is the number of times a algorithm incorrectly flags someone as being in the enrollment dataset. Algorithms that detect more faces or produce more tracks will generate more false positives. The presence of the 1 subscript in q indicates that a search produces a false positive if 1 or more (i.e. any) of the candidates are at or above threshold.

4.4 Uncertainty implied by incomplete ground truth

For a majority of the datasets, our evaluation methodology only checks that the correct person was found within a window, or in close proximity to the known ground truth. This is depicted in Figure 6. We do not test whether the subject was found in the correct location. This section describes the effect of this, and its mitigation.

Suppose we have search imagery where one actor and n non-actors appear concurrently, and the actor has a mate in an N person enrollment gallery. Suppose further we do not have spatiotemporal location information of the actor, i.e. we don't know where the actor is in the video. Suppose also that an algorithm detects all $n+1$ faces, tracks them, producing search templates and then candidate lists. There is some chance that a non-actor template incorrectly matches the actor's enrollment template randomly such that the accuracy computation counts a correct identifcation *for the wrong reason* - see Figure 5. The probability that any one non-actor is returned as a match within the top R candidates can be obtained from the hypergeometric distribution as R/N. This applies on binomial grounds to all n non-actors in the scene such that there is a systematic underestimation of the true FNIR as follows:

$$\text{FNIR}_{\text{OBSERVED}}(N, R, 0) = \text{FNIR}_{\text{TRUE}}(N, R, 0) \left(1 - \frac{R}{N}\right)^n \tag{4}$$

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

which has an approximation for small R[11].

To gauge the worst case magnitude of this error, we examine the case of N = 480 (the smallest gallery we used), R = 20 (the highest rank we used), and $n = 10$ (the approximate maximum number of people present in a clip). There the observed FNIR may understate the actual FNIR by a factor of 0.65. This is large. For rank 1 recognition, the factor is 0.98. For our largest gallery, N = 48 000, the factor becomes negligible even at rank 20 (0.996).

However, thus far the analysis has been about random association of a non-actor with an actor. But face recognition algorithms do not place items on candidate lists randomly[12], rather they nominate candidates in decreasing order of similarity score, rendering the hypergeometric model incorrect. This improves the situation markedly because the probability that a non-actor incorrectly matches the specific actor is related to the one-to-one false match rate, which reduces with threshold. This means that equation 4 is pessimistic, and particularly becomes irrelevant when the threshold is increased to limit false positives. A number of other aspects mitigate the problem further.

1. N is often much larger than our minimum here, N = 480.
2. For comparison of algorithms R = 1 is more appropriate.
3. We are mostly concerned with high threshold cases, particularly for the crowded surveillance datasets, H and T.
4. For datasets U, J, C, the number of people in the field of view, n, is often naturally near to 1.
5. For datasets J, L, H and T, the temporal ground truth is localized tightly such that the number of people present is usually below 5. Thus even in video clips lasting several minutes (dataset J) where footage of hundreds of indviduals appeared, we imposed the additional constraint that the matching software must report the subject over the correct time interval (See Figure 6).

Nevertheless we add the caveat to some tables in this report, for N = 480 and R = 20, to direct readers to the arguments of this section.

This issue has been discussed previously [8].

5 Experiments and results

5.1 Overview

The prior section gave formal exposition of the metrics. This section previews the experiment runs that support the results that follow. Algorithm performance is assessed over seven biometric datasets. The datasets differ with respect to the types of cameras used for collection, their placement and number, the background environment, and many unknown and difficult to document factors. Each dataset roughly imitates a particular scenario. In all cases, user cooperation is essentially non-cooperative, meaning the people in the videos were variously unaware, not-cognizant, or oblivious to the presence of cameras. In one dataset, the camera was mounted alongside an attractor (computer display). It was included with the intent to induce head elevation and thereby a more frontal pose. In all cases, subjects were not given instructions

[11] With $1-(1-p)^N \rightarrow pN$ for small pN, the approximate formula is $\text{FNIR}_{\text{OBSERVED}}(N,R,0) = \text{FNIR}_{\text{TRUE}}(N,R,0)\,(1-nR/N)$

[12] Except in cases where the image quality is very poor.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

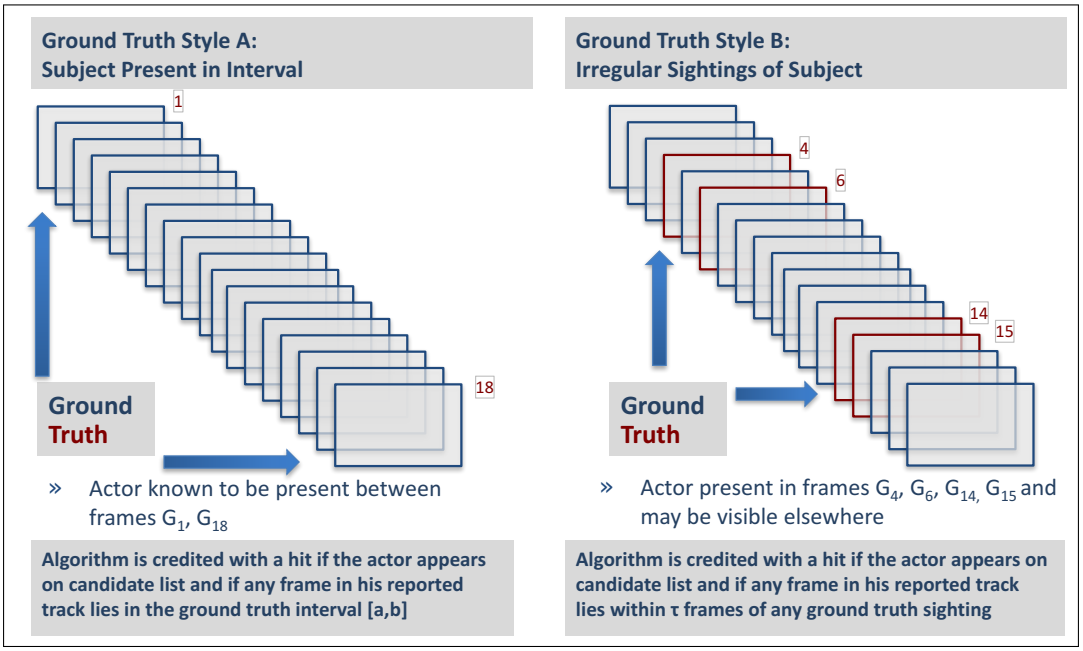


Figure 6: Algorithm evaluation based on ground truth methodology.

that would improve presentation of their faces to the camera. Non-cooperation renders recognition of persons a much more difficult problem compared to applications that involve cooperation of the subjects.

The properties of the data are summarized in Table 3. Further information about the image sets, associated metadata and ground truth are presented alongside the experimental design and the results in the following seven subsections. Each of these gives performance estimates for each evaluated algorithm. The term performance is a generic term covering recognition accuracy, computation duration, and storage requirements. Examples from the image sets are included alongside the recognition results.

Figure 7: DATASET U: PASSENGER GATE *video clip examples. **The face images in this figure are from the DHS/ S&T provided AEER dataset. The included subjects consented to release their images in public reports. Subject 79195746 (Perm Granted). Where consent for public release from individuals in the background was not obtained, their faces were masked (yellow circles).*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Code	P	H	T	L	J	U	C
Setting	Sports arena	Travel concourse	Travel concourse	Luggage rack	Aircraft loading bridge	Passenger boarding gate (chokepoint)	Photojournalism
Application I	Surveillance	Surveillance	Surveillance	Crime scene	Immigration Exit	Immigration Exit	Media search
Application II	Shoplifting detection	Re-identification	Re-identification	Legacy installation	Choke point	Access control	Catalog
Details and results	Sec. 5.5	Sec. 5.6	Sec. 5.7	Sec. 5.4	Sec. 5.3	Sec. 5.2	Sec. 5.8
Camera	Canon VIXIA HF R400	Avigilon B1	Various	Logitech C920	Vaddio PowerView (PTZ)	Logitech C920	Various pro + TV
Width, height (pixels), and frame rate (sec^{-1})	1920, 1080, 24	1920, 1080, 30	Variable	1920, 1080, 24	1920, 1080, 24	1920, 1080, 10	Variable, 24+
Number of cameras	11 in three locations	10 in banks of 4, 4, and 2 spanning width of different corridors, each with own geometry	10 each in separate locations	2, sharing FOV	3, sharing FOV	1	Many, 1 per scene
Camera height + elevation	Wall, 6ft + 8ft	Ceiling, 8ft	Unknown, low	Ceiling, 8ft	Ceiling 8ft Wall 6ft (x2)	Post, 5ft	Variable, inc. below
Subject motion	1. Queuing 2. Walking in hallway 3. Entry and exit via doors to outdoors	Walking mostly toward and below the cameras	Walking mostly toward and below the cameras	Picking up luggage from rack, then walking below cameras inspecting luggage	Usually single file toward and below the camera. Walking OR stopped in queue	Toward the camera, presenting boarding pass, then walking left of camera	Often standing at podium, or seated facing journalist. Seldom walking
Pitch of face	Benign, Moderate	Benign	Moderate	Adverse	Moderate	Good	Moderate
Yaw of face	Benign, Moderate	Benign	Moderate	Adverse	Good	Good (briefly)	Adverse, moderate
Typical number of subjects in view	[0,30]	[0,10]	[0,10]	[1,10]	[1,3] walk or [4,7] queue	[1,3]	[1,3] + variable
Number of actors	64	48	31	248	343	248	393, 1870
Impostor score generation	Figure 4 method A	Figure 4 method B	Figure 4 method B	Figure 4 method B	Figure 4 method B	Figure 4 method B	- (Closed universe searches only)

Table 3: The Datasets and their accompanying properties. Generally, each dataset is comprised of sets of reference cooperative stills and non-cooperative videos. Section 5 includes examples from each dataset ahead of the respective recognition results.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			


**Subject 79195747 (Perm Granted)


**Subject 79195746 (Perm Granted)


**Subject 79195743 (Perm Granted)

*Figure 8: Dataset U: Examples of enrollment images, collected with consumer SLR. **The face images in this figure are from the DHS / S&T provided AEER dataset. The included subjects consented to release their images in public reports.*

5.2 DATASET U: PASSENGER GATE

5.2.1 Overview

The DATASET U: PASSENGER GATE dataset contains videos of subjects walking toward an aircraft boarding pass reader, using it, then proceeding left across the optical axis passing the camera - see Figure 7. The subjects were queued in this process, and therefore the faces of other individuals are often present in the scene background.

The imagery was collected in a purely passive mode - the subjects are unaware of the camera, and make no attempt to look at it. The collection, therefore, is entirely non-cooperative. This diverges from the traditional use of biometrics for access control where the boarding pass presentation would form an identity claim and the traveler would be authenticated in a one-to-one process. In this concept of operations a gate might open as the result of this process, something that would require real-time operation. Alternatively, the result of the verification might simply be logged. Here, instead, the video data is searched in one-to-many mode against a dataset of individuals expected to board the aircraft. This single-factor authentication has the advantage of not delaying the existing boarding process at all, but has the cost of elevated recognition error rates over those achievable with one-to-one verification. Note that both of these processes could be conducted with or without cooperation. Even in the case of verification, there might not be any instruction to the traveler to look at the camera.

Property	Value
Camera	Logitech C920
Camera mounting	Attached to display observed by subject
Camera height	Approx. 1.75 meters, (5 feet 9 inches)
Range to subject	[0.7,4] meters
Frame rate	10 sec^{-1}
Width	1080
Height	1920
Chroma sampling	YUV420
Nominal bitrate	130 Mb sec^{-1}
Codec	WVC1 (advanced) 0x31435657

Table 4: Key imaging properties for DATASET U: PASSENGER GATE

Videos: The DATASET U: PASSENGER GATE videos are of subjects walking towards and using an aircraft boarding pass reader. Table 4 summarizes key imaging properties and Figure 7 shows examples from the single webcam device that observed this activity.

Enrolled still images: Video clips are matched against a set of enrolled still photographs collected with a consumer-grade SLR - see Figure 8. These are in good conformance to the ISO/IEC 19794-5 full frontal image type. These images are enrolled into three galleries of size $N = \{480, 4800, 48000\}$. These sizes are attained by including high quality frontal

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Quantity	Value or description
Mode	Video search to still enrollment
Number of actors	248
Number of non-actors	0
Number of cameras	1
Video duration with actors	18.4 minutes
Video duration no actors	0
Subject motion	Toward and then to left of camera
Number of clips	248, mean length 4.5 seconds
Clip sequencing	Main person in clip n is in background in clip $n-1$
Clip duration (frames)	Median 43; Min 18; Q25 37; Q75 48; Max 102
Number of enrolled subjects	480, 4800, 48000
Number of enrolled stills	1 per subject
Properties of enrolled stills	Frontal, close ICAO compliance; Mean IOD 106 pixels
FNIR estimation	Actors present video vs. enrolled gallery
FPIR estimation	Actors present video vs. separate non-actor gallery
Candidate list length	20
Number of persons in FOV	[0,3] usually; one dominant in foreground
Video ground truth	Style A: See Figure 6

Table 5: Key experimental design the DATASET U: PASSENGER GATE *results.*

portrait photographs from a disjoint background population. Exactly one image is enrolled for each person.

Experimental Design: Mated scores are generated by searching 248 video clips against the three enrolled dataset of still face images of subjects known to be in the search videos. **Nonmated scores** are collected by comparing the same 248 video clips against three enrollment datasets known not to contain actor images. These sets, which contain N = $\{480, 4800, 48000\}$ frontal images, are termed the **global nonmated enrollment dataset**, and are used to generate nonmated comparison scores for other datasets as well. A limitation of this approach is that by searching video of only 248 individuals we cannot calibrate thresholds that would yield low false positive identification rates (e.g. FPIR = 10^{-5}). To do that we'd need to run many more searches from many more people, something we do later for Dataset P.

Key experimental design details are summarized in Table 5.

5.2.2 Accuracy

Results for the DATASET U: PASSENGER GATE set are presented exhaustively in Tables 6 - 8, one for each enrolled population size. These tabulate false negative identification rates, FNIR(N, R, T), in the two special cases, one relevant to investigations, FNIR$(N, 1, 0)$, and the other to high volume identification, FNIR(N, L, T) where T is set to realize a fixed number of false positives, NFP(T) over all video clips. Extracts from these are then graphed in Figures 9 - 11.

The notable results are:

Absolute accuracy: The most accurate algorithms are those submitted by participant M (M30V, M31V, M32V). At the strictest threshold M30V achieves an FNIR of 0.056. The detailed interpretation of this is important. In an access control context, it says that when 248 subjects pass through a chokepoint, while executing their document scanning task, we expect to correctly identify 94.4% of those subjects in a gallery of size $N = 480$, without any explicit cooperation from them, while only producing 1 false positive. What is a false positive in this context? It is the failure to correctly reject an unauthorized (unenrolled) subject who attempts to gain access. Given that such an impostor might be rare, it may be tenable from a security perspective, to lower the threshold such that 10 false positives out of 248 people passing the chokepoint was acceptable. In that case, the same algorithm would then

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

correctly identify 96.4% of legitimate subjects.

Breadth of capability: Despite the favorable geometric and optical configuration of the installation, face recognition accuracy varies widely across developers. Four algorithms miss more than 90% of subjects at NFP$(T) = 10$, while nineteen algorithms achieve an FNIR below 25%, eleven achieve an FNIR below 15% at the same decision threshold, and five achieve an FNIR below 10% (M30V, M31V, M32V, G31V, G32V). On this basis, the marketplace capability is not broad.

Rank based accuracy: If for some reason, video from this geometrical and optical configuration were occasionally used in forensic searches, such as a human trafficking investigation, accuracy is better, and many more algorithms offer useful accuracy. Thus, even at N = 48 000, the best accuracy is a rank-20 hit rate of 99.2% (M31V) with 18 algorithms giving better than 90%.

Effect of threshold: Figure 9 shows accuracy for N = 480, at three decision thresholds corresponding to false positives counts of 1, 10, and 100. The latter two numbers are quite high: Ten false positives means that about one in every twenty-five (10/248) video clips would yield a false positive - the incorrect identification of a person with the enrolled subject. Because there is a tradeoff between the FNIR and the number of false positives, decision thresholds that elicit the fewest false positives, here 1, produce the highest FNIR(N, L, T).

Figure 10 plots FNIR(N, L, T) against the count of false positives. This plot is related to the error tradeoff characteristic that typically appears in biometric performance reports. It differs in that the x-axis is a *count* of false positives rather than a *rate* because the numbers of persons appearing in video imagery is (generally) not known. Such a plot is useful in that it supports cost benefit decisions: The y-axis, FNIR(N, L, T), is related to the benefit and the x-axis, NFP(T), drives the costs associated with subsequent consequences and resolution of false positives, typically via human adjudication and intervention.

Algorithm tuning: Some algorithms are configured toward giving better forensic investigational mode accuracy than in the high-threshold identification mode. This design feature is typical in biometrics, and suggests algorithm variants should be procured or parameterized properly, to emphasize discrimination between faces in large populations vs. invariance to facial appearance. Some algorithms are more robust to relaxation of the decision threshold than others. This sometimes makes it difficult to say one algorithm is more accurate than another in any absolute sense. For example, when $N = 480$, G32V and H30V produce similar FNIRs at FP=100 (0.048 vs. 0.044 respectively), but at FP=10, G32V achieves a much lower FNIR (0.056 vs. 0.137). Similarly, at FP=100, N32V achieves a lower FNIR than K31V (0.065 vs. 0.117) but at FP=10, K31V achieves a lower FNIR (0.145 vs. 0.185). The most accurate algorithm, M31V, achieves an FNIR at least as low as every other algorithm at all three decision thresholds.

Scaling to large gallery sizes: Figure 11 shows how FNIR is affected as the enrolled population increases from N = 480 to N = 48 000. As N increases, false positives occur more frequently and they displace rank one hits in mated searches, and necessitate higher decision thresholds to suppress false positives in non-mated searches. Generally, increasing N leads to elevated FNIR miss rate. For M30V, the FNIR(N, L, T10) increases from 0.032 to 0.097, roughly a factor of 3 increase for a 100-fold increase in the size of the gallery (from 480 to 48,000). For most algorithms, the FNIR tends to increase linearly with the log of the gallery size, paralleling the behavior of many still-face recognition algorithms [22, 23]. This dependence is benign and is the fundamental basis for the utility of face recognition in large population applications.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

The effect of enrolled gallery size is evident also in Figure 10, which replots the tabulated FNIR values. The expected behavior from binomial models, would be for the lines to be parallel and evenly spaced indicating that FNIR is conserved when a ten-fold increase in the gallery size is accompanied by a ten-fold increase in the number of tolerable false positives. Visually this applies only for some algorithms (e.g. I30V, J3xV).

Note that algorithm G30V and the four algorithms from provider N give better accuracy when N = 4 800 than at N = 480. This appears when the lines in Figure 10 cross, as NFP$\rightarrow 1$. This effect is unexpected. It may arise because gallery score normalization is more effective with larger N. Score normalization schemes are often used to stabilize the nonmate distribution. The rank-one miss rates do increase with N as expected.

Relevance to an access control application: The consequences of false negatives and positives differ by application. If this was a positive access control application using one-to-many identification to permit entry to a building, for example, then a false positive would correspond to an incorrect admission. This would occur if an unauthorized user matched any of the enrolled entries. Given the dataset and the results for it, can we conclude that accuracy would be fit-for-purpose? An access control application would require false positive identification rates to be somewhat lower than achieved here, i.e. FPIR $< 1/248$. Our best miss rate at this kind of false positive identification rate - the yellow column in Table 6 - is around 6%. However this result is pessimistic in that "white-list" applications have subjects who can engage the face camera in multiple cooperative attempts. Note that impostors, too, can make repeated attempts at recognition, unless some mechanism is implemented to impede that. The use of face recognition without an identity claim is an example of single-factor authentication and is therefore inherently weak from a security perspective. Acknowledging this, the value of video-based recognition here is in expediting two factor access control by using one-to-many identification of a subject as she approaches a document reader. This has potential to expedite the process, by doing face capture and feature extraction, prior to presentation of the token. Recognition too could be done as a one-to-many search of a database, if available, or against biometric data read from the identity token.

Relevance to an immigration exit application: If this was an immigration exit application using one-to-many identification to record biometric exit of in-scope visa holders boarding at an international departure gate, then the following categories of error need to be considered and addressed

- *Failure to enrol*: If an in-scope traveler's reference photograph is not enrolled into the gallery, then this counts toward the failure to enrol rate (FTE). This could occur because the reference photo was simply unavailable, or was of such poor quality that the algorithm could not, or would not, produce a template from it. This would essentially be an additive increment to FNIR(N, L, T). In this report, FTE = 0, by design. In operations, it may occur, for example when a new employee's photograph is not enrolled into an access control system.
- *Failure to acquire*: If an in-scope traveler is not detected by the face recognition algorithm, then the result is effectively a false negative. This could occur, for example, if the camera simply didn't cover the proper volume, or because the person's face was occluded by another's. If an explicit measurement of the failure to acquire rate, FTA, was available it could be combined with a false negative identification rate measured over a population appearing as intended, to give an overall statement of false negative error

$$\text{FNIR}_{\text{OVERALL}} = \text{FTA} + (1 - \text{FTA})\text{FNIR}_{\text{MEASURED}} \tag{5}$$

The point is that FTA is the proportion of travelers who would not be recognized even with a perfect recogni-

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

tion algorithm.

– *False negative*: If an algorithm does not identify an in-scope traveler against his enrollment template, then this counts toward elevated false negative identification rate, FNIR(N, R, T). This is a common occurence, and is the primary subject of this report.

– *False positive*: A false positive would occur when video from traveler A was incorrectly associated with an enrolled person, B, who is expected to board the aircraft. Person A could also be a) an in-scope traveler who just happens to match B rather than his own enrollment, or b) a traveler who is not enrolled at all. In this case, person A could instead be colluding with an enrolled subject, B, and trying to specifically impersonate him to record exit. This is very unlikely to be successful - probability $\sim 1/N$ without any effort, but higher with a dedicated presentation attack e.g. by using a sibling, or a face mask.

Further improving accuracy: With further trials it may be possible to improve accuracy. This could be achieved by improving temporal resolution (30 frames per second instead of 10), and (expensively) by adding another camera. However, it may also be possible by refining the position of the camera, in particular by colocating the camera with the boarding pass reader. While this would yield frontal and higher resolution frames, it may also impart some advserse distortion associated with the camera being too close to the subject. The key point here is that prospective deployers must engage in a deliberate optical, mechanical and environmental design effort.

5.2.3 Resolution

Optical resolution is highly influential on face recognition as it is necessary to resolve features that afford discrimination. Resolution in the recorded imagery is afforded by proximity of the camera to the subject, by design of specific optical properties of the camera, by lack of motion blur, and by benign application of modern compression algorithms, particularly in video.

We have no formal resolution measurements for the imagery used in FIVE. The only indicator we have is interocular distance (IOD) i.e. the distance between the eyes, as reported by face detection algorithm. This is measure of spatial sampling rate, and is a weak proxy for optical resolution (because low resolution images can be interpolated to produce high spatial sampling rates). Nevertheless it is a useful design parameter in face recognition, because it is assumed that the optical and compression specifications have been well designed.

For DATASET U: PASSENGER GATE Table 9 shows statistics of IOD reported while tracking individuals in video. It also reports track length statistics. The notable observations are:

Track lengths: Algorithms vary in the lengths of the tracks they report. Some algorithms (R, C, F) limit feature extraction to fewer than 7 frames corresponding to 0.7 seconds. Others consider longer tracks (K31V, G31V, G32V, N3xV, E30V) with extent approaching 30 frames (3 seconds).

Resolution: Algorithms vary in which part of the video clips they elect to use. The tabulated IOD values are averages of the minimum, mean, and maximum IOD reported for each track. The M3xV algorithms have tracks with mean IOD of 40 pixels. The J, D, R, and L algorithms extract information from subjects with mean IOD above 75 pixels. Thus, refering to Figure 7, some algorithms focus on subjects when they are relatively far from the camera, and others when they are close. In this dataset, the faces of subjects close to the camera usually exhibit larger yaw angles.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

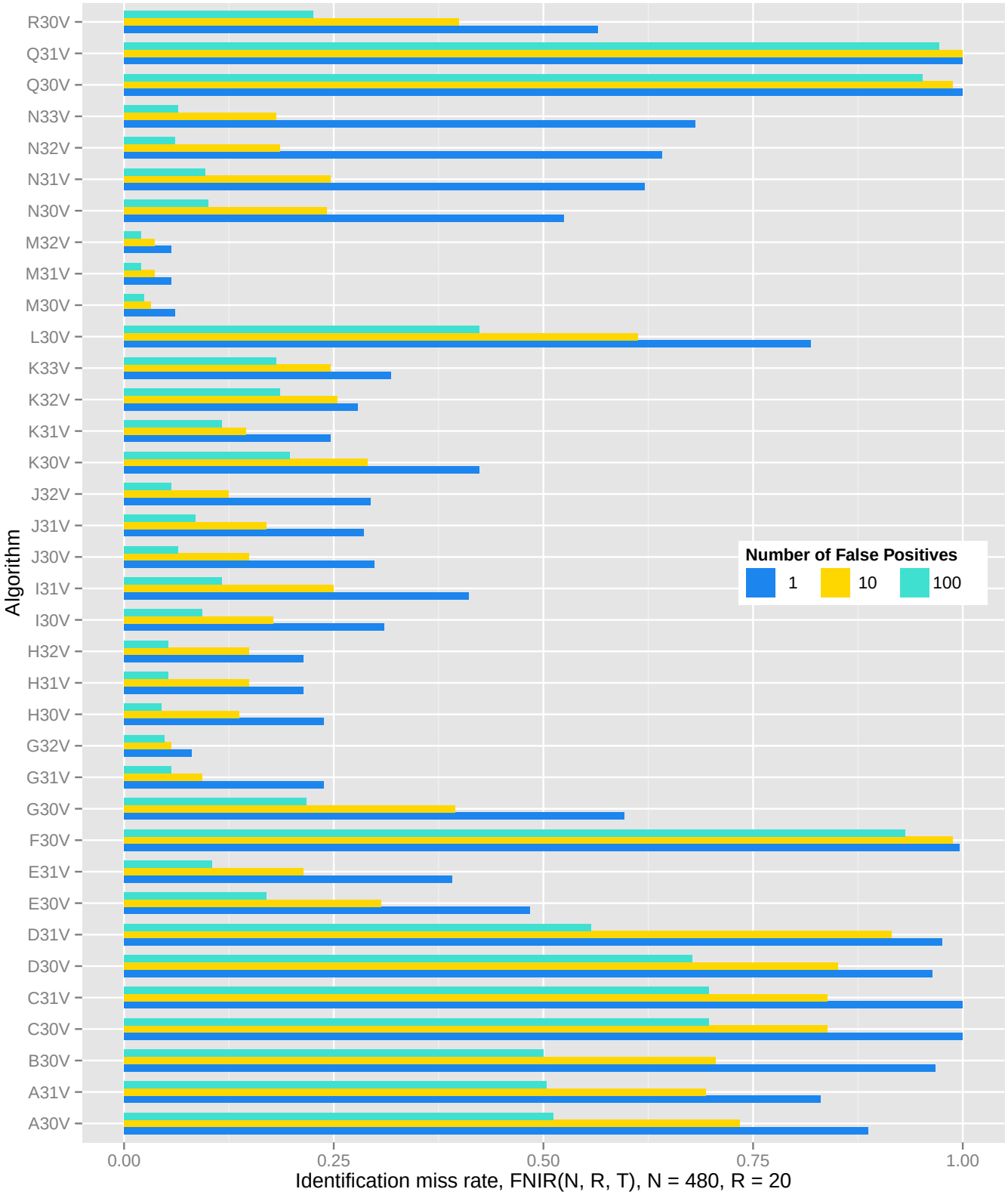


Figure 9: For DATASET U: PASSENGER GATE *with N = 480 subject enrolled, the bars show FNIR(N, L, T) for three different decision thresholds, T corresponding to 1, 10 and 100 false positives. High thresholds are necessary in applications to match the number of false positives to the human labor availability needed for their adjudication. False positive outcomes increase linearly with the number of faces appearing in video.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

N=480	NUM ACTORS 248		NUM FEEDS 1		NUM CLIPS 248		NUM FRAMES 11012		NUM MINUTES 18.4	
	DETECTIONS	THRESHOLD BASED AUTO WATCHLISTS					RANK BASED FORENSIC CASES			
ALG	NUM	FNIR(T), FP(T)=1		FNIR(T), FP(T)=10		FNIR(T), FP(T)=100		FNIR(R=1, T=0)		FNIR(R=5, T=0)		FNIR(R=20, T=0)	
A30V	1345	0.887	28	0.734	29	0.512	29	0.290	28	0.121	27	0.056	24
A31V	1345	0.831	27	0.694	27	0.504	28	0.282	27	0.117	26	0.056	23
B30V	714	0.968	32	0.706	28	0.500	27	0.274	26	0.153	30	0.060	25
C30V	2364	0.903	29	0.839	31	0.698	33	0.423	33	0.331	32	0.290	33
C31V	2661	0.907	30	0.839	30	0.698	32	0.419	32	0.335	33	0.282	32
D30V	2004	0.964	31	0.851	32	0.677	31	0.347	31	0.145	28	0.081	28
D31V	638	0.976	33	0.915	33	0.556	30	0.315	29	0.145	29	0.093	29
E30V	743	0.484	19	0.306	23	0.169	20	0.093	22	0.065	22	0.036	20
E31V	743	0.391	16	0.214	16	0.105	17	0.065	16	0.044	19	0.032	19
F30V	1502	0.996	34	0.988	35	0.931	34	0.875	34	0.710	34	0.476	34
G30V	493	0.597	22	0.395	24	0.218	24	0.173	25	0.113	24	0.105	30
G31V	691	0.238	8	0.093	5	0.056	8	0.056	15	0.044	20	0.040	21
G32V	691	0.081	4	0.056	4	0.048	5	0.056	14	0.048	21	0.044	22
H30V	668	0.238	7	0.137	7	0.044	4	0.036	4	0.012	4	0.012	6
H31V	668	0.214	5	0.149	9	0.052	6	0.040	5	0.020	5	0.012	4
H32V	668	0.214	6	0.149	10	0.052	7	0.040	7	0.020	6	0.012	5
I30V	1382	0.310	14	0.177	13	0.093	14	0.077	19	0.040	15	0.032	18
I31V	1382	0.411	17	0.250	20	0.117	19	0.077	20	0.040	18	0.028	17
J30V	441	0.298	13	0.149	11	0.065	12	0.048	10	0.024	8	0.012	8
J31V	441	0.286	11	0.169	12	0.085	13	0.069	18	0.036	13	0.012	7
J32V	441	0.294	12	0.125	6	0.056	9	0.044	9	0.024	10	0.016	10
K30V	1547	0.423	18	0.290	22	0.198	23	0.052	12	0.040	17	0.020	11
K31V	941	0.246	9	0.145	8	0.117	18	0.052	11	0.040	16	0.028	16
K32V	757	0.278	10	0.254	21	0.185	22	0.161	24	0.113	25	0.077	27
K33V	779	0.319	15	0.246	18	0.181	21	0.149	23	0.109	23	0.077	26
L30V	526	0.819	26	0.613	26	0.423	26	0.339	30	0.238	31	0.194	31
M30V	934	0.060	3	0.032	1	0.024	3	0.016	3	0.004	2	0.004	3
M31V	934	0.056	2	0.036	3	0.020	2	0.016	2	0.004	1	0.004	2
M32V	934	0.056	1	0.036	2	0.020	1	0.008	1	0.008	3	0.004	1
N30V	608	0.524	20	0.242	17	0.101	16	0.065	17	0.036	14	0.020	13
N31V	608	0.621	23	0.246	19	0.097	15	0.052	13	0.028	11	0.016	9
N32V	608	0.641	24	0.185	15	0.060	10	0.040	8	0.024	9	0.020	12
N33V	608	0.681	25	0.181	14	0.065	11	0.040	6	0.024	7	0.024	14
Q30V	501	1.000	35	0.988	34	0.952	35	0.907	35	0.819	35	0.714	35
Q31V	501	1.000	36	1.000	36	0.972	36	0.952	36	0.859	36	0.742	36
R30V	1472	0.565	21	0.399	25	0.226	25	0.077	21	0.032	12	0.024	15

Table 6: For the DATASET U: PASSENGER GATE *installation, with 480 subjects enrolled with a frontal still, the values are identification-mode FNIR(T) for each algorithm at three different decision thresholds corresponding to false positive counts of 1, 10, 100, and investigation-mode FNIR(R) for ranks 1, 5, 20. Each value is accompanied by an integer ranking across all algorithms. The shaded columns indicates the most important metric to watchlist applications. The green shaded cells indicates the most accurate algorithm. Caution: The last column give optimistically low error rates per the arguments of section 4.4.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

N=4800	NUM ACTORS 248		NUM FEEDS 1		NUM CLIPS 248		NUM FRAMES 11012		NUM MINUTES 18.4	
	DETECTIONS	THRESHOLD BASED AUTO WATCHLISTS					RANK BASED FORENSIC CASES			
ALG	NUM	FNIR(T), FP(T)=1		FNIR(T), FP(T)=10		FNIR(T), FP(T)=100		FNIR(R=1, T=0)		FNIR(R=5, T=0)		FNIR(R=20, T=0)	
A30V	1345	0.948	28	0.839	29	0.698	29	0.456	27	0.274	26	0.173	28
A31V	1345	0.879	27	0.774	27	0.637	27	0.448	26	0.278	27	0.173	27
B30V	714	0.968	29	0.782	28	0.641	28	0.460	28	0.323	30	0.222	30
C30V	2364	0.968	31	0.907	31	0.778	30	0.516	31	0.395	32	0.331	33
C31V	2661	0.968	30	0.907	30	0.782	31	0.512	30	0.399	33	0.331	32
D30V	2004	1.000	32	0.964	32	0.831	32	0.560	33	0.294	28	0.133	24
D31V	638	1.000	36	0.988	33	0.855	33	0.552	32	0.310	29	0.210	29
E30V	743	0.649	24	0.419	22	0.250	19	0.141	20	0.093	20	0.081	22
E31V	743	0.484	16	0.319	18	0.165	17	0.093	12	0.077	18	0.052	19
F30V	1502	1.000	35	0.996	34	0.968	34	0.952	34	0.895	34	0.782	34
G30V	493	0.577	21	0.411	20	0.274	21	0.206	23	0.169	23	0.125	23
G31V	691	0.294	4	0.202	8	0.097	7	0.077	10	0.060	16	0.052	20
G32V	691	0.379	8	0.194	4	0.109	9	0.077	8	0.056	11	0.052	18
H30V	668	0.327	7	0.194	7	0.089	6	0.077	9	0.040	7	0.024	6
H31V	668	0.327	5	0.194	5	0.089	4	0.073	6	0.040	4	0.024	4
H32V	668	0.327	6	0.194	6	0.089	5	0.073	7	0.040	6	0.024	5
I30V	1382	0.528	19	0.294	13	0.145	15	0.097	13	0.060	15	0.040	12
I31V	1382	0.472	14	0.298	14	0.181	18	0.117	16	0.056	14	0.044	14
J30V	441	0.403	9	0.319	17	0.149	16	0.121	17	0.056	13	0.032	8
J31V	441	0.407	10	0.302	15	0.137	12	0.125	18	0.056	12	0.036	11
J32V	441	0.435	11	0.258	11	0.129	11	0.105	15	0.044	9	0.032	10
K30V	1547	0.605	22	0.480	24	0.383	25	0.194	22	0.113	22	0.060	21
K31V	941	0.452	12	0.355	19	0.270	20	0.149	21	0.097	21	0.048	17
K32V	757	0.496	18	0.448	23	0.323	24	0.222	25	0.177	25	0.145	26
K33V	779	0.492	17	0.419	21	0.323	23	0.214	24	0.169	24	0.141	25
L30V	526	0.839	26	0.718	26	0.560	26	0.484	29	0.351	31	0.278	31
M30V	934	0.085	2	0.065	3	0.032	1	0.024	3	0.008	3	0.008	3
M31V	934	0.081	1	0.065	2	0.040	3	0.020	2	0.008	2	0.008	2
M32V	934	0.089	3	0.048	1	0.036	2	0.016	1	0.008	1	0.008	1
N30V	608	0.476	15	0.266	12	0.141	14	0.097	14	0.065	17	0.044	16
N31V	608	0.641	23	0.302	16	0.137	13	0.081	11	0.052	10	0.032	9
N32V	608	0.452	13	0.254	9	0.109	10	0.060	5	0.040	8	0.024	7
N33V	608	0.540	20	0.258	10	0.105	8	0.060	4	0.040	5	0.040	13
Q30V	501	1.000	33	1.000	35	0.988	36	0.972	35	0.923	35	0.891	35
Q31V	501	1.000	34	1.000	36	0.984	35	0.980	36	0.956	36	0.935	36
R30V	1472	0.758	25	0.496	25	0.319	22	0.125	19	0.077	19	0.044	15

Table 7: For the DATASET U: PASSENGER GATE *installation, with 4800 subjects enrolled with a frontal still, the values are identification-mode FNIR(T) for each algorithm at three different decision thresholds corresponding to false positive counts of 1, 10, 100, and investigation-mode FNIR(R) for ranks 1, 5, 20. Each value is accompanied by an integer ranking across all algorithms. The shaded columns indicates the most important metric to watchlist applications. The green shaded cells indicates the most accurate algorithm. Caution: The last column give optimistically low error rates per the arguments of section 4.4.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

