FaceTec 3D Face Matching Accuracy Report

Updated: May 2nd, 2022

Introduction

This report self-certifies the accuracy of FaceTec's 1:1 3D face matching algorithm. It reports the False Acceptance Rate (**FAR**) and False Rejection Rate (**FRR**) at various thresholds and compares them to other algorithms from research/testing organizations and biometric matching vendors.

3rd Party Testing & Certification

FaceTec is pleased to announce that A.C.C.S. Ltd., a UKAS-accredited conformity assessment body composed of auditors, certification specialists, and data protection experts that independently test and certify online and offline systems that check age and identity (such as passport scanners, biometric technology, and age verification software) has been contracted to perform a large-scale public test of FaceTec's 3D:3D Matching in an N:N context. This data collection from 10,000-20,000 unique subjects will begin in Q2, 2022, and we will publish the results in this report via an update after the testing is complete. FaceTec uses real-world data in its training and test datasets from volunteer users from 180+ countries, but still seeks to validate its accuracy in third-party testing.

Definitions

Unique Identity Number (UID#): Each person in the FaceTec dataset is assigned a unique numerical identifier; this is their UID#. If a person's face images are collected in two or more different capture sessions, the sessions will all be assigned to the same UID#.

Threshold (T): Given a pair of sessions (images, group of images, image data, or numerical representation of a face like FaceTec's 3D FaceVectors), a verification system outputs the probability (or a score) that the UID#s corresponding to the sessions are the same. This output probability is binarized based on a parameter called the "Threshold" (T). If the probability (score) is greater than T, the two UID#s are said to match. The threshold controls the tradeoff between the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) of the system.

False Acceptance Rate/False Rejection Rate (FAR/FRR): For a particular threshold, the FAR/FRR of a face verification/matching system is the probability that it will incorrectly match two sessions corresponding to two different UID#s (FAR), as well as the probability that two sessions with the same UID# are incorrectly marked as different users (FRR). Reporting FAR or FRR alone is an incomplete measure of accuracy in 1:1 face matching. Lower values for both FAR and FRR is the goal as lower values indicate higher match confidence and usability.

Reporting Methodology

Two common methods for reporting face matching algorithm accuracy that are used in industry and academia:

1. "All Combinations Method" -- **All possible pairs (both genuine and imposter) are tested**. FaceTec uses this method because it represents quite literally all possibilities that exist, and thus is the most real-world performance metric. This method is also most similar to the NIST FRVT testing method.
    a. This method is superior because it tests everything against everything else, instead of smaller random samples of the dataset.

2. "LFW Method" -- A different way to report face matching results in many academic papers as well as public dataset performance, like the "Labeled Faces in the Wild" (LFW) dataset. This method relies on random sampling and "10-fold cross-validation". This reporting method is often used because it outputs one single metric for overall accuracy. FaceTec does not use this method because:
    a. The LFW Method is based on a reporting method intended for use on identification (1:N) algorithms, not authentication/verification (1:1).
    b. The LFW dataset is intended for use on identification (1:N) algorithms, not authentication/verification (1:1).
    c. FaceTec's 1:1 3D Matching Algorithm is "too accurate" to report this metric. Random sampling generates a **significantly smaller dataset size. Because of this, accuracy (when measured on a sampled dataset) is very frequently 100%. This is not useful when comparing algorithms.**

FaceTec Dataset Properties

FaceTec strives to objectively test and report the accuracy of its algorithms.

FaceTec dataset properties:

1. 100% of the data was captured from real-world user devices running FaceTec SDKs.
2. The test set and training sets were obtained by randomly selecting UID#s and included all FaceTec 3D FaceMaps/FaceVectors for those UID#s. The individuals/UID#s that are in the test set are ensured to not have been in the training set.
3. The number of imposter comparisons is ~4,000,000,000 (10x the imposter comparisons in NIST FRVT Mugshot, which they report to be "10^8").
4. The dataset includes a wide spectrum of device, camera resolution, screen size, screen brightness, age, gender, ethnicity, and eyeglass-wearer combinations.
5. Environmental lighting is uncontrolled.
6. 3D FaceMaps/FaceVectors were evaluated from devices and users from 180+ different countries.
    a. Note: The NIST Mugshot database is US-only, and does not adequately represent the significant diversity in face phenotypes that exist across the globe.
7. The dataset consists of 100,000,000+ User Images from ~100,000 Unique User Identities, from which 600,000 3D FaceMaps are derived. FaceTec AI trains on 3D FaceMaps from every modern iOS device, over 10,000 unique Android device models, and thousands of webcam models, including those with very low resolutions, down to .3 megapixels.
8. The FaceTec dataset contains a high percentage of sessions where the user's face is not in ideal lighting. This means the face has uneven/directional light, glare in glasses, shadows, and low-light scenarios.
9. 3D FaceMap data collection from the same subject can span up to seven years and are included.
10. Users' ages are estimated to be between 18 and 95 years old.

