NISTIR 8430

Face Recognition Vendor Test (FRVT)

Part 4A: MORPH - Utility of 1:N Face Recognition Algorithms for Morph Detection

Mei Ngan
Patrick Grother
Kayee Hanaoka
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8430

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

NISTIR 8430

Face Recognition Vendor Test (FRVT)
Part 4A: MORPH - Utility of 1:N Face Recognition Algorithms for Morph Detection

Mei Ngan
Patrick Grother
Kayee Hanaoka
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8430

July 2022

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Undersecretary of Commerce for Standards and Technology

Acknowledgements

The authors would like to thank the Department of Homeland Security Science and Technology Directorate (S&T) and Office of Biometric Identity Management (OBIM) and the Federal Bureau of Investigation (FBI) for their contributions to this activity.

The authors are also grateful to the University of Bologna for the morph creation software that supported this activity and the staff in the NIST Biometrics Research Laboratory for infrastructure supporting rapid evaluation of algorithms.

And last but never the least, the authors would like to thank Nick Orlans, Jim Matey, and Arun Vemury for their time and candor in reviewing the contents of this report and for providing many meaningful suggestions to help make this a better publication.

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.

The data, protocols, and metrics employed in this evaluation were chosen to support morph detection research and should not be construed as indicating how well these systems would perform in applications. While changes in the data domain, or changes in the amount of data used to build a system, can greatly influence system performance, changing the task protocols could reveal different performance strengths and weaknesses for these same systems.

Institutional Review Board

The National Institute of Standards and Technology's Research Protections Office reviewed the protocol for this project and determined it is not human subjects research as defined in Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule for the Protection of Human Subjects (45 CFR 46, Subpart A).

Patents

The content of this document is covered, in whole or in part, by a United States patent application applied for by United States Government, as represented by the Secretary of Commerce. Licensing inquiries may be directed to the Technology Partnerships Office, NIST, Gaithersburg, MD, 20899; by phone at (301) 301-975-2573; or by email at tpo@nist.gov.

Executive Summary

Overview

This report is one of a series of reports on our investigation of face morphing, its relevance and implications as a vulnerability to automated face recognition, and methods to aid in detecting morphs. This report presents a methodology and quantitative results on the use of automated one-to-many (1:N) face recognition algorithms as a mechanism to potentially detect the presence of morphs, particularly in an ID renewal scenario. This report is intended to inform end-users and identity credential issuance entities, especially those that accept user-submitted photos, in understanding how a 1:N search against a centralized database might be used to flag suspicious activity related to face morphing. The algorithms used in our investigation were submitted to the NIST FRVT 1:N Identification Track in late 2021/early 2022.

Motivation

Face morphing is an image manipulation technique where two or more subjects' faces are blended together to form a single face in a photograph [1]. Morphed photos often look realistically like all contributing subjects. Morphing is easy to do and requires little to no technical experience given the vast quantity of tools available at little or no cost on the internet and mobile platforms. If an attacker is able to submit a morphed photo which is accepted and placed onto an identity credential, multiple, if not all constituents of the morph can use the same identity credential, because modern face recognition will often erroneously authenticate the morph with the different contributing subjects. Morphs can be used to fool both humans [2] [3] [4] and face recognition systems [5], which presents a vulnerability to current identity verification processes.

Results and Notable Observations

In the context of a potentially morphed image being used to apply for an identity credential, if the issuing organization maintains or has access to a centralized database of past applicant facial photos, there is an opportunity to search the application photo against the database with the goal of detecting morphs. We developed a methodology for doing morph detection using 1:N face recognition algorithms. Our proposed approach analyzes the rank 1 and rank 2 scores that are returned on candidate lists from searching morph and bona fide photos against both consolidated and unconsolidated galleries of 1.6 million unique subjects under new enrollment and renewal scenarios. Morph classifiers are trained using the rank 1 and 2 score pairs from several modern 1:N face recognition algorithms and evaluated to quantify the utility of these scores in detecting morphs.

- **Morph detection using 1:N face recognition is promising in a renewal scenario.** In a scenario where an identity credential is being renewed and there are **multiple** prior bona fide photos of the applicant in a database, the most accurate morph classifier successfully detected 83% of morphs at a threshold set to generate a false detection 1 out of every 1 000 bona fide searches (BPCER=0.001). Setting a less restrictive threshold that generates a false detection 1 out of every 100 bona fide searches (BPCER=0.01), the morph classifier successfully detected morphs 98% of the time. Section 6.5.3

  In a scenario where only a **single** bona fide photo of each subject is maintained in the database, the most accurate morph classifiers generated morph detection rates of 74% (BPCER=0.001) and 92% (BPCER=0.01) when both the applicant and the "hidden identity" exist in the gallery. When only the applicant exists in the gallery, morph detection rates were 65% (BPCER=0.001) and 77% (BPCER=0.01). Section 6.4.1

- **Reduced false detection rates are attainable.** While the prevalence of morphs in operations is not known, the assumption is that most operational transactions will be on legitimate photos that are not morphs. Therefore, it is important for morph detection technology to be able to operate at false detection rates low enough to support the level of resources available for secondary review, because any photo that is flagged as a morph will require additional resources to be adjudicated. We do not conceive of automated morph detection as being a lights-out operation. There will almost always be human review, investigation, and remediation of suspiciously flagged images, and the investigation process may be non-trivial.

  In a renewal scenario, for the most accurate morph classifiers tested in our study, morph detection rates at reduced false detection rates (i.e., BPCER=0.001) are better than many conventional differential morph detection algorithms evaluated in the FRVT MORPH activity that leverage an additional live image of the user or applicant to do morph detection. There will be errors associated with automated morph detection. The goal may be to establish thresholds such that false detection rates are at acceptable levels and even if morph detection rates are low, it would still yield gains in operations compared to not having any morph detection capability at all. Section 6.4.1, 6.5.3

- **Morph detection using 1:N face recognition is not effective in a new enrollment scenario.** In a scenario where a person is applying for a new identity credential where no prior photos of the applicant exist in the database, morph detection rates were very low. In the best cases, only 0.2% of morphs were detected at a false detection rate of 0.001. Section 7.3

Contents

List of Figures

1 Face Morphing - an Attack on Identity Credentials

Face morphing is an image manipulation technique where two or more subjects' faces are blended together to form a single face in a photograph [1]. Morphed photos often look realistically like all contributing subjects. If an attacker is able to submit a morphed photo which is accepted and placed onto an identity credential, multiple, if not all constituents of the morph can use the same identity credential. Morphs can be used to fool both humans [2] [3] [4] and current face recognition systems [1], which presents a vulnerability to current identity verification processes. Figure 1 illustrates the impact of morphed photos on current state-of-the-art face recognition algorithms submitted the NIST Ongoing FRVT 1:1 Verification test. The mated morph presentation match rate (MMPMR) [6] states how often both subjects erroneously authenticate against the morph and gives an indication of how vulnerable an algorithm is to morphs. The false non-match rate (FNMR) on non-morphed photos presents general face recognition accuracy.