N=48000	NUM ACTORS 248		NUM FEEDS 1		NUM CLIPS 248		NUM FRAMES 11012		NUM MINUTES 18.4	
	DETECTIONS	THRESHOLD BASED AUTO WATCHLISTS					RANK BASED FORENSIC CASES			
ALG	NUM	FNIR(T), FP(T)=1		FNIR(T), FP(T)=10		FNIR(T), FP(T)=100		FNIR(R=1, T=0)		FNIR(R=5, T=0)		FNIR(R=20, T=0)	
A30V	1345	0.992	31	0.923	28	0.819	28	0.718	31	0.536	28	0.423	28
A31V	1345	0.968	28	0.927	29	0.823	29	0.718	30	0.528	27	0.423	27
B30V	714	0.968	27	0.915	27	0.778	27	0.645	27	0.548	29	0.435	30
C30V	2364	0.972	30	0.956	31	0.871	30	0.677	29	0.577	31	0.456	33
C31V	2661	0.972	29	0.944	30	0.883	31	0.665	28	0.560	30	0.456	32
D30V	2004	1.000	32	1.000	33	0.964	32	0.786	32	0.605	32	0.355	26
D31V	638	1.000	36	1.000	36	0.988	33	0.790	33	0.637	33	0.440	31
E30V	743	0.847	24	0.569	22	0.391	20	0.246	18	0.149	20	0.117	20
E31V	743	0.770	23	0.444	14	0.270	13	0.153	11	0.109	13	0.081	16
F30V	1502	1.000	35	0.996	32	0.996	35	0.996	36	0.980	35	0.927	34
G30V	493	0.714	20	0.556	21	0.391	19	0.310	25	0.210	23	0.177	23
G31V	691	0.319	4	0.238	5	0.149	5	0.133	8	0.089	10	0.065	11
G32V	691	0.448	5	0.238	4	0.141	4	0.113	4	0.073	5	0.060	10
H30V	668	0.460	9	0.294	6	0.202	8	0.133	7	0.077	8	0.052	8
H31V	668	0.460	7	0.298	7	0.202	6	0.137	9	0.077	6	0.052	6
H32V	668	0.460	8	0.298	8	0.202	7	0.137	10	0.077	7	0.052	7
I30V	1382	0.718	21	0.480	16	0.274	14	0.181	14	0.109	12	0.081	14
I31V	1382	0.911	25	0.722	25	0.444	23	0.206	15	0.137	16	0.089	17
J30V	441	0.661	18	0.488	19	0.327	17	0.270	21	0.141	19	0.081	15
J31V	441	0.681	19	0.476	15	0.323	16	0.266	20	0.141	18	0.093	18
J32V	441	0.637	17	0.431	12	0.258	11	0.230	16	0.121	15	0.073	13
K30V	1547	0.617	16	0.581	23	0.484	25	0.254	19	0.190	22	0.145	22
K31V	941	0.460	6	0.444	13	0.347	18	0.234	17	0.141	17	0.133	21
K32V	757	0.560	12	0.524	20	0.427	22	0.294	24	0.242	25	0.194	25
K33V	779	0.560	11	0.484	17	0.411	21	0.294	23	0.238	24	0.194	24
L30V	526	0.956	26	0.839	26	0.669	26	0.645	26	0.516	26	0.435	29
M30V	934	0.250	2	0.097	3	0.069	3	0.044	2	0.020	2	0.008	2
M31V	934	0.262	3	0.093	2	0.065	2	0.048	3	0.016	1	0.008	1
M32V	934	0.177	1	0.085	1	0.052	1	0.044	1	0.024	3	0.012	3
N30V	608	0.573	14	0.427	11	0.266	12	0.161	12	0.113	14	0.073	12
N31V	608	0.573	13	0.484	18	0.274	15	0.169	13	0.097	11	0.056	9
N32V	608	0.593	15	0.371	10	0.246	10	0.129	6	0.077	9	0.040	4
N33V	608	0.516	10	0.331	9	0.210	9	0.117	5	0.056	4	0.048	5
Q30V	501	1.000	33	1.000	34	0.996	34	0.984	34	0.976	34	0.952	35
Q31V	501	1.000	34	1.000	35	1.000	36	0.996	35	0.988	36	0.980	36
R30V	1472	0.758	22	0.657	24	0.472	24	0.278	22	0.157	21	0.105	19

Table 8: For the DATASET U: PASSENGER GATE *installation, with 48000 subjects enrolled with a frontal still, the values are identification-mode FNIR(T) for each algorithm at three different decision thresholds corresponding to false positive counts of 1, 10, 100, and investigation-mode FNIR(R) for ranks 1, 5, 20. Each value is accompanied by an integer ranking across all algorithms. The shaded columns indicates the most important metric to watchlist applications. The green shaded cells indicates the most accurate algorithm. Caution: The last column give optimistically low error rates per the arguments of section 4.4.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

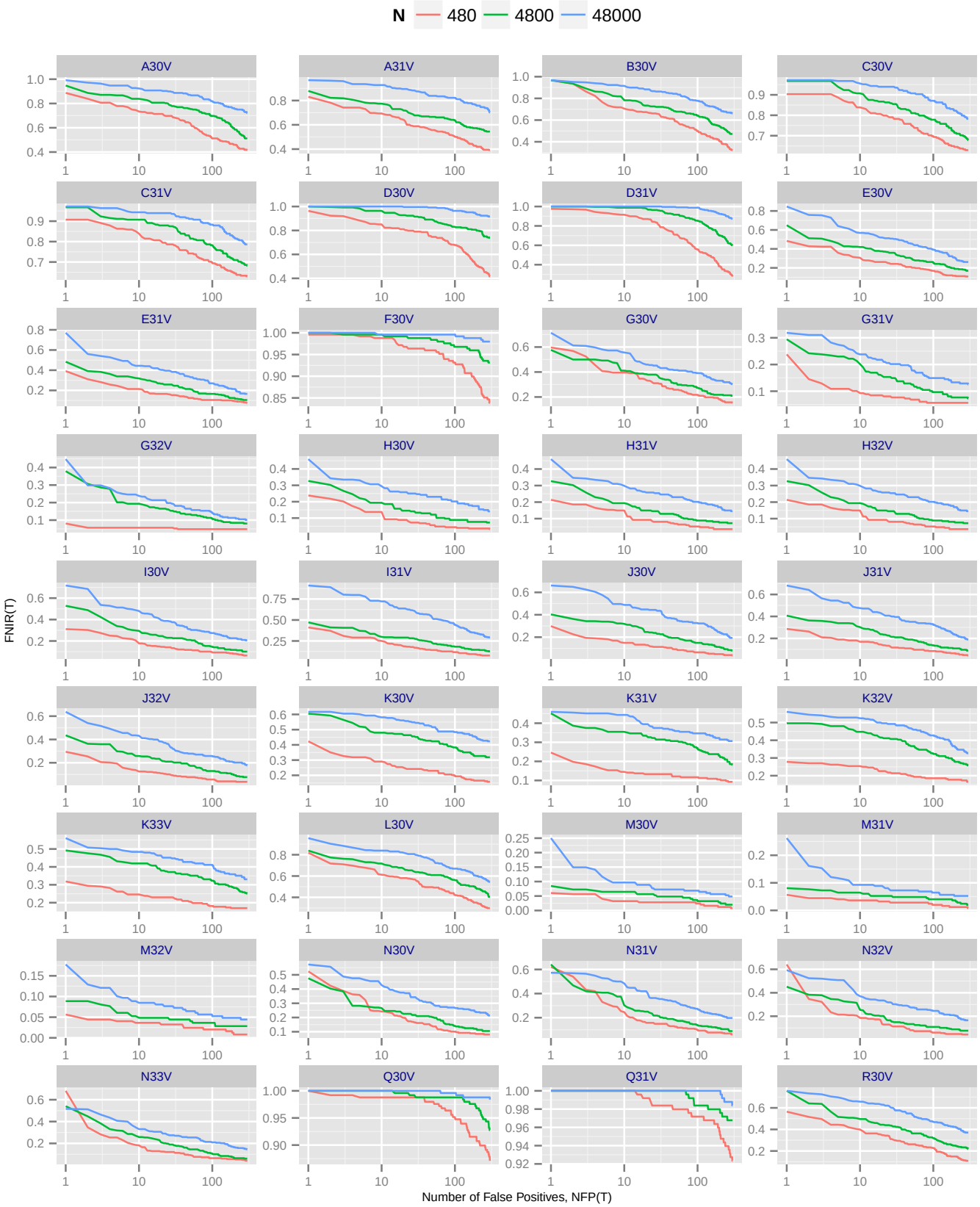


Figure 10: For DATASET U: PASSENGER GATE *, the panels show* $FNIR(N, L, T)$ *vs.* $NFP(T)$ *for each algorithm at three different gallery sizes. Each trace corresponds to an error tradeoff achieved by sweeping the threshold from low values, at right, to high values, at left.* **Note the different vertical scales.**

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

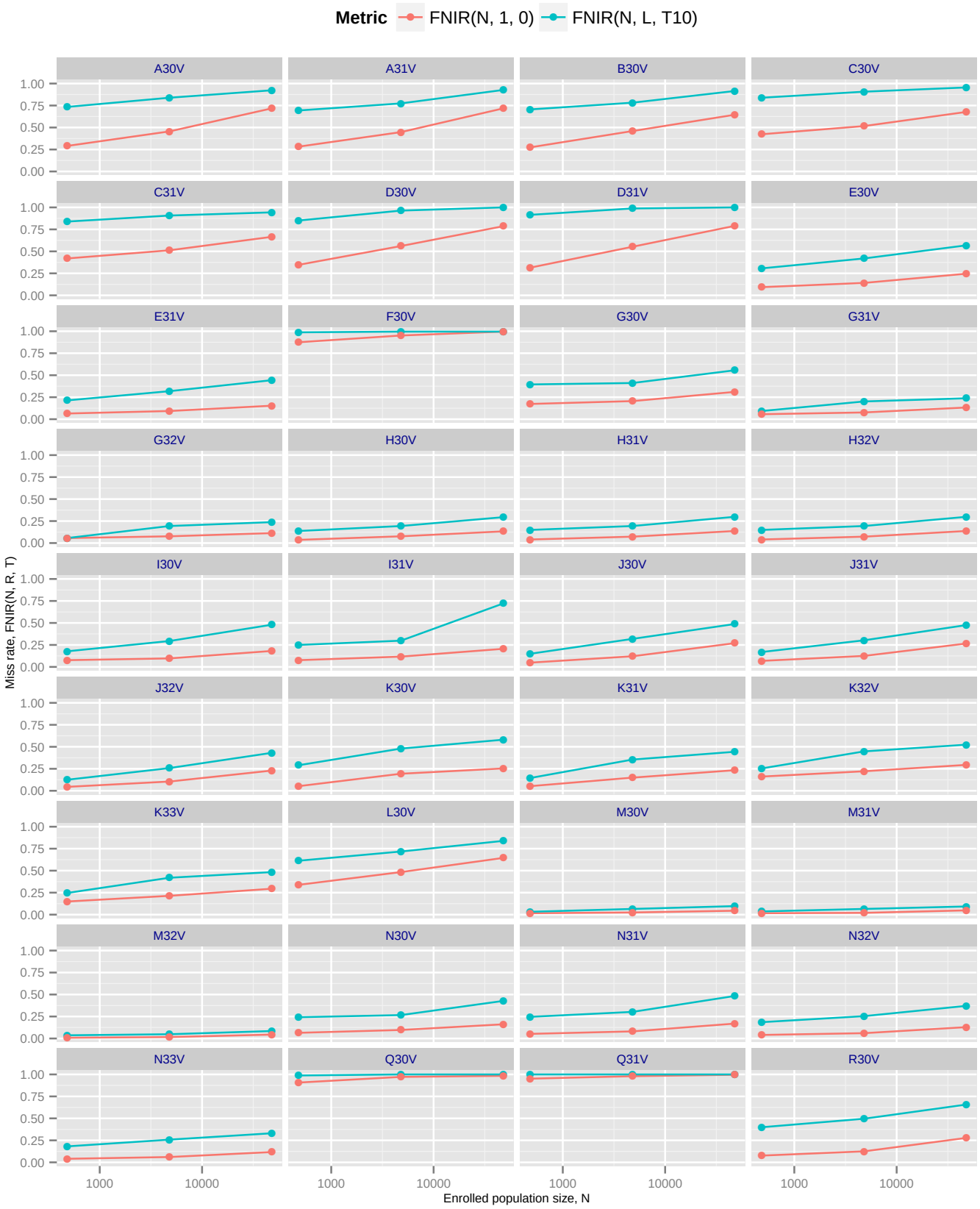


Figure 11: For DATASET U: PASSENGER GATE *, the panels show both identification mode high-threshold miss rates, FNIR(N, L, T), and investigation mode zero-threshold miss rates, FNIR$(N, 1, 0)$, as a function of enrolled gallery size, N. The threshold is set for each gallery size to elicit ten false positives over all searches for all 248 video clips.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

DATASET U	VIDEO SEARCH						STILL IMAGE	
	NUMBER OF	MEAN TRACK	MEAN TRACK	MEAN	MEAN	MEAN	IOD (PIXELS)	
ALGORITHM	TRACKS	LENGTH (FRAMES)	EXTENT (FRAMES)	MIN IOD (PX)	MEAN IOD (PX)	MAX IOD (PX)	MEAN	STD. DEV
A30V	1345	11	11	45	49	54	116	20
A31V	1345	11	11	45	49	54	116	20
B30V	714	23	26	45	58	97	118	21
C30V	2364	4	4	60	62	64	119	21
C31V	2661	4	4	55	57	59	119	21
D30V	2004	5	5	80	82	84	118	22
D31V	638	20	20	54	65	75	118	22
E30V	743	26	29	38	50	62	117	20
E31V	743	26	29	38	50	62	117	20
F30V	1502	5	7	54	62	71	116	20
G30V	487	14	20	57	67	76	118	21
G31V	662	22	27	42	54	63	117	21
G32V	662	22	27	42	54	63	117	21
H30V	668	19	21	53	66	77	119	21
H31V	668	19	21	53	66	77	119	21
H32V	668	19	21	53	66	77	119	21
I30V	1382	17	17	30	38	45	120	21
I31V	1382	17	17	30	38	45	120	21
J30V	434	19	23	66	82	95	121	21
J31V	434	19	23	66	82	95	121	21
J32V	434	19	23	66	82	95	121	21
K30V	1547	13	16	32	38	43	119	21
K31V	941	23	30	31	42	52	119	21
K32V	757	17	24	37	44	51	119	21
K33V	779	18	25	35	43	49	119	21
L30V	526	11	12	50	75	93	117	20
M30V	934	19	20	32	40	46	104	19
M31V	934	19	20	32	40	46	104	19
M32V	934	19	20	32	40	46	104	19
N30V	608	24	27	44	58	71	118	21
N31V	608	24	27	44	58	71	118	21
N32V	608	24	27	44	58	71	118	21
N33V	608	24	27	44	58	71	118	21
Q30V	501	12	13	42	48	53	116	20
Q31V	501	12	13	42	48	53	116	20
R30V	1472	1	3	77	78	80	116	20

Table 9: For DATASET U: PASSENGER GATE *and each video processing algorithm the table shows: a) the number of reported tracks; b) the mean number of frames reported within those tracks; c) the mean extent (first minus last frames indices plus one); d) the mean over all tracks of of the minimum interocular distance (IOD) reported; e) the mean of the mean IOD; f) the mean of the maximum IOD; g) the enrollment still image mean IOD; and h) its standard deviation. For some algorithms (F, J, L, R) the tracks don't include all consecutive frames, so the extent of the track can exceed the number of frames in it. For the D, F, and L algorithms, the reported still-image eye coordinates are erroneous.*

Still portrait IOD: The algorithms have good consensus on the IOD of faces in the enrollment still photographs. The M3xV algorithms, however, report systematically lower IOD values, perhaps because they are not following the ISO/IEC 19794-5 definition of eye centers.

5.2.4 Computational cost

In FIVE, algorithms were tasked with detecting and tracking individuals through video clips, and then producing a consolidated template from each detected track. This activity was placed behind a single function call invocation. This placed the responsibility for all image processing, pattern recognition, and feature extraction with the algorithm developer. It absolved the test laboratory (NIST) of making decisions such as about which frames to use, or on how to fuse scores from frame based matching.

Duration: Thus given an input video, the algorithm outputs zero or more templates. Each template has a variable size, which NIST logs. In some cases, zero is the correct response. The size of the templates produced has some operational significance as it can affect network bandwidth and storage requirements, and processing time.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

The plots of Figure 12 show the dependence on the number of input frames, and on the number of faces tracked. The durations are strongly algorithm specific and vary over at least an order of magnitude. In high volume applications like continuous video surveillance, this will have hardware cost implications. The J algorithms operate in fewer than 5 seconds, while the durations for the A algorithms exceeds 50 seconds on average. Each plot includes a text equation which models processing duration as a linear combination of the video length and the number of face tracks found. Its coefficients give the marginal duration increase associated with an additional frame or person. There is often considerable scatter arising, in part, because the algorithms vary in the minimum size of the face that they will detect and, thereby, the length of the track. This is evident in the track statistics of Table 9.

Template size: The size of a template extracted from video imagery is expected to depend on the amount of imagery that was available to algorithm, and on the mechanisms used to extract recognizable features from it. While historically, many researchers have extracted features separately from individual frames, there have long been attempts to integrate information over a track [28]. Indeed, research programs have identified goals to produce representations whose size is independent of the amount of available imagery [12]. This reflects the importance of bounded template size.

Figure 13 shows template size as a function of the length of the track the algorithm reports. The sizes and dependence are highly algorithm-specific. Some algorithms produce fixed template sizes (G3xV, J3xV, K32V, Q3xV). The implementations may be selecting a best-frame from the video, or integrating information temporally - in a black-box test, we do not know. Other algorithms tend to produce larger templates with a dependence on track length. Algorithms A3xV, B30V, D3xV, and R30V have a nearly perfect linear relationship. Other algorithms - K31V, K33V, M3xV - give the linear relationship but impose a hard limit on the template size above which additional features are not added even though the track is longer.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

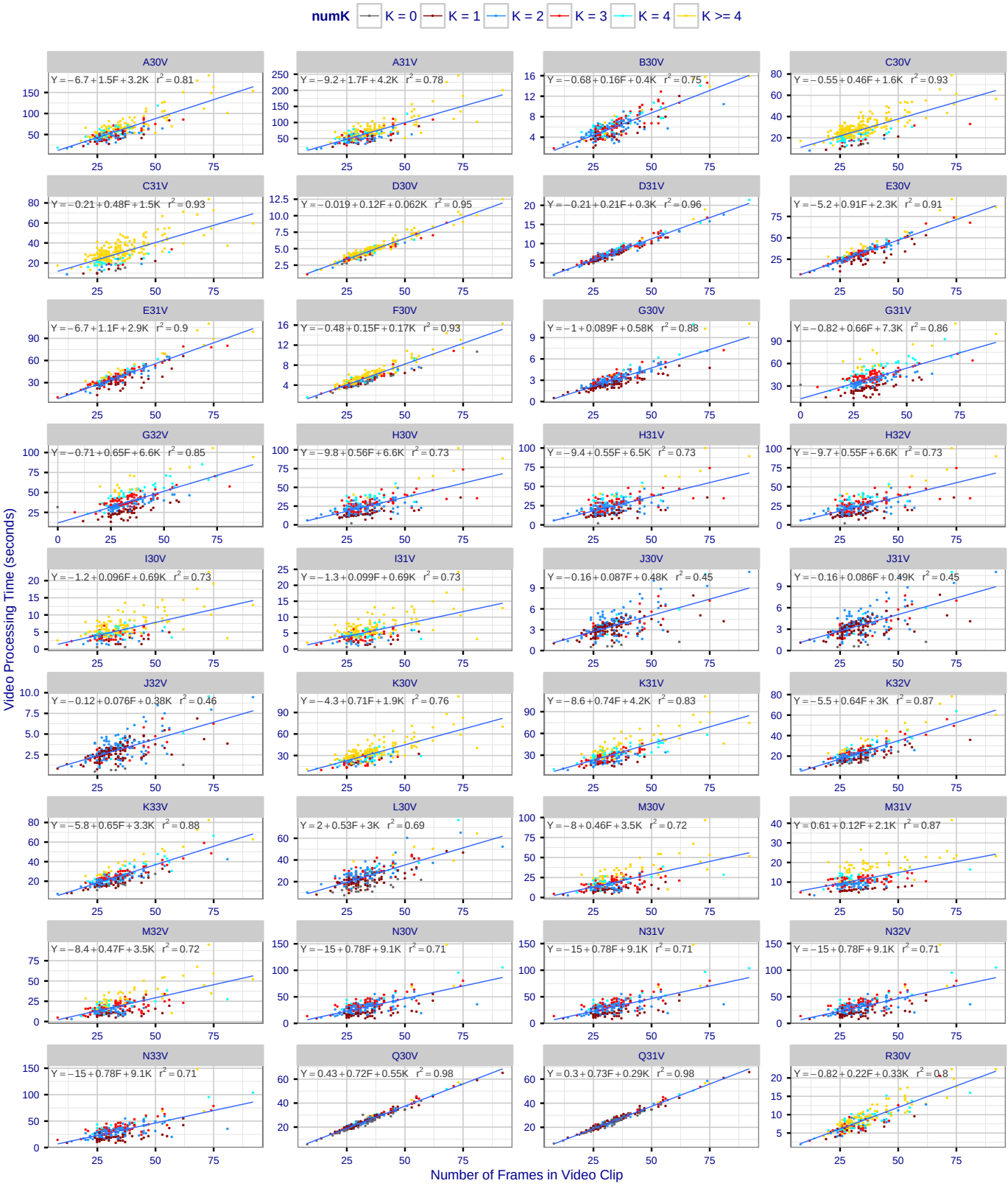


Figure 12: For DATASET U: PASSENGER GATE *, the plots show the duration of the video processing as a function of the number of frames passed to the algorithm. The duration varies also with the number of face tracks found in the clip, K, which is color coded. The regression formula is one of several simple models. Its coefficients can be interpreted as the marginal cost of adding one additional frame, or face, to the video.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

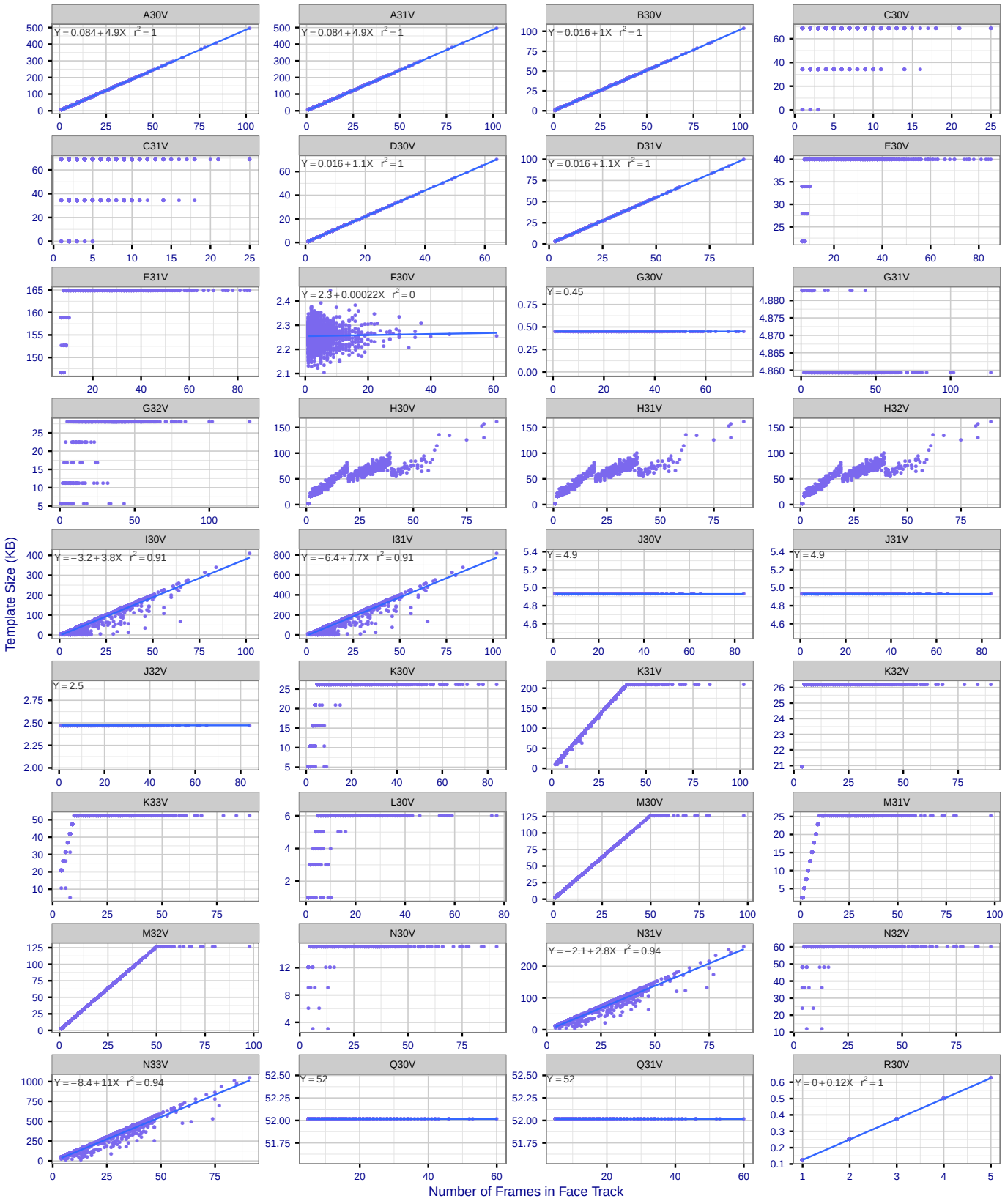


Figure 13: For DATASET U: PASSENGER GATE *, the plots show the template size for features extracted from video face tracks as a function of the length of the track reported by the algorithm, in frames. Note the R30V algorithm reports very few frames, and these are irregularly spaced throughout the clip. A best fit linear model is also plotted, when appropriate.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

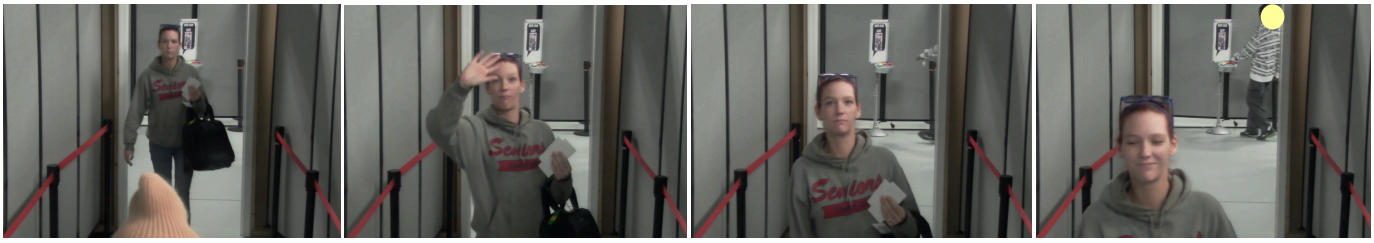


Figure 14: Example images from the ceiling mounted camera for the free movement scenarios from the DATASET J: PASSENGER LOADING BRIDGE *dataset. **The images in this table are from the subject S1115 in the DHS / S&T provided AEER dataset. The subject gave written opt-in permission to allow public release of all imagery. Where consent from individuals in the background was not obtained, their faces were masked (yellow circle).*

(a) Right Wall Mounted (b) Ceiling Mounted (c) Left Wall Mounted

Figure 15: Fields of view for the three cameras used in DATASET J: PASSENGER LOADING BRIDGE *. **The images in this table are from the DHS / S&T provided AEER dataset.*

5.3 DATASET J: PASSENGER LOADING BRIDGE

Overview: This dataset consists of videos of subjects walking along a purpose-built simulated passenger loading bridge (PLB). The PLB was equipped with three cameras, one ceiling mounted, with two more mounted on the walls symmetrically below it, in a vertical plane. All cameras were angled to observe subjects walking toward them - see Figure 14 - their orientation was selected to be favorable for face recognition with respect to the angle of the optical axis and the face normal - see Figure 15. Key imaging properties for this dataset are summarized in Table 10.

Property	Value
Camera	Vaddio PowerView (PTZ)
Camera mounting	Attached to display observed by subject
Camera height	2.44 meters (ceiling); and 1.83 meters (wall mounted cameras)
Camera orientation	Both elevation from ceiling and azimuth to sidewall below 15°
Range to subject	1m - 5m
Frame rate	30 sec^{-1}
Width	1920
Height	1080
Chroma sampling	YUV420
Nominal bitrate	
Codec	AVC H264

Table 10: Key imaging properties for DATASET J: PASSENGER LOADING BRIDGE

Experimental Design: The video clips show volunteer recruits acting as passengers walking along a PLB simulating an aircraft boarding process. There are 48 clips in total, with 16 from each of the three cameras. The 16 videos represent the appearance of 8 groups of people on two occasions, about 30 minutes apart. Each group has between 42 and 51 volunteer subjects. The groups were exposed to different experimental manipulations. To examine the effect of walking on recognition accuracy, an articial bottleneck was applied at the exit of the PLB. The result is that four groups of people walked freely past the cameras without stopping, and another four groups were "bottlenecked" mostly standing in a

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Quantity	Value or description
Mode	Video search to still enrollment
Number of actors	354, with subsets in different experiments
Number of non-actors	0
Number of cameras	3, one ceiling, two wall mounted, symmetrically
Video duration with actors	485 minutes
Video duration no actors	0
Subject motion	Usually single file toward and below the camera. Walking or stopped in queue.
Number of clips	48 = 2 (walk, queue) x 3 cameras x 2 attractor content x 2 repeats x 2 attractor on/off
Clip duration	Varies from min 7mins 15 seconds to 12 mins 48 seconds, mean 10.1 minutes
Number of enrolled subjects	480
Number of enrolled stills	1 per subject
Properties of enrolled stills	Frontal, close ICAO compliance; Mean IOD 106 pixels
FNIR estimation	Actors present video vs. enrolled gallery
FPIR estimation	Actors present video vs. separate non-actor gallery
Candidate list length	20
Number of persons in FOV	[0,3] free movement or [4,7] queued, approx.
Video ground truth	Style A: See Figure 6

Table 11: Key experimental design the DATASET J: PASSENGER LOADING BRIDGE *results.*

queue and occasionally moving forward. This factor represents the situation that often occurs when boarding twin-aisle vs. single-aisle aircraft. People in the latter scenario are typically in the camera's field of view for a longer period, but their behavior is less consistent - see Figure 14 - subjects would often look down or away from the camera. In addition, some faces are temporarily occluded by the people standing in front of them.

Each of the groups was also exposed, or not, to a video display attractor installed just above the ceiling-mounted camera. For four groups the attractor was off. For four others, the attractor was switched on. Two of those saw a "live agent" which was a rotating sequence of around six different people greeting the subjects and included both audio and video. The other two groups saw a "digital mirror" from the ceiling mounted camera showing their own live video fed back to them. This encouraged subjects to misbehave by exagerrating expressions - see Figure 14 - something not observed for the "live agent" content.

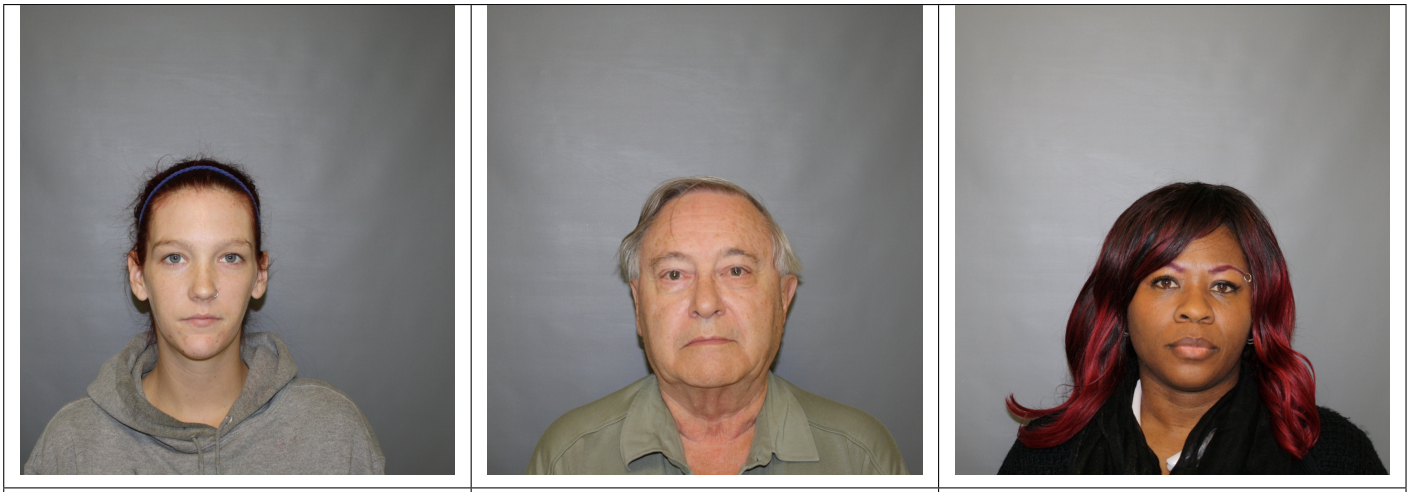
Mated scores are computed over long video clips from each of three cameras. The video clips of freely walking subjects last about from 8 to 10 minutes. The video clips of queued subjects last from 10 to 12 minutes. The clips were passed to the algorithms in their entirety. These are searched against a gallery of $N = 480$ images, one from each of the 354 actors, and an additional 126 from a disjoint background population. **Enrolled still images:** Enrollment images were collected cooperatively using a consumer-grade SLR - see Figure 16. These are in good conformance to the ISO/IEC 19794-5 full frontal image type, aside from some additional torso and background. Exactly one image is enrolled per subject.

Nonmated scores are computed by comparing templates generated from video clips against the *global nonmated enrollment dataset*. Thresholds are generally computed on a per-camera basis, i.e. using scores only from that camera's searches.

Key experimental design details are summarized in Table 11.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			



**Subject S1155 (Perm Granted)	**Subject S2880 (Perm Granted)	**Subject S1848 (Perm Granted)

Figure 16: DATASET J: PASSENGER LOADING BRIDGE . *Examples of enrollment images collected with Canon SLR camera. **The face images in this figure are from the DHS / S&T provided AEER dataset. The included subjects consented to release their images in public reports.*

Results: Figure 17 shows FNIR(N, L, T) identification-mode miss rates in bar form for convenient visual comparison. Similarly, Figure 18 shows FNIR$(N, 1, 0)$ investigation-mode miss rates. The decision threshold (Fig. 17) is set to elicit 10 false positives over all 354 subjects appearing in all clips from the respective camera (so both with and without the attractor enabled). While, this approximately corresponds to fewer than one false positive for every thirty four subjects walking down the passenger loading bridge, each subject walked along the PLB twice, and each algorithm might detect and search several templates from each person - see Figure 3.

Notable observations are:

Attractors are effective: Considering just the ceiling mounted camera, alongside which the TV display attractor is mounted, the lowest error rates are observed when the attractor as switched on, and when the subjects were waiting in line.

Ceiling mounted camera is superior: The overhead camera gives the best accuracy, outperforming the cameras mounted on the side walls of the passenger loading bridge. This comparison must be made with the attractor off (the first two columns of Figure 17) because the side-mounted cameras were not equipped with an attractor. For the five most accurate algorithm developers (M, H, J, I, N), the identification miss rates for the ceiling mounted camera are as much as half of those for the side mounted cameras. However this effect applies to subjects detained in a queue. It is much reduced for freely walking subjects. This latter result supports the assertion that elevation angle is important to face recognition, of a similar magnitude to the yaw angles inherent in the use of the wall-mounted cameras. The G30V algorithm gives better accuracy with wall-mounted cameras than with ceiling mounted when the attractor is disabled.

Uncertainty in error rates: While the lowest FNIR(T) value for a single camera is below 4% (algorithm M32V, ceiling camera, attractor enabled, queued), there is considerable uncertainty associated with this measurement, stemming from the small sample size - here just 88 subjects. This small population size imparts some uncertainty

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

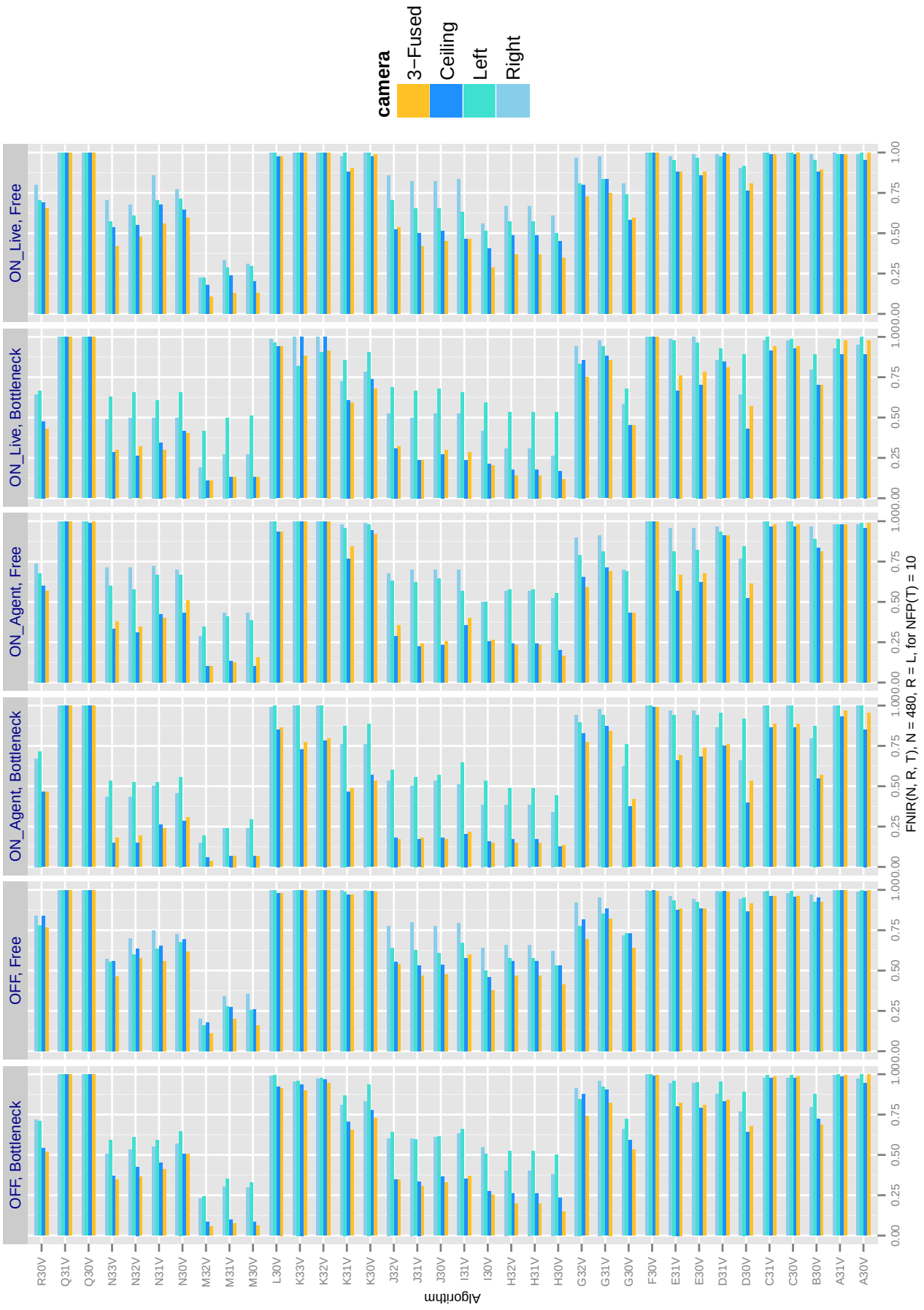


Figure 17: For each algorithm, the bars plot FNIR(N, L, T) for the boarding of 354 subjects before the cameras of DATASET J: PASSENGER LOADING BRIDGE *. Four bar charts appear for each algorithm, corresponding to the three cameras and then a max-score fusion of all three. The threshold T is set to give NFP = 10 false positives over all searches from the respective camera, except in the case of fusion, where it is computed over all searches from all three cameras. Fusion usually reduces FNIR, but can elevate it because the threshold increases to maintain NFP = 10. The six panels give accuracy for two covariates: a) Whether the queue was backed up (bottleneck) or walking freely; and b) TV display attractor (co-mounted with the ceiling mounted camera) was switched on, and then whether it was showing a generic guidance video, or a "digital mirror" video of the travelers themselves.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

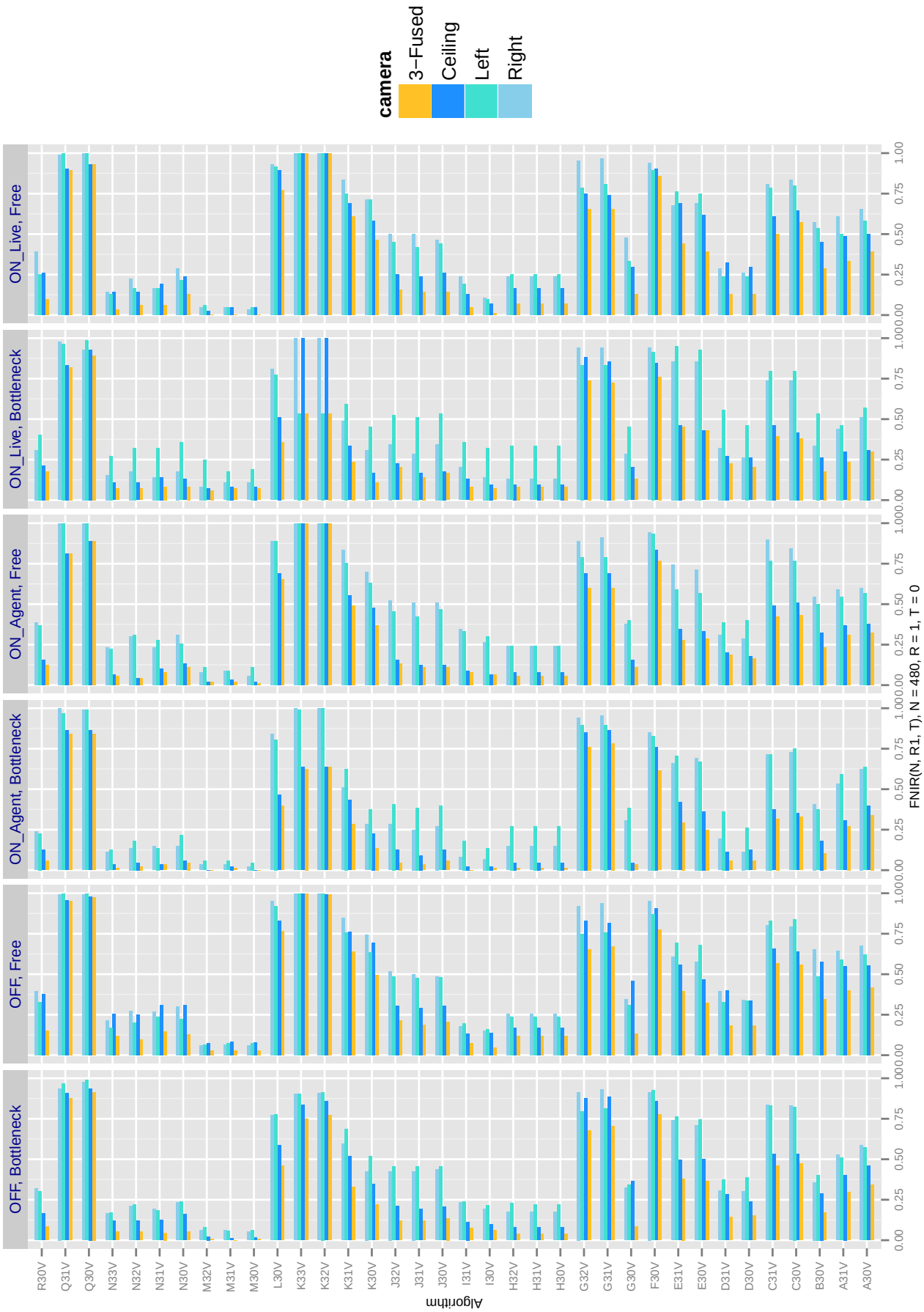


Figure 18: For each algorithm, the bars plot FNIR for the boarding of 354 subjects before the cameras of DATASET J: PASSENGER LOADING BRIDGE *. The threshold T is set to zero, and R = 1 in all cases. Four bar charts appear for each algorithm, corresponding to the three cameras and then fusion implemented by taking the minimum of the three ranks. This fusion, which ignores scores completely, is meaningful only if human reviewers adjudicate the three candidate lists in parallel rather than sequentially i.e. they all at all rank 1 candidates, then all rank 2, and so on. The six panels give accuracy for two covariates: a) Whether the queue was backed up (bottleneck) or walking freely; and b) TV display attractor (co-mounted with the ceiling mounted camera) was switched on, and then whether it was showing a generic guidance video, or a "digital mirror" video of the travelers themselves.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

on the results: Using a simple binomial assumption, this observed error rate only supports a claim that the error rate is below 10%.