No Observable Bias in Matching Errors

When errors are made in our testing by the FaceTec 3D Face Matching Algorithms (or the FaceTec 3D Liveness Algorithms, for that matter), the misidentified people do not appear in any way to our trained evaluators to have any pattern of error. While all systems that use visible light to capture biometric data will have some bias related to the capture of visible light, at the thresholds we have published in this report, the bias level for device, camera resolution, screen size, screen brightness, age, gender, device, country-of-origin, ethnicity, and eyeglass-wearer combinations are not noticed by our trained evaluators, and we consider them to be unobservable.

While developing and improving our algorithms, FaceTec assesses the results from many billions of match pairs. However, because the error rate is so low, the number of misidentifications is also very low, and our trained evaluators manually review 100% of the misidentifications.

FaceTec FAR/FRR Results

False Acceptance Rate (FAR)	False Rejection Rate (FRR)
1/12,800,00	0.0009 (0.09%)
1/25,000,000	0.0022 (0.13%)
1/50,000,000	0.0022 (0.18%)
1/70,00,000	0.0022 (0.29%)
1/95,000,000	0.0022 (0.25%)
1/125,000,000	< 0.0099 (0.99%)

Comparison to FaceTec's Previous 1/12,800,000 FAR Model

The Testing Dataset from the previously-released, 1/12,800,000 FAR Model was held out (individuals/UIDs from the Testing Dataset were ensured to not be in the Training Dataset) and subsequently tested against using the new 1/125,000,000 FAR Model. The previous 1/12.8M Model had exactly 61 False Accepts against its own Testing Dataset. The new 1/125M FAR Model had exactly 1 False Accept against this same Testing Dataset.

In summary, FaceTec's 1/125M Model had only ONE (1) error compared to the previous algorithm's 60 errors. (~98.4% reduction in error rate versus this specific Testing Dataset). This is irrefutable evidence that the new model is significantly more accurate on the previous test set and we expect this increased accuracy to extend to real-word performance.

That single False Accept is shown below:

Note: The above images are not of the same person, and the 125M Model incorrectly matched the two faces at the .99% FRR level. This was the only false accept out of billions of match pairs with the new 125M Model.

Building Confidence in Measured FAR/FRR Results - Internal Procedure/Reviews

FaceTec understands that 1/125,000,000 FAR @ <0.99% FRR is a bold claim, especially considering this is over 100x better than **both** the last reported FAR for Apple's Face ID and NIST's top 1:1 Face Matching Algorithms. To build confidence in our findings, the following procedures were followed and reviews were performed:

- Ongoing Training+Testing Runs, randomizing the Train/Test Splits for each run. Verification that FAR/FRR operating points remain roughly the same.
- Multiple test runs that hold out the previous 1/12.8M Testing Dataset and subsequently running the new Model against the previous Testing Dataset and verifying >10x improvements to FAR/FRRs against the 1/12.8M Testing Dataset.
- Performing a multitude of Training+Testing Runs utilizing different Training/Testing Splits – i.e., 60/40, 70/30, 80/20, 90/10. Verifying better results consistently as more data is incorporated into the Learned Model, and that both hard and easy samples do not skew the results heavily run-over-run.
- Running each Model against additional demographics-specific Hold Out Datasets that have not been seen by any Training+Testing Run and verifying >10 improvements to FAR/FRR operating points.
- Manual Review of all FARs and FRRs from a majority of Training+Testing Runs.
- Ongoing Review to ensure that no Individuals/Identities/UIDs are mislabeled.
- Manual ad-hoc inspection of Training+Testing Datasets run-over-run, to ensure true randomization of identities, that identities are not always in either the Training or Testing Dataset, and ensure that data from different Identities are not being somehow mixed and matched at runtime.
- FaceTec continues to build its Dataset of identities by ground-truthing new sessions daily. (FaceTec receives sessions from many hundreds, and sometimes thousands, of new volunteer Demo Users per day). These User are reviewed by FaceTec's in-house Data Management Team, verified not to be in the Training or Test Dataset already, and then are added to the Test Dataset in order to ensure stated error rates continue to hold true for new device models, and build confidence that rates as stated accurately.

Building Confidence in Measured FAR/FRR Results - Customer Validation

FaceTec works with myriad organizations that have extensive datasets ranging from hundreds-of-thousands to over 100 million unique users, and these organizations have reported real-world performance in line with the expectations of the 1/12.8M FAR Model. We expect the new 1/125M Model will perform in real-world usage >10x better than the previous 1/12.8M FAR Model. As of the publishing of this report, no regressions have been reported, and the accuracies shown in this report continue to be upheld.