The results in Figure 1 show the more accurate face recognition algorithms tend to also be more vulnerable to morphing. The face recognition algorithms with low false non-match rates (FNMR $\leq$ 0.003) are also prone to match against two partial strength images from genuine subjects. While this is a desirable capability for authentic, bona fide natural images, it is a vulnerability when manipulated images may be submitted and image authenticity is in question. The high proportion of morph comparisons that would successfully authenticate at operationally realistic thresholds provides the basis and motivation for research into how to detect this form of image manipulation.

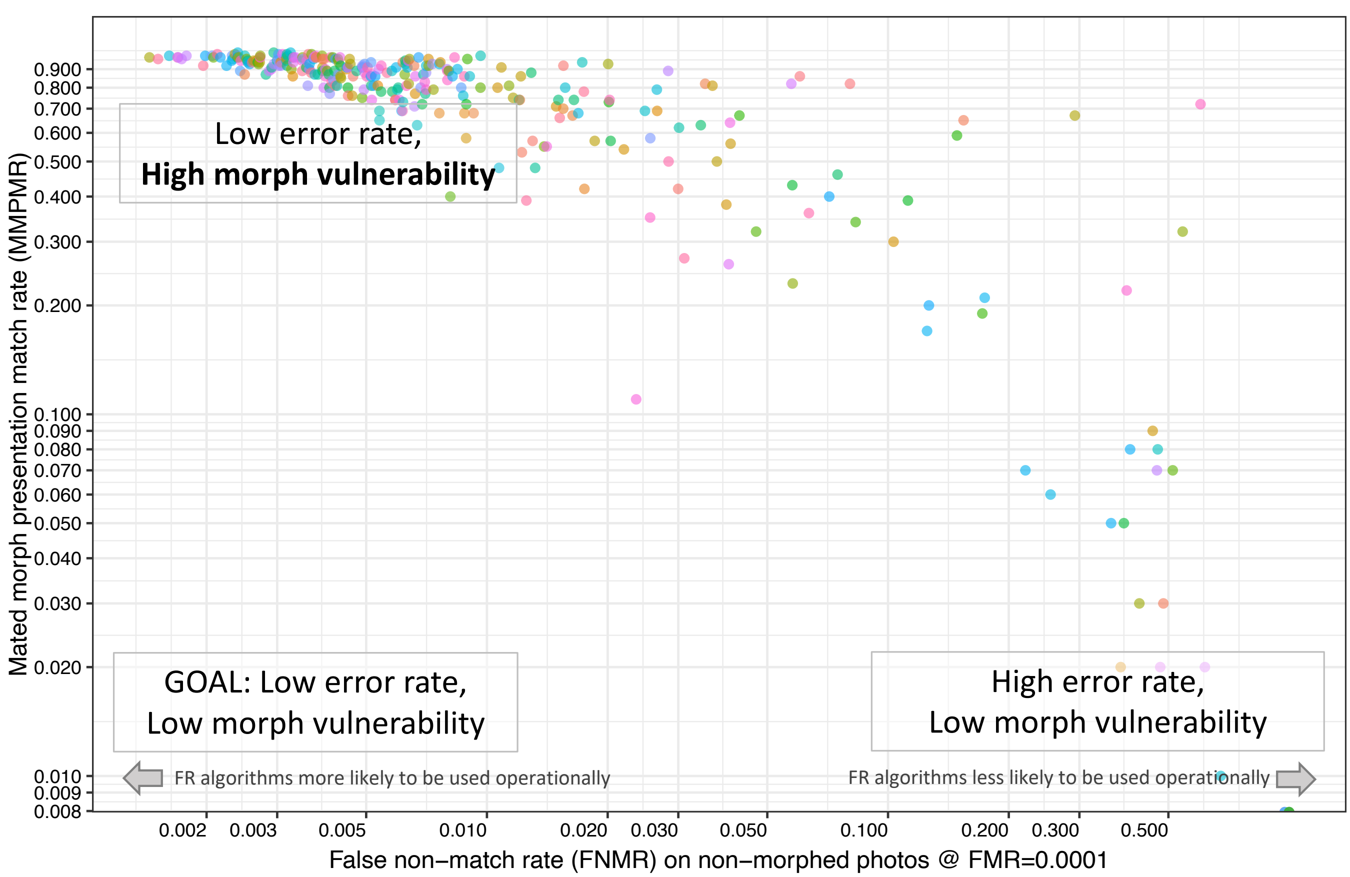


Figure 1: This graph plots face recognition algorithm vulnerability on morphs against general algorithm accuracy on non-morphed photos. Each point represents a face recognition algorithm recently submitted to the NIST FRVT 1:1 activity. The y-axis plots MMPMR, which is the fraction of morphs where both subjects incorrectly match to the morph. The x-axis plots FNMR or miss rate on regular photos, which provides an indication of general algorithm accuracy. Both MMPMR and FNMR are calculated with thresholds set to where the false match rate (FMR) is 0.0001. The morphs were generated with two people with equal contribution from each subject.

BPCER(T) False Detection Rate
APCER(T) Morph Miss Rate

2 Methodology

2.1 Test Environment

The evaluation was conducted offline at a NIST facility by applying algorithms to still photos that is sequestered on computers controlled by NIST. Offline evaluations are attractive because they allow uniform, fair, repeatable, and large-scale statistically robust testing. However, they do not capture all aspects of an operational system. Offline tests do not include a live image acquisition component or any interaction with real users. Our approach is adopted to allow evaluation on large datasets and to achieve repeatability. Testing was performed on high-end server-class blades running the CentOS Linux [7] operating system. The test harness used concurrent processing to distribute workload across dozens of computers.

2.2 Algorithms

The eight one-to-many face recognition algorithms used in our investigation are competitive algorithms submitted to the NIST FRVT 1:N Identification Track in the late 2021/early 2022 timeframe. The algorithms all report non-zero similarity scores, with larger values indicating higher likelihood that two samples are from the same person. The range of the scores is not regulated and will vary between algorithms.

3 Image Datasets

3.1 Portrait Images

The images are all high-quality frontal portraits of adult subjects collected in immigration offices and with a white background. As such, potential quality related drivers of high false match rates (such as blur) are expected to be absent. The images are collected in an attended interview setting using dedicated capture equipment and lighting. The images are of size 300x300 pixels, and the mean interocular distance of the subject in a photo is 61 pixels. The images are encoded as ISO/IEC 10918 i.e. JPEG. Over a random sample of 1000 images, the images have compressed file sizes (mean: 42KB, median: 58KB, 25-th percentile: 15KB, and 75-th percentile: 66KB). The implied bit-rates are mostly benign and superior to many e-Passports. Each image is accompanied by metadata, including subject age, sex, and place of birth.