Multicamera fusion: We model the use of multiple cameras by taking for each known actor the maximum of the scores produced by searches from the three cameras - this gives the best "hit". This is fusion over space. Identification error rates are often lower when score fusion is conducted across the three cameras. Fusion is sometimes counterproductive because our metric, FNIR(T), is computed at a threshold that is computed as the 10-th largest of $n_0 + n_1 + n_2$ impostor scores, coming from cameras $\{0, 1, 2\}$. This is done because a system operator would not want to process more false positive exceptions just because he had installed two additional cameras. Algorithms vary in this respect depending on whether the two side cameras, with more yaw and less pitch, give high impostor scores.

The gains are usually modest relative to using just the ceiling mounted camera alone. This is especially true when the attractor is in use, and when the queue is stalled. Thus, multiple cameras give most benefit when an attractor is absent. This may be important if cameras are installed covertly without the possibility to use an attractor.

For the best-rank fusion method (Fig. 18), the FNIR values reach zero indicating that the actors were all identified correctly by at least one of the three cameras. This applies to algorithms from developers M and I, with H and N almost there. Given limited sample size, we again cannot claim error rates below about 3.5% (via the rule of three). The use of best-rank fusion however is associated with increased labor costs, since for three cameras, three candidate lists will be produced (to first order) and the candidates will need to be interleaved and reviewed in rank-order.

Costs associated with fusion: Thus multiple cameras give most benefit when an attractor is absent, but their capital cost, and the costs associated with network transmission and computation, will increase linearly with the number of cameras. It is almost certainly less expensive therefore to deploy a capable attractor with eye catching and varied content, than it is to add additional cameras.

Caveats: Fielded accuracy will vary systematically from the numbers reported here. The equipment, illumination, and detailed installation details can have an effect. In addition the video here is being matched against still photographs collected on the same day. Such same-day matching is known to improve recognition accuracy. Note that some novel uses of face recognition do have a same-day concept of operations. One is to replace presentation of an airline boarding pass with one's face instead, using one-to-many recognition in a positive access control manner. This is only done after an initial identity check which includes a one-to-one verification of live imagery against an authoritative credential (e.g. passport).

5.3.1 Effect of reduced frame rate

The use of video data imparts a data size overhead. This arises because the pixel dimensions (e.g. 1920x1080) of the imagery are larger than face portraits (typically, 640x480) to support a useful field of view, and because video is typically collected at 24fps or 30fps. While video data is compressed using techniques that intelligently allow for interframe motion, the data rate is nevertheless large enough that a designer will need to make a dedicated computation of network bandwidth and latency needed to support operational goals.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

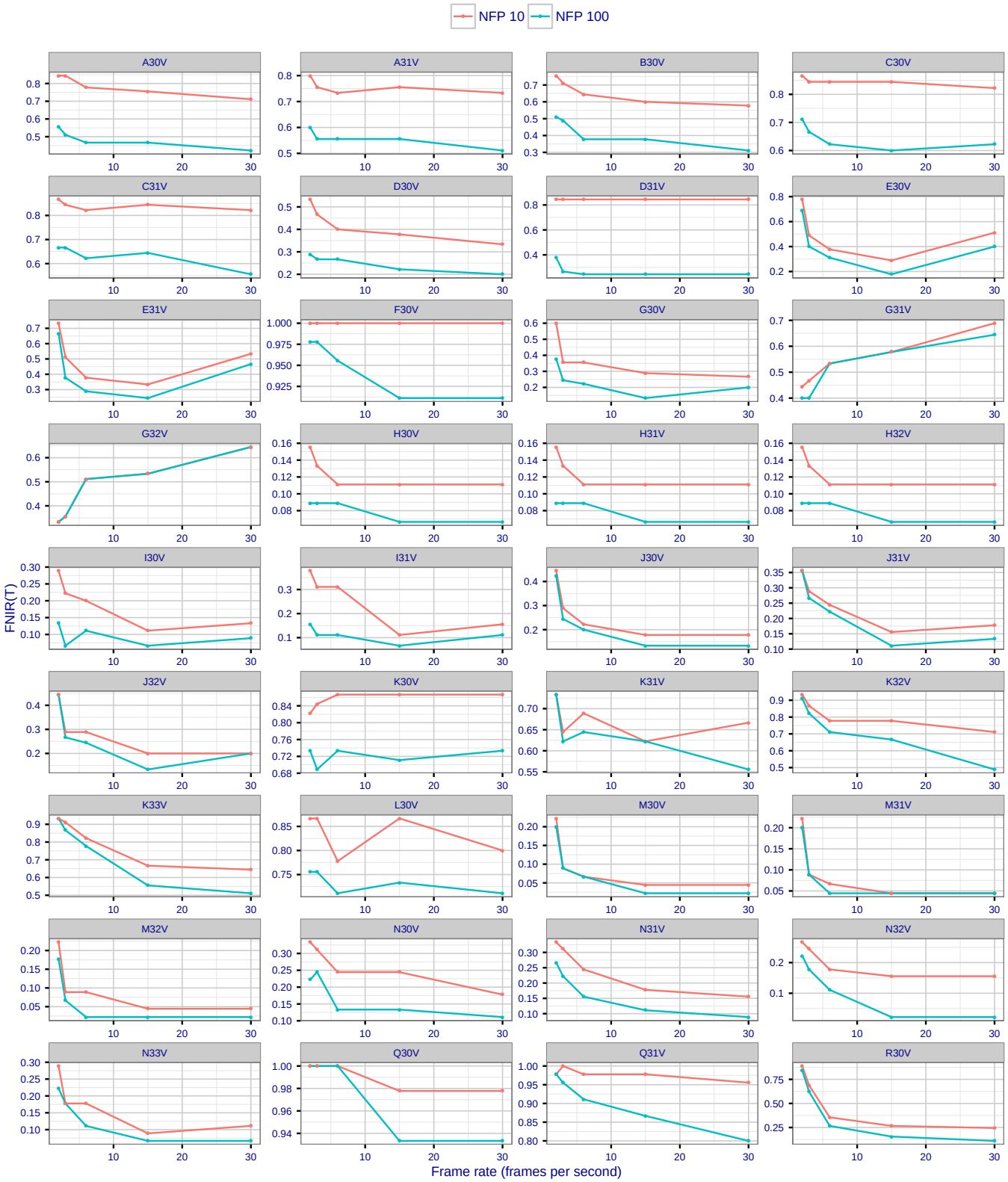
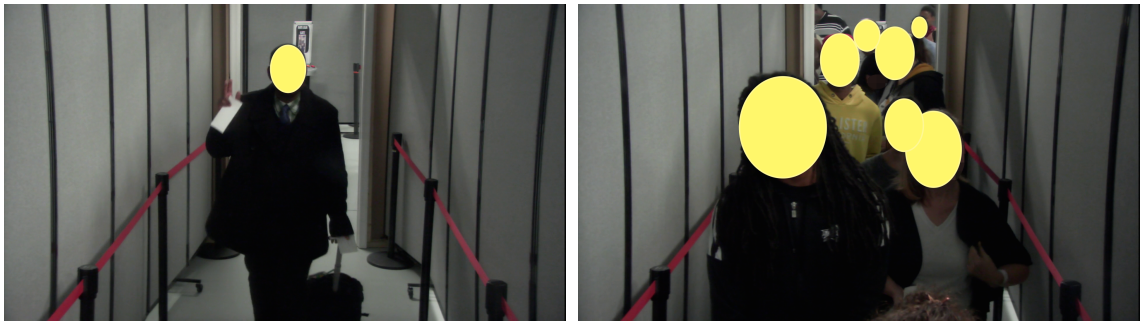


Figure 19: Using a subset of the DATASET J: PASSENGER LOADING BRIDGE *imagery, the panels show FNIR(N, L, T) against frame rate. The algorithms were given video sequences with progressively reduced temporal resolution. This was achieved by using every k-th frame as input, $k = \{1, 2, 5, 10, 15\}$, from an original frame rate of 30 fps. The recognition threshold, T, was set to achieve 10 and 100 false positives over identification searches into a gallery of size N = 480. Thresholds are set specifically for each frame rate. The camera is ceiling mounted. The subjects walk freely, and a video attractor was present.* **Note the different vertical scales.** *Note also this figure's use of lines is possibly erroneous as we did not measure accuracy at all possible frames rates, only those stated.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			



(a) Free Movement Scenario (b) Bottlenecked Scenario

Figure 20: Example frames for the Free Movement (left) and Bottlenecked (right) scenarios from the DATASET J: PASSENGER LOADING BRIDGE *dataset. The yellow ellipses are applied as these individuals did not consent for their faces to appear in the report.*

As video data is typically much larger than still imagery, the question of how to reduce the data size arises. This can be achieved in a number of ways. First, is to not transmit the video at all, and instead extract features in or near the camera. This "edge processing" mode of operations requires installation of some components of a face recognition system at, or near, the camera (rather than at a central server), and this in turn requires fielding of sufficient computation power also. It binds the camera to the algorithm and may make technology update more difficult. Second, is a hybrid solution where video of detected faces is transmitted. This requires face *detection* algorithms to be fielded. Third, is to reduce the bit-rate by using a different compression profile or (not exactly equivalently) by reducing the frame rate. This last aspect is analyzed here.

We presented results for DATASET J: PASSENGER LOADING BRIDGE at full frame rate above. We additionally searched the same video at reduced frame rates. This was done by passing only every k-th frame to the algorithms, with $k = \{15, 10, 5, 2, 1\}$, the last value representing the full 30 frames per second (fps). As algorithms use motion to detect, and track, and potentially integrate information over time, any reduction in frame rate can undermine accuracy.

Figure 19 shows the effect of reduced frame rate on FNIR(N, L, T), i.e. the proportion of actors not identified above a threshold T. The threshold T was set to give NFP$= \{10, 100\}$ false positives at each frame rate. The results are varied. Several algorithms - those from providers, B, D, H, I, M, N, Q, R - mostly give the expected behavior: better accuracy with more video data. However, some algorithms give entirely the opposite: G31V and G32V have substantially better accuracy with two frames per second than at higher frame rates. The G30V algorithm is usually superior but G32V is more effective at 2 frames per second. G31V and G32V produce larger templates more slowly. The algorithms from participants E, I, J mostly give best accuracy at 15 fps. Some other algorithms, from developers H and M, seem to operate at 15 fps naturally.

The data support a conclusion that operating at 15 fps is often lossless. A weakness of this study however is that we have not measured the attendant data size gains: While there are half as many frames, the compression algorithm may not realize such a reduction because interframe subject motion is larger at 15 fps than at 30 fps.

5.3.2 Face tracking behavior

A preliminary step in the identification process involves detecting and tracking the movement of a person across multiple frames. These *face tracks* are then processed into matchable templates. A single face track need not correspond to the

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

entire duration that a person is in the field of view of the camera. For example, if a person's face is only visible at sporadic moments (due to temporary occlusion or changes in head pose), the algorithm may decide to generate a separate face track for each continuous period over which the face is visible.

Figure 20 shows an example snapshot for two scenarios from DATASET J: PASSENGER LOADING BRIDGE . For the "Free Movement" scenario on the left, subjects walked unimpeded down the PLB, in most cases presenting their faces to the camera without interruption. For the "Bottlenecked" scenario on the right, a queue formed. People in the Bottlenecked scenario are typically in the field-of-view for a longer duration, but their behavior is less consistent. They often look down or away from the camera, and it is not uncommon for their faces to be temporarily occluded by the people waiting in front of them.

Figure 21 shows the number of face tracks detected for each algorithm and both scenarios. The number of detected face tracks varied widely from one algorithm to the next. Even for the free movement scenario, algorithm D30V reported about 13 times more face tracks than people who walked down the PLB. The algorithm is probably breaking each person's presentation into multuple face tracks, despite uninterrupted presentation of the face. One could argue that this gives the algorithm a greater number of opportunities to "hit" the person in the database, since we give an algorithm credit for a hit as long as it found the person at least once while he/she was walking down the aisle (more face tracks means more search templates means more opportunities to hit). On the other hand, it could lead to a greater number of false positives. Ideally, the algorithm would detect one face track per person walking down the PLB (although this behavior was not specifically requested in the API).

Algorithms G31V and G32V detect fewer faces in the bottlenecked scenario. Fewer, in fact, than there are people in the video, indicating either a failure to track certain individuals, or incorrect consolidation of several different people into a single face track.

Figure 22 supplements Figure 21 by showing the mean face track length (in seconds) compared to the number of face tracks found for each algorithm. This figure applies only to freely walking subjects. The notable results are:

As expected, there is an inverse relationship between face track length and number of face tracks.

Notably six of the more accurate algorithms are clustered right and below center. The majority of the more accurate algorithms (e.g. those from providers, H, I, J, M, N, G) track subjects for several seconds, and produce up to two times the number of tracks as there are people.

However for the long clip video over which these results were reported, the G31V and G32V algorithms report fewer tracks than there are people, which is fatal to FNIR.

Some algorithms (for example from providers A, C, D, F, L, R, Q) elect to track faces over short sub-second intervals but also report more tracks. Algorithm F30V finds many face tracks, but each face track spans an average of less than a fifth of a second. Since it typically takes a subject several seconds to walk down the PLB, the algorithm is likely breaking up the person's journey across several face tracks.

On the other hand, the K31V algorithm reports tracks much longer than the subjects are actually in view. The developer explored this tradespace with K30V producing many short tracks and K33V producing face tracks that span an average of 6.2 seconds, more in line with how long the person is actually in the field of view.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

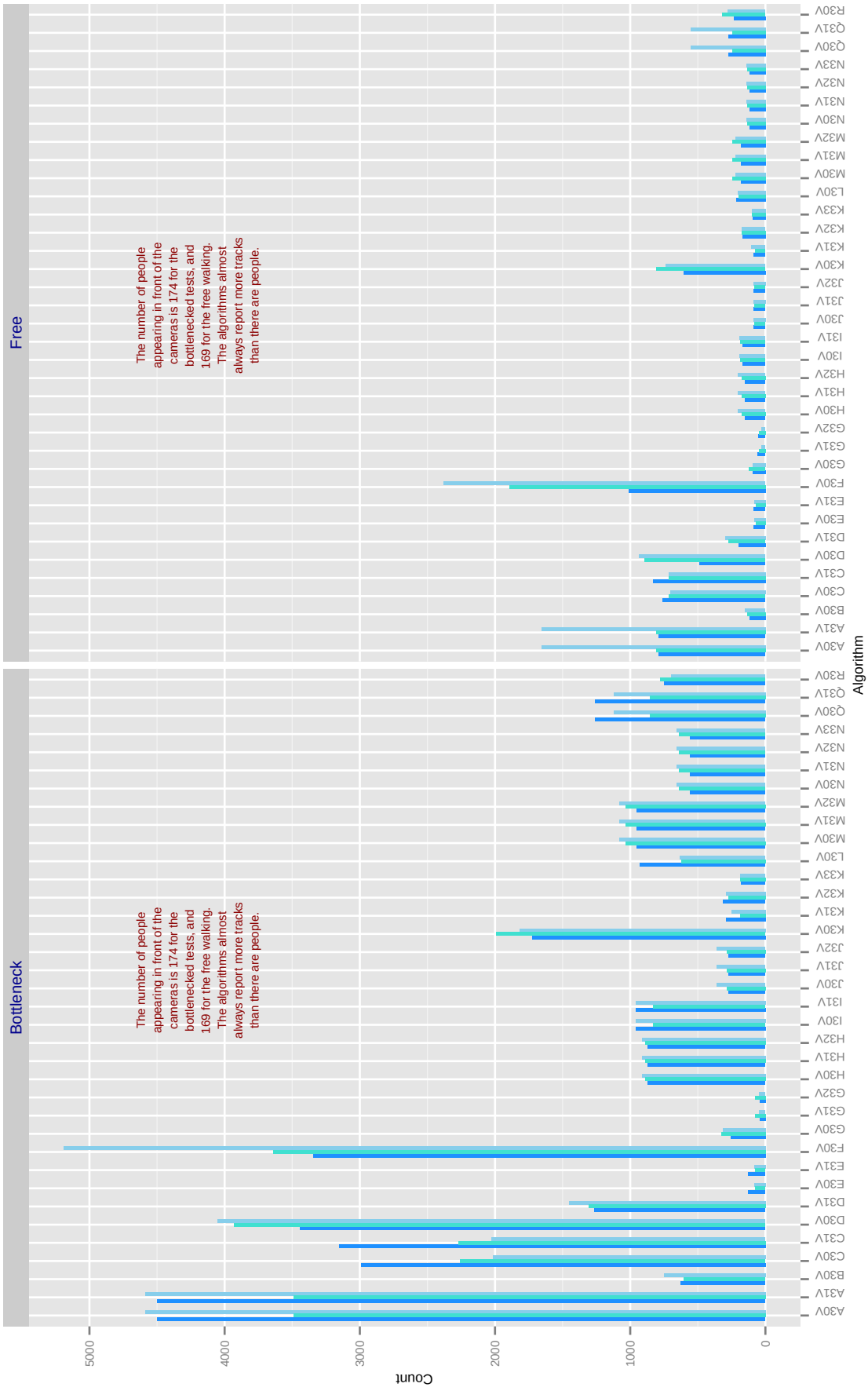


Figure 21: For each algorithm applied to the three cameras in DATASET I: PASSENGER LOADING BRIDGE*, the bars show the number of face tracks reported for people traversing the passenger loading bridge. Each reported track results in a face template and a one-to-many search. The right panel shows the case for travelers walking freely; the left panel shows higher counts for the case where a bottleneck was applied to force a queue, as occurs frequently when boarding single-aisle aircraft. In most cases the number of tracks is larger than the actual number of people who passed by the cameras - shown in the red annotation. This is a consequence of: a) false detections (of non-faces); b) imperfect tracking of each face through time; and c) possible deliberate algorithm design for example to match each frame separately.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

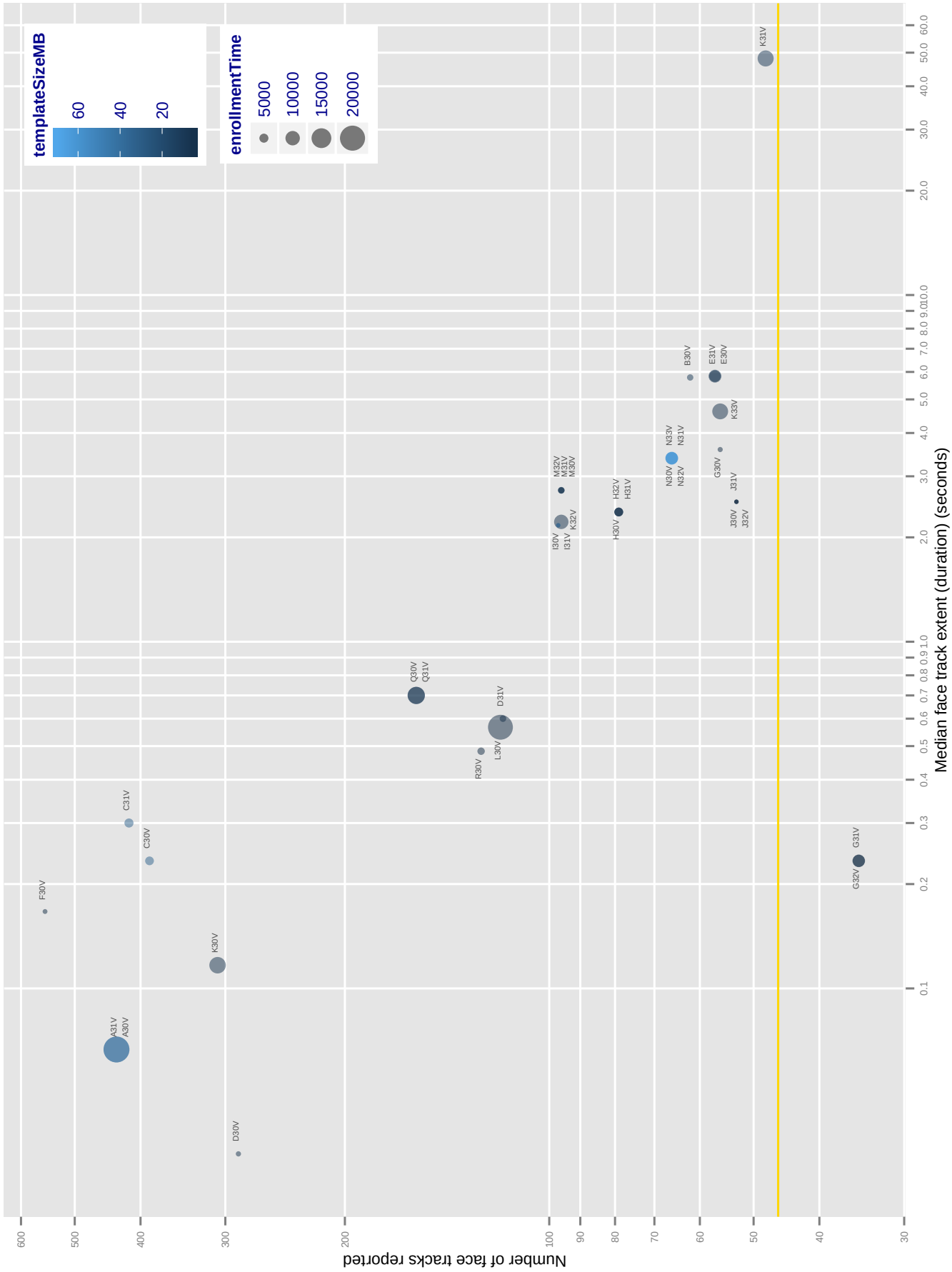


Figure 22: For DATASET J: PASSENGER LOADING BRIDGE, *the figure shows the median extent of a track*[a] *against the number of face tracks found. The input data is a single clip lasting 536 seconds during which 46 subjects (the gold line) walk freely toward and underneath a ceiling mounted camera. Some algorithms find more short tracks, some find fewer long tracks. However, some algorithms detect entirely false tracks (e.g. on moving clothing), and others miss tracks. The points are sized by the duration of the video enrollment function (normalized by the duration of the clip), and shaded according to the total size of all feature data produced.*

[a] This is the difference between the last and the first frame indices plus one, divided by the frame rate.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

5.3.3 Viable spatial resolution

Figure 23 leverages spatial track information reported by the algorithms to show at what resolutions[13] faces are being aquired and tracked through video clips. This was generated over 29118 seconds of video, collected at 30 frames per second, using three cameras yielding 48 clips.

Regarding the resolution at which algorithms first acquire subjects, at a minimum resolution, and then cease tracking, the following observations are notable.

Effect of walking: As is visible in Figure 23, most algorithms detect freely walking individuals earlier and track them longer than if the subject is standing in a queue. This occurs because bottlenecked individuals are often occluded.

Note the interocular distances given in Figure 23 are medians, and there could be considerable and differing variances around these figures.

Algorithms vary in minimum resolution requirement: One algorithm, L30V, acquires and reports faces at essentially zero resolution. Whether features are extracted from low resolution faces is not evident in a black box study. However, beyond this, some algorithms appear to be configured to only acquire faces with at least 20 (participants, A, K, M), 30 (C, D), 40 (D, R) or 50 (F) pixels between the eyes.

Algorithms vary in maximum resolution: Some algorithms (participants B, E, H, N) acquire and report faces from freely walking subjects with median 120 pixels or higher interocular distance. This high resolution occurs when subjects are close to the camera, usually with the most advserse pose angle (high pitch or yaw). Whether features are extracted from these frames can not be determined given this is a black box test. Other algorithms, notably the most accurate M algorithms cease to report face tracks with IOD beyond 80 pixels.

5.3.4 Template sizes

Figure 24 shows the size of template data extracted from 30 frames per second video data. Size is reported as a function of: a) whether subjects were queued or walking freely; b) camera placement (ceiling- vs. wall-mounted); c) whether the ceiling-mounted attractor was on or off.

The median amount of data extracted from a video track, the template size, varies massively between algorithms ranging over three orders of magnitude from hundreds of bytes up to more than a million. For most algorithms the median varies little with the three factors.

[13]The term resolution here follows common practice as being a synonym for spatial sampling rate - a measurement in pixels. It is more properly reserved for optical resolution which is measured by using dedicated tests and reported, often, by stating the modulation of intensity values at some spatial frequency. Resolution includes the effects of poor lenses, compression, and atmospheric distortion. It is possible to have many pixels on a poorly resolved target.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

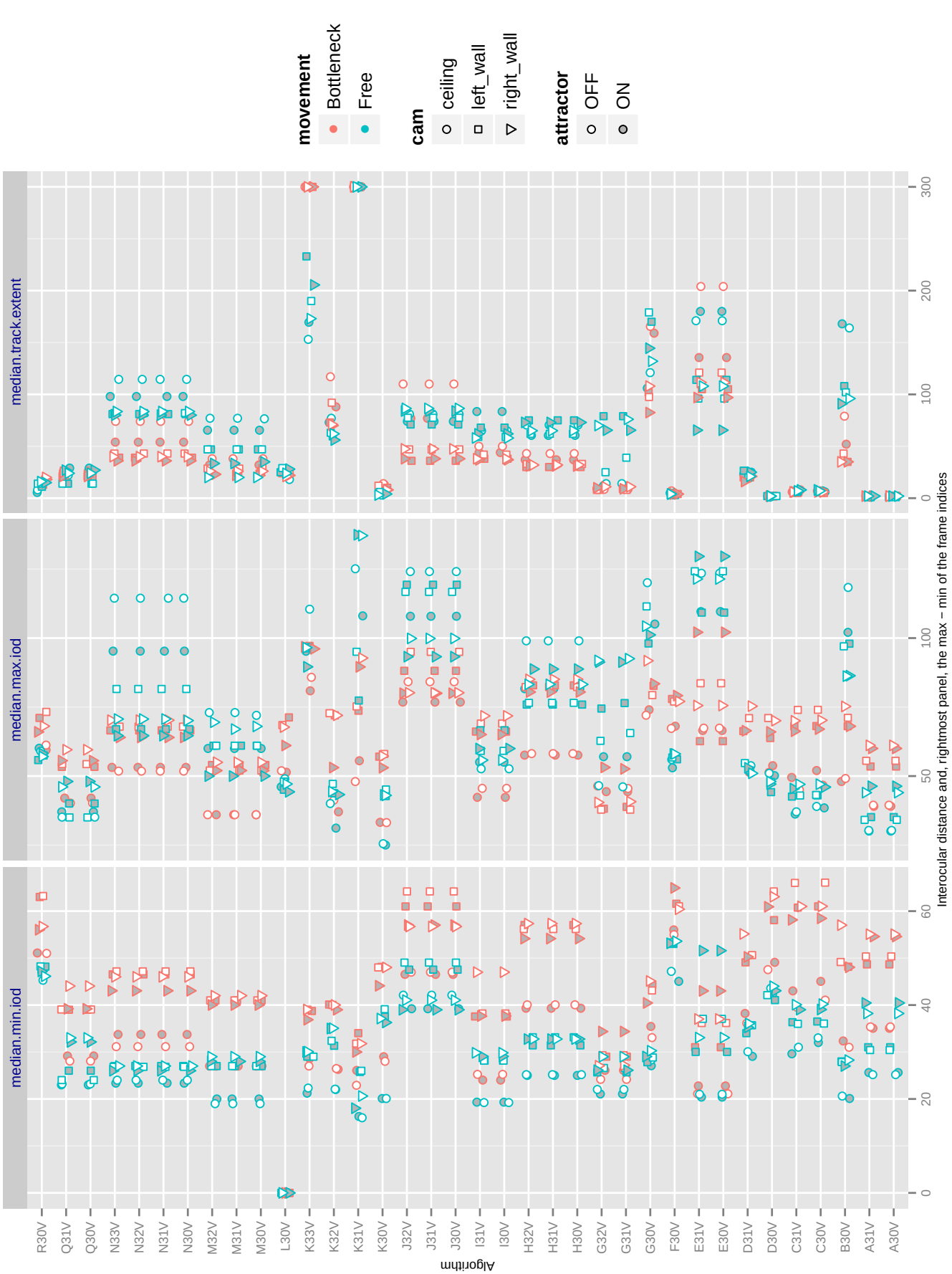


Figure 23: For each algorithm applied to DATASET J: PASSENGER LOADING BRIDGE *, the panels show: a) the median over all tracks of the minimum interocular distance reported in that track; b) the median over all tracks of the maximum interocular distance; and c) the median over all tracks of the index of the last frame minus the first frame, plus one, showing the time span over which subjects were tracked on average (frame rate is 30 s^{-1}). Anomalous high values above 300 (algorithms K31V and K33V) are capped at 300. The legend describes: a) whether the subjects were walking or stalled in a queue; b) whether the attractor display was showing a video of a talking agent or the live digital mirror; and c) which camera was used for recognition.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

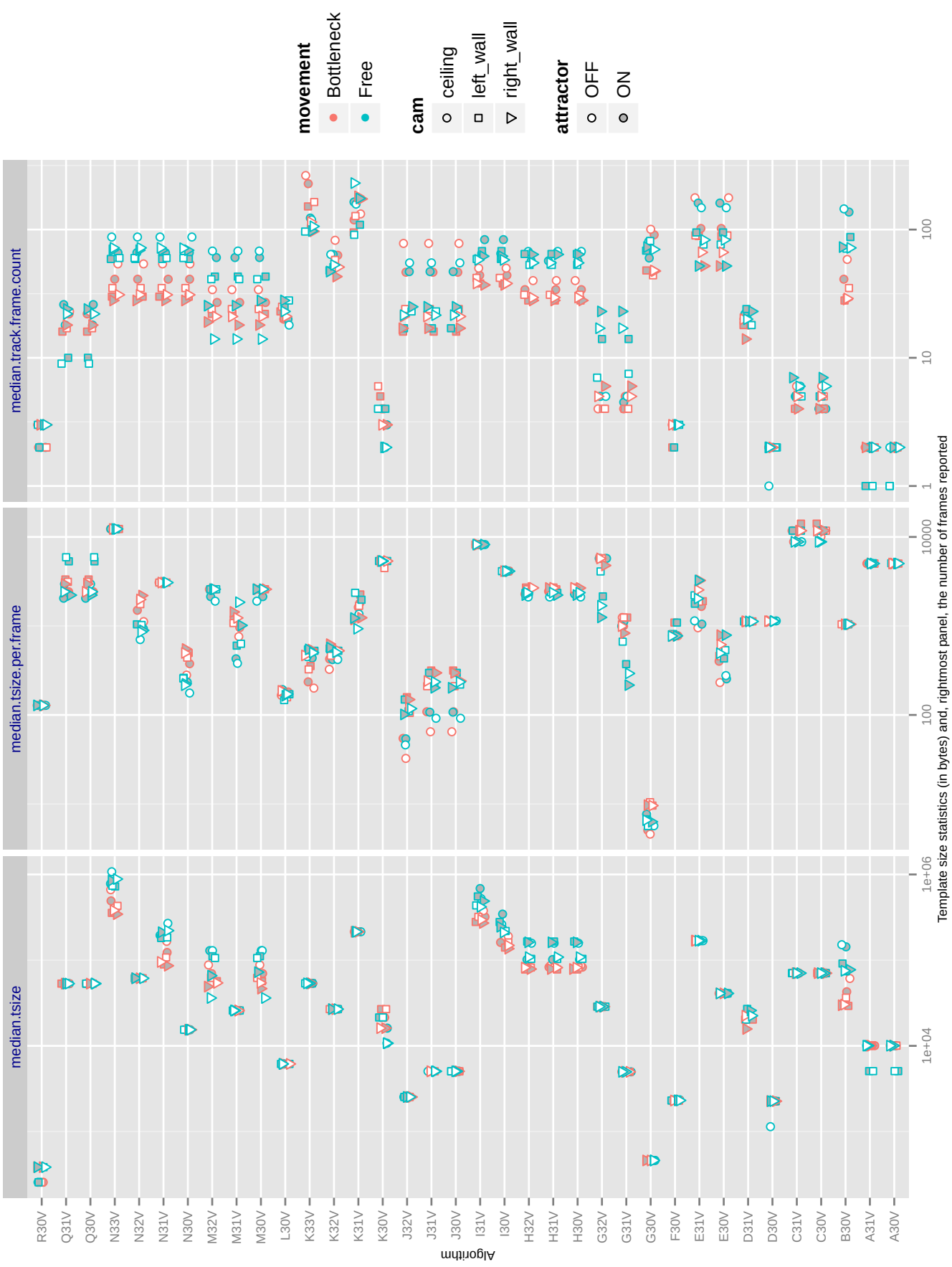


Figure 24: For each algorithm applied to DATASET J: PASSENGER LOADING BRIDGE *, the panels show: a) the median size in bytes for templates extracted from tracks detected in video; b) the median of template size divided by the number of frames reported for that track; and c) the the number of frames reported.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

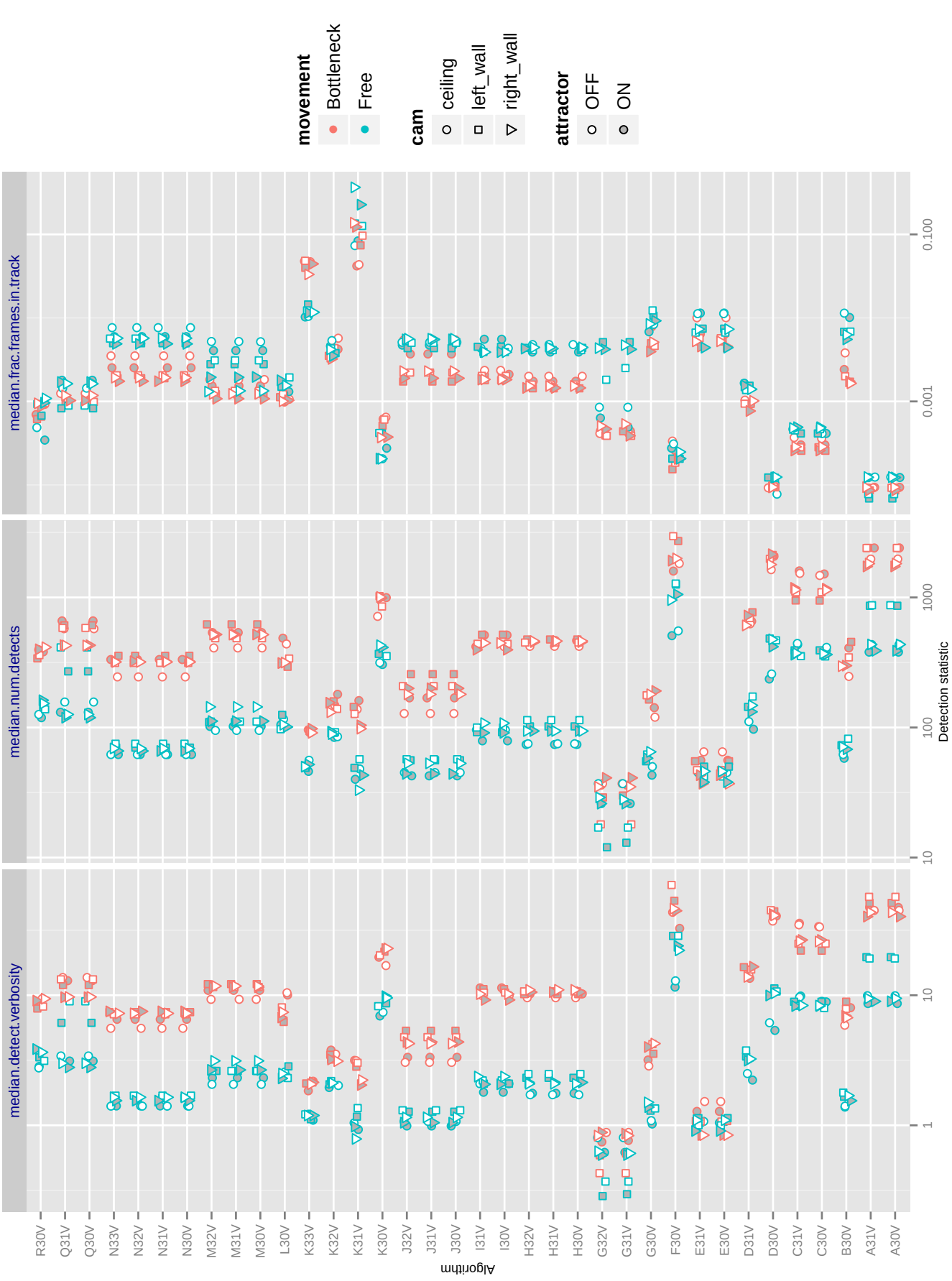


Figure 25: For each algorithm applied to DATASET J: PASSENGER LOADING BRIDGE *, the panels show: a) the ratio of tracks reported (faces detected) to the number of faces known to appear; b) the number of face tracks reported; and c) the median of the number of frames reported in a track as a proportion of the number of frames in the entire clip.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

DATASET J	VIDEO SEARCH						STILL IMAGE	
	NUMBER OF	MEAN TRACK	MEAN TRACK	MEAN	MEAN	MEAN	IOD (PIXELS)	
ALGORITHM	TRACKS	LENGTH (FRAMES)	EXTENT (FRAMES)	MIN IOD (PX)	MEAN IOD (PX)	MAX IOD (PX)	MEAN	STD. DEV
A30V	63317	15	15	51	54	57	116	20
A31V	63317	15	15	51	54	57	116	20
B30V	9533	117	139	50	65	131	116	17
C30V	37758	12	13	58	62	67	117	20
C31V	38860	12	13	57	61	66	117	20
D30V	54933	16	16	64	67	70	117	17
D31V	19168	56	56	53	63	73	117	17
E30V	2081	170	199	40	59	99	116	17
E31V	2081	170	199	40	59	99	116	17
F30V	69839	6	63	65	69	74	115	17
G30V	3947	169	288	46	64	95	116	18
G31V	1167	784	1004	37	50	78	116	18
G32V	1167	783	1003	37	50	78	116	18
H30V	12788	58	62	56	68	88	118	17
H31V	12788	58	62	56	68	88	118	17
H32V	12788	58	62	56	68	88	118	17
I30V	13170	93	93	38	56	72	118	20
I31V	13170	93	93	38	56	72	118	20
J30V	4677	87	140	62	78	99	120	17
J31V	4677	87	140	62	78	99	120	17
J32V	4677	87	140	62	78	99	120	17
K30V	30720	36	468	24	44	58	118	17
K31V	3964	305	3782	3	50	90	118	17
K32V	5590	167	1115	41	54	73	118	17
K33V	3388	290	3728	39	60	92	118	17
L30V	11206	49	50	11	34	67	116	18
M30V	14850	83	89	43	50	61	102	15
M31V	14850	83	89	43	50	61	102	15
M32V	14850	83	89	43	50	61	102	15
N30V	8977	103	122	46	59	77	117	18
N31V	8977	103	122	46	59	77	117	18
N32V	8977	103	122	46	59	77	117	18
N33V	8977	103	122	46	59	77	117	18
Q30V	17224	48	52	41	50	58	115	18
Q31V	17224	48	52	41	50	58	115	18
R30V	12198	3	116	62	67	73	116	17

Table 12: For DATASET J: PASSENGER LOADING BRIDGE *and each video processing algorithm the table shows: a) the number of reported tracks; b) the mean number of frames reported within those tracks; c) the mean extent (first minus last frames indices plus one); d) the mean over all tracks of of the minimum interocular distance (IOD) reported; e) the mean of the mean IOD; f) the mean of the maximum IOD; g) the enrollment still image mean IOD; and h) its standard deviation. For some algorithms (F, J, L, R) the tracks don't include all consecutive frames, so the extent of the track can exceed the number of frames in it. For the D, F, and L algorithms, the reported still-image eye coordinates are erroneous.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

5.4 DATASET L: PASSENGER LUGGAGE

Overview: This dataset is composed of video clips of 248 persons collected using two ceiling-mounted overhead cameras. The videos were collected to simulate use of inexpensive legacy cameras installed in a non-ideal location relative to the subjects who walk into the field of view, retrieve their luggage, and then proceed towards and underneath the cameras out of view - see Figure 26. The cameras are mounted at a height of about 2.13 meters (7 feet), above and to the left and right of the egress walkway. Each clip captures approximately 12-15 people walking toward the camera, and each person is typically in the field of view for 10-20 seconds. Key imaging properties for this dataset are summarized in Table 13.