FaceTec Dataset vs. NIST FRVT 1:1 Dataset (Mugshot)

	NIST Mugshot 2D Image Quality	FaceTec Dataset 3D FaceMap Quality
Best Quality Images From Dataset	 The entire NIST dataset consists of single, 2D images in ideal conditions. (2 photos per Identity on average.)	 Some 3D FaceMaps are created from ideal conditions, but most are not. And with an average of six (6) 3D FaceMaps per Identity, nearly every Identity contains data collected in non-ideal conditions
Average Quality	⚠️ The average quality of NIST Mugshot images were captured under ideal scenarios, so the average quality of the NIST Mugshot dataset is very high.	FaceTec's dataset only contains 3D FaceMaps collected from real users in real-world conditions, including significant variations in lighting, pose, shadows, expressions, glare, camera quality, and camera resolution. 

Dataset Comparison Remarks

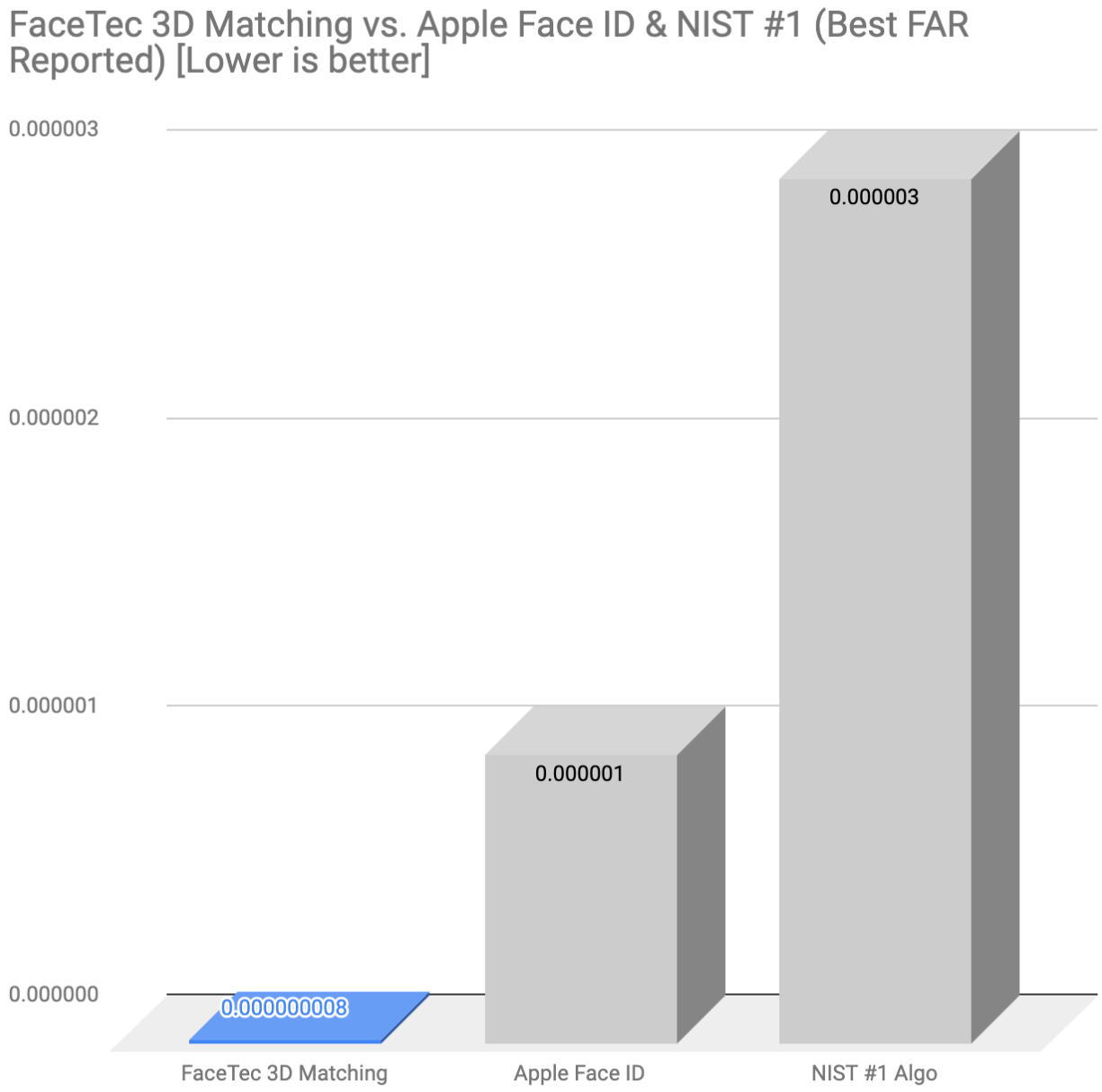
The dataset used to test FaceTec's Algorithm performance is representative of real-world users, usage, and environmental conditions because it was collected from real-world users, on their own devices, and in a wide spectrum of environments.