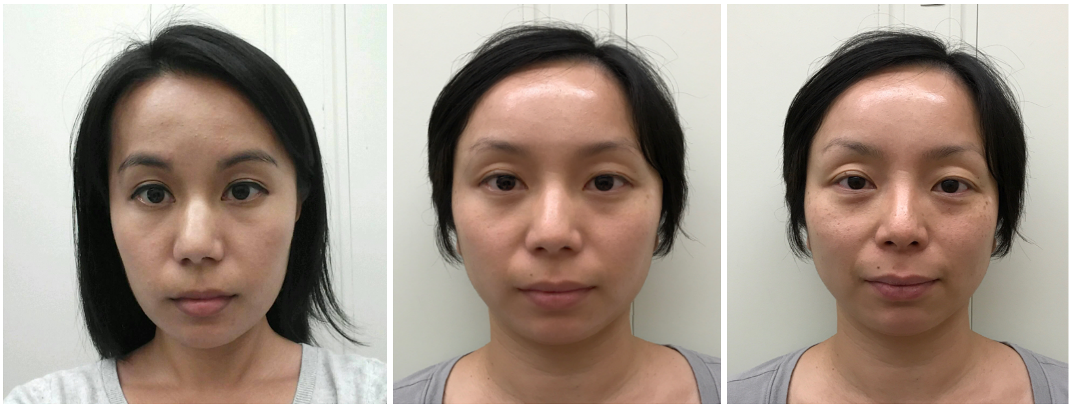
Figure 2: Illustrative portrait image examples. The subjects in the photos are all NIST employees.

3.2 Morphs

Morphed images were created from frontal portraits described in Section 3.1 using the University of Bologna's (UNIBO) v2.0 morphing tool [1,8–10] with two subjects. Subjects were demographically paired based on age (within one year), sex,

BPCER(T)	False Detection Rate
APCER(T)	Morph Miss Rate

and place of birth. The interior face regions from the two subjects were morphed with blending and warping factors of 0.50 (equal contributions from each subject to the morph), and one of the subjects provided the periphery (the head, hair, ears (if visible), body, and background). Using the methodology described above, an initial set of 40 799 morphs were generated from 81 598 unique subjects. All morphs were then validated for ”usefulness” with several one-to-one face matchers submitted to the NIST FRVT 1:1 evaluation, where the morph was compared with other photos of both subjects. We set a score threshold that corresponds to a false match rate of 0.001 for each matcher. Morphs where comparisons generated scores that were above threshold for all matchers (i.e., morphs that were able to fool all matchers) were included in the dataset, and those that were below threshold were discarded. This resulted in 21 393 (52.4%) usable morphs in our dataset.



(a) Subject A *(b) Morph (A+B)* *(c) Subject B*

Figure 3: Example of a morph generated with UNIBO’s v2.0 morphing tool. The subjects in the photos are both NIST employees.

4 Metrics

Consistent with the FRVT MORPH [11] evaluation and the wider developer community [6], we adopt terminology from the presentation attack detection testing standard [12] to quantify morph classification accuracy, namely Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER). APCER is defined as the proportion of morph attack samples incorrectly classified as bona fide (nonmorph) presentation. Similarly, BPCER is defined as the proportion of bona fide (nonmorph) samples incorrectly classified as morphed samples.

4.1 Detection Error Tradeoff (DET)

We assess morph detection accuracy by analyzing the confidence score returned by the morph detection algorithm. In this case, the higher the confidence value, the more likely the algorithm thinks it is a morph. A reasonable approach to the detection problem is to classify an image as either a morph or bona fide image by thresholding on its confidence value.

Given N detection scores on bona fide images and detection score b, from the i-th bona fide image, where i=1 ...N, BPCER is computed as the proportion of bona fide scores above some threshold, T. Similarly, given M detection scores on morphed images and detection score m, from the i-th morphed image, where i=1 ...M, APCER is computed as the proportion of morphed scores below some threshold, T. $H(x)$ is the unit step function [13], and $H(0)$ is taken to be 1.

$$\text{BPCER}(T) = \frac{1}{N}\sum_{i=1}^{N} H(b_i - T), \tag{1}$$

$$\text{APCER}(T) = 1 - \frac{1}{M}\sum_{i=1}^{M} H(m_i - T). \tag{2}$$

BPCER(T)	False Detection Rate
APCER(T)	Morph Miss Rate

5 1:N Search in Operational Scenarios

In the context of identity credentials with face portrait images, 1:N searches can be used for the purpose of detecting morphs. This assumes an issuing organization maintains or can access a database of past, trusted applicant face portraits, and there is a security requirement or reasonable basis to question the authenticity of newly submitted face images. 1:N searches for the purpose of morph detection should be done in conjunction with trained forensic human reviewers.

5.1 Enrollment Database Composition

In identity credentialing applications (e.g., passports, driver's licenses), collection and enrollment of biometric data from subjects often occur on more than one occasion. This might be done on a regular basis (once every 10 years for adult passports in the U.S.) or on an ad-hoc basis (re-issuance of a lost or stolen ID). Over time, images acquired during the ID re-issuance process are added to the database, and there are generally two approaches on handling multiple images collected of the same person.

Consolidated gallery (subject-based): Unique identities of people are maintained and a single representation of each subject exists in the database at any given time. New images acquired of a subject might be stored in a record that contains $K \geq 1$ images of the subject, and it is up to the face recognition algorithm on how the subject representation is modeled internally. Or, depending on data retention policies, only the most recent photo of a subject is retained and all previously collected photos are discarded from the database. The experiments and results from Section 6 assume a consolidated gallery with a single representation of each unique subject.

Unconsolidated gallery (event-based): Here, images are added to the database without regard for whether the person already exists or not. Under this model, there can be multiple images of the same person in the gallery. Templates or representations are generated from single images independently and are treated as different identities. Administratively, there might be record-keeping that associates same-person images, but the underlying face recognition algorithm is not aware of this. Currently, some U.S. government operations are running unconsolidated galleries. The experiment and results from Section 6.5 assume an unconsolidated gallery with multiple photos of the same subject stored in the database.