This dataset is challenging for recognition, because resolution is low and, particularly, pose is poor - subjects' head orientations are far from the optical axes. These problems exist in many legacy operational settings, such as shops and banks, that were never designed to support face recognition. Such data is nevertheless frequently used in investigations, because it is sometimes is the only evidence available to generate a lead.

Property	Value
Camera	Logitech C920
Camera mounting	Ceiling mounted, to left and right of subject motion
Camera height	2.1 meters
Range to subject	[0.7,5] meters
Frame rate	15 sec^{-1} or 30 sec^{-1}
Width	1080
Height	1920
Chroma sampling	YUV420
Nominal bitrate	130000 kb sec^{-1}
Codec	WVC1 (advanced) 0x31435657

Table 13: Key imaging properties for DATASET L: PASSENGER LUGGAGE

The enrollment images are from DATASET U: PASSENGER GATE and embedded into a set of background photographs such that the enrolled population size reaches $N = \{480, 4800\}$, with exactly one image per person. All images are high quality frontal portraits in approximate conformance to the ISO full frontal image type - see Figure 8.

Experimental Design: The videos contain footage of subjects collecting hand luggage and then walking toward and underneath two ceiling mounted cameras. **Mated scores** are computed by searching 34 long video clips against an enrolled dataset of still face images of subjects known to be in the search videos. **Nonmated scores** are collected by comparing the same 34 video clips against the *global nonmated enrollment dataset*.

Quantity	Value or description
Mode	Video search to still enrollment
Number of actors	248
Number of non-actors	0
Number of cameras	2 (Left and right of field of view)
Video duration with actors	47.8 minutes
Video duration no actors	0
Subject motion	Usually single file toward and below the camera. Walking or stopping to pick up luggage.
Number of clips	34
Clip duration (frames)	Median 1611; Min 997; Max 3777
Number of enrolled subjects	480, 4800
Number of enrolled stills	1 per subject
Properties of enrolled stills	Frontal, close ICAO compliance; Mean IOD 106 pixels
FNIR estimation	Actors present video vs. enrolled gallery
FPIR estimation	Actors present video vs. separate non-actor gallery
Candidate list length	20
Number of persons in FOV	[4,7] usually
Video ground truth	Style A: See Figure 6

Table 14: Key experimental design for the DATASET L: PASSENGER LUGGAGE *results.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Key experimental design details are summarized in Table 14.

Results: Table 16 presents FNIR for two gallery sizes, $N = \{480, 4800\}$, at two decision thresholds, corresponding to false positives counts of $\text{NFP}(T) = \{1, 10\}$, and at two ranks $R = \{1, 10\}$. Given the camera installation is not representative of a face recognition deployment, the rank-based metrics are probably more important, as they better represent the error rates that a forensic investigator would have to accept in a law enforcement investigation.

The table shows generally high error rates, making this dataset the worst among those in this report. The cameras are installed such that there is a high downward view angle (pitch) and high side-to-side angle (yaw). Additionally the optical resolution of the camera is low: The algorithms report mean interocular distances of 32 pixels - see Table 15. Illumination also is non-uniform. The result is that even for a small gallery, $N = 480$, the best high-threshold miss rate is 28% (M30V), and only two developers have algorithms capable of FNIR below 50%. For investigations, this improves to 11% when we consider the rank 1 error rate (M30V), and then eight developers have algorithms capable of producing $\text{FNIR}(480, 1, 0)$ below 50%. Given that law enforcement investigations are often willing to follow any lead these error rates are still well below 100% and therefore low enough that investigators would continue to use face recognition as an investigative tool.

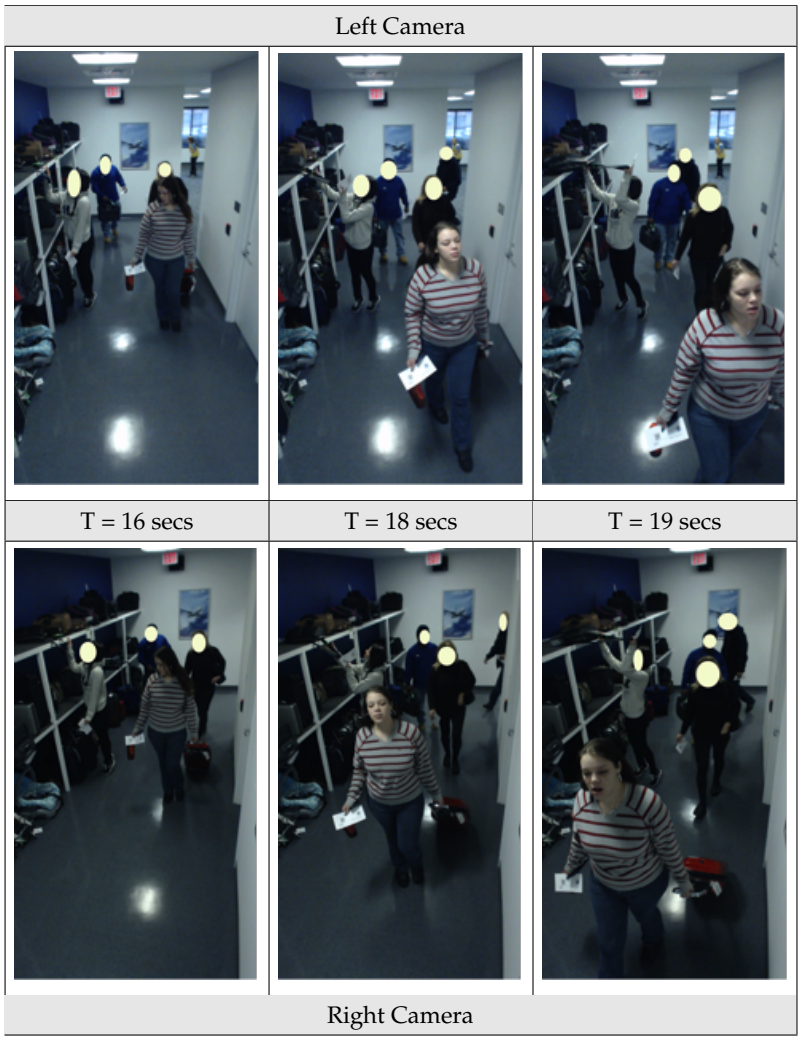


Figure 26: This example clip from DATASET L: PASSENGER LUGGAGE *has the subject in view for around 20 seconds and contains 300 frames. Note the differences in illumination between the left and right camera positions. **The face images in this figure are from the DHS/ S&T provided AEER dataset. The included subject consented to release their images in public reports. Subject 79195743 (Perm Granted). Where consent for public release from individuals in the background was not obtained, their faces were masked (yellow circles).*

That said the enrolled population size here is only $N = 480$. As N increases, low image quality is expected to cause rank one miss rates to increase as false positives displace some mates from rank one. This gives the usual decline in face recognition accuracy as more individuals are nominated to watch-lists. However in this dataset, the effect is larger. Thus, for the most accurate algorithm (M30V), $\text{FNIR}(4800, L, T) = 0.53$ is almost double $\text{FNIR}(480, L, T) = 0.28$ for fixed T. For the best algorithm from the next most accurate developer on this set (G32V), error rates increase to 0.94 from 0.42.

An interesting effect is that algorithms differ when a 10-fold increase in N is accompanied by a 10-fold increase in the number of tolerable false positives. Some algorithms (G30V, J31V, N3xV, R30V) conserve FNIR, as is expected from binomial models; some give moderately (M3xV, J30V, J32V, N3xV) or much (G32V) worse FNIR; and still others improve

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

(H3xV, I3xV, N32V). The conclusion is that a system operator cannot rely on an assertion that gallery size increases can be offset by proportional increases of human-reviewed false positives. This is addressed later, in the discussion of Figure 36 for Dataset H.

DATASET L	VIDEO SEARCH						STILL IMAGE	
	NUMBER OF	MEAN TRACK	MEAN TRACK	MEAN	MEAN	MEAN	IOD (PIXELS)	
ALGORITHM	TRACKS	LENGTH (FRAMES)	EXTENT (FRAMES)	MIN IOD (PX)	MEAN IOD (PX)	MAX IOD (PX)	MEAN	STD. DEV
A30V	4071	7	7	31	33	36	116	20
A31V	4071	7	7	31	33	36	116	20
B30V	764	34	42	28	48	185	118	21
C30V	2062	5	6	39	42	46	119	21
C31V	2482	5	5	35	38	42	119	21
D30V	3794	3	3	58	60	61	118	22
D31V	887	22	23	37	46	58	118	22
E30V	664	41	53	27	37	56	117	20
E31V	664	41	53	27	37	56	117	20
F30V	4213	3	5	35	36	38	116	20
G30V	463	13	31	35	43	57	118	21
G31V	678	27	87	32	41	54	117	21
G32V	678	27	87	32	41	54	117	21
H30V	714	29	31	35	45	62	119	21
H31V	714	29	31	35	45	62	119	21
H32V	714	29	31	35	45	62	119	21
I30V	1218	39	39	12	23	34	120	21
I31V	1218	39	39	12	23	34	120	21
J30V	445	19	24	44	54	68	121	21
J31V	445	19	24	44	54	68	121	21
J32V	445	19	24	44	54	68	121	21
K30V	4093	6	49	16	21	25	119	21
K31V	440	81	967	8	29	63	119	21
K32V	562	25	65	31	37	46	119	21
K33V	310	50	325	31	41	58	119	21
L30V	463	15	16	9	32	53	116	20
M30V	1094	29	33	26	32	42	104	19
M31V	1094	29	33	26	32	42	104	19
M32V	1094	29	33	26	32	42	104	19
N30V	598	35	46	26	39	58	118	21
N31V	598	35	46	26	39	58	118	21
N32V	598	35	46	26	39	58	118	21
N33V	598	35	46	26	39	58	118	21
Q30V	1242	24	26	25	32	40	116	20
Q31V	1242	24	26	25	32	40	116	20
R30V	1148	2	51	50	53	57	116	20

Table 15: For DATASET L: PASSENGER LUGGAGE *and each video processing algorithm the table shows: a) the number of reported tracks; b) the mean number of frames reported within those tracks; c) the mean extent (first minus last frames indices plus one); d) the mean over all tracks of of the minimum interocular distance (IOD) reported; e) the mean of the mean IOD; f) the mean of the maximum IOD; g) the enrollment still image mean IOD; and h) its standard deviation. For some algorithms (F, J, L, R) the tracks don't include all consecutive frames, so the extent of the track can exceed the number of frames in it. For the D, F, and L algorithms, the reported still-image eye coordinates are erroneous.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

N=480	NUM CLIPS 34	NUM FEEDS 2		NUM ACTORS 248		NUM FRAMES 62845		NUM MINUTES 47.8	
N=4800	DETECTIONS	THRESHOLD BASED IDENTIFICATION				RANK BASED INVESTIGATION (T = 0)			
ALG	NUM	FNIR(480, T), NFP(T)=1		FNIR(4800, T), NFP(T)=1		FNIR(480, T), NFP(T)=10		FNIR(480, R=1)		FNIR(480, R=10)		FNIR(4800, R=10)	
A30V	4071	0.955	22	0.988	20	0.951	21	0.619	26	0.769	25	0.190	16	0.518	24
A31V	4071	0.984	24	0.984	19	0.976	25	0.575	22	0.765	24	0.166	12	0.506	22
B30V	764	0.943	21	0.992	22	0.964	22	0.611	24	0.777	26	0.300	25	0.575	26
C30V	2062	0.992	29	1.000	35	0.976	26	0.773	28	0.850	28	0.433	28	0.603	28
C31V	2482	0.988	26	0.996	24	0.972	24	0.773	27	0.846	27	0.425	27	0.595	27
D30V	3794	0.984	23	1.000	26	0.988	28	0.583	23	0.713	22	0.287	24	0.514	23
D31V	887	1.000	35	1.000	32	1.000	34	0.615	25	0.692	21	0.401	26	0.518	25
E30V	664	0.927	20	1.000	31	0.927	20	0.377	13	0.603	19	0.170	13	0.332	11
E31V	664	0.984	25	1.000	36	0.964	23	0.401	16	0.713	23	0.154	10	0.389	16
F30V	4213	0.992	28	0.996	25	0.988	29	0.911	34	0.964	31	0.555	30	0.789	29
G30V	463	0.713	5	0.992	23	0.737	8	0.433	17	0.583	17	0.223	18	0.368	14
G31V	755	0.911	19	0.915	12	0.866	18	0.368	12	0.619	20	0.223	19	0.409	17
G32V	755	0.417	4	0.943	16	0.818	15	0.397	15	0.551	12	0.190	15	0.352	13
H30V	714	0.769	9	0.773	4	0.664	4	0.247	5	0.453	7	0.089	4	0.287	7
H31V	714	0.794	11	0.777	5	0.680	5	0.279	7	0.453	6	0.142	7	0.296	8
H32V	714	0.794	12	0.777	6	0.680	6	0.279	8	0.457	8	0.142	9	0.287	6
I30V	1218	0.826	15	0.935	14	0.696	7	0.239	4	0.368	4	0.113	5	0.186	4
I31V	1218	0.883	18	0.960	17	0.781	11	0.259	6	0.405	5	0.117	6	0.198	5
J30V	476	0.725	7	0.862	9	0.765	10	0.441	18	0.571	15	0.255	21	0.437	21
J31V	476	0.729	8	0.842	8	0.737	9	0.445	19	0.559	13	0.275	23	0.433	20
J32V	476	0.721	6	0.891	10	0.785	12	0.445	20	0.563	14	0.255	22	0.409	18
K30V	4093	0.996	31	1.000	34	1.000	36	0.785	29	0.960	30	0.522	29	0.854	32
K31V	440	1.000	32	1.000	27	1.000	30	0.866	33	0.984	34	0.700	34	0.911	34
K32V	562	1.000	36	1.000	33	1.000	35	0.798	30	0.972	32	0.595	32	0.850	31
K33V	310	0.992	27	1.000	29	1.000	32	0.834	32	0.980	33	0.668	33	0.895	33
L30V	463	0.996	30	0.992	21	0.984	27	0.834	31	0.931	29	0.575	31	0.802	30
M30V	1094	0.275	1	0.530	2	0.340	1	0.105	1	0.182	2	0.065	3	0.113	2
M31V	1094	0.304	3	0.547	3	0.344	2	0.126	3	0.206	3	0.061	2	0.117	3
M32V	1094	0.300	2	0.445	1	0.356	3	0.117	2	0.170	1	0.053	1	0.109	1
N30V	598	0.810	14	0.927	13	0.838	16	0.385	14	0.587	18	0.202	17	0.417	19
N31V	598	0.806	13	0.972	18	0.862	17	0.356	11	0.543	11	0.158	11	0.348	12
N32V	598	0.830	16	0.838	7	0.806	14	0.300	9	0.457	9	0.174	14	0.308	10
N33V	598	0.773	10	0.939	15	0.794	13	0.316	10	0.466	10	0.142	8	0.304	9
Q30V	1242	1.000	33	1.000	28	1.000	31	0.984	36	1.000	35	0.899	36	0.988	36
Q31V	1242	1.000	34	1.000	30	1.000	33	0.968	35	1.000	36	0.749	35	0.947	35
R30V	1148	0.870	17	0.911	11	0.870	19	0.490	21	0.575	16	0.227	20	0.385	15

Table 16: For the DATASET L: PASSENGER LUGGAGE *installation, with 480 and 4800 subjects enrolled with a frontal still, the values are identification-mode FNIR(T) for each algorithm at two different decision thresholds corresponding to false positive counts of 1, 10, and investigation-mode FNIR(R) for ranks 1 and 10.Each value is accompanied by an integer ranking across all algorithms. The shading indicates the most important metric to forensic investigation applications. Two cameras were used. The detections are summed over all of them. The accuracy values are aggregated over all sightings of all subjects in the field of view of those cameras. Caution: The N=480, R=10 column is unreliable per the arguments in section 4.4.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

5.5 DATASET P: SPORTS ARENA

Overview: This dataset is composed of videos from eleven cameras mounted in the indoor access areas surrounding a sports arena. Video was collected on five evenings corresponding to real sporting events. The collection includes some known actors and many more incidental unknown persons. The cameras collected imagery for several hours each evening, with sunset occuring early in the evenings.

The cameras are visible to the subjects. Therefore, the subjects are in-principle aware of the cameras, but they largely ignore them. The number of video clips is very large, comprised of clips known to contain actors, and clips known *not* to contain actors. In all cases, there are multiple faces visible in the clips. The duration of the subjects' appearance is generally 10 to 30 seconds, but there are instances where a subject is only visible for a much briefer interval. Key imaging properties for this dataset are summarized in Table 17.

Property	Value
Cameras	Canon VIXIA HF R400
Camera mounting	Fixed to wall or door, no attractor
Camera ht. hallway	1.83 and 2.44 meters
Camera ht. door	2.44 meters
Range to subject	[1,10] meters
Frame rate	24 sec^{-1}
Width	1920
Height	1080
Chroma sampling	YUV420
Nominal bitrate	24 Mbit sec^{-1}
Codec	AVC H264

Table 17: Key imaging properties for DATASET P: SPORTS ARENA

Enrolled still images: Images were collected under controlled lighting, background, and pose conditions. The FIVE API [20] supports multiple still image input along with nominal pose (yaw and pitch) values to the algorithm software for template generation, which enables performance analysis when the algorithm is provided with multiple images/poses of the subject.

Full-frontal: This gallery is composed of N=480 subjects, with exactly one full front image per subject.

Multi-pose: This gallery is composed of N=480 subjects, with three images per subject as in Figure 33.

Videos: The DATASET P: SPORTS ARENA videos were collected from eleven cameras, in three groups:

Doors: Figure 28 shows example frames collected from three cameras mounted over three entry-exit doors. These are numbered 5, 6 and 7. Here the subjects walk into the building toward the cameras at the beginning of the sporting event and, after reversing the cameras viewpoint, from subjects walking toward the cameras as they exit the venue. Table 36 in Appendix 1 include statistics for intercoular distances this group, as reported by each algorithm.

Queue: Figure 27 shows images collected from two groups of cameras. The first includes images from three cameras (labelled 2, 3, 4) that are mounted in a concession stand queue where subjects walk mostly toward and transverse to the cameras. These cameras are designated as "near" in that subjects are close to the camera. Tables 39 and 40 show interocular distance statisics for, respectively, low and high mounted cameras, with the low cameras giving, on average, somewhat smaller detected faces.

Hallway: Figure 27 also shows examples from the second group of cameras. The first group is comprised of five cameras mounted at the end of a hallway that observe subjects walking both toward and away from the cameras. These cameras are numbered 1, 9, 10, 11, 12 and designated as "far". The faces are typically far from the camera, although in some cases, particularly camera 1, the subjects are closer. The cameras are mounted at a height of 1.83 meters (6 feet, "low") or 2.44 meters (8 feet, "high"). Tables 37 and 38 include interocular distance statistics

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

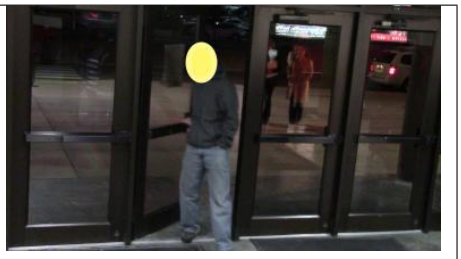
Quantity	Value or description
Mode	Video search to still enrollment
Number of actors	64
Number of non-actors	Many
Number of cameras	11
Video duration with actors	7995 mins over four evenings
Video duration no actors	2883 mins over one different evening
Number of clips actors	16460
Number of clips no actors	5809
Subject motion	Mostly toward, but many laterial or away from cameras
Clip duration (frames)	Median 634; Min 24; Q25 369; Q75 774; Max 2881
Number of enrolled subjects	480
Number of enrolled stills	1 FF (also separately 1FF + 1QR + 1QL) per subject
Properties of enrolled stills	Frontal, close ICAO compliance; Mean IOD 111 pixels
FNIR estimation	Actors present video vs. enrolled gallery
FPIR estimation	Actors absent video vs. same enrolled gallery
Candidate list length	20
Number of persons in FOV	[0,30] across FOV, none dominant
Video ground truth	Style A: See Figure 6

Table 18: Key experimental design for the DATASET P: SPORTS ARENA *results.*

for, respectively, low and high mounted cameras. The reported interocular distances are significantly lower, on average, than in the queue.

CAMERA 5 (8FT)

CAMERA 6 (8FT)



CAMERA 7 (8FT)

Figure 28: DATASET P: SPORTS ARENA *: Examples from public-area surveillance video clips (entry and exit). **The face images in this figure are from a DHS/ S&T provided dataset. Written consent from DHS / S&T to use these images in public reports was obtained. Actors AS, AY (Perm Granted). Lacking individual consent, faces are masked (yellow circle).*

Experimental Design: The videos contain footage of people in various places inside a sports arena including the main entrance and exit doors, the hallway, and in the queue to a concession stand.

Mated scores are computed over 16460 video clips searched against an enrollment dataset of size 480 of still face images of subjects known to be in the search videos. The enrollment database is padded with FERET and mugshot images to attain the desired size.

Nonmated scores scores are computed by running many actors-absent video clips. This is done for each camera.

Key experimental design details are summarized in Table 18.

Detection Results: Table 19 includes total detection counts over 2883 minutes of video. As with other datasets, the algorithms vary widely in the number of reported face tracks. The most accurate algorithm M32V reports 144311 tracks from all cameras, corresponding to 54 detections per minute, on average. The G30V algorithm reports just 36630 tracks

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

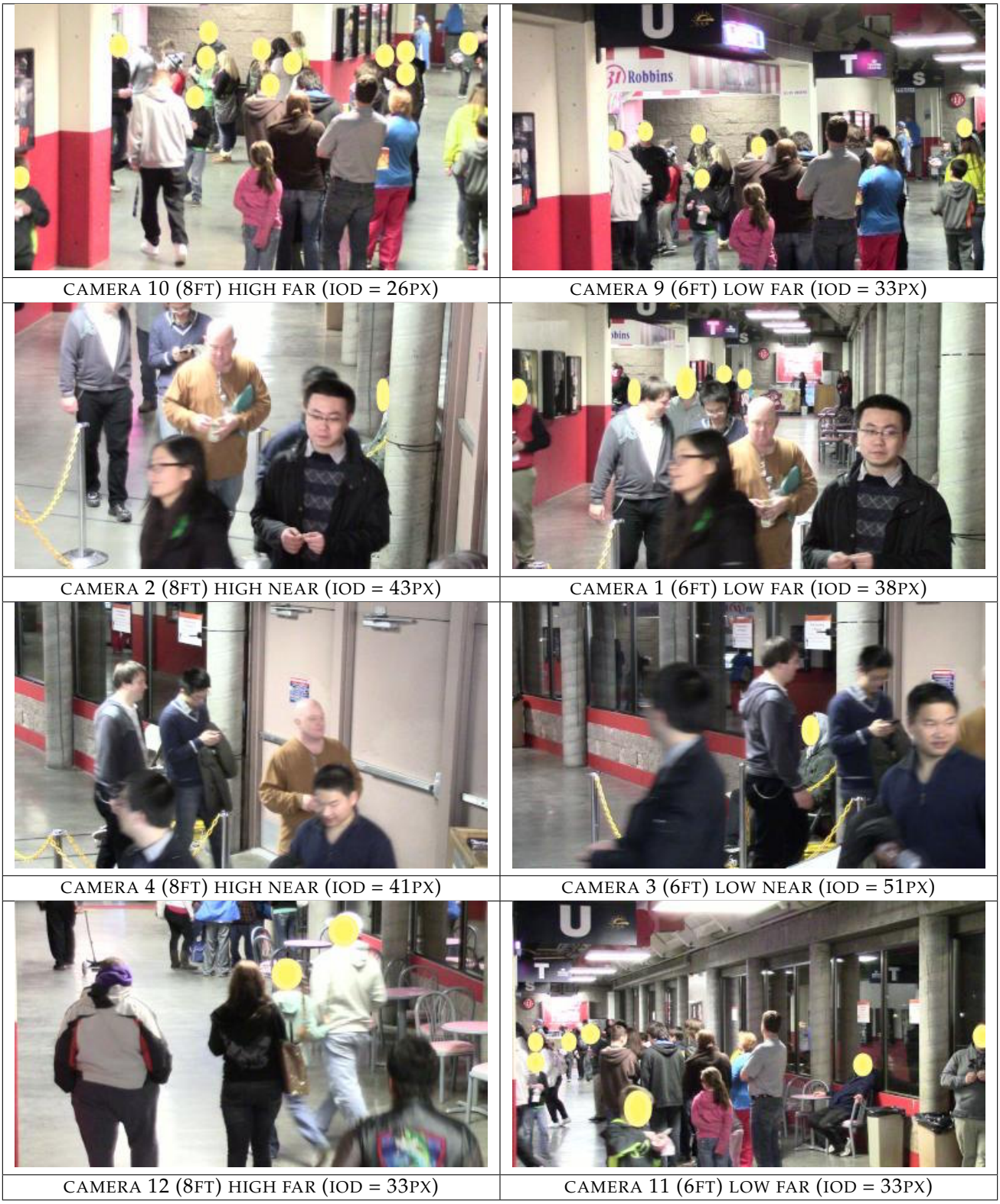


Figure 27: DATASET P: SPORTS ARENA *: Examples from public-area surveillance video clips (hallway and queue). Results are grouped by the labels low—high near—far. For camera 1 the FAR label is imperfect in that subjects are aquired both near and far, and accuracy is more inline with far acquisition. The interocular distances* (IOD) *values are global averages from four algorithms over all reported tracks; peak* IOD *values will be larger. See tables in the Appendix. The left and right images are contemporaneous (except in the last row). **The face images in this figure are from a DHS/ S&T provided dataset. Written consent from DHS / S&T to use these images in public reports was obtained on August 12, 2016. Actors AS, AV, AW, AY, I, V (Perm Granted). Lacking individual consent, faces are masked (yellow circle).*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

(14 min^{-1}), F30V gives 386103 (148 min^{-1}). This variability arises because detection and tracking is a non-trivial task, and is a tunable parameter. For example, the K3xV algorithms have a 20-fold spread in the number of tracks reported. A verbose detector is only less accurate if its detected faces yield high-scoring false positive candidates.

A parsimonious detector is less accurate only if it misses faces (of actors). In any case, the hardware cycles needed to execute searches rise linearly with detection rates.

Figure 30 plots detection counts alongside false positive counts. It does this for the five camera groups. It uses a log scale because, while the number of detections is in the tens of thousands, we set the threshold to produce just 800 false positives. This was achieved by setting the threshold globally, over *all* impostor video clips. Operationally, it would be more advisable to set the threshold on a per camera or per location basis. The reason for this is because false positives vary by location, or more specifically, with the imaging environment. Notably high mounted cameras give fewer false positives, expecially when close to the subject. The most false positives are produced in the LOW-FAR camera configuration where detection rates are also highest.

Recognition Results: Detection and accuracy results are tabulated in Table 19. The dotplots of Figure 29 are included to visualize both identifcation-mode and investigation-mode miss rates i.e. FNIR$(480, L, T)$ and FNIR$(480, 1, 0)$. In the former case the decision threshold, T, is set to correspond to 200 false positives over all tracks detected in the 2883 minutes of video footage that does not contain actors (see Table 18). The number 200 may imply a significant level of human adjudication but it corresponds to one false positive for every 11.7 minutes of video footage on average. Thus the decision threshold is more stringent than that used in the DATASET U: PASSENGER GATE analysis because the footage is much longer and contains many more people, on average.

Overall accuracy: Given the challenging nature of the video in terms of illumination and the use of an inexpensive imaging system, the identification miss rates are generally higher than with other datasets.

Door cameras: As shown in Table 19, the FNIR values at the door entry and exit are much worse than for the hallway mounted cameras. The most accurate algorithm fails to place the actor at rank 1 fully 40% of the time. When the threshold is raised, the best value ascends to 64%. This is due to the high elevation (pitch) angle, adverse lighting - see Figure 28 - and the short duration that a subject is in view, corresponding to a high radial component of subject motion. In the surveillance mode when threshold is set to limit false positives, only one manufacturer is capable of missing fewer than 80% of the subjects. This essentially says that face recognition is not viable without deliberate tested improvements to this imaging environment.

Queue cameras: The best FNIR values occur for subjects in the queue. There, rank 1 recognition miss rates are below 6%, and when the threshold is raised, remain below 13% (algorithm M32V). Two factors are at play. First is resolution, as discussed in the next bullet. Second is duration of imaging: As subjects approaching the queue cameras come to a standstill, their track lengths (durations) are longer, affording more viewss of the face. Tracking then is helped by increased resolution.

Hallway cameras: FNIR values are markedly worse for subjects imaged at greater distances along the hallway. This is especially true for identification-mode high-threshold FNIR estimates, where FNIR is often double the value for the queue. Recognition accuracy is likely driven by resolution: From the eye coordinates reported by the algorithms - see Appendix 1 - it is clear that the near view interocular distances are almost double those of faces in the far field hallway cameras.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Camera height: Cameras mounted at 2.44 meters give worse miss rates than those mounted lower, at 1.83 meters. This effect is more pronounced when imaging subjects near to the camera - as the look down angles are larger - and at higher recognition thresholds. The effect is highly algorithm dependent, with the I and M having more immunity to pitch angle.

5.5.1 Effect of fusing results over cameras

This report mostly addresses the question of how well algorithms can identify an individual in a video clip from a single camera, and the accuracy figures are very useful for comparing algorithms. In settings where several cameras are used to observe a volume, accuracy can be improved - this was addressed in section 5.3 - essentially by spatially fusing recognition outcomes from a single appearance of a subject. Here we address fusion over time, answering the question of how well persons are recognized at any point during an event, in this case over an evening. Our study is not well controlled in that the actors were not instructed to appear in a location a fixed number of times during the event. Instead we have imbalance where the number of appearances varies across subjects.

The results of Figure 31 plot identification error rates with fusion against those without. Fusion is implemented by taking the highest scoring candidate entry over all sightings of a subject by all cameras in camera group over the course of an event. Referencing Figures 27 and 28 and the camera group definitions (door cameras 5-7, high mounted far cameras 10,12, high near 2,4, low mounted far 9, 11, 1, and near 3. Camera 1 was installed as a near-field camera but is grouped with the other far cameras because algorithms acquire many far faces and give worse accuracy accordingly. Fusion substantially improves accuracy values, often FNIR is reduced by factors of three or more. This is directly related to the number of appearances of a subject and would be mostly ineffective for applications where a subject passes a single camera exactly once. This is evident for the door group - fusion is less effective simply because subjects usually only appeared in that vicinity twice, at the beginning and end of an evening event. The technique also has reduced operational relevance, because the fusion is necessarily "after action" meaning it is no longer a real-time operation. The technique may be useful for applications that seek to determine whether an individual appeared at any time. For example, did aircraft maintenance staff board the aircraft at any point during the three hour stopover.

Figure 32 takes the fusion further, by fusing over all cameras in addition to all appearances. The result is that the most accurate algorithm correctly identify all the actors present in the dataset. This generally supports the conclusion that more cameras support better identification rates. While effective, it's efficiency in terms of cost and time, is questionable. Instead of equipping a building with many cameras, it is likely more worthwhile to construct a volume through which all subjects pass, and to install cameras and illumination there, paying attention to pose angles and optical specifications including resolution, depth of field, and field of view.