In contrast, NIST states that their Mugshot dataset contains ideal facial images in an ideal pose, neutral expression, no glare in eyeglasses, no shadows, no bright spots, no significant lighting variation, and generally a very consistent capture environment.

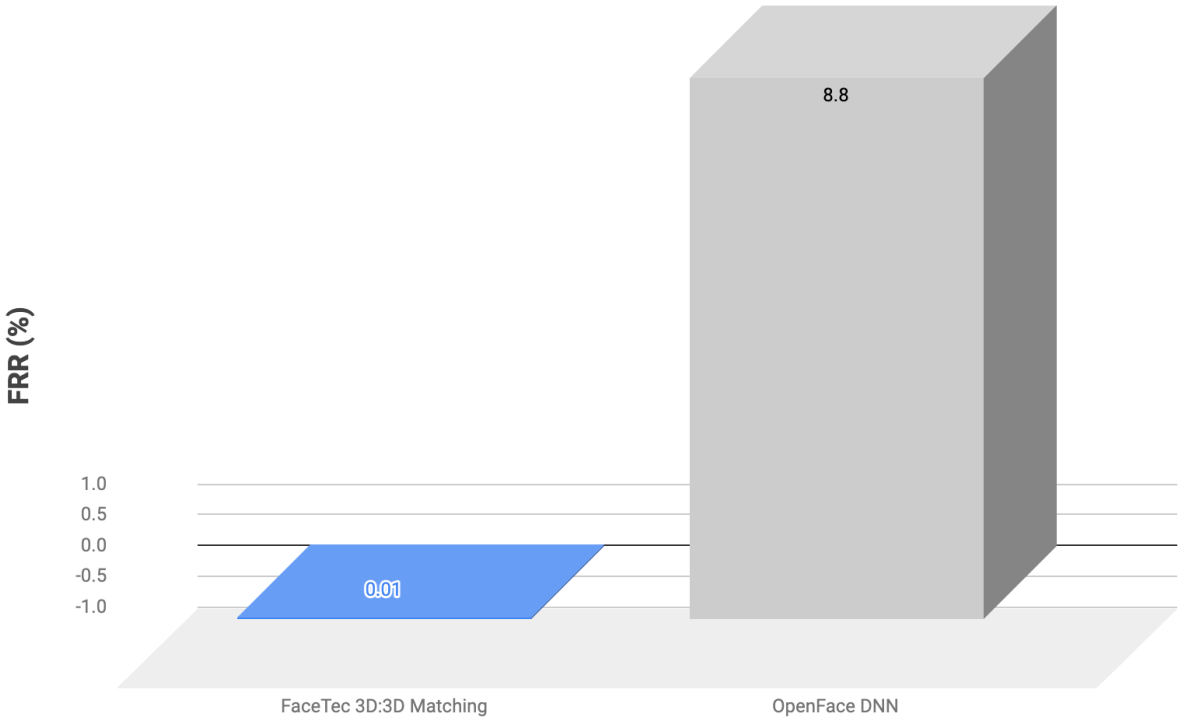
Based on these well-documented properties of the NIST dataset, as well as FaceTec's knowledge of its own dataset, we believe it is reasonable to conclude:

1. The FaceTec 1:1 3D Matching Algorithm would perform ***better*** than the FARs/FRRs stated in this report if tested on a 3D FaceMap version of the NIST Mugshot dataset (due to the requirement of ideal capture conditions).
2. NIST-tested algorithms **would perform even worse** than the operating points stated in the NIST report if tested against a real-world dataset like FaceTec's.