5.2 Search Scenarios

We consider two operational scenarios for leveraging a 1:N search algorithm for morph detection, including

- Renewal of an existing ID credential (Section 6)
  - The application photo is a legitimate bona fide photo of the applicant, and prior bona fide photo(s) of the applicant exist in the database, or
  - The application photo is a morphed photo of the applicant and a ”hidden identity” and
    - Prior bona fide photo(s) of the applicant exist in the database, or
    - Prior bona fide photo(s) of both the applicant and the ”hidden identity” exist in the database
- New enrollment/application for ID credential (Section 7)
  - The application photo is a legitimate bona fide photo of the applicant, and there are no prior photos of the applicant in the database, or
  - The application photo is a morphed photo of the applicant and a ”hidden identity”, and no prior photo(s) of the contributing subjects exist in the database

6 Renewal of an Existing ID Credential

In a scenario where an applicant is renewing an existing ID credential, prior photo(s) of the applicant would exist in the database. We could reasonably expect that a 1:N search (with a modern face recognition algorithm) against a database of frontal, portrait photos to return the applicant at the top of the candidate list (i.e., at rank 1). For a mated search of a legitimate bona fide photo (which we expect will be the majority of transactions operationally), in addition to the applicant being returned at rank 1, we would also expect a very high similarity score to be associated with the match. In the case that the application photo is a morph and assuming only the applicant's photo(s) exist in the database, we expect the applicant to be returned at rank 1, but with a reduced similarity score. This is due to the fact that the morphed photo contains a reduced amount of the applicant's identity information (in our case, 50%). Now due to the fact that morphs contain identity information of two different people, it may be advantageous to also look at the rank 2 candidate that is returned. For simplicity in illustration, we assume a consolidated database where there's exactly one representation of each subject in the gallery and no threshold is set such that a search always returns the top K candidates. A discussion on morph detection on unconsolidated galleries is presented in Section 6.5.

6.1 Morph vs. Bona Fide Searches

As illustrated in Figure 4, a bona fide search would presumably retrieve the mate at rank 1 with a very high score, and a different person at rank 2 with a very low score. In the case that the probe is a morph, two scenarios could exist. First, if only the applicant exists in the gallery, then the mate at rank 1 would return a reduced but relatively high score and a different person at rank 2 with a very low score. If the probe is a morph and both subjects exist in the gallery, then we have a scenario where either subject could be retrieved at rank 1 with a reduced (but relatively high) score, and the subject at rank 2 would also match with a relatively high score due to the fact that the morph also contains identity information for that individual.

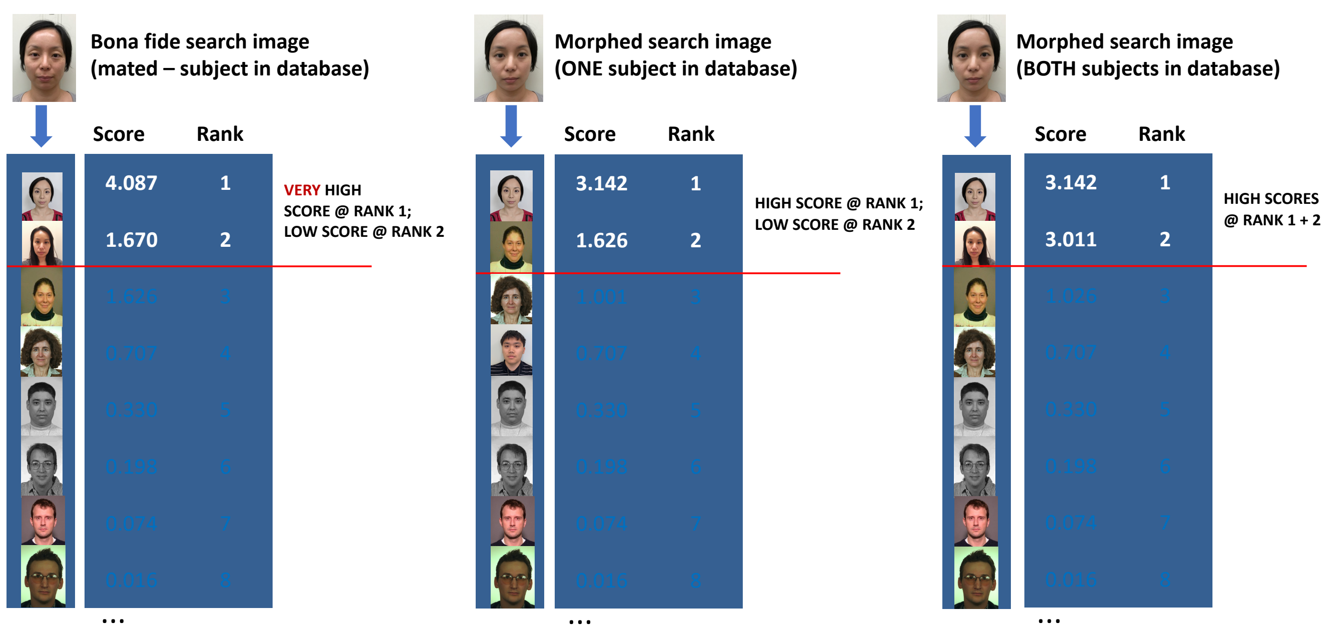


*Figure 4: Notional examples of similarity scores that are returned on a candidate list for a **mated** search against a **consolidated** gallery based on 1) the type of probe (bona fide or morph) and 2) the subject(s) in the database. The images in this figure are of NIST employees + subjects from the public NIST FERET Database [14].*

BPCER(T) False Detection Rate
APCER(T) Morph Miss Rate

6.2 Experiment: Consolidated Gallery

We generated 21 393 morphs from 42 786 images of unique subjects using the methodology documented in Section 3.2. We then generated two different galleries - gallery 1 contains a prior photo of one of the subjects from any particular morph, and gallery 2 contains a prior photo of both of the subjects in the morph. Both galleries also include photos of other people to achieve a size of 1.6 million unique people. The 42 786 bona fides were searched against gallery 2, and the 21 393 morphs were searched against both gallery 1 and 2. We used eight state-of-the-art 1:N face recognition algorithms submitted to the NIST FRVT 1:N Identification Evaluation.

6.3 Analysis of Rank 1 and Rank 2 Scores: Consolidated Gallery

Figure 5 shows the distribution of the algorithm's native rank 1 and rank 2 scores for each of the bona fide and morph search scenarios discussed. Additionally, the plot divides the bona fide search scores by searches where the probe and gallery images were collected less than 5 years apart and between 5 to 10 years apart. This gives an indication of how ageing may impact the rank 1 and 2 scores in our analysis of separability between bona fide and morph searches.

Consistent with our illustration from Figure 4, we observe that mated bona fide searches where the probe and gallery images are less than 5 years apart consistently generate rank 1 scores that are generally higher than when the probe is a morph. But operationally, ID credentials such as passports or driver's licenses are typically renewed once every 5 to 10 years depending on the organization/country's renewal policy, which introduces a factor that must be accounted for, which is ageing. We do observe reduced rank 1 similarity scores from bona fide probes when the time elapsed between the probe and gallery images is between 5 to 10 years, which increases the overlap between bona fide and morph similarity scores. We observe reduced but relatively high rank 1 similarity scores for morph searches against a gallery containing just one of the contributing subjects and very low scores at rank 2. But morph searches when both contributing subjects exist in the gallery yield reduced (but relatively high) scores at both rank 1 and 2 due to the fact that the morph contains identity information for two individuals in the gallery. Figure 5 presents score distributions for one algorithm. Results for a number of other 1:N algorithms we conducted this analysis on are presented in the Appendix.