5.5.2 Effect of enrolling multiple pose views

To test whether algorithms are capable of exploiting multiple views of a face, we executed the same sets of genuine and impostor video searches against two galleries. In the first, subjects were enrolled with a single full frontal still image. In the second, subject were enrolled with three images, as shown in Figure 33. This is achieved via the FIVE API [20] which supports providing multiple still images with their nominal pose (yaw and pitch) values to the algorithm software in a single template generation function call. This allows the software to exploit the set of images in whatever manner it

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

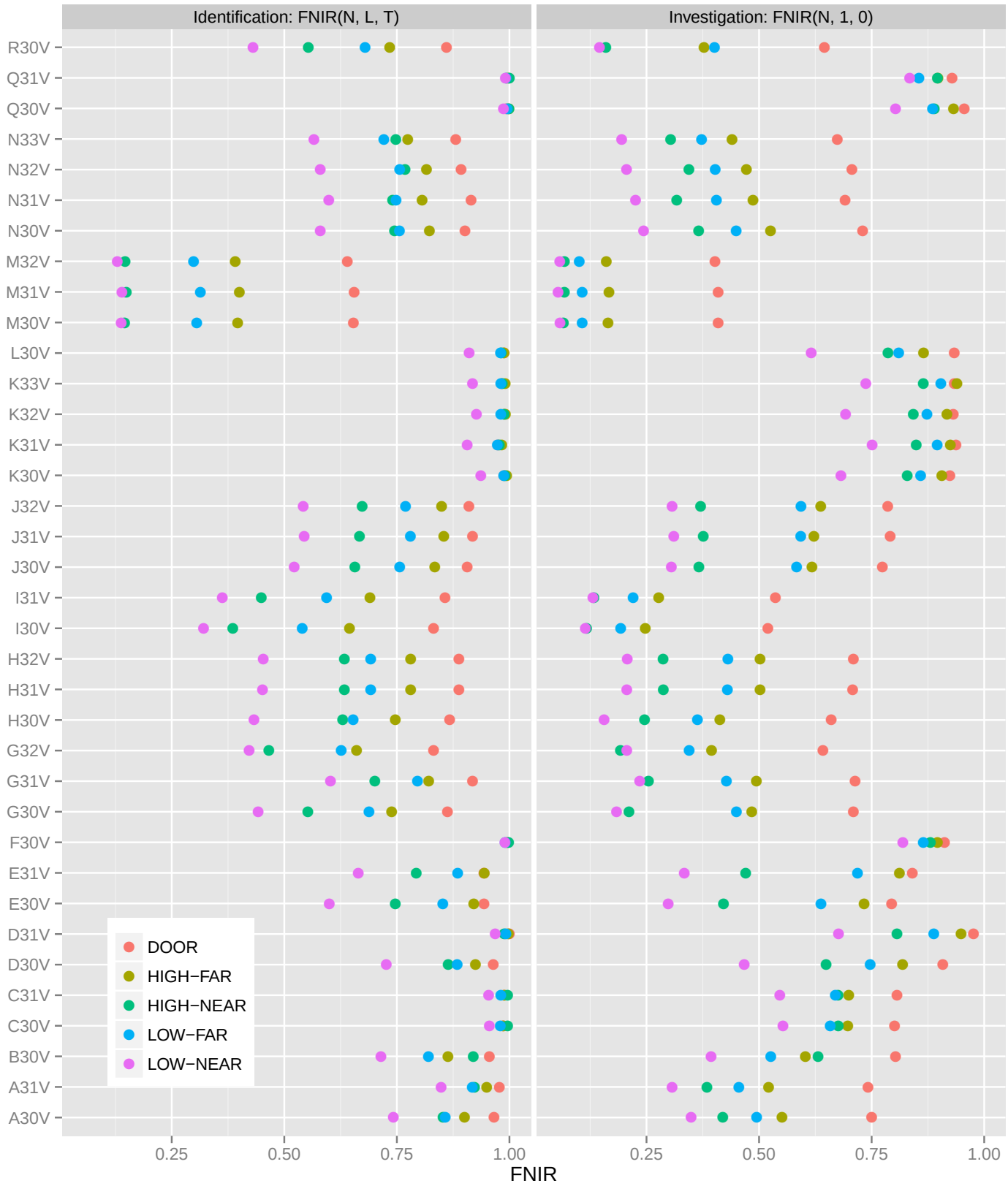


Figure 29: In the left panel, threshold-based FNIRs for each algorithm over the DATASET P: SPORTS ARENA *dataset with threshold set to yield NFP = 20). At right, is the investigation mode, rank 1, miss rates. The colored dots indicate the camera mounting height and range.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

N=480	FRAMES 3758691		IDENTIFICATION, FNIR(N, L, T)									INVESTIGATION, FNIR(N, 1, 0)					
	DETECTIONS			FAR FIELD OF VIEW			NEAR FIELD OF VIEW					FAR FIELD OF VIEW			NEAR FIELD OF VIEW		
ALG	COUNT	MIN^{-1}	T	HIGH 8FT		LOW 6FT	HIGH 8FT		LOW 6FT		DOOR 8FT	HIGH 8FT		LOW 6FT	HIGH 8FT	LOW 6FT	DOOR 8FT
A30V	530585	203	0.779	0.900	21	0.858 22	0.853	22	0.742	24	0.966 24	0.551	18	0.495 18	0.420 21	0.350 23	0.750 18
A31V	530585	203	0.685	0.949	25	0.918 25	0.922	25	0.848	25	0.978 26	0.521	16	0.455 17	0.385 20	0.307 19	0.742 17
B30V	95404	36	75.16	0.863	20	0.820 20	0.920	24	0.715	22	0.956 22	0.603	19	0.526 19	0.631 24	0.394 24	0.803 24
C30V	266896	102	56.74	0.986	27	0.980 27	0.996	33	0.956	32	0.986 28	0.697	23	0.658 24	0.676 27	0.553 27	0.801 23
C31V	330825	126	56.71	0.989	29	0.981 29	0.996	32	0.954	31	0.988 30	0.700	24	0.669 25	0.676 26	0.546 26	0.806 25
D30V	329042	126	0.858	0.925	23	0.884 23	0.864	23	0.727	23	0.964 23	0.819	27	0.747 27	0.649 25	0.467 25	0.908 27
D31V	134847	51	0.830	0.998	34	0.992 34	0.989	30	0.969	33	1.000 36	0.949	36	0.888 34	0.807 29	0.676 29	0.976 36
E30V	71679	27	3928	0.921	22	0.852 21	0.747	18	0.600	19	0.944 20	0.733	25	0.638 23	0.421 22	0.298 17	0.795 22
E31V	71713	27	4641	0.945	24	0.885 24	0.794	21	0.664	21	0.944 21	0.812	26	0.719 26	0.471 23	0.334 22	0.840 26
F30V	386103	148	0.480	0.995	33	0.991 33	0.999	34	0.990	35	0.995 35	0.896	29	0.865 31	0.881 34	0.820 35	0.912 28
G30V	36630	14	1716	0.738	8	0.688 9	0.552	7	0.442	9	0.862 8	0.484	11	0.450 16	0.211 8	0.184 8	0.710 13
G31V	74197	28	2094	0.821	15	0.796 19	0.702	15	0.603	20	0.919 19	0.494	13	0.428 12	0.255 10	0.235 15	0.713 15
G32V	74174	28	1682	0.661	5	0.627 6	0.466	6	0.422	6	0.832 4	0.394	7	0.345 6	0.193 7	0.206 11	0.642 6
H30V	85474	32	3132	0.747	9	0.653 7	0.630	9	0.433	8	0.868 9	0.413	8	0.363 7	0.246 9	0.156 7	0.660 8
H31V	85474	32	3159	0.781	11	0.692 10	0.633	10	0.452	10	0.888 11	0.502	15	0.430 13	0.287 12	0.206 12	0.708 12
H32V	85474	32	3159	0.781	12	0.692 11	0.633	11	0.454	11	0.888 12	0.502	14	0.431 14	0.287 11	0.207 13	0.710 14
I30V	161790	62	1280	0.645	4	0.540 4	0.386	4	0.321	4	0.832 5	0.247	4	0.193 4	0.117 4	0.114 4	0.520 4
I31V	161790	62	1577	0.690	6	0.594 5	0.449	5	0.363	5	0.857 6	0.277	5	0.221 5	0.133 5	0.131 5	0.537 5
J30V	30772	11	0.477	0.834	17	0.757 16	0.657	12	0.522	12	0.907 15	0.618	20	0.584 20	0.367 17	0.305 18	0.774 19
J31V	30772	11	0.475	0.854	19	0.780 18	0.667	13	0.545	14	0.919 18	0.622	21	0.593 21	0.376 19	0.310 21	0.791 21
J32V	30772	11	0.490	0.850	18	0.770 17	0.673	14	0.542	13	0.910 16	0.637	22	0.593 22	0.370 18	0.307 20	0.786 20
K30V	539503	206	0.715	0.994	32	0.988 32	0.991	31	0.937	30	0.988 31	0.906	31	0.859 30	0.829 30	0.682 30	0.924 29
K31V	87169	33	0.723	0.984	26	0.974 26	0.978	26	0.906	26	0.975 25	0.925	33	0.896 35	0.849 32	0.751 33	0.937 34
K32V	38863	14	0.647	0.991	31	0.981 31	0.988	29	0.927	29	0.989 32	0.918	32	0.873 32	0.843 31	0.692 31	0.932 31
K33V	27544	10	0.639	0.991	30	0.981 30	0.984	28	0.918	28	0.984 27	0.939	35	0.904 36	0.865 33	0.737 32	0.933 32
L30V	99667	38	0.588	0.988	28	0.981 28	0.982	27	0.911	27	0.988 29	0.866	28	0.810 28	0.786 28	0.616 28	0.934 33
M30V	143311	54	0.552	0.397	2	0.306 2	0.146	1	0.138	2	0.654 2	0.165	2	0.107 2	0.066 1	0.058 3	0.409 3
M31V	143311	54	0.552	0.401	3	0.314 3	0.149	3	0.140	3	0.655 3	0.167	3	0.107 3	0.068 3	0.054 1	0.409 2
M32V	143311	54	0.548	0.392	1	0.299 1	0.147	2	0.129	1	0.640 1	0.161	1	0.100 1	0.068 2	0.057 2	0.402 1
N30V	57592	22	0.607	0.823	16	0.756 14	0.745	17	0.580	17	0.902 14	0.526	17	0.449 15	0.366 16	0.244 16	0.730 16
N31V	57592	22	0.594	0.806	13	0.748 13	0.741	16	0.599	18	0.915 17	0.487	12	0.406 11	0.317 14	0.225 14	0.691 10
N32V	57592	22	0.605	0.816	14	0.757 15	0.768	20	0.580	16	0.893 13	0.472	10	0.403 10	0.344 15	0.206 10	0.706 11
N33V	57592	22	0.596	0.774	10	0.722 12	0.748	19	0.566	15	0.881 10	0.440	9	0.373 8	0.303 13	0.195 9	0.674 9
Q30V	207337	79	0.232	0.999	36	0.995 36	0.999	35	0.987	34	0.993 33	0.932	34	0.885 33	0.889 35	0.803 34	0.956 35
Q31V	207337	79	0.260	0.998	35	0.993 35	1.000	36	0.991	36	0.995 34	0.897	30	0.855 29	0.896 36	0.834 36	0.929 30
R30V	84117	32	0.672	0.735	7	0.680 8	0.554	8	0.431	7	0.861 7	0.378	6	0.401 9	0.160 6	0.146 6	0.645 7

Table 19: For DATASET P: SPORTS ARENA *, the accuracy values are FNIR "miss rates" for actors in video clips against a gallery of N = 480 individuals, by camera mounting height and algorithm. The miss rate, FNIR$(N, L, T200)$, a moderate-threshold surveillance metric where the threshold is set to the 200-th highest non-mate score for the gallery composed of frontal images. The right hand columns give FNIR$(N, 1, 0)$, i.e. rank-one miss rate. This metric is more appropriate for forensic applications. The metrics differ in that one requires hits to be strong, the other allows hits to be weak on the basis that a human reviewer will adjudicate candidates. The detection count in column 2 is summed over 237 minutes of video from each of 11 cameras operating at 24 frames per second. The detection rates in columns 3 are averages for one camera.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

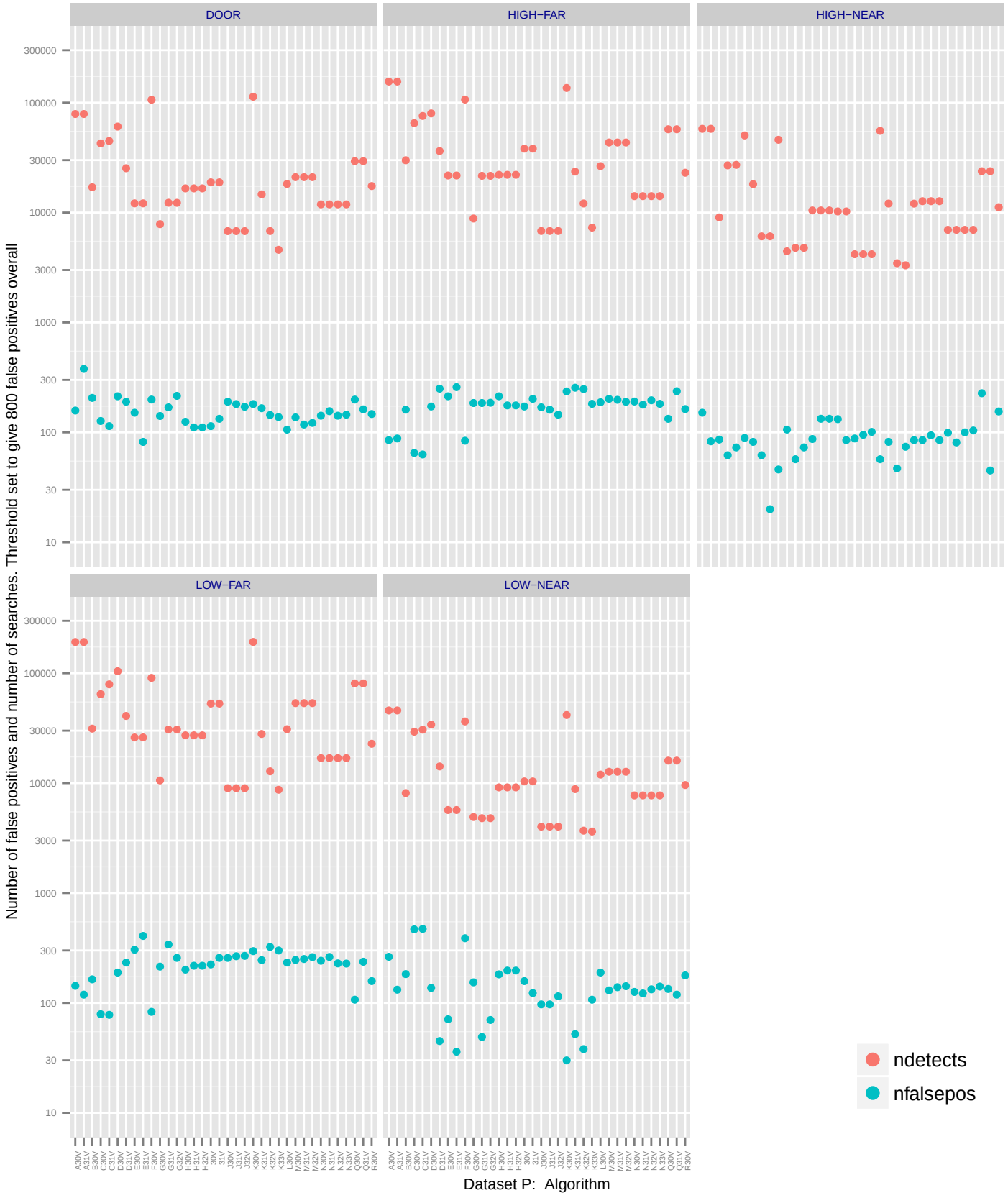


Figure 30: False positives and detection counts for the DATASET P: SPORTS ARENA *dataset with threshold set to give 800 false positives over all impostor searches. The panel shows results by camera mount location. Generally more people are detected in the door and "far" field views. The highest false positive counts occur here too, except at the doors, despite the detection counts, possibly due to steeper elevation angles.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

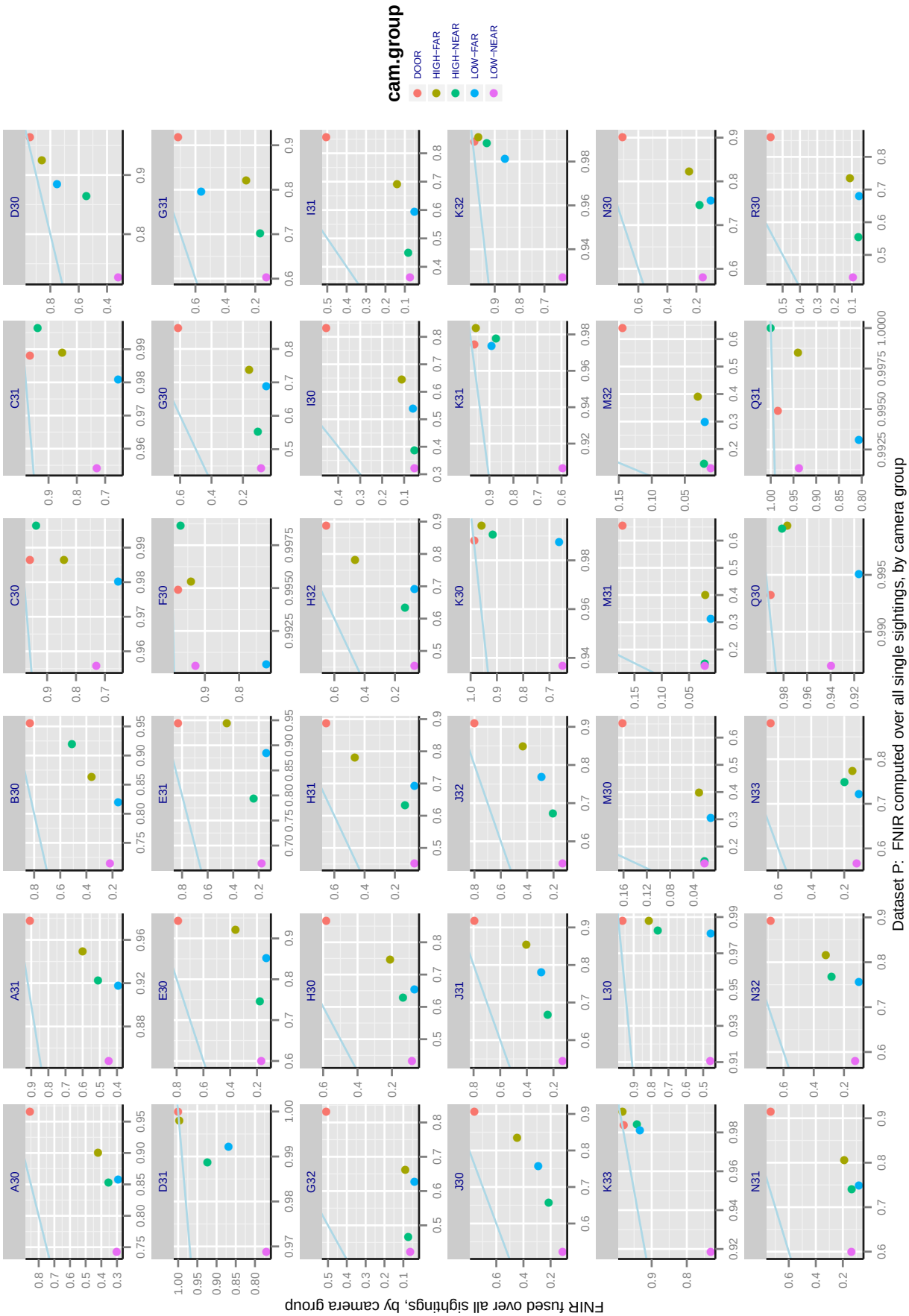


Figure 31: The effect of max-score fusion over time in DATASET P: SPORTS ARENA *. The panels show fused FNIR(N, L, T) against un-fused FNIR(N, L, T), with T fixed to give 200 false positives over all impostor sightings i.e. impostor sightings are not fused. By fusing scores over all sightings, the y-axis shows the rate at which actors are missed over an entire event. While this gives substantial accuracy gains (points far below the line), the operational relevance is reduced because fusion removes the real-time actionability of a correct detection and recognition. The error rates are more useful in a retroactive context to indicate how often a person of interest was missed at an event. This may have application in non-repudiation where, given sufficiently high score, a subject cannot claim to have not been in a particular building or location.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

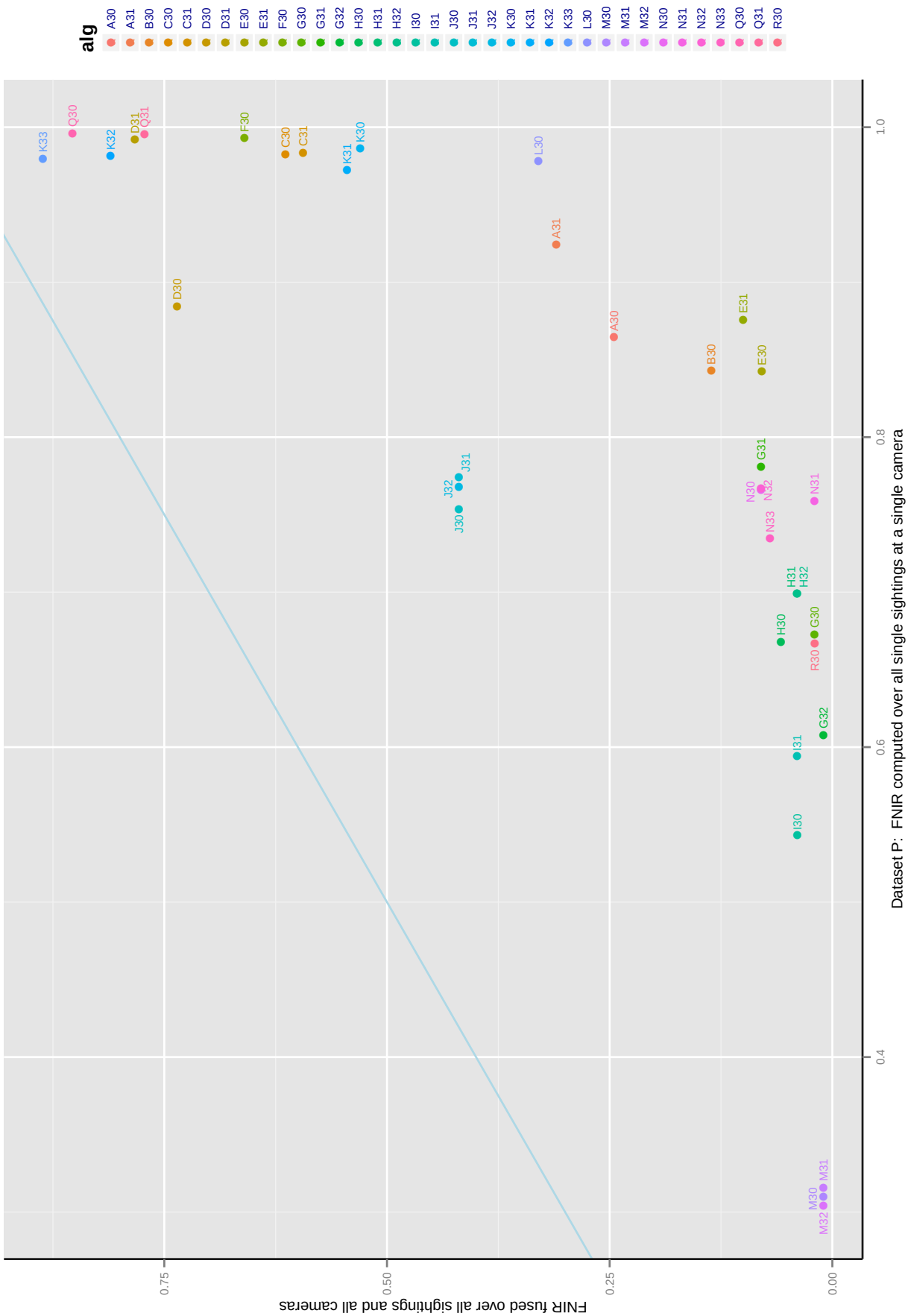


Figure 32: For DATASET P: SPORTS ARENA *the figure shows the effect of max-score fusion over space and time. The x-axis records FNIR(N, L, T) in the normal case, without fusion. The y-axis records FNIR(N, L, T) with fusion i.e. taking the best score from an actor from any camera and at point during the day of capture. The threshold is identical in both cases. Values below the diagonal line ($y = x$) indicate improvement, and this is clearly substantial.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

(-22,0) (0,0) (22,0)

Figure 33: DATASET P: SPORTS ARENA *: Examples of mugshot-like enrollment images, with nominal pose values (yaw, pitch) provided to the algorithm software. **The face images here are from a Dataset P. Written consent from DHS / S&T to use these images in public reports was obtained on August 12, 2016. Actor 011414X (Perm Granted).*

sees fit. For example, the literature describes methods for synthesizing a single representation from multiple views. This enables comparative performance analysis as follows.

Table 20 shows identification error rates for the "rich" gallery vs. the traditional single still. Figures 34 and 35 show the change in the rank and score of the mate, respectively. The notable observations are:

Value across providers: There are broad accuracy improvements realized by using additional enrollment images. In the investigative mode, rank one miss rates are improved for thirty five algorithms with only one algorithm, R30V, offering essentially unchanged performance. In the identification mode, FNIR(N, L, T) is more modestly improved with only N30V giving significantly worse accuracy.

Different value across providers: For some algorithms (providers K, Q, E, D, L) the use of three images per subject gives better mate rankings in many searches. For other algorithms the gains are confined to a smaller number of searches, particularly those from providers M, I, G, H, N. Rich enrollment gives substantial FNIR(N, 1, 0) error rate reductions: algorithms H31V and H32V give a fully 41% fewer errors than with a single full frontal image. For algorithm A31V the reduction is 37%, and for K32V, 32%.

The most accurate algorithms, M3xV, already place many mates at rank 1: For M30V, FNIR$(N, 1, 0)$ drops from 0.079 to 0.070 a gain of just 11%. This is evident in Figure 34 which shows the mean change in rank of the mate for M30V is -0.1, but is -7.7 for K31V.

More value in forensics: The accuracy gains are larger in forensics than in watch list surveillance, that is FNIR(N, R=1, T=0) reductions are larger than those for FNIR(N, R = L, T). Furthermore gains are better still if an investigator is able to review R = 20 candidates. Here some algorithms (e.g. from developers K, N, H, A, G, L, Q) give fewer than half as many errors, with K33V producing almost one quarter as many errors. This result implies that the effect of three-image enrollment is to improve the rank of the mate without greatly increasing its score - it produces more hits but the hits have modest scores. This is associated also with the two different thresholds for the two galleries. Except for algorithms G30V and G31V, the thresholds for three-image enrollment are higher than those for single-image enrollment, and this harms FNIR to preserve FPIR.

Growth in computational cost: As shown later in Table 35, computational cost increases about linearly in the number of images passed to the template generation function. This cost is incurred once, at enrollment time, and is amortized over searches. See section 6.4.

The benefits are substantial enough that it suggests traditional enrollment processes could be extended to capture additional views of a subject. While the mainline face recognition industry has grown around deduplication, identification

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

and verification of formally standardized[14] frontal images, the forensics community who, after all, are often left to adjudicate the possibly erroneous results emanating from a face recognition engine, see benefit in having alternative views of subjects' faces. Indeed the result here, that accuracy gains are more substantial in investigative searches (to rank 10 or higher) supports this conclusion. The exact details of what views afford the most benefit is a topic for additional research.

N=480	IDENTIFICATION, FNIR(N, L, T)				INVESTIGATION, FNIR(N, R, 0)					
	THRESHOLD		FNIR(T), NFP(T) = 200		FNIR(R = 1)		FNIR(R = 10)		FNIR(R = 20)	
	FF	FF QL QR	FF	FF QL QR	FF	FF QL QR	FF	FF QL QR	FF	FF QL QR
A30V	0.779	0.786	0.847	0.815	0.480	0.342	0.162	0.076	0.066	0.024
A31V	0.685	0.687	0.916	0.886	0.445	0.281	0.131	0.048	0.066	0.024
B30V	75.16	76.40	0.836	0.834	0.552	0.486	0.270	0.193	0.169	0.109
C30V	56.74	56.74	0.979	0.976	0.643	0.617	0.259	0.225	0.090	0.064
C31V	56.71	56.71	0.979	0.977	0.649	0.619	0.259	0.224	0.087	0.059
D30V	0.858	0.870	0.871	0.811	0.704	0.569	0.610	0.408	0.553	0.328
D31V	0.830	0.849	0.993	0.995	0.879	0.839	0.810	0.738	0.763	0.679
E30V	3928	3961	0.818	0.739	0.586	0.482	0.278	0.201	0.192	0.149
E31V	4641	4655	0.852	0.771	0.653	0.524	0.318	0.194	0.219	0.124
F30V	0.480	0.481	0.992	0.992	0.849	0.835	0.500	0.445	0.267	0.205
G30V	1715	1569	0.657	0.519	0.348	0.253	0.192	0.097	0.153	0.067
G31V	2093	2051	0.761	0.629	0.375	0.309	0.171	0.110	0.129	0.081
G32V	1682	1812	0.588	0.609	0.291	0.231	0.137	0.093	0.104	0.064
H30V	3132	3151	0.601	0.527	0.302	0.226	0.116	0.085	0.080	0.060
H31V	3159	3161	0.655	0.528	0.383	0.227	0.183	0.085	0.137	0.060
H32V	3159	3161	0.655	0.528	0.383	0.227	0.183	0.086	0.137	0.061
I30V	1279	1290	0.456	0.409	0.167	0.127	0.054	0.040	0.034	0.029
I31V	1577	1584	0.501	0.436	0.199	0.145	0.061	0.040	0.033	0.025
J30V	0.477	0.507	0.728	0.640	0.491	0.390	0.328	0.245	0.278	0.204
J31V	0.475	0.492	0.754	0.663	0.505	0.416	0.342	0.278	0.298	0.233
J32V	0.490	0.518	0.738	0.649	0.494	0.391	0.327	0.256	0.276	0.210
K30V	0.715	0.720	0.975	0.933	0.792	0.536	0.564	0.245	0.416	0.163
K31V	0.723	0.729	0.950	0.873	0.834	0.584	0.678	0.279	0.593	0.170
K32V	0.647	0.657	0.965	0.902	0.812	0.555	0.643	0.280	0.540	0.191
K33V	0.639	0.649	0.962	0.898	0.833	0.604	0.691	0.319	0.610	0.221
L30V	0.588	0.798	0.972	0.963	0.785	0.590	0.407	0.177	0.257	0.041
M30V	0.552	0.554	0.235	0.202	0.079	0.070	0.036	0.031	0.026	0.022
M31V	0.552	0.555	0.242	0.210	0.080	0.069	0.033	0.030	0.025	0.022
M32V	0.548	0.550	0.236	0.204	0.081	0.070	0.033	0.027	0.024	0.020
N30V	0.607	0.621	0.734	0.833	0.382	0.364	0.211	0.113	0.159	0.065
N31V	0.594	0.594	0.727	0.678	0.342	0.271	0.175	0.124	0.136	0.088
N32V	0.605	0.608	0.725	0.650	0.351	0.225	0.187	0.097	0.147	0.066
N33V	0.596	0.596	0.706	0.653	0.325	0.259	0.174	0.121	0.140	0.088
Q30V	0.232	0.232	0.990	0.986	0.834	0.753	0.487	0.303	0.274	0.120
Q31V	0.260	0.261	0.992	0.991	0.838	0.777	0.440	0.328	0.239	0.134
R30V	0.672	0.673	0.601	0.611	0.272	0.279	0.119	0.121	0.082	0.079

Table 20: For DATASET P: SPORTS ARENA *, the accuracy values are FNIR "miss rates" for actors present in a gallery of individuals all of whom enrolled with a single full-frontal (FF) image, or in a separate gallery where individuals are enrolled with three images: one full frontal, and one "quarter left" (QL) and one "quarter right" (QR) as in Figure 33. Both galleries hold n = 480 individuals. The actor video clips are approximately one quarter of the full Dataset P set. The impostor videos are the full set as used in the other Dataset P tables. Cells are shaded red when the richer gallery increases error, and shaded progressively more green when the fractional reduction in FNIR is better than* $\{0.8, 0.67, 0.5\}$*. Note the higher thresholds in columns 3 are needed to limit the number of false positives to the same number, 200, over all searches from the non-actor video clips.*

[14] See ISO/IEC 19794-5:2005, and ICAO's Portrait Quality specification [34].

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

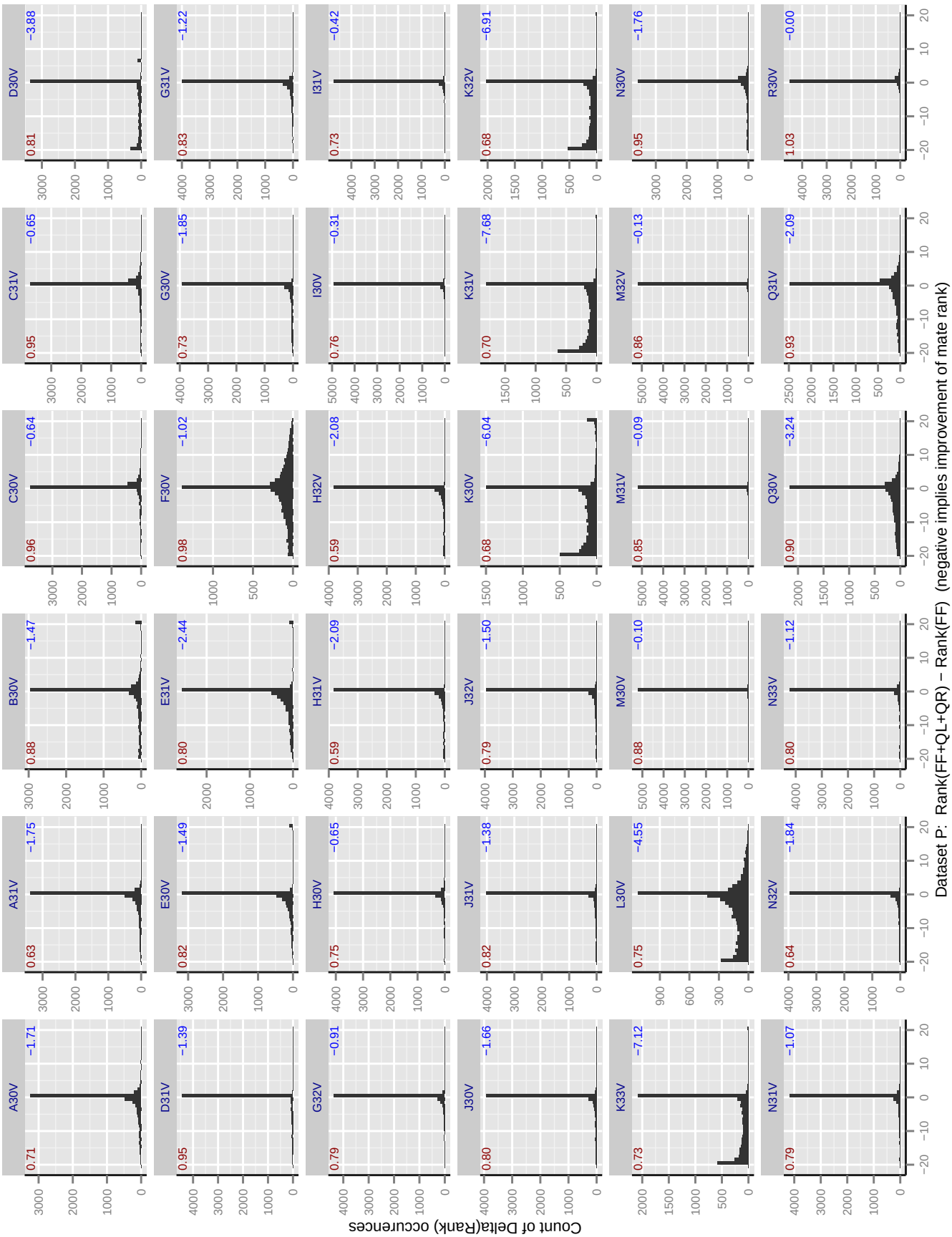


Figure 34: The figures show the effect of enrolling three images per person in the gallery vs. just one full frontal image. The three images vary in yaw angle of the face relative to the cameras: They are quarter-left, quarter-right, and full frontal. Each panel shows the histogram of the change in rank of the correct mate so negative values indicate that a search yielded an improvement in the rank e.g. from rank 5 to rank 1 would be a gain of 4. If a mate was not on the candidate list at all then rank was set to L+1, with L the fixed candidate list length, L = 20. The value appearing in red text is the ratio of rank one hit rates, i.e. FNIR(N, R=1, T=0) for the three image enrollment divided by that for one image only. Values below 1 indicate improvement. The blue text at top right is the the mean difference in ranks.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

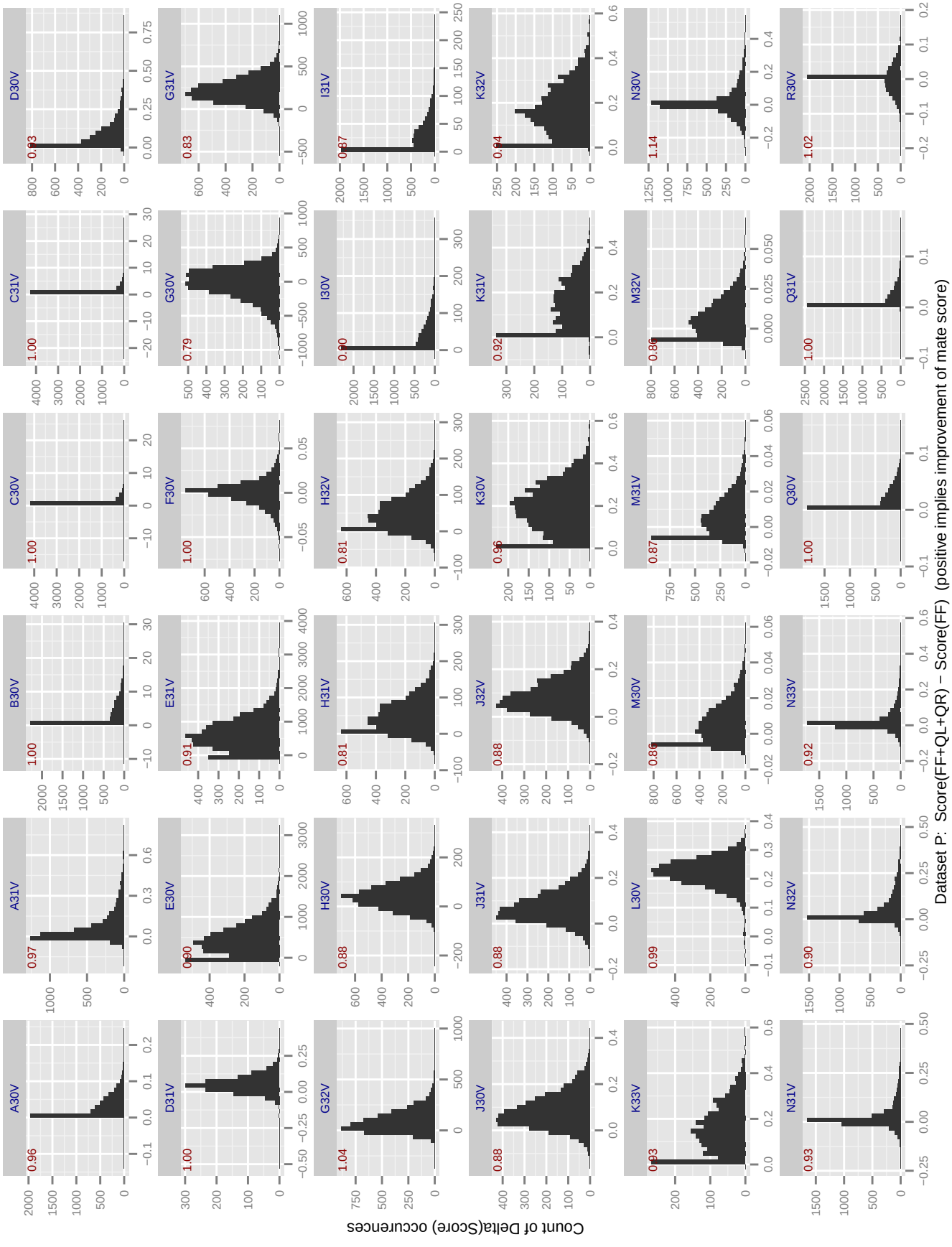


Figure 35: The figures show the effect of enrolling three images per person in the gallery vs. just one full frontal image. The three images vary in yaw angle of the face relative to the cameras: They are quarter-left, quarter-right, and full frontal. Each panel shows the histogram of the change in similarity score of the correct mate so positive values indicate that a search yielded an improvement in the score The histogram includes only those searches for which the mate was present on both candidate lists. The value appearing in red text is ratio of moderate-threshold miss rates, i.e. FNIR(N, R=L, T) for the three image enrollment divided by that for one image only. The threshold is set to give 200 false positives over all impostor searches. The threshold is set for the two galleries separately. Values below 1 indicate improvement. The FNIR ratio computation includes cases where mates were absent from candidate lists - these are represented with a score of 0.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

5.6 DATASET H: TRAVEL WALKWAY

Overview: This dataset is composed of videos collected using professional grade cameras mounted on the ceiling in a busy travel concourse. The cameras are mounted and configured specifically for the purpose of video surveillance and face recognition. The cameras are placed in three locations, in banks of 4, 4 and 2 respectively. Within any one bank, the cameras are mounted next to each other, transverse to their optical axes, which are parallel. Subjects usually walk towards and underneath the cameras. Occasionally, individuals walk in various transverse directions, including away from the camera. The cameras are visible to the subjects but are almost always ignored. Videos are comprised of clips known to contain actors and clips known not to contain actors. In all cases there are multiple faces visible in the clip. The amount of time a face is fully visible in a scene can vary from approximately 0 to 5 seconds. Key imaging properties for this dataset are summarized in Table 21.

Property	Camera bank F	Camera bank R	Camera bank S
Cameras	Avigilon 2.0MP-HD-H264-B1		
Image width	1920		
Image height	1080		
Frame rate	30 sec^{-1}		
Nominal bitrate	24 Mbit sec^{-1}		
Codec	H264		
Camera mounting	Ceiling		
Number of cameras	4	2	4
Max. subjects in view of all cams	20	14	25
Camera height (meters)	2.65	2.3	2.10
Range to subject (meters)	[2,7]	[3,8]	[2,4]
Camera elevation to frontal face (degrees)	[28,8]	[12,4]	[12,6]

Table 21: Key imaging properties for DATASET H: TRAVEL WALKWAY *results.*

Enrolled still images: Enrollment images are mugshot-like photos collected under controlled lighting, background, and pose conditions.