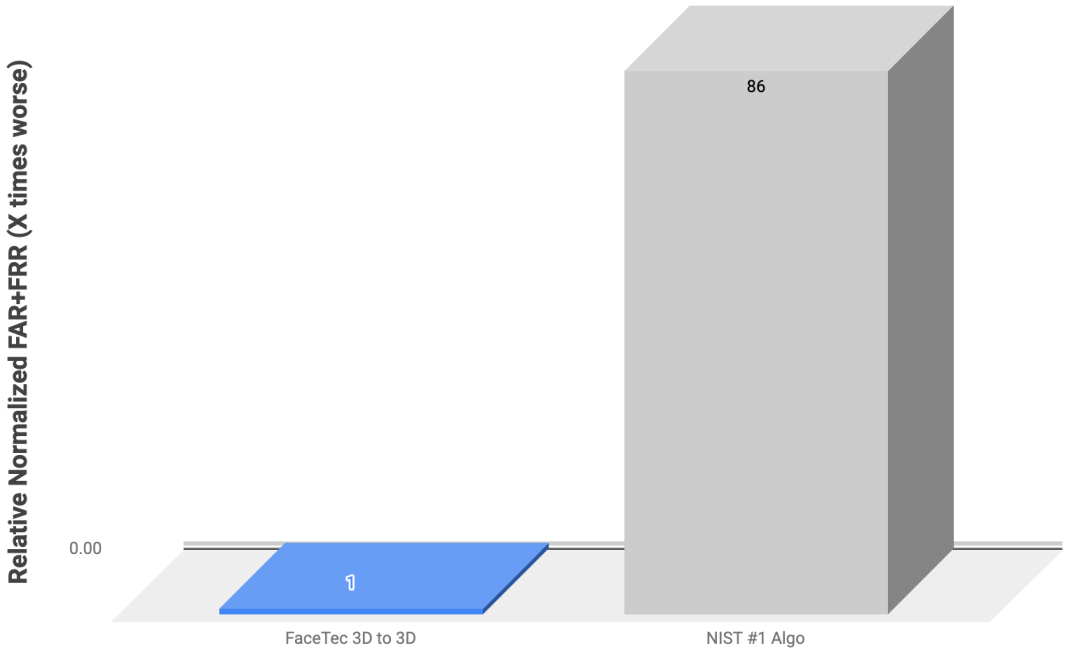
Analysis Versus Other Algorithms and Standards



FaceTec vs. OpenFace - FRR @ 1/10,000 FAR [Lower is better]

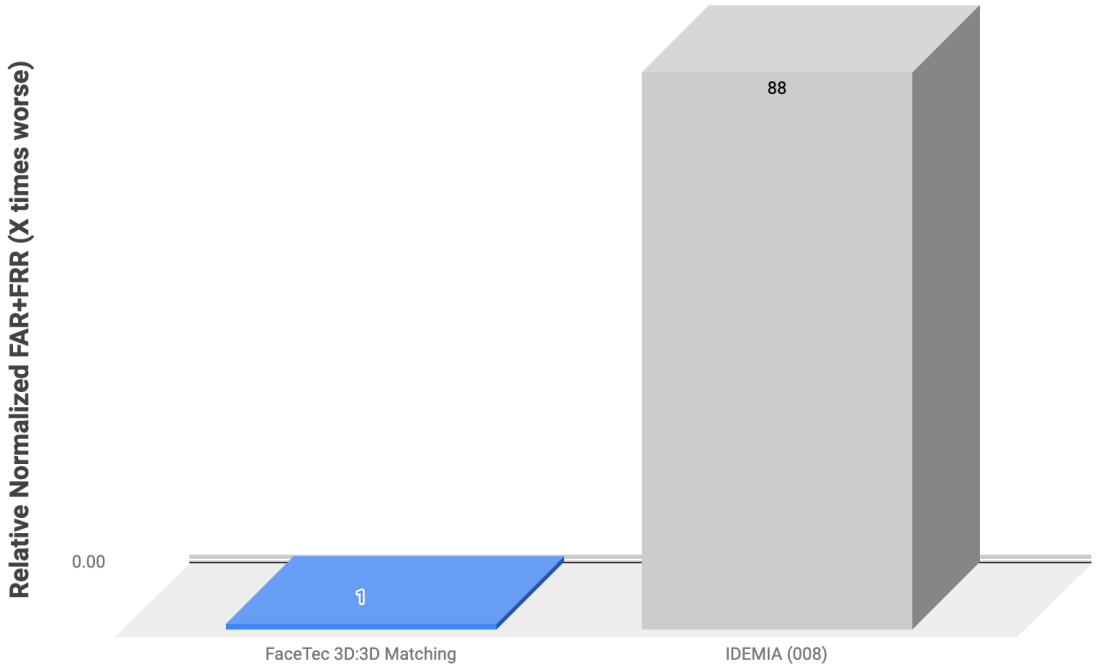


FaceTec vs. NIST #1 - Relative Normalized FAR & FRR (MUGSHOT) [Lower is better]