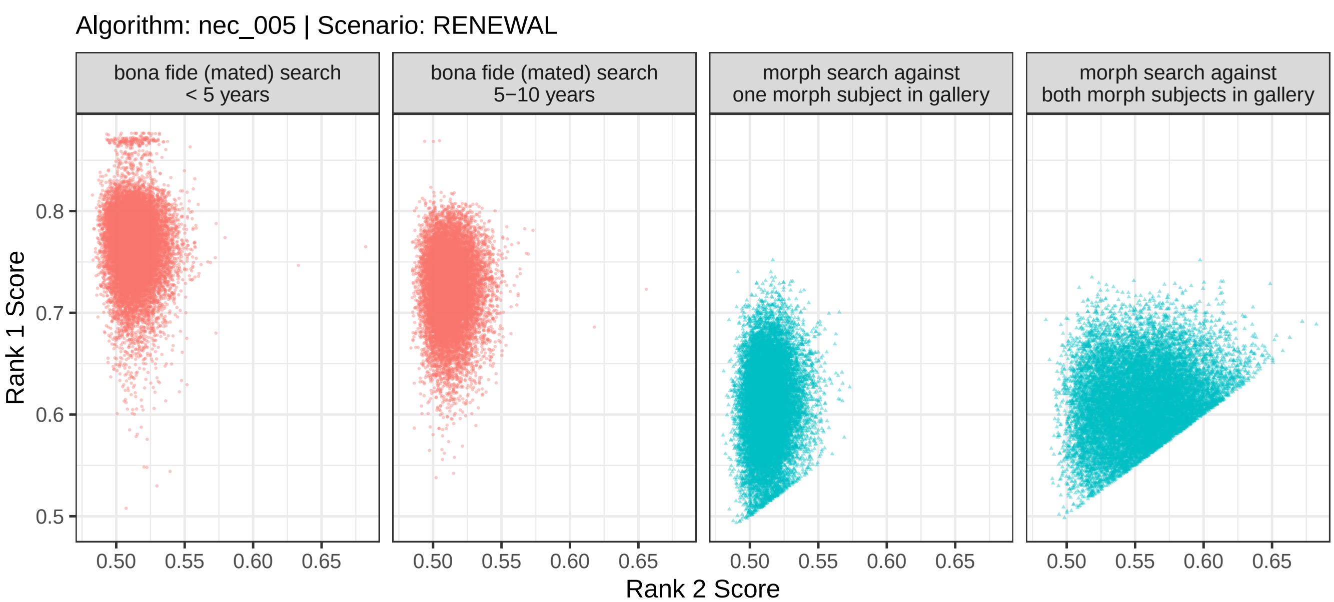


*Figure 5: The position of each point in the scatter plot represents the native rank 1 and 2 similarity scores for bona fide or morph probes searched against a **consolidated** gallery that contains a prior photo of the subject(s). For bona fides, the probes are broken out by the age difference between the probe and the gallery image (< 5 years and 5 - 10 years). For morphs, scores are shown for galleries that include either one subject or both subjects that went into the morph. All galleries contain 1.6 million unique people.*

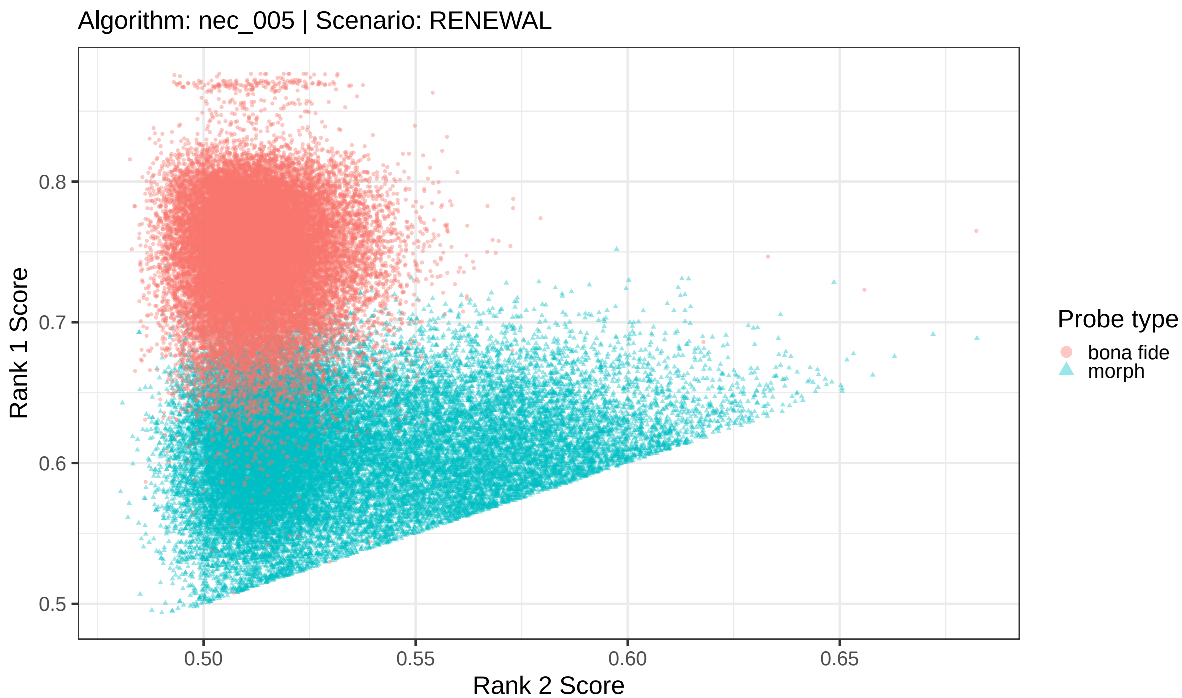


*Figure 6: This plot combines the rank 1 and 2 scores from the various bona fide and morph search scenarios against a **consolidated** gallery from Figure 5.*

6.4 Training a Morph Classifier

Through visual inspection, we can observe appreciable separation between the rank 1 and 2 scores for bona fide versus morph searches from a gallery that contains a prior photo of the subject(s) in the probe. To quantitatively express how useful these scores might be in detecting the presence of morphing, as a proof of concept, we trained a simple neural network to do morph classification using the rank 1 and 2 scores. For each 1:N algorithm used in our investigation, rank 1 and 2 score pairs from bona fide and morph searches are fed into a simple neural network (Figure 7). All scores are normalized using min-max normalization (Equation 3) to adjust the data to a common scale between 0 and 1 prior to input into the network. This resulted in eight different morph classifiers created from the eight 1:N algorithm score pairs.

$$score_{norm} = \frac{score - min(rank2\, scores)}{max(rank1\, scores) - min(rank2\, scores)} \tag{3}$$

We followed a 5-fold cross validation method where a random sample of 80% of the data (68 458 score pairs) is used for training, and the remaining 20% (17 114 score pairs) is used for testing, repeated over five iterations. The same exact partitions of data were used to train and test each morph detector. This was accomplished by setting a different random seed in each iteration, then reusing the same seed to reproduce the random sampling of scores across the different algorithms.