Videos: The DATASET H: TRAVEL WALKWAY videos were collected from ten cameras placed on ceilings in a public facility similar to a passenger terminal. The cameras are more expensive and of higher quality than those used for DATASET P: SPORTS ARENA . The videos are also subject to less compression.

Experimental design: Mated scores are computed by searching 439 video clips against three enrolled dataset of portrait face images of subjects known to be in the search videos. The size of the enrollment datasets are $N = \{480, 4800, 48000\}$. The enrollment database is extended to these sizes by adding high quality frontal portrait photographs from a disjoint background population. **Nonmated scores** are produced by searching the same videos against the three *global nonmated enrollment datasets*, of the same N values.

Key experimental design details are summarized in Table 22.

Results: The results are presented in four tables and one figure.

Tables 23, 24 25 give detection counts and recognition accuracy results for enrolled gallery sizes of N = 480, 4800, and 48000, respectively. The tables report both identification mode FNIR(N, L, T) and investigation mode FNIR$(N, R, 0)$, for, three thresholds and three ranks of interest, respectively. Note that in high flow, high volume surveillance application, where it will be necessary to minimize false positives, the FNIR(N, L, T) metric is more relevant than FNIR$(N, R, 0)$.

Table 26 shows accuracy comparing the three camera banks, corresponding to three different imaging locations and geometries.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Quantity	Value or description
Mode	Video search to still enrollment
Number of actors	56
Number of non-actors	Many
Number of cameras	10, all ceiling mounted
Video duration with actors	2883 mins
Video duration no actors	8093 mins (but not used)
Number of clips actors	439
Number of clips no actors	0
Subject motion	Mostly toward and parallel to optical axis.
Clip duration (frames)	4800 fixed
Frame rate (s^{-1})	30
Number of enrolled subjects	480, 4800, 48000
Number of enrolled stills	1 per subject
Hroperties of enrolled stills	Frontal, close ICAO compliance; Mean IOD 124 pixels
FNIR estimation	Actors present video vs. enrolled gallery
FPIR estimation	Actors present video vs. separate non-actor gallery
Candidate list length	20
Number of persons in FOV	[0,5] typical across FOV
Video ground truth	Style B: See Figure 6

Table 22: Key experimental design for the DATASET H: TRAVEL WALKWAY *results.*

Figure 36 shows how FNIR and NFP vary at fixed threshold when gallery size changes.

Notable results are as follows.

Detection: Algorithms vary by almost a factor of 20 in the total number of detections they report. Thus, given nearly 20 hours of video, from 10 cameras, the number of detections varies from around 8 per minute to nearly 150 per minute. This variation exists within the K3xV variants and to a slightly lesser extent across the entire set of algorithms. The reasons for this are to do with:

1. Algorithmic false negatives from failed face detection;
2. Algorithmic false positives from non-faces being reported as faces;
3. Tracking integrity in which algorithms may lose track of an individual over time;
4. Detection policy - minimum spatial resolutions at which to accept a face for processing (see Figure 23)
5. Algorithms may legitimately choose to break a person's track into several parts and generate templates from each - see Figure 3.

While the last two of these are under the control of the algorithm designer, the first three are not readily so. Thus it appears that face detection remains a non-trivial task with a diversity of approaches [13]. We would like to produce a "verbosity" index by normalizing the detection counts by the number of faces actually present. However, per the discussion in section 4.1, that number is unknown and unknowable. Without any assumptions we can only observe the greatly varying numbers of detections. However, in Figure 25, we address this issue using a different dataset for which we do know the actual numbers of people present. There the detection verbosities vary widely too, and increase markedly for subjects who are standing still.

Absolute miss rates: The M3xV algorithms give the lowest FNIR(N, L, T) values, for $N = 480, 4800, 48000$. This holds for three operating thresholds corresponding to false positive counts of NFP = $10, 100, 1000$. There is a very large range in accuracy across the 36 algorithms evaluated - this is typical in independent biometric evaluations.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

The most accurate algorithms give FNIR$(48000, L, T)$, NFP$(T) = 100$, as low as 0.314 (M32V) and 0.599 (G32V) both of which represent useful numbers of identifications in an operational context.

Effect of population size: For the M3xV algorithms, the error rates for NFP$(T) = 10$ climb very slowly from N = 480 to 4 800, but more substantially by N = 48 000. This behavior is exhibited by all algorithms, and thus it becomes imperative, even in a high-end purpose-built video surveillance installation equipped with ISO-standard enrolled face images, to maintain N as small as possible. This should be done by establishing a curation function that limits the number of individuals enrolled, and also removes poor quality imagery.

Comparing rank-based and threshold-based accuracy: The general miss rate metric, FNIR(N, R, T) is the proportion of actors returned outside the top R ranks or below threshold T. As with all other datasets, the rank-based investigational accuracy FNIR$(N, 1, 0)$ is better than the threshold-based identification metric FNIR(N, L, T) essentially because it allows rank-based hits to be "weak", i.e. they are at rank 1, but with a score that is below the high thresholds that are required in surveillance applications. The rank-based metric is useful in "forensic" style searches where there is adequate human labor available to adjudicate candidates produced in manageable low daily volume of searches. Many of the algorithms will produce valuable hits even with N = 48 000. For example, the 18th most accurate algorithm has FNIR just less than twice that of the most accurate algorithm.

Accuracy relative to Dataset P: The two datasets H and P differ primarily in terms of the expense and properties of the imaging systems, and secondarily in terms of the lighting environment. The interocular distances (see the means tabulated in Appendix I) are higher for H than P, due to lens configuration and narrower field of view. The accuracy values behave accordingly. Rank 1 accuracy, FNIR$(480, 1, 0)$, for dataset H are two or three times lower than dataset P except for the M3xV and I3xV algorithms which work about as well.

Reasoning about scaling: A primary concern when operating a biometric identification system is about setting the threshold correctly as enrolled population size grows. Typically it is necessary to raise the threshold to maintain a fixed rate of false positives. Practitioners often conceive of false positive outcomes increasing linearly with enrolled population size. From binomial theory, with a fixed threshold T, FPIR(T) = N FMR(T) where FMR is a fixed one-to-one false match rate. Also FNIR is usually considered to be independent of N, at fixed T, at least if the genuine score computation doesn't depend on the other gallery entries. Figure 36 shows simple binomial theory does not hold. This is not unexpected as some biometric identification systems do not compute the N scores independently of one another. As such, this consideration of binomial models is therefore naive and moot. However, as is evident in Figure 36 the number of false positives, NFP, does scale linearly with N for most algorithms, the exceptions being those from developers G, H, M, N. In all cases FNIR varies with N, at fixed threshold. In several cases accuracy with small galleries, $N = 480$ is inferior to that with larger N. The conclusion should be that system owners should monitor false positive rates, particularly as N changes, perhaps by embedding tests into the operational system.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

N=480	NUM ACTORS 48		NUM FEEDS 41		NUM CLIPS 439				NUM FRAMES 2107200		NUM MINUTES 1170.7	
	DETECTIONS		THRESHOLD BASED IDENTIFICATION						RANK BASED INVESTIGATIONS			
ALG	NUM	MIN^{-1}	FNIR(T), NFP(T)=10		FNIR(T), NFP(T)=100		FNIR(T), NFP(T)=1000		FNIR(R=1, T=0)		FNIR(R=5, T=0)		FNIR(R=20, T=0)	
A30V	81195	69.4	0.689	29	0.541	28	0.396	25	0.230	20	0.173	14	0.133	11
A31V	81195	69.4	0.650	28	0.523	27	0.388	24	0.218	18	0.166	13	0.133	10
B30V	16585	14.2	0.630	27	0.495	24	0.375	22	0.222	19	0.187	16	0.142	14
C30V	53549	45.7	0.827	31	0.759	31	0.641	32	0.324	29	0.277	26	0.240	24
C31V	50980	43.5	0.840	32	0.782	32	0.665	33	0.369	30	0.323	30	0.281	31
D30V	136909	116.9	0.351	12	0.329	16	0.317	18	0.243	21	0.234	22	0.185	21
D31V	37923	32.4	0.609	26	0.579	29	0.556	30	0.310	27	0.281	28	0.224	23
E30V	10323	8.8	0.511	23	0.391	19	0.326	19	0.286	23	0.265	23	0.243	25
E31V	10323	8.8	0.471	20	0.396	20	0.336	20	0.287	24	0.277	27	0.249	28
F30V	89883	76.8	0.988	35	0.966	36	0.907	36	0.671	36	0.559	36	0.376	36
G30V	11746	10.0	0.449	19	0.339	17	0.268	15	0.213	17	0.201	20	0.185	22
G31V	11514	9.8	0.421	16	0.397	21	0.353	21	0.299	26	0.273	25	0.246	27
G32V	11075	9.5	0.342	11	0.327	15	0.317	17	0.313	28	0.296	29	0.273	30
H30V	15142	12.9	1.000	36	0.964	35	0.619	31	0.161	8	0.142	5	0.123	3
H31V	15142	12.9	0.298	8	0.256	7	0.201	5	0.160	6	0.144	6	0.123	1
H32V	15142	12.9	0.298	9	0.256	8	0.201	6	0.160	7	0.144	7	0.123	2
I30V	14660	12.5	0.289	6	0.240	4	0.197	4	0.153	4	0.141	3	0.132	8
I31V	14660	12.5	0.310	10	0.264	10	0.207	7	0.154	5	0.144	8	0.126	6
J30V	12459	10.6	0.284	4	0.252	5	0.230	10	0.203	14	0.193	19	0.163	18
J31V	12459	10.6	0.295	7	0.255	6	0.230	9	0.210	16	0.191	17	0.164	19
J32V	12459	10.6	0.287	5	0.261	9	0.234	14	0.209	15	0.191	18	0.161	17
K30V	172566	147.4	0.584	25	0.521	26	0.467	28	0.250	22	0.225	21	0.181	20
K31V	10974	9.4	0.476	21	0.430	22	0.378	23	0.293	25	0.271	24	0.246	26
K32V	12178	10.4	0.550	24	0.498	25	0.434	27	0.385	32	0.363	32	0.332	33
K33V	9670	8.3	0.505	22	0.456	23	0.409	26	0.379	31	0.364	33	0.347	34
L30V	21112	18.0	0.748	30	0.652	30	0.544	29	0.400	33	0.335	31	0.259	29
M30V	17334	14.8	0.206	2	0.188	2	0.173	3	0.145	1	0.141	4	0.132	9
M31V	17334	14.8	0.207	3	0.191	3	0.170	1	0.148	3	0.138	1	0.130	7
M32V	17334	14.8	0.201	1	0.185	1	0.172	2	0.148	2	0.139	2	0.124	4
N30V	14594	12.5	0.410	15	0.292	11	0.234	13	0.178	12	0.159	12	0.142	15
N31V	14594	12.5	0.427	17	0.316	14	0.234	11	0.178	11	0.156	11	0.136	13
N32V	14594	12.5	0.396	14	0.311	13	0.234	12	0.172	10	0.154	10	0.135	12
N33V	14594	12.5	0.387	13	0.302	12	0.227	8	0.169	9	0.148	9	0.124	5
Q30V	30436	26.0	0.972	34	0.954	34	0.884	35	0.633	35	0.520	35	0.375	35
Q31V	30436	26.0	0.941	33	0.889	33	0.761	34	0.530	34	0.407	34	0.281	32
R30V	30257	25.8	0.441	18	0.356	18	0.292	16	0.193	13	0.173	15	0.148	16

Table 23: For the DATASET H: TRAVEL WALKWAY *installation, camera bank all, with 480 subjects enrolled with a frontal still, the values are identification-mode FNIR(N, L, T) for each algorithm at three different decision thresholdscorresponding to false positive counts of 10, 100, 1000, and investigation-mode FNIR$(N, R, 0)$ for ranks 1, 5, 20. Each value is accompanied by an integer ranking across all algorithms. The shading indicates the most important metric to watchlist applications. Ten cameras were used. The detections are summed over all of them. The accuracy values are aggregated over all sightings of all subjects in the field of view of those cameras. Caution: The R=20 column is unreliable per the arguments in section 4.4.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

N=4800	NUM ACTORS 48		NUM FEEDS 41		NUM CLIPS 439		NUM FRAMES 2107200		NUM MINUTES 1170.7	
	DETECTIONS		THRESHOLD BASED IDENTIFICATION				RANK BASED INVESTIGATIONS			
ALG	NUM	MIN^{-1}	FNIR(T), NFP(T)=10		FNIR(T), NFP(T)=100		FNIR(T), NFP(T)=1000		FNIR(R=1, T=0)		FNIR(R=5, T=0)		FNIR(R=20, T=0)	
A30V	81195	69.4	0.735	27	0.639	25	0.510	26	0.290	21	0.252	21	0.194	15
A31V	81195	69.4	0.781	28	0.643	26	0.496	25	0.289	20	0.246	19	0.194	14
B30V	16585	14.2	0.656	24	0.567	23	0.459	22	0.298	22	0.247	20	0.201	16
C30V	57311	49.0	0.893	32	0.840	32	0.735	33	0.375	28	0.336	28	0.277	25
C31V	54744	46.8	0.905	33	0.843	33	0.735	32	0.393	30	0.350	29	0.290	28
D30V	136909	116.9	0.496	17	0.415	19	0.350	19	0.267	18	0.241	18	0.221	21
D31V	37923	32.4	0.708	25	0.658	27	0.593	28	0.342	26	0.302	26	0.287	27
E30V	10323	8.8	0.855	30	0.690	29	0.422	21	0.304	23	0.296	24	0.277	24
E31V	10323	8.8	0.889	31	0.834	31	0.677	31	0.319	25	0.299	25	0.286	26
F30V	89883	76.8	0.994	35	0.987	35	0.948	36	0.756	35	0.655	35	0.544	35
G30V	11746	10.0	0.508	19	0.387	17	0.290	16	0.231	16	0.222	17	0.207	19
G31V	11514	9.8	0.498	18	0.406	18	0.359	20	0.307	24	0.274	23	0.252	23
G32V	11075	9.5	0.388	12	0.353	16	0.321	17	0.277	19	0.261	22	0.243	22
H30V	15142	12.9	0.320	4	0.271	4	0.225	4	0.182	8	0.169	8	0.159	8
H31V	15142	12.9	0.323	5	0.274	5	0.233	6	0.182	6	0.169	6	0.159	6
H32V	15142	12.9	0.323	6	0.274	6	0.233	7	0.182	7	0.169	7	0.159	7
I30V	14660	12.5	0.382	11	0.307	9	0.227	5	0.166	4	0.157	4	0.148	4
I31V	14660	12.5	0.418	13	0.314	14	0.249	10	0.169	5	0.157	5	0.150	5
J30V	12459	10.6	0.357	7	0.292	7	0.252	12	0.230	15	0.219	16	0.203	17
J31V	12459	10.6	0.363	8	0.295	8	0.252	11	0.225	14	0.219	15	0.204	18
J32V	12459	10.6	0.379	10	0.308	11	0.261	15	0.234	17	0.216	14	0.213	20
K30V	172566	147.4	0.717	26	0.681	28	0.634	29	0.381	29	0.357	30	0.319	30
K31V	10974	9.4	0.607	22	0.556	21	0.490	24	0.356	27	0.329	27	0.308	29
K32V	12178	10.4	0.640	23	0.603	24	0.533	27	0.430	32	0.410	32	0.385	33
K33V	9670	8.3	0.601	21	0.556	22	0.489	23	0.419	31	0.403	31	0.376	31
L30V	21112	18.0	0.809	29	0.736	30	0.639	30	0.489	33	0.440	33	0.381	32
M30V	17334	14.8	0.209	3	0.194	2	0.182	3	0.157	1	0.153	3	0.144	1
M31V	17334	14.8	0.209	2	0.200	3	0.181	2	0.160	3	0.153	2	0.145	3
M32V	17334	14.8	0.209	1	0.191	1	0.178	1	0.160	2	0.153	1	0.145	2
N30V	14594	12.5	0.418	14	0.310	12	0.252	13	0.188	11	0.181	11	0.164	11
N31V	14594	12.5	0.455	16	0.323	15	0.255	14	0.196	12	0.182	12	0.169	12
N32V	14594	12.5	0.379	9	0.307	10	0.246	9	0.182	9	0.181	10	0.164	10
N33V	14594	12.5	0.430	15	0.313	13	0.239	8	0.185	10	0.176	9	0.160	9
Q30V	30436	26.0	1.000	36	1.000	36	0.947	35	0.761	36	0.680	36	0.578	36
Q31V	30436	26.0	0.973	34	0.930	34	0.849	34	0.637	34	0.570	34	0.461	34
R30V	30068	25.7	0.563	20	0.428	20	0.338	18	0.222	13	0.210	13	0.185	13

Table 24: For the DATASET H: TRAVEL WALKWAY *installation, camera bank all, with 4800 subjects enrolled with a frontal still, the values are identification-mode FNIR(N, L, T) for each algorithm at three different decision thresholdscorresponding to false positive counts of 10, 100, 1000, and investigation-mode FNIR$(N, R, 0)$ for ranks 1, 5, 20. Each value is accompanied by an integer ranking across all algorithms. The shading indicates the most important metric to watchlist applications. Ten cameras were used. The detections are summed over all of them. The accuracy values are aggregated over all sightings of all subjects in the field of view of those cameras.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

N=48000	NUM ACTORS 48		NUM FEEDS 41		NUM CLIPS 439		NUM FRAMES 2107200		NUM MINUTES 1170.7	
	DETECTIONS		THRESHOLD BASED IDENTIFICATION				RANK BASED INVESTIGATIONS			
ALG	NUM	MIN^{-1}	FNIR(T), NFP(T)=10		FNIR(T), NFP(T)=100		FNIR(T), NFP(T)=1000		FNIR(R=1, T=0)		FNIR(R=5, T=0)		FNIR(R=20, T=0)	
A30V	81195	69.4	0.960	27	0.921	27	0.868	28	0.744	30	0.674	30	0.610	30
A31V	81195	69.4	1.000	34	0.994	33	0.967	33	0.733	29	0.667	29	0.610	29
B30V	16585	14.2	0.930	26	0.907	26	0.843	26	0.701	28	0.628	28	0.554	27
C30V	59390	50.7	0.996	30	0.987	32	0.960	31	0.821	32	0.796	32	0.761	32
C31V	56856	48.6	0.999	31	0.987	31	0.961	32	0.841	33	0.815	33	0.791	33
D30V	136909	116.9	0.828	17	0.724	16	0.587	12	0.416	10	0.379	10	0.361	11
D31V	37923	32.4	0.911	25	0.819	22	0.741	23	0.450	11	0.412	12	0.394	15
E30V	10323	8.8	0.999	32	0.973	29	0.612	13	0.412	9	0.370	9	0.327	9
E31V	10323	8.8	0.990	29	0.973	30	0.871	29	0.526	20	0.443	15	0.354	10
F30V	89883	76.8	1.000	36	0.999	35	0.996	35	0.956	35	0.936	36	0.917	36
G30V	11746	10.0	0.791	11	0.716	13	0.625	16	0.481	16	0.452	17	0.427	19
G31V	11514	9.8	0.710	5	0.613	8	0.545	9	0.470	14	0.436	14	0.385	14
G32V	11075	9.5	0.670	4	0.599	4	0.532	8	0.476	15	0.450	16	0.409	17
H30V	15142	12.9	0.763	10	0.604	5	0.434	4	0.289	6	0.267	5	0.244	4
H31V	15142	12.9	0.763	8	0.612	6	0.443	5	0.289	4	0.267	4	0.244	2
H32V	15142	12.9	0.763	9	0.612	7	0.443	6	0.289	5	0.268	6	0.244	3
I30V	14660	12.5	0.739	6	0.656	10	0.572	10	0.502	18	0.492	21	0.480	24
I31V	14660	12.5	0.884	22	0.821	23	0.701	21	0.545	23	0.501	23	0.481	25
J30V	12459	10.6	0.828	18	0.757	18	0.659	19	0.532	21	0.495	22	0.447	22
J31V	12459	10.6	0.827	16	0.753	17	0.646	18	0.538	22	0.489	20	0.434	20
J32V	12459	10.6	0.796	13	0.719	15	0.628	17	0.489	17	0.459	18	0.406	16
K30V	172566	147.4	0.898	24	0.879	25	0.853	27	0.601	26	0.563	26	0.511	26
K31V	10974	9.4	0.846	20	0.813	21	0.756	24	0.597	25	0.542	25	0.447	21
K32V	12178	10.4	0.887	23	0.846	24	0.776	25	0.636	27	0.606	27	0.569	28
K33V	9670	8.3	0.841	19	0.787	19	0.701	20	0.564	24	0.517	24	0.470	23
L30V	21112	18.0	0.966	28	0.945	28	0.899	30	0.813	31	0.767	31	0.704	31
M30V	17334	14.8	0.350	2	0.317	3	0.292	3	0.259	3	0.252	2	0.246	6
M31V	17334	14.8	0.347	1	0.314	2	0.292	2	0.258	2	0.255	3	0.246	5
M32V	17334	14.8	0.353	3	0.314	1	0.280	1	0.258	1	0.249	1	0.243	1
N30V	14594	12.5	0.757	7	0.641	9	0.514	7	0.366	7	0.336	7	0.298	7
N31V	14594	12.5	0.803	15	0.719	14	0.612	14	0.452	13	0.413	13	0.372	13
N32V	14594	12.5	0.794	12	0.689	11	0.575	11	0.397	8	0.344	8	0.299	8
N33V	14594	12.5	0.801	14	0.713	12	0.616	15	0.452	12	0.410	11	0.366	12
Q30V	30436	26.0	1.000	33	1.000	36	1.000	36	0.957	36	0.930	34	0.898	34
Q31V	30436	26.0	1.000	35	0.997	34	0.990	34	0.941	34	0.930	35	0.905	35
R30V	29040	24.8	0.881	21	0.810	20	0.708	22	0.507	19	0.461	19	0.409	18

Table 25: For the DATASET H: TRAVEL WALKWAY *installation, camera bank all, with 48000 subjects enrolled with a frontal still, the values are identification-mode FNIR(N, L, T) for each algorithm at three different decision thresholdscorresponding to false positive counts of 10, 100, 1000, and investigation-mode FNIR$(N, R, 0)$ for ranks 1, 5, 20. Each value is accompanied by an integer ranking across all algorithms. The shading indicates the most important metric to watchlist applications. Ten cameras were used. The detections are summed over all of them. The accuracy values are aggregated over all sightings of all subjects in the field of view of those cameras.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

N=4800	THRESHOLD	NUM ACTORS 48		NUM FEEDS 41		NUM FRAMES ALL 2107200				NUM FRAMES F 456000				NUM FRAMES R 571200				NUM FRAMES S 1080000			
		DETECTION COUNTS				THRESHOLD BASED IDENTIFICATION								RANK BASED INVESTIGATION CASES							
	ALL					FNIR(T) WITH T SET FOR NFP=10								FNIR(R) WITH RANK R=1							
ALG		ALL	F	R	S	ALL		F		R		S		ALL		F		R		S	
A30V	0.856	81195	15034	27730	38431	0.735	27	0.836	24	0.865	26	0.644	27	0.290	21	0.314	23	0.372	22	0.248	22
A31V	0.665	81195	15034	27730	38431	0.781	28	0.893	26	0.904	29	0.689	28	0.289	20	0.314	22	0.372	21	0.245	21
B30V	86.92	16585	3669	2771	10145	0.656	24	0.893	25	0.833	24	0.496	22	0.298	22	0.379	25	0.442	25	0.208	15
C30V	63.32	57311	14813	9107	33391	0.893	32	0.950	30	1.000	36	0.828	32	0.375	28	0.386	26	0.474	30	0.330	29
C31V	63.37	54744	14951	9114	30679	0.905	33	0.964	32	1.000	34	0.844	33	0.393	30	0.400	27	0.468	29	0.359	30
D30V	0.944	136909	15384	71048	50477	0.496	17	0.614	19	0.667	18	0.383	17	0.267	18	0.279	21	0.365	20	0.222	18
D31V	0.895	37923	5057	15932	16934	0.708	25	0.907	27	0.885	27	0.562	25	0.342	26	0.421	28	0.481	32	0.256	23
E30V	9105	10323	2680	1831	5812	0.855	30	0.936	29	0.949	30	0.786	30	0.304	23	0.200	15	0.385	23	0.309	27
E31V	12255	10323	2680	1831	5812	0.889	31	0.993	33	0.974	31	0.815	31	0.319	25	0.229	19	0.404	24	0.317	28
F30V	0.501	89883	31324	13676	44883	0.994	35	1.000	36	0.994	32	0.992	35	0.756	35	0.843	35	0.763	35	0.720	36
G30V	2571	11746	2899	1915	6932	0.508	19	0.507	13	0.692	21	0.433	18	0.231	16	0.150	12	0.276	13	0.243	20
G31V	2825	11514	1998	2537	6979	0.498	18	0.571	18	0.590	15	0.433	19	0.307	24	0.321	24	0.327	19	0.293	26
G32V	2345	11075	2213	2537	6325	0.388	12	0.429	11	0.423	4	0.359	15	0.277	19	0.257	20	0.282	15	0.282	25
H30V	3035	15142	3277	2564	9301	0.320	4	0.400	6	0.436	5	0.243	6	0.182	8	0.136	10	0.224	8	0.182	9
H31V	3210	15142	3277	2564	9301	0.323	5	0.407	7	0.442	6	0.243	4	0.182	6	0.136	8	0.224	6	0.182	5
H32V	3210	15142	3277	2564	9301	0.323	6	0.407	8	0.442	7	0.243	5	0.182	7	0.136	9	0.224	7	0.182	8
I30V	1450	14660	3374	2715	8571	0.382	11	0.414	9	0.519	11	0.314	11	0.166	4	0.093	4	0.192	3	0.182	3
I31V	1686	14660	3374	2715	8571	0.418	13	0.529	15	0.545	12	0.325	14	0.169	5	0.093	5	0.212	5	0.179	2
J30V	0.607	12459	2250	2015	8194	0.357	7	0.364	4	0.500	8	0.296	9	0.230	15	0.214	17	0.301	17	0.206	14
J31V	0.590	12459	2250	2015	8194	0.363	8	0.386	5	0.506	9	0.296	8	0.225	14	0.214	16	0.288	16	0.203	13
J32V	0.659	12459	2250	2015	8194	0.379	10	0.414	10	0.545	13	0.298	10	0.234	17	0.221	18	0.308	18	0.208	16
K30V	0.851	172566	32537	46737	93292	0.717	26	0.957	31	0.833	25	0.580	26	0.381	29	0.693	33	0.449	27	0.237	19
K31V	0.846	10974	1928	2749	6297	0.607	22	0.836	23	0.744	23	0.467	20	0.356	27	0.429	29	0.481	31	0.277	24
K32V	0.793	12178	3150	1996	7032	0.640	23	0.821	22	0.712	22	0.544	24	0.430	32	0.457	31	0.442	26	0.414	32
K33V	0.783	9670	2252	1472	5946	0.601	21	0.771	21	0.667	19	0.512	23	0.419	31	0.429	30	0.462	28	0.398	31
L30V	0.627	21112	4828	3090	13194	0.809	29	0.921	28	0.897	28	0.731	29	0.489	33	0.550	32	0.609	33	0.417	33
M30V	0.560	17334	4736	2912	9686	0.209	3	0.157	1	0.276	3	0.201	3	0.157	1	0.057	1	0.186	2	0.182	6
M31V	0.557	17334	4736	2912	9686	0.209	2	0.164	3	0.276	2	0.198	1	0.160	3	0.064	3	0.186	1	0.185	11
M32V	0.555	17334	4736	2912	9686	0.209	1	0.164	2	0.269	1	0.201	2	0.160	2	0.064	2	0.192	4	0.182	4
N30V	0.623	14594	2917	2274	9403	0.418	14	0.514	14	0.564	14	0.322	13	0.188	11	0.143	11	0.244	9	0.182	10
N31V	0.614	14594	2917	2274	9403	0.455	16	0.557	17	0.596	17	0.359	16	0.196	12	0.157	13	0.250	11	0.187	12
N32V	0.620	14594	2917	2274	9403	0.379	9	0.500	12	0.506	10	0.282	7	0.182	9	0.107	6	0.256	12	0.179	1
N33V	0.616	14594	2917	2274	9403	0.430	15	0.557	16	0.596	16	0.314	12	0.185	10	0.121	7	0.250	10	0.182	7
Q30V	0.408	30436	4825	8796	16815	1.000	36	1.000	34	1.000	33	1.000	36	0.761	36	0.907	36	0.821	36	0.683	35
Q31V	0.321	30436	4825	8796	16815	0.973	34	1.000	35	1.000	35	0.953	34	0.637	34	0.843	34	0.731	34	0.522	34
R30V	0.773	30068	7843	5170	17055	0.563	20	0.679	20	0.667	20	0.478	21	0.222	13	0.186	14	0.276	14	0.214	17

Table 26: For the three installations in the DATASET H: TRAVEL WALKWAY *dataset, with 4800 subjects enrolled with a frontal still, the values are identification-mode FNIR(T) for each algorithm at a decision thresholdcorresponding to a false positive count of 10 over all cameras, and investigation-mode FNIR(R) for rank 1. Each value is accompanied by an integer ranking across all algorithms. Four cameras were used in banks F and S, with two in bank R. The detections are summed over the cameras in the identified groups. The accuracy values are aggregated over all sightings of all subjects in the field of view of those cameras. Given fixed equipment, the variation across columns is indicative of environmental differences.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

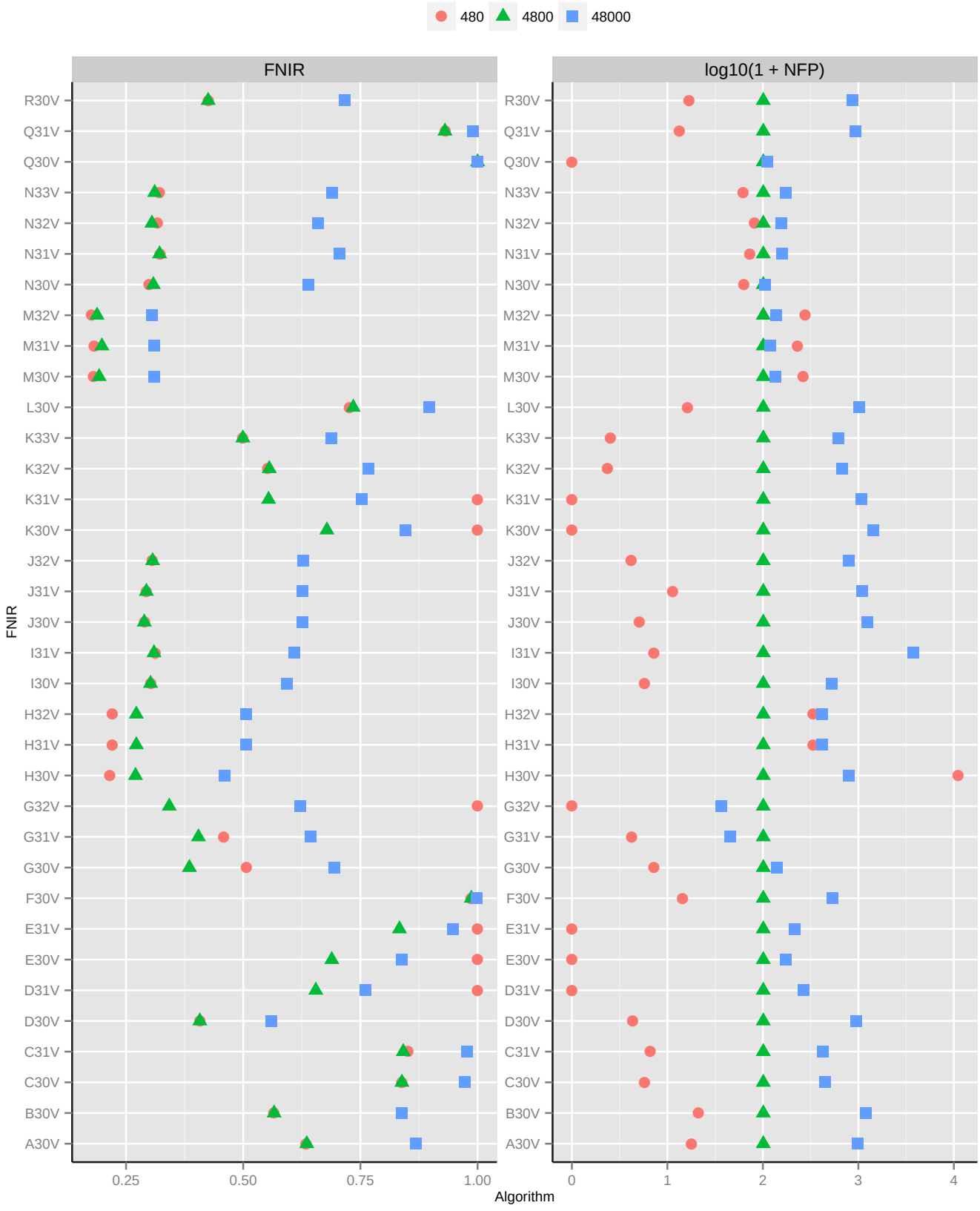


Figure 36: Over all cameras in the DATASET H: TRAVEL WALKWAY *collection, the dots show accuracy for N= $\{48000, 4800, 480\}$, at a single global decision threshold set to produce NFP(T) = 100 false positives over all searches of video templates against impostor galleries of size $N = 4800$. The left panel shows FNIR(N, L, T). The right panel shows $\log_{10}(1+NFP(T))$. Simple binomial theory would dictate linear growth NFP with N, and FNIR independent of N.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

N = 480	NUM ACTORS 31		NUM FEEDS 10		NUM CLIPS 329			NUM FRAMES 40737		NUM MINUTES 154.8		
	DETECTIONS		THRESHOLD BASED AUTO WATCHLISTS					RANK BASED FORENSIC CASES				
ALG	NUM	MIN^{-1}	FNIR(T), FP(T)=10		FNIR(T), FP(T)=100		FNIR(T), FP(T)=1000	FNIR(R=1, T=0)		FNIR(R=5, T=0)		FNIR(R=20, T=0)	
A30V	17587	113.6	0.846	27	0.756	27	0.566	24	0.425	24	0.178	12	0.078	4
A31V	17587	113.6	0.831	26	0.723	24	0.581	25	0.419	23	0.172	10	0.078	3
B30V	6679	43.1	0.807	24	0.687	23	0.509	22	0.431	25	0.289	24	0.160	13
C30V	16558	106.9	0.958	33	0.907	33	0.780	32	0.605	31	0.455	28	0.307	28
C31V	18251	117.9	0.949	32	0.904	32	0.777	31	0.617	33	0.467	30	0.322	29
D30V	31670	204.5	0.711	22	0.654	22	0.557	23	0.283	15	0.247	21	0.208	22
D31V	6630	42.8	0.919	31	0.880	31	0.783	33	0.533	27	0.497	31	0.431	33
E30V	4094	26.4	0.575	17	0.440	14	0.295	10	0.274	13	0.223	15	0.166	14
E31V	4094	26.4	0.593	18	0.440	15	0.304	12	0.271	12	0.217	14	0.169	16
F30V	19479	125.8	1.000	36	0.961	34	0.852	34	0.843	34	0.687	34	0.479	34
G30V	3295	21.3	0.551	14	0.437	13	0.337	15	0.307	18	0.232	18	0.184	20
G31V	4890	31.6	0.524	10	0.416	10	0.307	13	0.256	10	0.172	11	0.108	12
G32V	4890	31.6	0.286	4	0.232	4	0.184	4	0.235	9	0.166	9	0.093	9
H30V	7333	47.4	0.440	7	0.298	6	0.193	5	0.160	6	0.111	5	0.084	5
H31V	7333	47.4	0.461	8	0.319	8	0.220	7	0.205	7	0.142	7	0.102	10
H32V	7333	47.4	0.461	9	0.319	9	0.220	8	0.205	8	0.142	8	0.102	11
I30V	180141	1163.4	0.398	6	0.286	5	0.199	6	0.102	1	0.057	1	0.048	2
I31V	180141	1163.4	0.389	5	0.298	7	0.223	9	0.102	2	0.057	2	0.042	1
J30V	4748	30.7	0.551	15	0.479	20	0.380	19	0.352	21	0.283	23	0.232	24
J31V	4748	30.7	0.542	13	0.476	19	0.377	18	0.343	20	0.274	22	0.238	25
J32V	4748	30.7	0.530	12	0.458	18	0.383	20	0.364	22	0.298	25	0.253	26
K30V	17418	112.5	0.895	30	0.855	30	0.759	30	0.515	26	0.386	26	0.211	23
K31V	4903	31.7	0.816	25	0.729	25	0.636	27	0.551	28	0.449	27	0.301	27
K32V	3346	21.6	0.858	28	0.783	28	0.645	28	0.584	30	0.518	33	0.425	32
K33V	2999	19.4	0.786	23	0.738	26	0.620	26	0.569	29	0.518	32	0.416	31
L30V	8210	53.0	0.870	29	0.804	29	0.708	29	0.614	32	0.467	29	0.343	30
M30V	8875	57.3	0.241	3	0.190	2	0.151	1	0.133	4	0.108	4	0.090	8
M31V	8875	57.3	0.235	2	0.193	3	0.154	2	0.136	5	0.108	3	0.087	7
M32V	8875	57.3	0.229	1	0.187	1	0.157	3	0.130	3	0.114	6	0.087	6
N30V	3737	24.1	0.530	11	0.452	17	0.352	17	0.298	17	0.244	20	0.178	19
N31V	3737	24.1	0.557	16	0.446	16	0.346	16	0.295	16	0.235	19	0.187	21
N32V	3737	24.1	0.611	19	0.428	12	0.304	11	0.262	11	0.208	13	0.175	17
N33V	3737	24.1	0.630	20	0.422	11	0.313	14	0.277	14	0.226	16	0.169	15
Q30V	5773	37.3	0.997	34	0.997	36	0.967	36	0.940	36	0.852	36	0.699	36
Q31V	5773	37.3	0.997	35	0.985	35	0.928	35	0.898	35	0.804	35	0.599	35
R30V	10518	67.9	0.699	21	0.566	21	0.425	21	0.316	19	0.226	17	0.175	18

Table 27: For the DATASET T: TRAVEL WALKWAY *installation, with 480 subjects enrolled with a frontal still, the values are detection counts summed over 10 cameras, the rate over all cameras, identification-mode FNIR(T) for each algorithm at three different decision thresholds corresponding to false positive counts of 10, 100, 1000, and investigation-mode FNIR(R) for ranks 1, 5, 20. Each value is accompanied by an integer ranking across all algorithms. The shading indicates the most important metric to watchlist applications. 10 cameras were used. Their frames rates varied from 1.5 fps to 20 fps. The detections are summed over all of them. The accuracy values are aggregated over all sightings of all subjects in the field of view of those cameras. NB: This dataset is not sequestered - it has been made available to some developers. Their ability to tune and train on this data may mean that accuracy values may be optimistic.*

5.7 DATASET T: TRAVEL WALKWAY **Surveillance**

Experimental design: The dataset is composed of videos collected in a transit terminal/passenger environment. As with Dataset H, actors and members of the general public walk underneath ceiling mounted cameras. Extracts from dataset has been made available to certain face recognition algorithm developers. **Thus, the dataset is not sequestered, and should not therefore be relied upon when comparing algorithms.**

Mated scores are computed by searching 329 video clips against an enrolled dataset of still face images of subjects known to be in the search videos. The size of the enrollment dataset was 480 or 4 800. The enrollment database is extended to these sizes by adding high quality frontal portrait photographs from a disjoint background population.