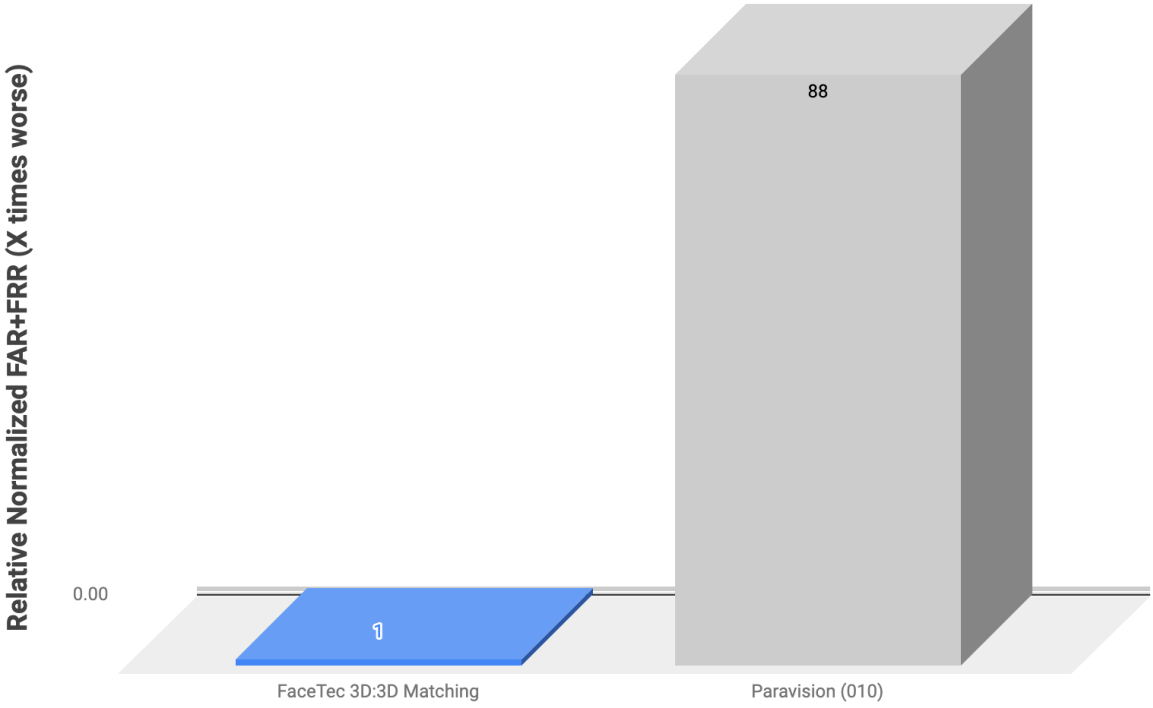


* Estimated from DET curve and operating points reported by NIST.
Please see the "FaceTec Overall Performance" section of the performance table below for an explanation of this metric.

FaceTec vs. IDEMIA - Relative Normalized FAR & FRR (MUGSHOT) [Lower is better]



FaceTec vs. Paravision - Relative Normalized FAR & FRR (MUGSHOT) [Lower is better]



Analyzing Overall Performance

Algorithm or Standard	FAR	FRR (%)	FaceTec Overall Performance (Normalized FAR & FRR, Times Better)
FaceTec 3D	**1/125,000,000**	**0.99%**	-
Face ID	1/1,000,000	*See Note 1	> 125x
NIST #1	1/333,333	0.27%	> 86x
Android P Recommendations	1/50,000	< 10%	> 244x
FIDO Standard	1/10,000	3%	> 361x
Dlib Pretrained DNN	1/100,000	35.4%	> 4336x
Department of Justice DEA EPCS	1/1,000	**See Note 2	> 125,000x

* Note 1 - Apple does not report FRR for Face ID making their "1/1,000,000" claims only partially comparable to other algorithms.
** Note 2 - There is no FRR requirement for DEA EPCS certification.

Other Notes:

- For NIST tested algorithms, Mugshot is the most similar and thus the most comparable dataset to FaceTec's as it is the only set that is frontal face + live captures + 100% adult subjects, but it is an ideal dataset, not real-world.
- The NIST Mugshot database contains *only* images captured in the United States.
    - The FaceTec 3D Face Matching Algorithm is trained and tested against sessions from over 180 different countries.
- The NIST Mugshot database is, by design, "ideal scenario" captures of faces: i.e., faces are shown in near-perfect lighting, with no shadows, no glare in glasses, and the image capture apparatuses are standardized per ISO 19794-5, as noted in the NIST report.
- The NIST Mugshot test methodology is a modified "All Combinations." For undisclosed reasons, NIST separates its dataset into males and females and generates genuine/imposter pairs from these gender-separated sets.
- "FaceTec Overall Performance (% Better)" -- This is a custom metric intended to show the relative strength of FaceTec's matching algorithm while normalizing for differences in scale reported in other tests and/or by other industry standards. We must call out that this metric is intended to be approximate -- not exact -- as we understand that FAR/FRR performance curves are always non-linear.
- FaceTec tested OpenFace and Dlib Pretrained DNN using the same dataset used against the FaceTec 3D Matching Algorithm.
- While the specifics are lacking, the NIST report states that the number of identities in its set is the same order of magnitude as the number of identities in its set; exactly two images per identity.