For testing, the inputs to the trained morph classifier are the normalized rank 1 and 2 score for a particular search, and the output is a confidence score between 0 and 1, with a score of 1 representing certainty that the score pairs correspond to a morph search, and 0 indicating certainty of a bona fide search. The 85 570 morph prediction scores (42 970 bona fide, 42 600 morph) across the five iterations of testing are used to measure classifier accuracy by plotting attack presentation classification error rate (APCER) or morph miss rate against bona fide classification error rate (BPCER) or false detection rate, as shown in Figure 8.

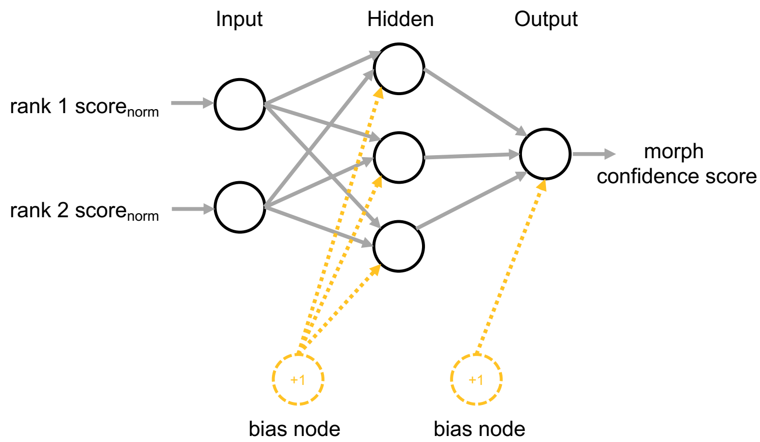


Figure 7: Neural network architecture for training a morph classifier with 1:N algorithm rank 1 and 2 scores. A sigmoid function is used in the the hidden and output layers.

6.4.1 Morph Detection Performance: Consolidated Gallery

From the results shown in Figure 8, we observe that the rank 1 and 2 score pairs generated by 1:N algorithm searches of bona fide and morph probes appear to have utility in morph detection. In the best cases, we were able to train a morph classifier with the score pairs to detect morphs at performance levels that might be operationally useful. Figure 8 decomposes the trained morph classifier accuracy by whether there is only one subject or both subjects in the gallery. For the most accurate 1:N algorithm when one subject exists in the gallery (idemia-009), its rank 1 and 2 score pairs generated a morph classifier where morph miss rate is 0.351 at a false detection rate of 0.001, meaning the classifier successfully detected 65% of morphs when the threshold is set to generate a false detection 1 out of every 1 000 bona fide searches. Reducing the threshold to produce a false detection rate of 0.01 (one false detection out of every 100 bona fide searches), the trained idemia-009 morph classifier would successfully detect morphs 77% of the time. For the most accurate algorithm when both subjects exist in the gallery (nec-005), the morph miss rate is 0.26 (74% of morphs successfully detected) at a false detection rate of 0.001, and relaxing the false detection rate to 0.01 would yield a morph miss rate of 0.08 (92% of morphs successfully detected).

We observe that there is a range of morph detection performance across the different 1:N algorithm scores that we trained on. Algorithms exhibit different levels of separability when it comes to the similarity scores it generates on morph versus bona fide searches, and morph detection performance appears broadly correlated with general algorithm accuracy. Many classifiers are able to detect morphs with lower error rates when both subjects exist in the gallery. This is likely due to the fact that there is larger separability between the bona fide and morph scores when both subjects are in the gallery, as visualized in Figure 5. But operationally, the prior probability of whether only the applicant or the applicant **and** the ”hidden identity” exist in the gallery is not well known. In the case of an ID renewal, the applicant will almost always exist in the gallery in order to pass typical identity confirmation checks, but whether the ”hidden identity” will also exist in the gallery can depend on any number of factors.

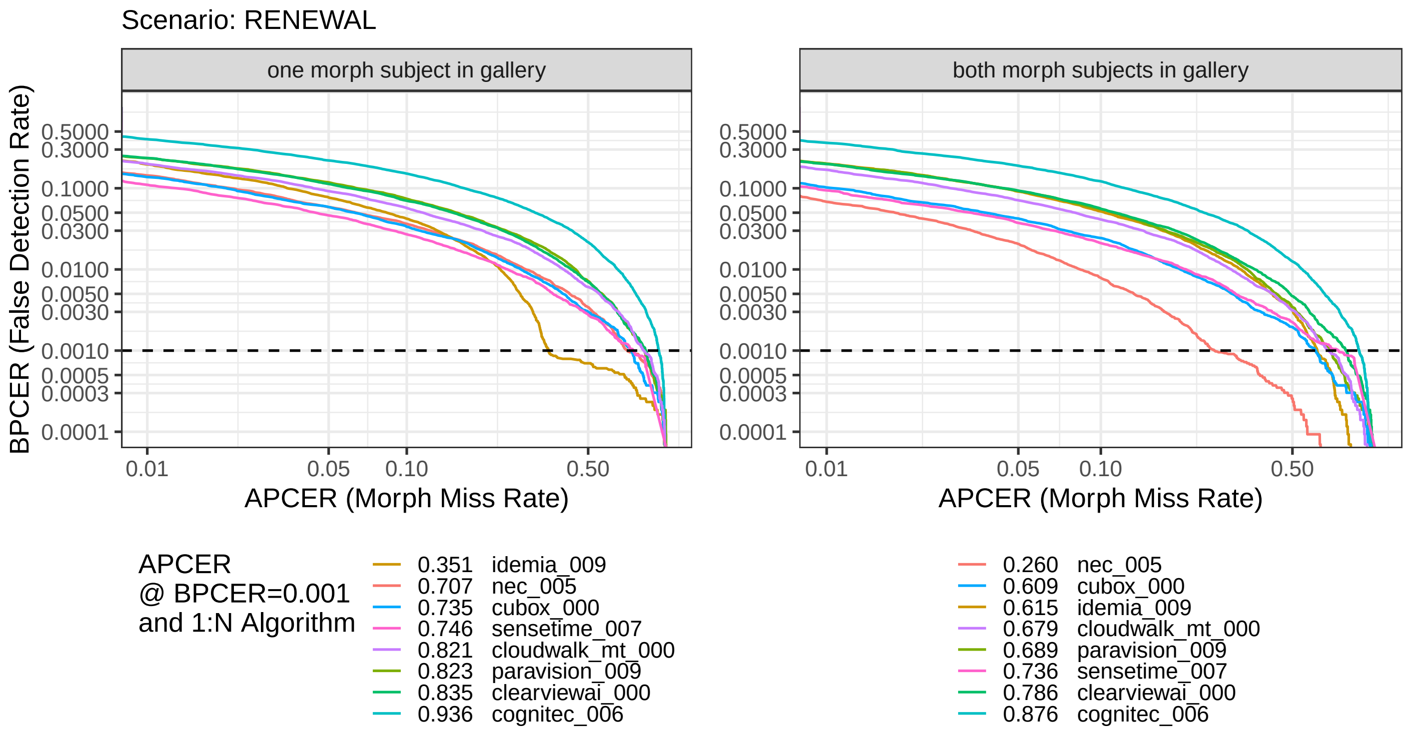


*Figure 8: Accuracy of morph classifiers trained on different 1:N algorithm rank 1 + 2 scores for morphed searches (one or both subjects in gallery) and mated bona fide searches against a **consolidated** gallery.*