Nonmated scores are computed by replacing the gallery, which normally contains frontal images of known actors, with the *global nonmated enrollment dataset*.

Results: Detection and recognition error rates are tabulated in Tables 27 and 28. Notably the detection counts are 2-5

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

N = 4800	NUM ACTORS 31		NUM FEEDS 10		NUM CLIPS 329		NUM FRAMES 40737		NUM MINUTES 154.8	
	DETECTIONS		THRESHOLD BASED AUTO WATCHLISTS				RANK BASED FORENSIC CASES			
ALG	NUM	MIN^{-1}	FNIR(T), FP(T)=10		FNIR(T), FP(T)=100		FNIR(T), FP(T)=1000		FNIR(R=1, T=0)		FNIR(R=5, T=0)		FNIR(R=20, T=0)	
A30V	17587	113.6	0.919	26	0.834	23	0.678	24	0.581	24	0.401	23	0.256	17
A31V	17587	113.6	1.000	35	1.000	35	1.000	36	1.000	36	1.000	36	1.000	36
B30V	6670	43.1	0.837	22	0.747	22	0.596	22	0.518	23	0.422	24	0.331	24
C30V	16558	106.9	0.961	29	0.946	29	0.852	29	0.660	27	0.536	26	0.440	26
C31V	18251	117.9	0.967	30	0.949	31	0.858	31	0.681	31	0.536	25	0.431	25
D30V	31643	204.4	0.795	20	0.699	21	0.608	23	0.377	18	0.316	18	0.283	20
D31V	6622	42.8	0.961	28	0.934	28	0.855	30	0.593	25	0.578	28	0.563	32
E30V	4094	26.4	0.714	18	0.542	17	0.401	14	0.325	13	0.259	12	0.223	11
E31V	4094	26.4	0.994	32	0.946	30	0.470	19	0.316	12	0.247	11	0.226	12
F30V	19487	125.8	1.000	36	1.000	36	0.928	33	0.904	33	0.825	33	0.693	33
G30V	3311	21.4	0.642	16	0.503	12	0.386	13	0.355	16	0.283	14	0.238	13
G31V	4858	31.4	0.714	19	0.545	18	0.410	15	0.307	11	0.238	10	0.190	10
G32V	4890	31.6	0.608	12	0.512	14	0.352	10	0.286	9	0.223	9	0.175	9
H30V	7333	47.4	0.262	1	0.229	3	0.223	4	0.235	8	0.196	8	0.166	6
H31V	7333	47.4	0.452	5	0.334	5	0.280	5	0.235	6	0.193	7	0.169	7
H32V	7333	47.4	0.452	6	0.334	6	0.280	6	0.235	7	0.190	6	0.169	8
I30V	180141	1163.4	0.491	7	0.370	7	0.286	7	0.142	1	0.093	2	0.072	2
I31V	180141	1163.4	0.548	9	0.410	8	0.292	8	0.145	2	0.087	1	0.060	1
J30V	4748	30.7	0.633	14	0.536	16	0.473	20	0.413	21	0.355	21	0.304	22
J31V	4748	30.7	0.645	17	0.524	15	0.455	18	0.407	20	0.349	20	0.304	21
J32V	4748	30.7	0.620	13	0.545	19	0.449	17	0.419	22	0.370	22	0.322	23
K30V	17418	112.5	0.967	31	0.955	32	0.913	32	0.657	26	0.566	27	0.470	27
K31V	4903	31.7	0.892	23	0.877	25	0.801	27	0.672	28	0.590	29	0.500	28
K32V	3346	21.6	0.934	27	0.892	27	0.801	28	0.678	30	0.627	31	0.548	30
K33V	2999	19.4	0.904	24	0.855	24	0.783	25	0.675	29	0.593	30	0.551	31
L30V	8202	53.0	0.907	25	0.880	26	0.789	26	0.738	32	0.648	32	0.545	29
M30V	8875	57.3	0.268	2	0.226	2	0.175	1	0.157	5	0.130	4	0.114	3
M31V	8875	57.3	0.274	3	0.232	4	0.181	3	0.157	4	0.133	5	0.117	4
M32V	8875	57.3	0.298	4	0.217	1	0.181	2	0.151	3	0.127	3	0.123	5
N30V	3737	24.1	0.554	10	0.449	10	0.361	11	0.331	14	0.292	16	0.256	18
N31V	3737	24.1	0.633	15	0.506	13	0.413	16	0.377	19	0.319	19	0.271	19
N32V	3737	24.1	0.527	8	0.434	9	0.349	9	0.307	10	0.277	13	0.244	14
N33V	3737	24.1	0.590	11	0.467	11	0.373	12	0.346	15	0.286	15	0.253	16
Q30V	5773	37.3	1.000	34	1.000	34	0.997	35	0.964	35	0.934	35	0.867	35
Q31V	5773	37.3	0.997	33	0.994	33	0.970	34	0.946	34	0.886	34	0.786	34
R30V	10517	67.9	0.822	21	0.663	20	0.491	21	0.367	17	0.307	17	0.244	15

Table 28: For the DATASET T: TRAVEL WALKWAY *installation, with 4800 subjects enrolled with a frontal still, the values are detection counts summed over 10 cameras, the rate over all cameras, identification-mode FNIR(T) for each algorithm at three different decision thresholds corresponding to false positive counts of 10, 100, 1000, and investigation-mode FNIR(R) for ranks 1, 5, 20. Each value is accompanied by an integer ranking across all algorithms. The shading indicates the most important metric to watchlist applications. 10 cameras were used. Their frames rates varied from 1.5 fps to 20 fps. The detections are summed over all of them. The accuracy values are aggregated over all sightings of all subjects in the field of view of those cameras. NB: This dataset is not sequestered - it has been made available to some developers. Their ability to tune and train on this data may mean that accuracy values may be optimistic.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

times higher than with Dataset H, depending on algorithm, and this reflects the somewhat wider fields of view and denser crowds in T vs. H. There is one exception to this, algorithms I3xV, both report enormous numbers of detections, almost 100 times more than with Dataset H. We have no explanation for this anomalous behavior, but note that the same algorithms produce the best rank-based miss rates. While the I algorithms are competitive on other datasets, it is possible that by producing many templates, the search accuracy is artificially improved by sheer volume of searches, some of which fortuitously place the correct actor at rank 1. This applies with both N = 480 and 4 800. This observation is consistent with I algorithms having relatively lower performance when the threshold is raised to produce only small numbers of false positives. Whether the developers of the I algorithms experimented with Dataset T is not known.

While the M and I algorithms give similar rank one miss rates for Datasets T and H, other algorithms give generally higher rate on the T data. The cause of this is unknown, but may be associated with higher crowd densities. The high-threshold FNIR(N, L, T) rates are not easily comparable because the thresholds set to achieve, for example, NFP(T) = 100 are different given that the datasets differ in the total number of people appearing in the videos. The Dataset H error rates are generally lower, and again we'd like to attribute this to lower volumes (travelers per minute) with Dataset H, but a large number of other factors come into play. The only solid conclusion is that while Datasets T and H are both nominally professional installations of video surveillance cameras, the observed differences in identification miss rates are essentially a measure of uncertainty such that a deployer cannot know a priori precisely how well face recognition will work *in their environment*.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Property	Value
Cameras	Many, mostly professional still + television
Camera mounting	Handheld or fixed mount
Camera height	Near head height
Camera declination to face	Adverse yaw, only modest pitch
Frame rate	Variable, usually $\geq 24\text{sec}^{-1}$
Width	Variable
Height	Variable

Table 29: Key imaging properties for DATASET C: PHOTOJOURNALISM *results.*

5.8 DATASET C: PHOTOJOURNALISM

Overview: This section documents two experiments, one video-to-still (as in the rest of this report) and one video-to-video. Both experiments use imagery comprised of unconstrained photographs and videos of celebrities, actors, politicians, and diplomats. The images are quite different in character compared to the others used in this report. First, they were not collected with any notion that face recognition would be applied. Second, as they were mostly acquired by professional photo journalists, they are quality-biased, in the sense that they have survived a selection process in which, mostly, they do not exhibit poor focus and poor exposure. Instead, the images are selected to be engaging to a human viewer. Thus, neutral expressions and fully frontal views are *not* the norm. Some faces are partially occluded, and there are wide variations in head pose and expression. While the term "in the wild" has been used to describe such data [24], it is a misnomer in the sense that professional photographers collect better constrained data than entirely amateur "wild" and un-constrained data. A slightly better description might be "in the limelight"! Note that the imagery used here was not selected on the basis that a face detector found the faces and selected it for inclusion. Note also that the imagery is in the public domain, and could, in-principle have been used in training the algorithms submitted to FIVE. The still images and video frames from this dataset have recently been released [26][15].

Figure 37: DATASET C: PHOTOJOURNALISM *: Examples of "in the wild" photojournalism stills. **The images in this figure are of celebrities and politicians. They are freely available on the internet under Creative Commons licenses.*

The type of equipment is not known. Given the photojournalism origins of most of the data, we can assume that the cameras were of professional grade.

[15] https://www.nist.gov/itl/iad/image-group/ijba-dataset-request-form

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Figure 38: DATASET C: PHOTOJOURNALISM *: Examples of "in the wild" photojournalism video clips. **The video frames in this figure are taken from videos of celebrities and politicians that are available on the internet under Creative Commons licenses.*

Experimental design: The individuals are often giving speeches or are being interviewed. There are sometimes many other people in the scene - see, for example, Figure 38. Two experiments were conducted.

In the first **video-to-still** experiment, a common set of 7194 videos were searched against two still galleries, one of size 940 (S1), and the other of size 930 (S2) persons. Both galleries are composed of unconstrained face photographs, one per subject, examples of which are shown in Figure 37. The probeset contains both mated and nonmated search videos: A subset denoted P1 forms the mated searches for S1 and the non-mated searches for S2. The disjoint subset P2 forms the nonmated searches for S1 and mated searches for S2. Example videos appear in Figure 38.

In the second **video-to-video experiment**, 1356 mated video clips are searched against templates extracted from each of 393 gallery video clips. Enrollment of this kind generally produces more templates than there are videos, because multiple faces appear in some videos. As all searches contain at least one face known to be in the gallery, this experiment is "closed universe" and is therefore atypical operationally.

Key experimental design details are summarized in Table 30.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Quantity	Value or description
Mode	Video search to video enrollment
Number of enrolled subjects	393
Number of enrolled videos	393
Number of enrolled stills	0
Number of actors	393
Number of non-actors	Many
Number of search clips actors	1356
Number of search clips no actors	0
Video duration with actors (minutes)	699
Video duration no actors	0
Gallery video duration (frames)	Variable: 433 median, 765 mean
Probe video duration (frames)	Variable: 460 median, 847 mean
Properties of enrolled videos	Unconstrained mostly photojournalism
FNIR estimation	Actors present video vs. enrolled gallery
FPIR estimation	Not done
Mode	Video search to still enrollment
Number of enrolled subjects	Two disjoint galleries, S1-940, S2-930
Number of enrolled videos	0
Number of enrolled stills	Multiple per subject
Number of actors	1870
Number of search clips	7195 about half actors, half non-actors
Properties of enrolled stills	Unconstrained mostly photojournalism
FNIR estimation	Actors present video vs. enrolled gallery: P1-S1, P2-S2
FPIR estimation	Actors absent video vs. same enrolled gallery: P1-S2, P2-S1
Probe video frame rate (per second)	Variable, many 24
Probe video total duration (frames)	4 493 284
Probe video clip duration (frames)	Variable: 390 median, 625 mean
	Factors in common
Candidate list length	20
Subject motion	Often on podium or seated facing journalist
Number of persons in FOV	Variable, typically 1, 2, few
Number of cameras	Many, often professional
Video ground truth	Style A: See Figure 6

Table 30: Key experimental design for the DATASET C: PHOTOJOURNALISM *results.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

N=930 OR	NUM ACTORS 1870		NUM FEEDS 7194		NUM CLIPS 7194		NUM FRAMES 4493284		NUM MINUTES 2781.5			
N=940	DETECTIONS	THRESHOLD BASED IDENTIFICATION					RANK BASED INVESTIGATION					
ALG	NUM	FNIR(T), FP(T)=1, FPIR(T)=0.0001		FNIR(T), FP(T)=10, FPIR(T)=0.001		FNIR(T), FP(T)=100, FPIR(T)=0.01		FNIR(R=1, T=0)		FNIR(R=5, T=0)		FNIR(R=20, T=0)	
A30V	431026	0.993	26	0.975	26	0.955	27	0.816	28	0.696	28	0.138	10
A31V	431026	0.992	23	0.988	29	0.980	29	0.816	27	0.696	27	0.138	9
B30V	52954	0.975	3	0.936	14	0.898	22	0.706	25	0.571	25	0.325	31
C30V	297740	0.994	29	0.992	31	0.985	31	0.913	31	0.798	31	0.233	28
C31V	341668	0.993	28	0.991	30	0.985	30	0.916	32	0.798	32	0.228	27
D30V	219672	0.998	31	0.998	33	0.998	33	0.954	35	0.810	33	0.426	34
D31V	91868	0.982	5	0.975	27	0.964	28	0.879	29	0.697	29	0.395	32
E30V	53578	0.975	4	0.954	23	0.873	19	0.550	10	0.412	12	0.215	25
E31V	53540	0.971	1	0.960	24	0.932	24	0.575	16	0.433	16	0.216	26
F30V	507530	1.000	36	1.000	36	1.000	36	0.994	36	0.960	36	0.476	36
G30V	44229	0.992	24	0.925	12	0.841	16	0.621	22	0.476	22	0.260	29
G31V	46406	0.992	22	0.914	6	0.705	2	0.445	5	0.307	4	0.135	8
G32V	46191	0.991	21	0.841	1	0.624	1	0.359	1	0.245	1	0.108	6
H30V	80856	0.999	33	0.954	22	0.849	18	0.596	19	0.465	21	0.190	18
H31V	80856	0.991	19	0.951	20	0.815	11	0.588	17	0.459	19	0.186	16
H32V	80856	0.991	20	0.951	21	0.822	14	0.593	18	0.465	20	0.190	19
I30V	131712	0.993	25	0.922	11	0.797	7	0.494	6	0.346	6	0.095	4
I31V	131712	0.993	27	0.983	28	0.937	25	0.561	12	0.401	10	0.099	5
J30V	46554	0.990	18	0.910	4	0.813	10	0.570	15	0.426	15	0.213	24
J31V	46554	0.989	17	0.915	7	0.803	8	0.566	13	0.422	13	0.210	23
J32V	46554	0.986	12	0.941	16	0.817	12	0.557	11	0.409	11	0.188	17
K30V	219018	0.988	14	0.968	25	0.943	26	0.712	26	0.587	26	0.164	13
K31V	82280	0.983	6	0.938	15	0.896	21	0.639	23	0.493	24	0.195	20
K32V	53494	0.986	11	0.943	18	0.900	23	0.605	21	0.459	18	0.207	22
K33V	49348	0.985	10	0.935	13	0.886	20	0.601	20	0.441	17	0.201	21
L30V	159850	0.999	32	0.997	32	0.993	32	0.891	30	0.766	30	0.276	30
M30V	120034	0.987	13	0.922	10	0.713	4	0.419	2	0.305	2	0.065	3
M31V	120034	0.989	16	0.911	5	0.722	5	0.425	3	0.307	5	0.064	1
M32V	120034	0.984	9	0.943	17	0.710	3	0.426	4	0.305	3	0.064	2
N30V	50806	0.983	8	0.902	3	0.777	6	0.514	7	0.387	8	0.167	14
N31V	50806	0.988	15	0.920	9	0.824	15	0.537	9	0.400	9	0.146	12
N32V	50806	0.994	30	0.891	2	0.809	9	0.568	14	0.423	14	0.182	15
N33V	50806	0.973	2	0.916	8	0.821	13	0.529	8	0.381	7	0.141	11
Q30V	125234	1.000	35	1.000	35	1.000	35	0.930	34	0.830	34	0.474	35
Q31V	125234	1.000	34	1.000	34	0.999	34	0.925	33	0.831	35	0.412	33
R30V	111832	0.983	7	0.945	19	0.845	17	0.641	24	0.477	23	0.125	7

Table 31: For the DATASET C: PHOTOJOURNALISM *installation, with 1870 subjects enrolled with one or more unconstrained stills, the values are identification-mode* FNIR(N, L, T) *for each algorithm at three different decision thresholds corresponding to false positive counts of 1, 10, 100, and investigation-mode* FNIR(N, R, 0) *for ranks 1, 5, 20. Each value is accompanied by an integer ranking across all algorithms. The shading indicates arguably the most important metric to operator-led media searching applications (e.g. broadcast news) where there is only mild intolerance for false positives. Note very high miss rates throughout, especially at low false positive rates. The* FPIR(T) *values are computed as the the number of false positives divided by the number of impostor videos used. This is an unconventional definition because the number of faces found in any given clip may exceed one, and each of these produces a template which is searched. This means* FPIR *here is an upper bound greater than or equal to its value if the number of faces present was known exactly.*

Results V2S: The results are summarized in Table 31. It shows FNIR values aggregated over both trials by concatenating all scores and ranks. This step is defensible only because the two galleries have almost the same size, and are sampled randomly from the same parent image population.

The error rates are much higher than in other experiments documented in this report, reflecting the lack of geometric constraints on the photography, particularly in allowing and selecting highly variable head poses. Moreover, this factor applies to both the gallery and search imagery. All other tests in this report use standards-conformant frontal enrollments.

At high thresholds, FNIR is essentially 100% with NFP = 1. Even allowing NFP = 100, the lower threshold still yield a best FNIR(N, L, T) = 0.62 from algorithm G32V. If such data was to be used with a human review, then the mate is not at rank one still quiet often: FNIR(N, 1, 0) = 0.36 (G32V) and not within the top 20 ranks at best with FNIR(N, 20, 0) = 0.064 (M31V). This dataset is easier at rank 20, than is Dataset H, despite the latter being easier at rank 1 i.e. the cumulative match characteristics cross.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Results V2V: Figure 39 summarizes accuracy for the video-to-video case by showing FNIR$(N, R, 0)$ i.e. miss rates at three ranks. We do not show high-threshold FNIR(N, L, T) as we did not run mateless video-to-video searches over which we could compute false positive outcomes. The algorithms given the best rank-1 miss rates are from providers M, G, I, H, N. Notably for rank-20 the K, E and B algorithms give low error rates, but N does not.

As with detection on prior datasets, the number of persons detected and enrolled into the gallery varies considerably. Algorithm E30V finds fewer people than are actually present. The most accurate algorithms, M3xV, find 2688 faces and A30V finds 11635. We do not have complete ground truth on the actual number of faces present, per the discussion in section 4.1.

Some notable observations from Figure 39 are as follows. Some developers, N, M, J, I, H, appear use the same detection and tracking algorithms for all their submissions, producing identical numbers of tracks. However the feature extraction code sometimes differs. Thus, M31V produces accuracy better than M30V with enrollment templates three times smaller. N30V similarly has accuracy better than N32V but with a template four times smaller. I3xV doubles its template size between submissions, with the smaller template giving better accuracy.

Refering to the text annotations in Figure 39, over all algorithms, there is a massive variation in the amount of template data extracted from video: G30V extracts less than a 1MB from the entire video set, representing each of the 1078 detected faces with just 0.4 kilobytes. R30V is almost as parsimonious detecting slightly few faces (937) but encoding each with an average of 0.6 kilobytes. At the other end of the spectrum, the N33V enrollments average 1830 kilobytes, i.e. three orders of magnitude more per face than the R30V and G30V algorithms. The G31V, J3xV, and L30V algorithms are very economical also. Size may have implications for speed of searching large media collections.

6 Computational resource requirements

Face recognition engines detect and track faces in video clips, then produce searchable templates. While search duration is fast for small enrollment databases, the image processing and template generation stage is expensive. The time and memory resources it takes for an algorithm to generate a template and search it against a database determines hardware requirements. Furthermore, the size of templates produced can impact network bandwidth and disk space requirements.

6.1 Test environment

Software environment: The algorithms were submitted to NIST as pre-compiled libraries that implemented a NIST-specified C++ API [20]. The API declares the functions necessary to find and extract features from faces in still and video imagery. Source code is not provided to NIST: As such, this is a black box evaluation. NIST has no knowledge of how the implementations effect recognition.

Configuration data: Each algorithm was accompanied by configuration data.

Hardware: Testing was performed on high-end server-class blades, most of which were 6-core machines with dual processors running at 3.47 GHz with 192 GB of main memory. GPUs were not used. IPP - Intel Integrated Performance Primitives was permitted if supporting libraries were delivered with submission.

Operating system: All processing was done on the CentOS 7.0 64-bit Linux variant. The test harness was built on top of the NIST Biometric Evaluation Framework [17] and used concurrent processing to distribute workload across dozens of

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

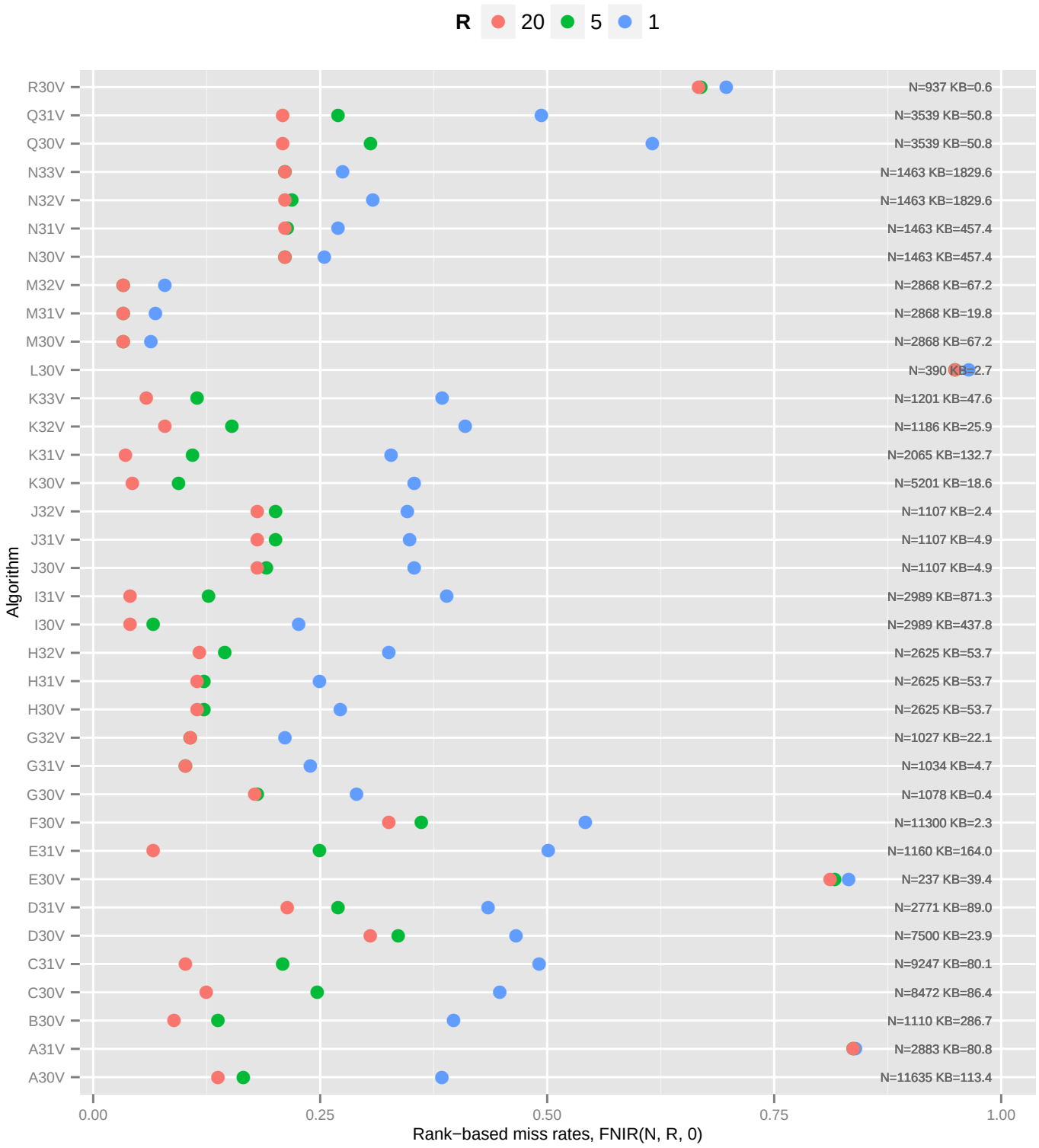


Figure 39: Video-to-video recognition: The dots show miss rates, FNIR$(N, R, 0)$, for $R = \{1, 5, 20\}$ for each algorithm applied to DATASET C: PHOTOJOURNALISM *imagery. The algorithms enroll variable numbers of templates from faces detected in the enrollment video clips. These numbers appear as text, and vary under the influence of false positive and false negative detection rates, tracking integrity rates, and image quality acceptance criteria. The template size is computed as the size of the finalized enrollment data divided by the number of persons detected. Note that the number of enrolled templates is generally higher than the 393 input videos because, in addition to the 393 known individuals, there are additionally an unknown number of other persons present. All of the 1356 search video clips contain at least one, and very rarely more, of the actors in the gallery. The probes also contain unknown faces.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Algorithm		Algorithm		Algorithm		Algorithm	
A30V	104	A31V	104				
B30V	1						
C30V	173	C31V	173				
D30V	95	D31V	95				
E30V	99	E31V	99				
F30V	122						
G30V	1646	G31V	1647	G32V	1647		
H30V	155	H31V	155	H32V	155		
I30V	746	I31V	745				
J30V	388	J31V	433	J32V	249		
K30V	684	K31V	684	K32V	684	K33V	684
L30V	604						
M30V	284	M31V	284	M32V	284		
N30V	317	N31V	317	N32V	317	N33V	317
Q30V	47	Q31V	47				
R30V	119						

Table 32: For each algorithm, sizes in megabytes, of static read-only configuration data supplied with the algorithm, as reported by the unix command "du -sm".

blades.

6.2 Configuration directory size

Participants were permitted to provide static, read-only configuration data with their algorithm submission. The location of such configuration data is provided to the implementation during algorithm initialization for template generation and search. Configuration data typically includes trained feature extraction models, but here, the content is entirely unregulated. Table 32 reports the size of the configuration data for each algorithm.

6.3 Video processing time

Face recognition in video sequences implies computational cost over recognition using still images. A portrait photograph can be enrolled in less than a second [21]. In video however, two factors imply slower processing. First, video imagery typically has larger width and height, as cameras often have wider fields of view. Second, there are many frames, and processing times scale, to first order, linearly with the frame rate. Given 24 or 30 frames per second the costs are considerable. This section details cost.

Video processing time refers to the amount of time that elapsed while a video sequence was processed by a recognition algorithm. The primary purpose of processing a video is to produce matchable templates. Reported times do not include any pre-processing steps performed by the testing harness such as loading the video from disk. The timing machine ran with an Intel Xeon E5-2695 v3 @ 2.30 GHz, 56 logical CPUs, 528 GB RAM.

Generating a template from a video sequence is computationally intensive and generally scales with the number of frames in the video sequence but depends on other factors as well. For example, Table 33 demonstrates that all algorithms require greater processing time as the number of people in the videos increases. This component of the duration includes the time taken for any fine localization and alignment, and feature extraction.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

DATASET	DATASET P				DATASET J							
	EXACTLY 0 SUBJECTS				AT ANY POINT, 1 TO 3 SUBJECTS				AT ANY POINT, 4 TO 7 SUBJECTS			
	CLIP LEN.	NUMBER OF	PROC. DURATION	PROC. RATE	CLIP LEN.	NUMBER OF	PROC. DURATION	PROC. RATE	CLIP LEN.	NUMBER OF	PROC. DURATION	PROC. RATE
ALG	SECONDS	TEMPLATES	MINUTES	$[TT^{-1}]$	SECONDS	TEMPLATES	MINUTES	$[TT^{-1}]$	SECONDS	TEMPLATES	MINUTES	$[TT^{-1}]$
A30V	50	1	19.87 35	23.89	547	456	364.21 36	39.95	649	2034	670.34 31	61.97
A31V	50	1	19.87 34	23.88	547	456	362.29 35	39.74	649	2034	666.65 30	61.63
B30V	50	0	1.96 12	2.36	547	67	42.88 11	4.70	649	322	76.10 10	7.03
C30V	50	0	10.49 28	12.61	547	413	82.64 17	9.06	649	1446	137.51 14	12.71
C31V	50	0	10.50 29	12.62	547	443	86.93 18	9.53	649	1536	146.24 16	13.52
D30V	50	0	1.83 11	2.20	547	258	25.49 8	2.80	649	1647	49.96 4	4.62
D31V	50	0	3.20 20	3.85	547	95	41.00 10	4.50	649	577	69.55 9	6.43
E30V	50	0	6.34 22	7.62	547	44	140.54 23	15.41	649	55	226.56 17	20.94
E31V	50	0	6.51 23	7.83	547	44	164.30 27	18.02	649	55	271.53 24	25.10
F30V	50	0	0.80 6	0.96	547	624	17.30 6	1.90	649	1594	87.17 11	8.06
G30V	50	1	1.26 10	1.52	547	50	23.33 7	2.56	649	137	52.95 5	4.89
G31V	50	1	10.95 30	13.16	547	48	144.76 25	15.88	649	26	244.63 21	22.62
G32V	50	1	15.16 33	18.22	547	48	146.08 26	16.02	649	26	253.50 22	23.43
H30V	50	0	1.16 8	1.39	547	75	82.46 16	9.04	649	440	234.15 18	21.65
H31V	50	0	1.16 9	1.40	547	75	81.93 15	8.99	649	440	239.09 19	22.10
H32V	50	0	1.15 7	1.39	547	75	87.40 19	9.59	649	440	242.62 20	22.43
I30V	50	0	0.36 1	0.43	547	96	10.09 3	1.11	649	490	31.29 1	2.89
I31V	50	0	0.36 2	0.43	547	96	9.89 2	1.08	649	490	31.33 2	2.90
J30V	50	0	0.39 3	0.47	547	45	10.32 5	1.13	649	157	62.05 7	5.74
J31V	50	0	0.40 5	0.48	547	45	10.10 4	1.11	649	157	59.32 6	5.48
J32V	50	0	0.39 4	0.47	547	45	8.30 1	0.91	649	157	46.90 3	4.34
K30V	50	3	8.97 26	10.78	547	340	255.69 33	28.04	649	843	316.99 26	29.30
K31V	50	3	8.91 25	10.71	547	54	227.32 32	24.93	649	134	466.87 29	43.16
K32V	50	0	8.88 24	10.67	547	77	187.13 29	20.52	649	173	345.78 27	31.97
K33V	50	0	9.32 27	11.20	547	56	176.63 28	19.37	649	91	391.14 28	36.16
L30V	50	0	20.80 36	25.00	547	116	347.95 34	38.16	649	427	675.22 32	62.42
M30V	50	17	3.03 18	3.64	547	95	44.01 12	4.83	649	469	123.19 13	11.39
M31V	50	17	3.21 21	3.86	547	95	35.38 9	3.88	649	469	67.04 8	6.20
M32V	50	17	3.00 17	3.60	547	95	46.35 13	5.08	649	469	119.23 12	11.02
N30V	50	0	2.90 15	3.49	547	62	139.82 22	15.34	649	280	925.16 34	85.53
N31V	50	0	2.89 14	3.47	547	62	141.12 24	15.48	649	280	878.32 33	81.20
N32V	50	0	2.87 13	3.45	547	62	129.09 20	14.16	649	280	942.37 35	87.12
N33V	50	0	2.96 16	3.56	547	62	138.26 21	15.17	649	280	947.41 36	87.58
Q30V	50	0	14.44 31	17.36	547	157	224.99 31	24.68	649	556	308.34 25	28.50
Q31V	50	0	14.77 32	17.75	547	157	219.79 30	24.11	649	556	260.75 23	24.11
R30V	50	0	3.09 19	3.71	547	131	55.96 14	6.14	649	380	137.93 15	12.75

Table 33: There are three groups each with five columns. The first group gives the expense of processing video imagery containing no faces and no motion. The second and third groups apply to video clips containing, respectively, walking and queued subjects. Within each group the columns are: a) the duration of the input clip; b) the number of templates reported by the algorithm; c) the processing duration; d) its rank over all algorithms; and e) the processing rate. This last quantity is dimensionless (denoted by TT^{-1}) and interpretable as the duration of computation, in seconds, per second of input video using one core on a c. 2015 server. This last column is shaded yellow as it is the most important column; Values below 1 indicate realtime processing is possible on one core. While the numbers are rounded to two decimal places, precision probably extends to only one place. Computations are done in full precision.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Alg	Template Size (bytes)	Size Rank	Template Generation Time (sec)	Time Rank
A30V	5070	22	0.81 (σ=0.12)	30
A31V	5070	22	0.87 (σ=0.16)	31
B30V	1060	4	0.08 (σ=0.02)	2
C30V	109150 (σ=33810)	36	1.32 (σ=0.35)	35
C31V	109150 (σ=33810)	36	1.29 (σ=0.34)	34
D30V	1140	6	0.06 (σ=0.01)	1
D31V	1140	6	0.09 (σ=0.03)	3
E30V	10251	29	0.64 (σ=0.12)	23
E31V	42219	32	0.79 (σ=0.17)	29
F30V	2268 (σ=111)	10	0.12 (σ=0.01)	5
G30V	459 (σ=20)	2	0.19 (σ=0.01)	8
G31V	4990 (σ=227)	18	1.49 (σ=0.08)	36
G32V	5736 (σ=261)	27	1.22 (σ=0.06)	32
H30V	2021	8	0.67 (σ=0.04)	24
H31V	2021	8	0.71 (σ=0.08)	28
H32V	2021	8	0.68 (σ=0.06)	25
I30V	4098 (σ=375)	17	0.68 (σ=0.09)	26
I31V	8152 (σ=747)	28	0.68 (σ=0.07)	27
J30V	5048	20	0.15 (σ=0.03)	6
J31V	5048	20	0.15 (σ=0.02)	7
J32V	2532	11	0.12 (σ=0.02)	4
K30V	5368	24	0.30 (σ=0.05)	10
K31V	5368	24	0.30 (σ=0.04)	11
K32V	5368	24	0.30 (σ=0.05)	13
K33V	5368	24	0.30 (σ=0.05)	12
L30V	1032	3	1.26 (σ=0.15)	33
M30V	2585	13	0.34 (σ=0.03)	16
M31V	2585	13	0.32 (σ=0.02)	15
M32V	2585	13	0.32 (σ=0.01)	14
N30V	3134	16	0.58 (σ=0.06)	22
N31V	3134	16	0.58 (σ=0.05)	21
N32V	12350	30	0.57 (σ=0.05)	20
N33V	12350	30	0.57 (σ=0.05)	19
Q30V	53264	34	0.46 (σ=0.17)	18
Q31V	53264	34	0.41 (σ=0.15)	17
R30V	128	1	0.21 (σ=0.09)	9

Table 34: Mean template size and generation time for each submission over 480 still face images from DATASET J: PASSENGER LOADING BRIDGE *. Standard deviations in parenthesis when there is variation. The mean interocular distance is 118 pixels.*

6.4 Still face template size

The time it takes to generate templates for still face images is less operationally relevant since it is only a factor when the database is being generated or altered. Its computation time is tiny compared to the time it takes to generate templates for video face tracks (see Table 34). The size of the templates will affect disk space requirements. For templates generated from a single still-face image, sizes range from 128 bytes (R30V) to 109,150 bytes (C30V, and C31V). Many systems operate by loading the entire database into memory to expedite matching. In this case, a system would need 109 MB of memory for a database of size 1,000 and 10.3 GB for a database of size 100,000.

Back in section 5.5.2 the effect of enrolling $K = 3$ images per subject was examined. Table 35 demonstrates that in such cases template generation time usually scales about linearly with the number of input faces. This is true in most cases, but with the exceptions of F30V, G30V, G31V, J30V, J31V, and J32V which maintain approximately constant size. This may occur because only one image is used, for example the "best" one, or because information is being fused across all K inputs, either at the image level (modeling) or at the feature level.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

Table 35: Mean template sizes in bytes, and mean template generation times in milliseconds, for each submission over 56 still face images from the DATASET H: TRAVEL WALKWAY *dataset. The column heading numbers give the number of still images passed to the template generation function. The numbers in light grey are the factor increase compared to the 1-pose column. The mean interocular distance is 199 pixels. The timing numbers are sometimes slower than the formal estimates over the larger set given in Table 34 due to higher resolution.*

Alg	Size1	Size3		Size5		Size7		Time1	Time3		Time5		Time7	
A30V	5070	14857	2.9	24553	4.8	33016	6.5	1340	3924	2.9	6551	4.9	9106	6.8
A31V	5070	14857	2.9	24553	4.8	33016	6.5	1371	4006	2.9	6516	4.8	8904	6.5
B30V	1060	3148	3.0	5236	4.9	7324	6.9	55	165	3.0	267	4.9	378	6.9
C30V	120593	373575	3.1	621151	5.2	866837	7.2	1263	3952	3.1	6528	5.2	9021	7.1
C31V	120593	373575	3.1	621151	5.2	866837	7.2	1292	3888	3.0	6542	5.1	9592	7.4
D30V	1140	3388	3.0	5616	4.9	7802	6.8	103	301	2.9	496	4.8	692	6.7
D31V	1140	3388	3.0	5616	4.9	7782	6.8	237	678	2.9	866	3.7	1163	4.9
E30V	10251	30713	3.0	51175	5.0	71637	7.0	660	1791	2.7	4080	6.2	4191	6.3
E31V	42219	126617	3.0	211015	5.0	295413	7.0	783	2189	2.8	3764	4.8	5151	6.6
F30V	2267	2308	1.0	2896	1.3	3142	1.4	111	314	2.8	503	4.5	656	5.9
G30V	460	436	0.9	436	0.9	436	0.9	206	596	2.9	1000	4.9	1484	7.2
G31V	5000	4976	1.0	4976	1.0	4976	1.0	1400	4316	3.1	7347	5.2	9887	7.1
G32V	5748	17220	3.0	28692	5.0	40062	7.0	1236	3682	3.0	6160	5.0	8938	7.2
H30V	2021	5991	3.0	9925	4.9	13858	6.9	707	1928	2.7	3334	4.7	4541	6.4
H31V	2021	5991	3.0	9925	4.9	13858	6.9	640	1925	3.0	3231	5.0	4556	7.1
H32V	2021	5991	3.0	9925	4.9	13858	6.9	782	1908	2.4	3328	4.3	4562	5.8
I30V	4058	12241	3.0	20128	5.0	27280	6.7	633	1947	3.1	3206	5.1	4365	6.9
I31V	8073	24359	3.0	40058	5.0	54290	6.7	770	1987	2.6	3214	4.2	4409	5.7
J30V	5048	5048	1.0	5048	1.0	5048	1.0	208	526	2.5	842	4.0	1163	5.6
J31V	5048	5048	1.0	5048	1.0	5048	1.0	222	540	2.4	860	3.9	1265	5.7
J32V	2532	2532	1.0	2532	1.0	2532	1.0	185	501	2.7	819	4.4	1161	6.3
K30V	5368	15809	2.9	26154	4.9	26345	4.9	600	1766	2.9	2927	4.9	4040	6.7
K31V	5368	16000	3.0	26537	4.9	37073	6.9	581	1769	3.0	2931	5.0	4076	7.0
K32V	5368	15809	2.9	26154	4.9	26345	4.9	578	1761	3.0	2921	5.1	4057	7.0
K33V	5368	16000	3.0	26537	4.9	37073	6.9	578	1765	3.1	2948	5.1	4078	7.1
L30V	1032	4104	4.0	6115	5.9	6152	6.0	1744	4647	2.7	8291	4.8	9865	5.7
M30V	2585	7755	3.0	12740	4.9	17818	6.9	371	1082	2.9	1800	4.9	2526	6.8
M31V	2585	7755	3.0	12740	4.9	17818	6.9	362	1096	3.0	1786	4.9	2515	6.9
M32V	2585	7755	3.0	12740	4.9	17818	6.9	371	1088	2.9	1804	4.9	2528	6.8
N30V	3134	9278	3.0	15422	4.9	21566	6.9	657	1985	3.0	3680	5.6	5385	8.2
N31V	3134	9278	3.0	15422	4.9	21566	6.9	659	2011	3.1	3727	5.7	5377	8.2
N32V	12350	36926	3.0	61502	5.0	86078	7.0	656	2008	3.1	3662	5.6	5299	8.1
N33V	12350	36926	3.0	61502	5.0	86078	7.0	659	2020	3.1	3666	5.6	5298	8.0
Q30V	53264	146476	2.8	245395	4.6	316731	5.9	1091	3254	3.0	5435	5.0	7602	7.0
Q31V	53264	146476	2.8	245395	4.6	316731	5.9	1040	3179	3.1	5215	5.0	7294	7.0
R30V	128	144	1.1	190	1.5	208	1.6	340	942	2.8	1531	4.5	2100	6.2

6.5 Memory usage during video processing

The FIVE approach to video processing is to pass an entire video clip to the algorithm via a single function call invocation. The size of the input data is equal to the product of the width of the image, the height, the number of color channels (always 3 here) and the number of frames. The largest single video clips in this study were those from the simulated aircraft boarding of DATASET J: PASSENGER LOADING BRIDGE , extending to about 12 minutes in length. Given 1920x1080 frames, 30 frames per second, the uncompressed data requires about 125 gigabytes of storage. Total memory requirements exceed that because algorithms generally allocate working memory. This is recorded for two somewhat shorter video clips in Figure 40. The plot shows, as vertical lines, the baseline amount of memory used by a dummy NIST implementation which did no computation.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

The algorithms generally consume very litle additional memory. The exceptions are the algorithms from the J, K, L and I developers. The K algorithms use more than double the input data size consistent with making a copy of the input data. We assume this is unnecessary and therefore not material to algorithm selection.