Sources:

- iPhone X keynote
- pages.nist.gov/frvt/reports/11/frvt_11_report.pdf
- source.android.com/compatibility/android-cdd
- fidoalliance.org/specs/biometric/Biometrics-Requirements-v1.0-wd-20180830.html
- www.deadiversion.usdoj.gov/21cfr/cfr/1311/subpart_c100.htm

Appendix 1: Technology Discussion

Results Highlight a Significant 3D Breakthrough

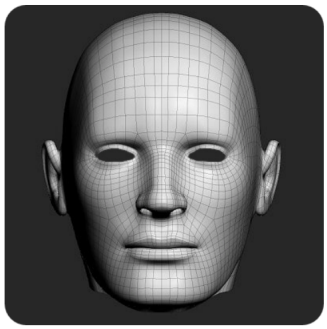
Intrinsically, we all know a real 3D human face contains more unique data than a 2D photo, or even a video, of that same face. This is because when a 3D face is flattened into a single 2D layer, the true relational depth data is lost, and consistency of appearance issues become apparent. In the real world, capture distance, camera position, and lens diameter all contribute to how well we perceive that a derivative 2D photo represents the original 3D face. See examples of 2D photo/perspective distortion here.

We can all agree that 3D is the higher-quality and more consistent derivative: it has more data and can be used to better differentiate individual people. While there's no doubt about it, there has been one big problem: In the past, capturing 3D face scans always required special hardware. Today, FaceTec solves that problem by measuring perspective distortion and reverse-engineering the 3D face from 2D video frames captured on any smartphone or webcam, making it ideal for 1:1 and 1:1N face matching.

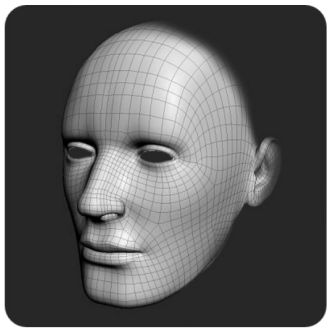
Four Dimensions - X, Y, Z & Time

2D Images - Shows flat data on the X & Y axes, presumably gleaned from a 3D subject.

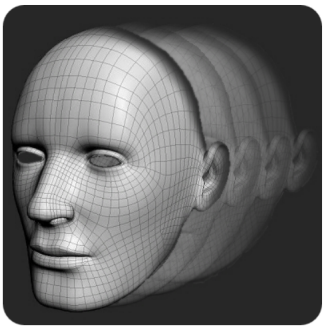
3D Data - Digital representation of a 3D object, which may include images for texture mapping and depth data of the relative distance between features on X,Y & Z axes.



2D (X,Y)
Legacy 2D Matching Algorithms



Typical **3D** (X,Y,Z)
Apple Face ID & 3D Hardware



FaceTec **3D** (X,Y + Time)
Any Smartphone or Webcam

3D FaceMaps - FaceTec creates 3D FaceMaps with any 2D camera from the 60-180 frontal frames it captures as the user and the camera are brought closer together. If the subject is 3D, the camera observes perspective distortion, and the way the facial features interact throughout the observed motion is unique to every person. By analyzing the face feature depth from the extent of perspective distortion observed, FaceTec's AI can create a consistent 3D model of the user's face.

Time as the 4th Dimension - Using X & Y + time, FaceTec captures numerous 2D video frames over a known period and uses AI to interpolate the 3D object from the 2D images it has observed.

Beyond the NIST FRVT Test Datasets

The NIST 2D Mugshot Dataset is the closest thing to FaceTec's 3D FaceMap they have/can test with. Thus, we use the results of the NIST 2D Mugshot testing for comparison, even though the FaceTec 3D FaceMap dataset is representative of the spectrum of real-world capture and the NIST 2D Mugshot. We would prefer that NIST also conduct the testing on FaceTec's 3D Face Matching Algorithms, but unfortunately that has not happened yet because **NIST does not have a 3D FaceMap dataset.** FaceTec's proprietary method to capture 2D images and interpolate 3D face data from them gives FaceTec an undeniable advantage over 2D matching algorithms, but ultimately only the results matter. The reality is this level of performance will never be achieved from a 2D algorithm because there just isn't enough differentiating data in a 2D image, and 3D Faces that are flattened into 2D images contain perspective distortion, lens distortion, and depending on the capture distance the same individual can appear significantly different to the camera.

It should be noted that at the time of this writing, the NIST FRVT's top two companies' algorithms each have been submitted for testing 5+ times, yet they have not gotten much more accurate over the last few submissions. FaceTec believes that this is an indicator that **2D Face Matching has stalled out**, and that 3D Face Matching is the only viable option left to achieve any "order of magnitude" accuracy gains.

In addition, FaceTec has observed that some of the top algorithms on the NIST FRVT regress to as much as 6X lower accuracy, only to have their next algorithm submission be better than its previous best result. Vendors have a lot to gain by "gaming the test" to get better results (and then trumpeting marketing claims like "Top NIST Algorithm"). And in this case, it is quite obvious that many of these vendors are using NIST's "unlimited submissions allowed" rule to learn how to tune their algorithms **to increase their accuracy in the test, while their real-world accuracy is likely hindered**.