6.4.2 Generalizability

In our experiment, as a proof of concept, we followed a 5-fold cross validation approach by splitting the rank 1 and 2 score pairs generated from the same "dataset" into randomly selected training and test sets, and repeated the process over five iterations. If the goal is to develop a generalizable morph detection algorithm that would work across different types of morphs (and bona fides), it would be due diligence to test and validate against different sets of score pairs generated using different types of morphs and bona fides and galleries composed of images of different qualities. For future work, NIST may assess the generalizability of using 1:N rank 1 and 2 score pairs for morph detection across different datasets and scenarios.

But, the initial goal of our investigation is primarily to present a methodology by which organizations might leverage a 1:N face recognition system in their operational pipeline to flag suspicious activity related to face morphing. Practically, an organization may only be concerned with their own data without needing to generalize, so training and testing with their own data would be a reasonable thing to do initially.

6.5 Consolidated vs. Unconsolidated Galleries

Whether a database is consolidated or unconsolidated will impact what gets returned on the candidate list. When a morph or bona fide probe is searched against an unconsolidated gallery under a renewal scenario, if **multiple** images of the subject(s) exist in the gallery, possible outcomes for what gets returned on the candidate list (see Figure 9) include:

- For a bona fide search, prior photos of the subject would be returned at rank 1 and 2 with very high similarity scores
- For a morph search

- If only one of the contributing subjects exists in the gallery, prior photos of that subject would be returned at rank 1 and 2 with high but reduced similarity scores
- If both subjects exist in the gallery, any combination of subject A-only, subject-B only, or subject A and B could be returned at rank 1 and 2, but in all cases, with high but reduced similarity scores

If only a **single** image of the subject(s) exists in the gallery, the expected behavior would be the same as described in Section 6.1.

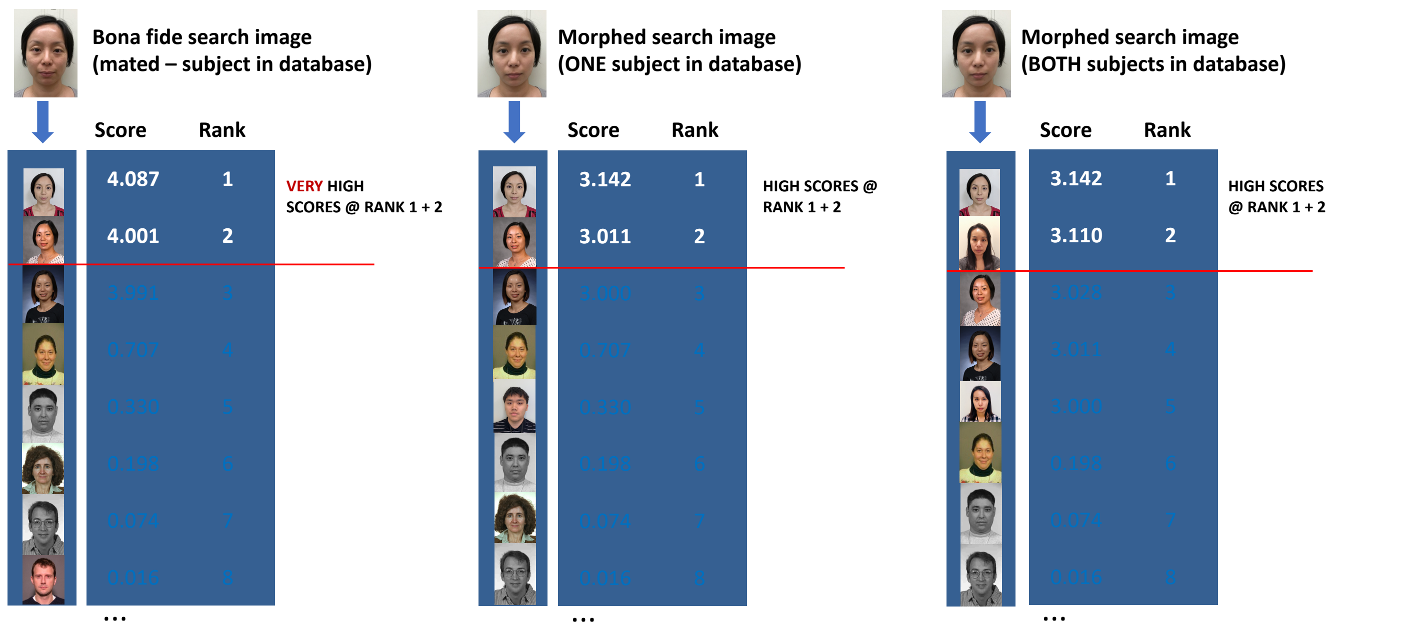


*Figure 9: Notional examples of similarity scores that are returned on a candidate list for a mated search against an **unconsolidated** gallery where two or more images of each subject exist in the gallery. Examples show possible candidate list outcomes based on 1) the type of probe (bona fide or morph) and 2) the subject(s) in the database. The images in this figure are of NIST employees + subjects from the public NIST FERET Database [14].*

6.5.1 Experiment: Unconsolidated Gallery

A subset of the probes from Section 6.2 were used - morphs and bona fides with two or more enrollment photos of each subject available were extracted from the probe set, and all corresponding enrollment photos were enrolled into each gallery. This resulted in 3 883 morphs and 7 808 bona fides that were searched against databases under the following scenarios: 1) bona fides were searched against a gallery where two or more prior photos of the subject exists, 2) morphs were searched against a gallery where two or more prior photos of one of the subjects in the morph exists, or 3) morphs were searched against a gallery where two or more prior photos of both subjects in the morph exist. All galleries contained one or more photos of 1.6 million unique subjects.

6.5.2 Analysis of Rank 1 and 2 Scores: Unconsolidated Gallery

Figure 10 shows the distribution of rank 1 and rank 2 scores when bona fide and morph probes are searched against unconsolidated galleries of different subject composition. Consistent with the results observed in Section 6.3, mated bona fide searches generate rank 1 scores that are often higher than when the probe is a morph. Figure 10 presents score distributions for one algorithm. Results for a number of other 1:N algorithms we conducted this analysis on are presented in the Appendix.