The FIVE measurements of memory use are likely representative to applications which do offline processing of video. For cases where video is streamed continuously to a face recognition system, programmers will need to architect a solution that operates on a first-in first-out buffer that, to first order, is sized about the length of the subject appearance, and is long enough to afford good tracking, noise suppression, and feature extraction.

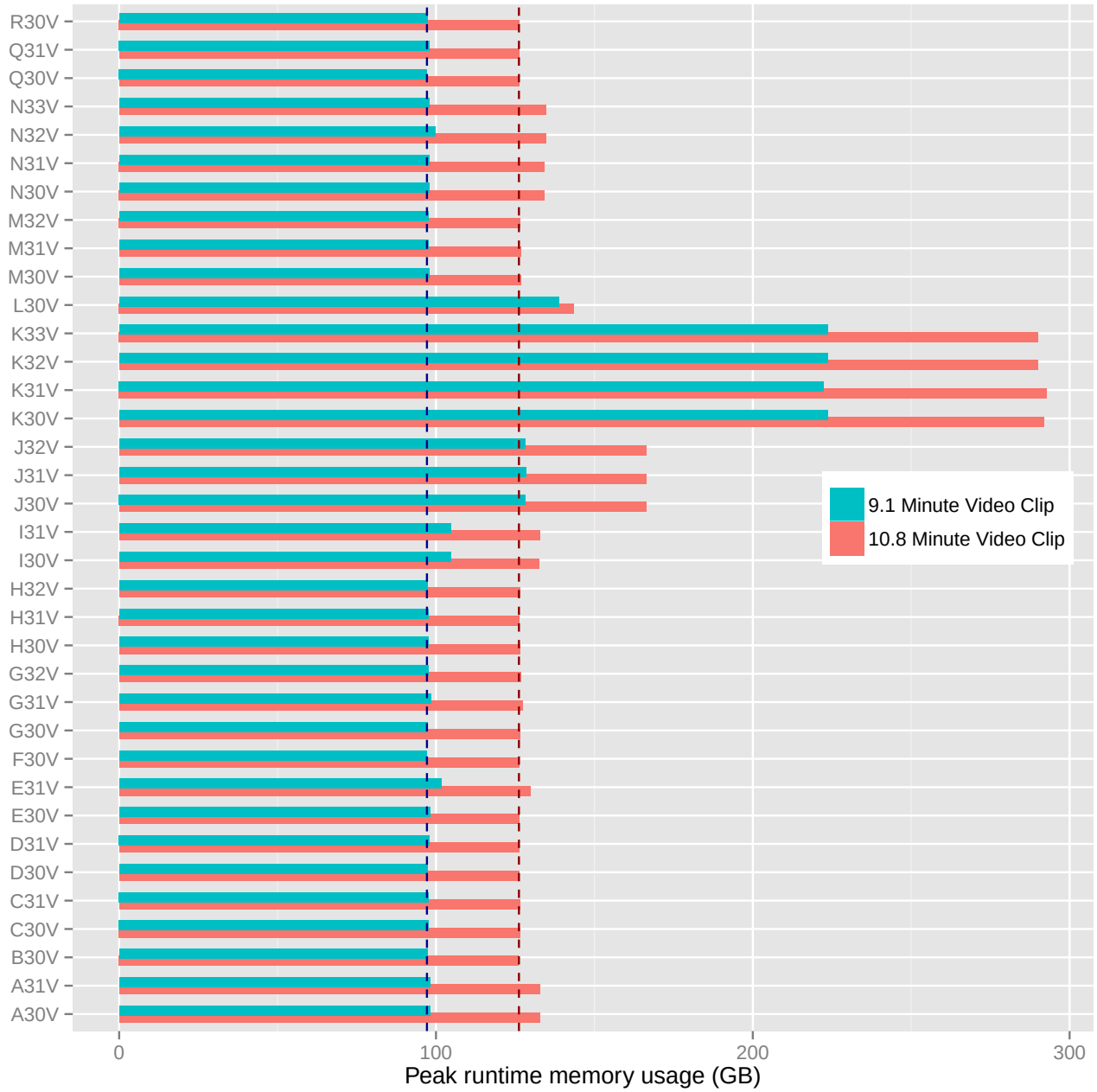


Figure 40: Maximum memory usage during video processing (template generation) using shorter and longer video clips from DATASET J: PASSENGER LOADING BRIDGE . *The dotted lines mark how much memory was required to load the raw video data (independent of the matching software).*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

References

[1] United states of america vs. dow chemical. Technical Report 84-1259. 476 U.S. 227, Supreme Court of the United States, May 1986. https://www.law.uh.edu/faculty/thester/courses/Emerging

[2] Kyllo v. united states. Technical Report 99-8508. 533 U.S. 27, Supreme Court of the United States, June 2001. https://supreme.justia.com/cases/federal/us/533/27/.

[3] Face recognition as a search tool foto-fahndung. Technical report, Bundeskriminamt (BKA), Thaerstrasse 11, 65193, Wiesbaden, Germany, February 2007.

[4] Striking the balance - a government approach to facial recognition privacy and civil liberties. Technical report, Federal Bureau of Investigation, FBI National Academy, Quantico, Virginia, March 2012. https://www.fbi.gov/file-repository/about-us-cjis-fingerprints_biometrics-biometric-center-of-excellences-forum_3_minutes.pdf.

[5] United states of america vs. leonel michel vargas. Technical Report 13-6025-EFS, United States District Court Eastern District of Washington, December 2014. https://www.eff.org/files/2014/12/15/vargas_order.pdf.

[6] A national surveillance camera strategy for england and wales [draft]. Technical report, Surveillance Camera Commissioner, First Floor, Peel Building, 2 Marsham Street, London, SW1P 4DF, October 2016. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/561520/NSCS_Strategy_FINAL.pdf.

[7] United states of america vs. rocky joe houston. Technical Report 14-5800, United States Court of Appeals for the Sixth Circuit, February 2016. http://www.opn.ca6.uscourts.gov/opinions.pdf/16a0031p-06.pdf.

[8] V. Biaud, C. Herold, V. Despiegel, and S. Gentric. Semi-supervised evaluation of face recognition in videos. In *Proc. Third International Biometrics Performance Conference, Gaithersburg (MD)*, 4 2014. https://www.nist.gov/itl/iad/image-group/ibpc-2014-presentations.

[9] Marc Jonathan Blitz. The fourth amendment future of public surveillance: Remote recording and other searches in public space. *American University Law Review*, 63(Issue 1 Article 2), 2013.

[10] Blumstein, Cohen, Roth, and Visher, editors. *Random parameter stochastic models of criminal careers*. National Academy of Sciences Press, 1986.

[11] Jeremy Brown. Pan, tilt, zoom: Regulating the use of video surveillance of public places. *Berkeley Technology Law Journal*, 23(Issue 1 Article 33), January 2008. http://scholarship.law.berkeley.edu/btlj/vol23/iss1/33.

[12] Mark Burge. Janus program broad agency announcement. Technical Report IARPA-BAA-13-07, Intelligence Advanced Research Projects Activity, 12 2013. https://www.fbo.gov/index?s=opportunity&id=9af4574d0fcf89a2fbcd9eb2d8c8921f.

[13] Jordan Cheney, Ben Klein, Anil K. Jain, and Brendan F. Klare. Unconstrained face detection: State of the art baseline and challenges. In *Proc. Eighth IAPR International Conference on Biometrics (ICB)*, May 2015.

[14] White D., Kemp R. I., Jenkins R., Matheson M, and Burton A. M. Passport officers errors in face matching. *PLoS ONE*, 9(8), 2014. e103510. doi:10.1371/journal. pone.0103510.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

[15] Working Group 3. Ed. D. D'Amato. *ISO/IEC 19794-5 Amendment 1 - Biometric data interchange formats - Part 5: Face image data - Conditions for taking photographs for face image data*. JTC1 :: SC37, 1 edition, 12 2007. http://webstore.ansi.org.

[16] Stephance Gentric et al. Border control: From technical to operational evaluation. In *Proceedings of the Second International Biometrics Performance Conference*, NIST, Gaithersburg, MD, March 2012.

[17] Gregory Fiumara, Wayne Salamon, and Craig Watson. Towards Repeatable, Reproducible, and Efficient Biometric Technology Evaluations. In *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*, pages 1–8, Sept 2015.

[18] Working Group 3. Ed. P. Griffin. *ISO/IEC 19794-5 Information Technology - Biometric Data Interchange Formats - Part 5: Face image data*. JTC1 :: SC37, 1 edition, 2005. http://webstore.ansi.org.

[19] P. Grother, G. W. Quinn, and P. J. Phillips. Evaluation of 2d still-image face recognition algorithms. NIST Interagency Report 7709, National Institute of Standards and Technology, 8 2010. http://face.nist.gov/mbe as MBE2010 FRVT2010.

[20] Patrick Grother and Mei Ngan. Face In Video Evaluation (FIVE) Concept, Evaluation Plan, and API, November 2014. http://biometrics.nist.gov/cs_links/face/five/NIST_FIVE_API_2014Nov19.pdf.

[21] Patrick Grother and Mei Ngan. Interagency report 8009, performance of face identification algorithms. *Face Recognition Vendor Test (FRVT)*, May 2014.

[22] Patrick Grother and Jonathon Phillips. Models of large population recognition performance. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC*, volume 2, pages 68–75, June 2004.

[23] Patrick Grother, George W. Quinn, and Mei Ngan. Face recognition vendor test - still face image and video concept, evaluation plan and api. Technical report, National Institute of Standards and Technology, 7 2013. http://biometrics.nist.gov/cs_links/face/frvt/frvt2012/NIST_FRVT2012_api_Aug15.pdf.

[24] Gary B. Huang and Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, 4 2014.

[25] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. *CoRR*, abs/1512.00596, 2015.

[26] Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, Mark J. Burge, and Anil K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark A. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1931–1939, 2015. http://dx.doi.org/10.1109/CVPR.2015.7298803.

[27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436 – 444, 5 2015.

[28] Kuang-Chih Lee, Jeffrey Ho, Ming-Hsuan Yang, and David J. Kriegman. Visual tracking and recognition using probabilistic appearance manifolds. *Computer Vision and Image Understanding*, 99(3):303–331, 2005.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

[29] Joseph N. Onek and Sharon Bradford Franklin. Guidelines for public video surveillance - a guide to protecting communities and preserving civil liberties. Technical report, The Constitution Project. A report of TCP's Liberty and Security Committee, 1200 18th Street, NW Suite 1000 Washington DC, 20036, 2007. http://www.constitutionproject.org/wp-content/uploads/2012/09/54.pdf.

[30] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[31] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[32] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security*, October 2016. http://dx.doi.org/10.1145/2976749.2978392.

[33] David White, James D. Dunn, Alexandra C. Schmid, and Richard I. Kemp. Error rates in users of automatic face recognition software. *PLoS ONE*, October 2015.

[34] Andreas Wolf. Reference facial images for mrtds - portrait quality. Technical report, International Civil Aviation Organization (ICAO), 2017. Technical Report.

[35] Galit Yovel and Alice J. OToole. Recognizing people in motion. *Trends in Cognitive Sciences*, 20(5):383–395, May 2016.

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

APPENDIX I: INTEROCULAR DISTANCES

DATASET P DOOR	VIDEO SEARCH						STILL IMAGE	
	NUMBER OF	MEAN TRACK	MEAN TRACK	MEAN	MEAN	MEAN	IOD (PIXELS)	
ALGORITHM	TRACKS	LENGTH (FRAMES)	EXTENT (FRAMES)	MIN IOD (PX)	MEAN IOD (PX)	MAX IOD (PX)	MEAN	STD. DEV
A30V	24383	10	10	35	37	39	123	25
A31V	24383	10	10	35	37	39	123	25
B30V	6014	44	53	34	46	142	125	25
C30V	18552	7	8	47	51	55	126	26
C31V	19962	6	7	45	49	53	126	26
D30V	22065	8	8	55	57	59	125	27
D31V	8741	29	30	42	48	55	125	28
E30V	4297	74	100	37	49	62	123	25
E31V	4297	74	100	37	49	62	123	25
F30V	35204	2	9	43	45	47	122	24
G30V	2673	52	78	41	49	62	125	25
G31V	3623	43	85	38	44	52	125	25
G32V	3623	43	85	38	44	52	125	25
H30V	5654	40	45	40	48	62	125	25
H31V	5654	40	45	40	48	62	125	25
H32V	5654	40	45	40	48	62	125	25
I30V	7629	50	50	16	28	40	127	25
I31V	7629	50	50	16	28	40	127	25
J30V	2564	43	63	46	55	64	127	25
J31V	2564	43	63	46	55	64	127	25
J32V	2564	43	63	46	55	64	127	25
K30V	34354	6	20	16	20	23	126	25
K31V	5022	61	358	7	27	45	126	25
K32V	2537	52	102	44	50	56	126	25
K33V	1680	89	228	42	51	61	126	25
L30V	6385	27	28	12	32	49	123	24
M30V	6993	35	40	34	38	43	111	22
M31V	6993	35	40	34	38	43	111	22
M32V	6993	35	40	34	38	43	111	22
N30V	4148	43	60	36	44	53	124	25
N31V	4148	43	60	36	44	53	124	25
N32V	4148	43	60	36	44	53	124	25
N33V	4148	43	60	36	44	53	124	25
Q30V	9750	27	29	30	37	44	121	24
Q31V	9750	27	29	30	37	44	121	24
R30V	6578	5	43	48	51	56	123	24

Table 36: For DATASET P: SPORTS ARENA *and each video processing algorithm the table shows: a) the number of reported tracks; b) the mean number of frames reported within those tracks; c) the mean extent (first minus last frames indices plus one); d) the mean over all tracks of of the minimum interocular distance (IOD) reported; e) the mean of the mean IOD; f) the mean of the maximum IOD; g) the enrollment still image mean IOD; and h) its standard deviation. For some algorithms (F, J, L, R) the tracks don't include all consecutive frames, so the extent of the track can exceed the number of frames in it. For the D, F, and L algorithms, the reported still-image eye coordinates are erroneous.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

DATASET P LOW NEAR	VIDEO SEARCH						STILL IMAGE	
	NUMBER OF	MEAN TRACK	MEAN TRACK	MEAN	MEAN	MEAN	IOD (PIXELS)	
ALGORITHM	TRACKS	LENGTH (FRAMES)	EXTENT (FRAMES)	MIN IOD (PX)	MEAN IOD (PX)	MAX IOD (PX)	MEAN	STD. DEV
A30V	45875	11	11	42	45	48	123	25
A31V	45875	11	11	42	45	48	123	25
B30V	8117	77	96	40	55	178	125	25
C30V	33188	9	10	55	59	64	126	26
C31V	36176	9	10	52	56	61	126	26
D30V	34092	11	11	60	63	65	125	27
D31V	14196	35	35	47	55	63	125	28
E30V	5697	95	129	37	47	55	123	25
E31V	5697	95	129	37	47	55	123	25
F30V	36430	5	32	59	62	66	122	24
G30V	4793	83	116	45	55	69	125	25
G31V	4403	106	201	38	47	57	125	25
G32V	4403	106	201	38	47	57	125	25
H30V	9148	55	60	47	58	74	125	25
H31V	9148	55	60	47	58	74	125	25
H32V	9148	55	60	47	58	74	125	25
I30V	12869	62	62	15	31	47	127	25
I31V	12869	62	62	15	31	47	127	25
J30V	4021	74	98	53	66	77	127	25
J31V	4021	74	98	53	66	77	127	25
J32V	4021	74	98	53	66	77	127	25
K30V	41675	12	31	15	23	27	126	25
K31V	8820	73	344	9	33	50	126	25
K32V	3694	102	225	50	56	62	126	25
K33V	3629	115	279	50	57	65	126	25
L30V	11966	33	34	18	40	59	123	24
M30V	12700	46	52	36	42	47	111	22
M31V	12700	46	52	36	42	47	111	22
M32V	12700	46	52	36	42	47	111	22
N30V	7733	63	79	42	52	61	124	25
N31V	7733	63	79	42	52	61	124	25
N32V	7733	63	79	42	52	61	124	25
N33V	7733	63	79	42	52	61	124	25
Q30V	16025	32	35	34	45	53	121	24
Q31V	16025	32	35	34	45	53	121	24
R30V	9614	4	59	59	65	71	123	24

Table 37: For DATASET P: SPORTS ARENA *and each video processing algorithm the table shows: a) the number of reported tracks; b) the mean number of frames reported within those tracks; c) the mean extent (first minus last frames indices plus one); d) the mean over all tracks of of the minimum interocular distance (IOD) reported; e) the mean of the mean IOD; f) the mean of the maximum IOD; g) the enrollment still image mean IOD; and h) its standard deviation. For some algorithms (F, J, L, R) the tracks don't include all consecutive frames, so the extent of the track can exceed the number of frames in it. For the D, F, and L algorithms, the reported still-image eye coordinates are erroneous.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

DATASET P HIGH NEAR	VIDEO SEARCH						STILL IMAGE	
	NUMBER OF	MEAN TRACK	MEAN TRACK	MEAN	MEAN	MEAN	IOD (PIXELS)	
ALGORITHM	TRACKS	LENGTH (FRAMES)	EXTENT (FRAMES)	MIN IOD (PX)	MEAN IOD (PX)	MAX IOD (PX)	MEAN	STD. DEV
A30V	21953	6	6	39	41	42	123	25
A31V	21953	6	6	39	41	42	123	25
B30V	3353	70	99	40	61	269	125	25
C30V	7852	7	8	56	60	64	126	26
C31V	8072	6	8	55	59	63	126	26
D30V	23794	6	6	63	64	66	125	27
D31V	7954	22	23	50	56	63	125	28
E30V	1919	95	127	43	52	63	123	25
E31V	1919	95	127	43	52	63	123	25
F30V	9980	4	23	55	57	58	122	24
G30V	1125	97	164	48	58	76	125	25
G31V	1538	68	165	44	51	60	125	25
G32V	1538	68	165	44	51	60	125	25
H30V	3700	43	50	51	61	79	125	25
H31V	3700	43	50	51	61	79	125	25
H32V	3700	43	50	51	61	79	125	25
I30V	4529	60	60	15	28	44	127	25
I31V	4529	60	60	15	28	44	127	25
J30V	1618	54	86	52	60	68	127	25
J31V	1618	54	86	52	60	68	127	25
J32V	1618	54	86	52	60	68	127	25
K30V	28308	5	22	12	16	19	126	25
K31V	5128	42	353	6	24	40	126	25
K32V	979	107	229	48	55	61	126	25
K33V	984	135	323	46	54	62	126	25
L30V	3034	26	27	15	31	57	123	24
M30V	4216	36	44	44	48	53	111	22
M31V	4216	36	44	44	48	53	111	22
M32V	4216	36	44	44	48	53	111	22
N30V	2064	55	89	44	53	62	124	25
N31V	2064	55	89	44	53	62	124	25
N32V	2064	55	89	44	53	62	124	25
N33V	2064	55	89	44	53	62	124	25
Q30V	10421	19	23	30	37	42	121	24
Q31V	10421	19	23	30	37	42	121	24
R30V	3653	3	50	57	61	65	123	24

Table 38: For DATASET P: SPORTS ARENA *and each video processing algorithm the table shows: a) the number of reported tracks; b) the mean number of frames reported within those tracks; c) the mean extent (first minus last frames indices plus one); d) the mean over all tracks of of the minimum interocular distance (IOD) reported; e) the mean of the mean IOD; f) the mean of the maximum IOD; g) the enrollment still image mean IOD; and h) its standard deviation. For some algorithms (F, J, L, R) the tracks don't include all consecutive frames, so the extent of the track can exceed the number of frames in it. For the D, F, and L algorithms, the reported still-image eye coordinates are erroneous.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

DATASET P LOW FAR	VIDEO SEARCH						STILL IMAGE	
	NUMBER OF	MEAN TRACK	MEAN TRACK	MEAN	MEAN	MEAN	IOD (PIXELS)	
ALGORITHM	TRACKS	LENGTH (FRAMES)	EXTENT (FRAMES)	MIN IOD (PX)	MEAN IOD (PX)	MAX IOD (PX)	MEAN	STD. DEV
A30V	65205	5	5	27	28	30	123	25
A31V	65205	5	5	27	28	30	123	25
B30V	10230	24	34	32	43	125	125	25
C30V	23004	4	5	37	40	42	126	26
C31V	32672	4	5	31	34	36	126	26
D30V	24399	5	5	60	63	66	125	27
D31V	12360	15	16	44	50	57	125	28
E30V	8802	41	63	24	29	36	123	25
E31V	8802	41	63	24	29	36	123	25
F30V	27367	2	5	41	42	44	122	24
G30V	3023	23	38	39	45	54	125	25
G31V	9440	16	43	28	31	36	125	25
G32V	9440	16	43	28	31	36	125	25
H30V	9088	23	28	36	42	51	125	25
H31V	9088	23	28	36	42	51	125	25
H32V	9088	23	28	36	42	51	125	25
I30V	19650	34	34	12	18	25	127	25
I31V	19650	34	34	12	18	25	127	25
J30V	3178	20	31	48	53	61	127	25
J31V	3178	20	31	48	53	61	127	25
J32V	3178	20	31	48	53	61	127	25
K30V	62966	3	18	15	18	20	126	25
K31V	7744	53	332	5	23	38	126	25
K32V	3488	29	72	32	36	39	126	25
K33V	2211	59	226	29	36	44	126	25
L30V	9640	17	18	6	21	38	123	24
M30V	16696	19	24	24	26	29	111	22
M31V	16696	19	24	24	26	29	111	22
M32V	16696	19	24	24	26	29	111	22
N30V	5924	26	40	34	41	49	124	25
N31V	5924	26	40	34	41	49	124	25
N32V	5924	26	40	34	41	49	124	25
N33V	5924	26	40	34	41	49	124	25
Q30V	34187	16	20	19	22	26	121	24
Q31V	34187	16	20	19	22	26	121	24
R30V	7128	3	41	65	70	77	123	24

Table 39: For DATASET P: SPORTS ARENA *and each video processing algorithm the table shows: a) the number of reported tracks; b) the mean number of frames reported within those tracks; c) the mean extent (first minus last frames indices plus one); d) the mean over all tracks of of the minimum interocular distance (IOD) reported; e) the mean of the mean IOD; f) the mean of the maximum IOD; g) the enrollment still image mean IOD; and h) its standard deviation. For some algorithms (F, J, L, R) the tracks don't include all consecutive frames, so the extent of the track can exceed the number of frames in it. For the D, F, and L algorithms, the reported still-image eye coordinates are erroneous.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

DATASET P HIGH FAR	VIDEO SEARCH						STILL IMAGE	
	NUMBER OF	MEAN TRACK	MEAN TRACK	MEAN	MEAN	MEAN	IOD (PIXELS)	
ALGORITHM	TRACKS	LENGTH (FRAMES)	EXTENT (FRAMES)	MIN IOD (PX)	MEAN IOD (PX)	MAX IOD (PX)	MEAN	STD. DEV
A30V	92654	6	6	22	23	24	123	25
A31V	92654	6	6	22	23	24	123	25
B30V	15079	24	32	28	36	96	125	25
C30V	40243	4	5	30	33	35	126	26
C31V	60916	5	6	25	28	30	126	26
D30V	29462	4	4	60	62	63	125	27
D31V	15773	16	18	38	42	47	125	28
E30V	13747	57	83	21	25	31	123	25
E31V	13747	57	83	21	25	31	123	25
F30V	40336	2	4	36	38	39	122	24
G30V	4209	22	43	32	37	46	125	25
G31V	12635	23	58	25	28	33	125	25
G32V	12635	23	58	25	28	33	125	25
H30V	11421	32	39	30	35	45	125	25
H31V	11421	32	39	30	35	45	125	25
H32V	11421	32	39	30	35	45	125	25
I30V	32034	42	42	9	15	20	127	25
I31V	32034	42	42	9	15	20	127	25
J30V	2937	22	36	41	46	53	127	25
J31V	2937	22	36	41	46	53	127	25
J32V	2937	22	36	41	46	53	127	25
K30V	73715	6	20	13	18	21	126	25
K31V	13124	58	217	2	18	31	126	25
K32V	6657	32	81	24	27	30	126	25
K33V	4198	69	266	21	26	33	126	25
L30V	14776	19	21	5	19	33	123	24
M30V	29021	24	30	19	21	24	111	22
M31V	29021	24	30	19	21	24	111	22
M32V	29021	24	30	19	21	24	111	22
N30V	7108	32	52	27	33	39	124	25
N31V	7108	32	52	27	33	39	124	25
N32V	7108	32	52	27	33	39	124	25
N33V	7108	32	52	27	33	39	124	25
Q30V	34001	23	26	20	23	27	121	24
Q31V	34001	23	26	20	23	27	121	24
R30V	8551	3	33	43	45	50	123	24

Table 40: For DATASET P: SPORTS ARENA *and each video processing algorithm the table shows: a) the number of reported tracks; b) the mean number of frames reported within those tracks; c) the mean extent (first minus last frames indices plus one); d) the mean over all tracks of of the minimum interocular distance (IOD) reported; e) the mean of the mean IOD; f) the mean of the maximum IOD; g) the enrollment still image mean IOD; and h) its standard deviation. For some algorithms (F, J, L, R) the tracks don't include all consecutive frames, so the extent of the track can exceed the number of frames in it. For the D, F, and L algorithms, the reported still-image eye coordinates are erroneous.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

DATASET H F	VIDEO SEARCH						STILL IMAGE	
	NUMBER OF	MEAN TRACK	MEAN TRACK	MEAN	MEAN	MEAN	IOD (PIXELS)	
ALGORITHM	TRACKS	LENGTH (FRAMES)	EXTENT (FRAMES)	MIN IOD (PX)	MEAN IOD (PX)	MAX IOD (PX)	MEAN	STD. DEV
A30V	251	11	11	35	38	41	123	30
A31V	251	11	11	35	38	41	123	30
B30V	51	73	77	40	71	485	125	30
C30V	205	9	10	50	53	57	127	32
C31V	207	9	10	49	53	57	127	32
D30V	192	15	15	51	55	60	126	36
D31V	66	50	51	42	54	67	126	36
E30V	41	81	92	39	55	78	123	31
E31V	41	81	92	39	55	78	123	31
F30V	283	5	28	51	53	56	121	30
G30V	34	55	86	41	56	82	125	31
G31V	36	73	297	37	49	70	125	31
G32V	36	73	297	37	49	70	125	31
H30V	70	48	53	62	80	108	126	31
H31V	70	48	53	62	80	108	126	31
H32V	70	48	53	62	80	108	126	31
I30V	62	74	74	22	43	64	126	31
I31V	62	74	74	22	43	64	126	31
J30V	43	57	68	49	63	82	127	32
J31V	43	57	68	49	63	82	127	32
J32V	43	57	68	49	63	82	127	32
K30V	917	4	189	15	17	18	127	31
K31V	59	87	2460	10	28	52	127	31
K32V	45	51	163	43	52	65	127	31
K33V	36	64	302	43	54	71	127	31
L30V	63	35	37	21	47	66	123	30
M30V	56	60	66	36	47	60	111	28
M31V	56	60	66	36	47	60	111	28
M32V	56	60	66	36	47	60	111	28
N30V	52	56	68	38	53	71	125	31
N31V	52	56	68	38	53	71	125	31
N32V	52	56	68	38	53	71	125	31
N33V	52	56	68	38	53	71	125	31
Q30V	72	40	41	39	50	61	122	30
Q31V	72	40	41	39	50	61	122	30
R30V	120	2	96	47	51	56	123	30

Table 41: For DATASET H: TRAVEL WALKWAY *and each video processing algorithm the table shows: a) the number of reported tracks; b) the mean number of frames reported within those tracks; c) the mean extent (first minus last frames indices plus one); d) the mean over all tracks of of the minimum interocular distance (IOD) reported; e) the mean of the mean IOD; f) the mean of the maximum IOD; g) the enrollment still image mean IOD; and h) its standard deviation. For some algorithms (F, J, L, R) the tracks don't include all consecutive frames, so the extent of the track can exceed the number of frames in it. For the D, F, and L algorithms, the reported still-image eye coordinates are erroneous.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

DATASET H R	VIDEO SEARCH						STILL IMAGE	
	NUMBER OF	MEAN TRACK	MEAN TRACK	MEAN	MEAN	MEAN	IOD (PIXELS)	
ALGORITHM	TRACKS	LENGTH (FRAMES)	EXTENT (FRAMES)	MIN IOD (PX)	MEAN IOD (PX)	MAX IOD (PX)	MEAN	STD. DEV
A30V	2020	5	5	68	71	75	123	30
A31V	2020	5	5	68	71	75	123	30
B30V	170	61	70	45	80	427	125	30
C30V	695	6	7	60	63	66	127	32
C31V	696	6	7	60	63	66	127	32
D30V	7118	2	2	50	52	54	126	36
D31V	1354	36	31	42	49	57	126	36
E30V	117	79	102	43	58	81	123	31
E31V	117	79	102	43	58	81	123	31
F30V	731	3	12	51	53	55	121	30
G30V	117	47	76	51	67	92	125	31
G31V	167	46	236	51	63	81	125	31
G32V	167	46	236	51	63	81	125	31
H30V	166	61	66	48	61	82	126	31
H31V	166	61	66	48	61	82	126	31
H32V	166	61	66	48	61	82	126	31
I30V	182	72	72	29	51	73	126	31
I31V	182	72	72	29	51	73	126	31
J30V	141	52	62	52	65	83	127	32
J31V	141	52	62	52	65	83	127	32
J32V	141	52	62	52	65	83	127	32
K30V	3334	3	196	13	15	18	127	31
K31V	266	63	2062	7	32	56	127	31
K32V	133	55	98	48	60	77	127	31
K33V	101	81	371	46	62	86	127	31
L30V	204	26	27	16	40	67	123	30
M30V	181	54	59	41	51	64	111	28
M31V	181	54	59	41	51	64	111	28
M32V	181	54	59	41	51	64	111	28
N30V	154	60	68	42	57	79	125	31
N31V	154	60	68	42	57	79	125	31
N32V	154	60	68	42	57	79	125	31
N33V	154	60	68	42	57	79	125	31
Q30V	695	29	34	33	40	46	122	30
Q31V	695	29	34	33	40	46	122	30
R30V	364	2	113	63	68	75	123	30

Table 42: For DATASET H: TRAVEL WALKWAY *and each video processing algorithm the table shows: a) the number of reported tracks; b) the mean number of frames reported within those tracks; c) the mean extent (first minus last frames indices plus one); d) the mean over all tracks of of the minimum interocular distance (IOD) reported; e) the mean of the mean IOD; f) the mean of the maximum IOD; g) the enrollment still image mean IOD; and h) its standard deviation. For some algorithms (F, J, L, R) the tracks don't include all consecutive frames, so the extent of the track can exceed the number of frames in it. For the D, F, and L algorithms, the reported still-image eye coordinates are erroneous.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

DATASET H S	VIDEO SEARCH						STILL IMAGE	
	NUMBER OF	MEAN TRACK	MEAN TRACK	MEAN	MEAN	MEAN	IOD (PIXELS)	
ALGORITHM	TRACKS	LENGTH (FRAMES)	EXTENT (FRAMES)	MIN IOD (PX)	MEAN IOD (PX)	MAX IOD (PX)	MEAN	STD. DEV
A30V	1155	13	13	59	63	67	123	30
A31V	1155	13	13	59	63	67	123	30
B30V	271	65	75	61	86	307	125	30
C30V	1442	7	8	91	97	103	127	32
C31V	1447	7	8	91	97	102	127	32
D30V	1708	9	9	64	66	69	126	36
D31V	524	32	33	54	63	72	126	36
E30V	167	74	83	58	74	98	123	31
E31V	167	74	83	58	74	98	123	31
F30V	1516	6	24	66	69	73	121	30
G30V	196	48	73	58	72	92	125	31
G31V	175	83	301	53	67	85	125	31
G32V	175	83	301	53	67	85	125	31
H30V	257	62	66	60	73	93	126	31
H31V	257	62	66	60	73	93	126	31
H32V	257	62	66	60	73	93	126	31
I30V	253	77	77	34	59	80	126	31
I31V	253	77	77	34	59	80	126	31
J30V	224	59	68	62	74	89	127	32
J31V	224	59	68	62	74	89	127	32
J32V	224	59	68	62	74	89	127	32
K30V	4715	5	323	14	16	18	127	31
K31V	218	124	2508	20	46	76	127	31
K32V	209	74	150	57	71	87	127	31
K33V	184	85	239	58	74	92	127	31
L30V	329	34	36	38	62	80	123	30
M30V	248	61	65	49	62	76	111	28
M31V	248	61	65	49	62	76	111	28
M32V	248	61	65	49	62	76	111	28
N30V	262	53	58	56	70	87	125	31
N31V	262	53	58	56	70	87	125	31
N32V	262	53	58	56	70	87	125	31
N33V	262	53	58	56	70	87	125	31
Q30V	413	34	37	51	62	71	122	30
Q31V	413	34	37	51	62	71	122	30
R30V	537	2	74	63	66	71	123	30

Table 43: For DATASET H: TRAVEL WALKWAY *and each video processing algorithm the table shows: a) the number of reported tracks; b) the mean number of frames reported within those tracks; c) the mean extent (first minus last frames indices plus one); d) the mean over all tracks of of the minimum interocular distance (IOD) reported; e) the mean of the mean IOD; f) the mean of the maximum IOD; g) the enrollment still image mean IOD; and h) its standard deviation. For some algorithms (F, J, L, R) the tracks don't include all consecutive frames, so the extent of the track can exceed the number of frames in it. For the D, F, and L algorithms, the reported still-image eye coordinates are erroneous.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			

DATASET C	VIDEO SEARCH						STILL IMAGE	
	NUMBER OF	MEAN TRACK	MEAN TRACK	MEAN	MEAN	MEAN	IOD (PIXELS)	
ALGORITHM	TRACKS	LENGTH (FRAMES)	EXTENT (FRAMES)	MIN IOD (PX)	MEAN IOD (PX)	MAX IOD (PX)	MEAN	STD. DEV
A30V	215513	18	18	35	37	39	55	79
A31V	215513	18	18	35	37	39	55	79
B30V	26477	149	179	34	45	145	97	92
C30V	148870	15	17	45	48	52	56	81
C31V	170834	15	17	41	44	47	56	81
D30V	109836	19	19	55	57	59	5679	10451
D31V	45934	60	61	44	50	55	6576	10734
E30V	26789	161	187	30	36	42	88	87
E31V	26771	160	186	30	36	42	88	87
F30V	253765	9	84	35	37	39	3502	8332
G30V	20905	148	179	40	48	59	58	85
G31V	20831	171	250	35	41	50	62	81
G32V	20731	172	250	35	41	50	62	81
H30V	40428	84	89	37	44	54	64	85
H31V	40428	84	89	37	44	54	64	85
H32V	40428	84	89	37	44	54	64	85
I30V	65856	95	95	14	23	31	51	79
I31V	65856	95	95	14	23	31	51	79
J30V	21972	122	151	44	52	58	53	86
J31V	21972	122	151	44	52	58	53	86
J32V	21972	122	151	44	52	58	61	87
K30V	109509	43	87	16	26	33	67	68
K31V	41140	127	326	12	28	39	68	69
K32V	26747	155	266	32	36	40	67	67
K33V	24674	184	363	31	36	41	68	69
L30V	80041	42	44	19	29	40	1226	5345
M30V	60017	81	87	26	29	33	50	64
M31V	60017	81	87	26	29	33	50	64
M32V	60017	81	87	26	29	33	50	64
N30V	25403	112	135	37	45	53	81	81
N31V	25403	112	135	37	45	53	81	81
N32V	25403	112	135	37	45	53	81	81
N33V	25403	112	135	37	45	53	81	81
Q30V	62617	46	48	29	36	41	36	64
Q31V	62617	46	48	29	36	41	36	64
R30V	55916	4	86	40	43	47	55	82

Table 44: For DATASET C: PHOTOJOURNALISM *and each video processing algorithm the table shows: a) the number of reported tracks; b) the mean number of frames reported within those tracks; c) the mean extent (first minus last frames indices plus one); d) the mean over all tracks of of the minimum interocular distance (IOD) reported; e) the mean of the mean IOD; f) the mean of the maximum IOD; g) the enrollment still image mean IOD; and h) its standard deviation. For some algorithms (F, J, L, R) the tracks don't include all consecutive frames, so the extent of the track can exceed the number of frames in it. For the D, F, and L algorithms, the reported still-image eye coordinates are erroneous.*

PARTICIPANT KEY			
A = DIGITAL BARRIERS	E = NEUROTECHNOLOGY	I = EYEDEA	M = NEC
B = HBINNO	F = VAPPLICA	J = HISIGN	N = TOSHIBA
C = VIGILANT	G = MORPHO	K = COGNITEC	Q = IMAGUS
D = AYONIX	H = 3M COGENT	L = CYBEREXTRUDER	R = RANK ONE

SET	SCENE	CAMERA	SET	SCENE	CAMERA
C	PHOTOJOURNALISM	PRO	T	CONCOURSE	PRO
J	PASSENGER LOADING	PRO	H	CONCOURSE	PRO
P	SPORTS ARENA	CONSUMER	U	CHOKEPOINT	WEBCAM
L	LUGGAGE RACK	WEBCAM			