2022 Update: FaceTec has observed that in recent versions of the report, NIST has ceased to include submissions older than the last two submissions per entity/company. FaceTec interprets the removal of past testing results as further demonstration of the fallout of the policy of allowing unlimited submissions with unlimited frequency in order for companies to abuse the "get to the top of the NIST list," rather than building a test that rewards making Algorithms that work better in the real world, where the matching accuracy actually matters.

The NIST submission system and Leaderboard rewards solutions that fit into the long-established NIST mold, and does not reward outside-the-box innovation and ingenuity. We agree that FaceTec's 3D FaceMaps and 2D images are not exactly apples-to-apples (actually, they are more like a 3D printed apple to a photo of an apple), but the

matching performance should be compared because FaceTec is capturing the 3D data with a standard 2D camera. In fact, any user with a $40 smartphone can access FaceTec's 3D tech. So instead of the procrustean view of forcing vendors into the NIST 2D mold, organizations looking to utilize cutting-edge face-matching tech should be willing to collect new data to test innovative methods as long as they run on widely distributed devices.

Why 3D Matching Helps Solve the "Twins Problem"

Identical twins constitute .3% of the world's population, so in a random database of 1,000,000 users, there will be about 3,000 individuals who may share a likeness with another person. These twins are indeed different people, but will highly match with each other and often give a false positive for the other individual.

Though identical twins are a challenge for *all* Face Matching algorithms, FaceTec's proprietary 3D algorithms differentiate identical twins much better than 2D algorithms can. They also have a lower FRR when matching the same user with/without glasses, with changes in makeup, facial hair, or after signs of aging. A better FAR means fewer False Accepts for the entire system, resulting in better differentiation of identical twins in the real world.

Sources:

- www.researchgate.net/publication/260712434_Double_Trouble_Differentiating_Identical_Twins_by_Face_Recognition
- en.wikipedia.org/wiki/Twin#Monozygotic_(identical)_twins

Appendix 2: FAQ

Question: *"My company/country has a "facial recognition" algorithm, and the vendor we bought it from promised it was state-of-the-art, and it's even been listed on the NIST Leaderboard! So why can't we just use FaceTec for 3D liveness and use the new 2D algorithm that we just bought for the matching?"*

Answer: 2D matching is used in surveillance and law enforcement scenarios because the match results list can be kept secret, and it's all they have. It's not chosen because it works that well. 2D face matching has been around for about 50 years and has gotten a lot better over time, but it's not good enough to use in real world scenarios where the match results are communicated to real users. 2D is insufficient when matching, and liveness must be reliable, like for 1:1 account security or 1:N duplicate prevention.

In the wild, 2D matchers cannot maintain a high enough FAR while keeping the FRR usable to run 1:N on large databases. See the FIDO and DEA EPCS standards, which require a meager 1/10,000 (@ 3% FRR) and 1/1,000 (no FRR requirement) respectively. If they demanded anything higher it would disqualify too many vendors. Every 2D "facial recognition" company has this problem, and is why you may have heard about the "one-to-few" strategy. 2D doesn't work well on large databases (en.wikipedia.org/wiki/Birthday_problem).

Question: *"I see the NIST list and those numbers look great! Why can't I expect the same results in the real world?"*

Answer: The "great" performance you see on the NIST Leaderboard is the result of a couple of things: #1. The datasets are near-ideal: they are not real-world (i.e., random users in random real scenarios) and they do not contain even moderately difficult lighting conditions or challenging scenarios. #2. The algorithm creators optimize their performance for these sets and have submitted algorithms to NIST many times in order to "tailor" their algorithms based on past submission performance. *The creators of the current #1 algorithm have submitted algorithms 10 times.* Any vendor that has submitted multiple times has had the opportunity to glean information

about the NIST “blackbox” datasets and experiment with tuning their algorithm to evaluate the effect in the next iteration of testing. This specialization essentially games the system.

Question: *“Why not compare against the NIST VISA set?”*

Answer: The NIST VISA contains 2nd-generation images (pictures of pictures), and is overall a very different set than 100% real-world, live frontal-face captures. Note: NIST states that Mugshot is 100% from live captures.

Question: *“Who personally attests to these results?”*

Answer: FaceTec’s algorithm scientists attest that the results were achieved honestly, that no data from the test sets is ever in the training sets, and that the test set data was randomly selected from a dataset that is representative of data that FaceTec observes in real-world scenarios.

FaceTec’s CTO - Josh Rose - LinkedIn
Chief Scientist - John Bernhard - LinkedIn
Senior Algorithm Development Engineer - Jase Kurasz - LinkedIn