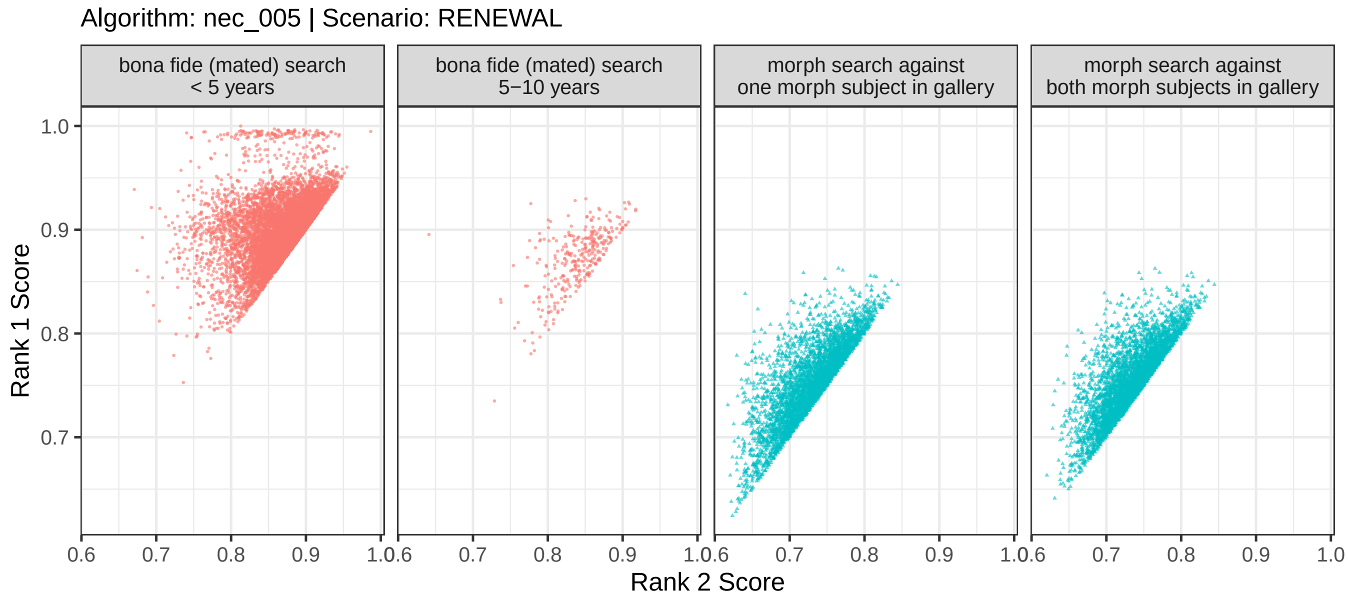


*Figure 10: The position of each point in the scatter plot represents the native rank 1 and 2 similarity scores for bona fide or morph probes searched against an **unconsolidated** gallery that contains two or more prior photos of the subject(s). For bona fides, the probes are broken out by the age difference between the probe and the gallery images (< 5 years and 5 - 10 years). For morphs, scores are shown for galleries that include either one subject or both subjects that went into the morph. All galleries contain one or more photos of 1.6 million unique people.*

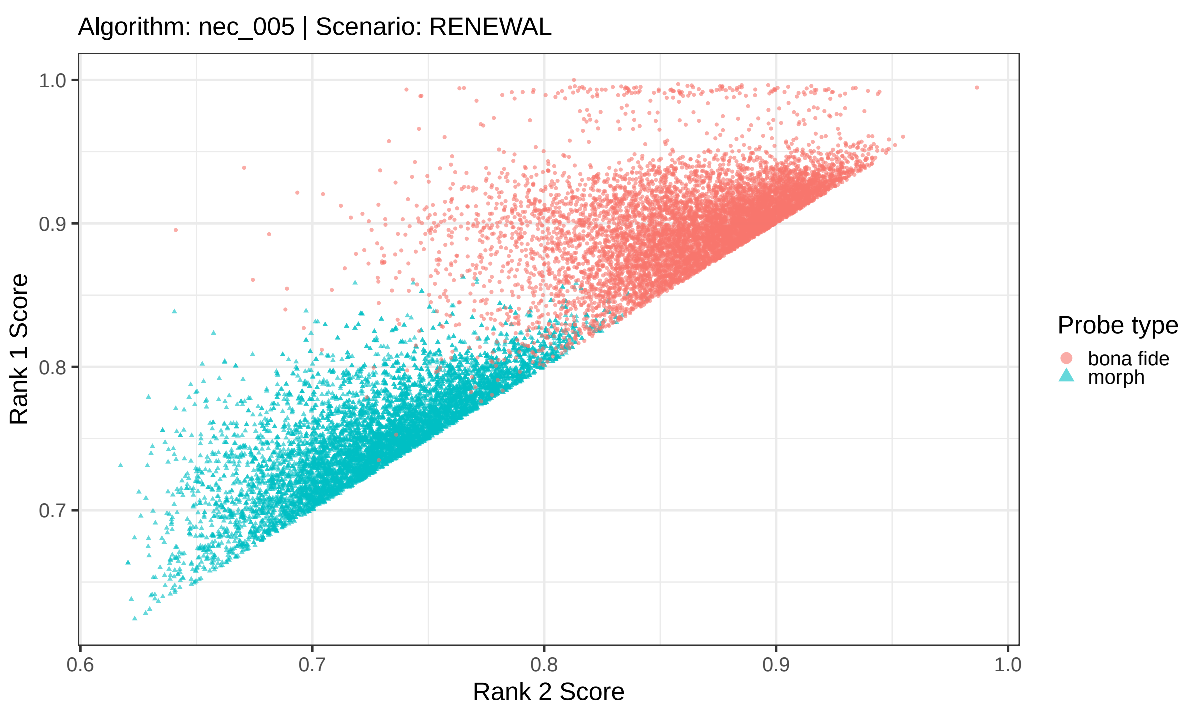


*Figure 11: This plot combines the rank 1 and 2 scores from the various bona fide and morph search scenarios against an **unconsolidated** gallery from Figure 10.*

6.5.3 Morph Detection Performance: Unconsolidated Gallery

Rank 1 and 2 score pairs generated by the same 1:N algorithms from Figure 8 were used to train and test morph classifiers under the scenario where the galleries are unconsolidated and contain multiple images of the same person. Following the same 5-fold cross validation training and testing methodology from Section 6.4, a random sample of 80% of the data (12 460 score pairs) is used for training, and the remaining 20% (3114 score pairs) is used for testing, repeated over five iterations. The 15 570 morph prediction scores (7 794 bona fide and 7 776 morph) across the five iterations of testing are used to measure classification accuracy.

From the results shown in Figure 12, we observe that the approach of using rank 1 and 2 score pairs for doing morph detection is often more effective when multiple images of the subject(s) exist in the gallery. A reduction in morph miss rates is observed when the morph classifiers are trained with scores from searching an unconsolidated database when compared to searching a consolidated database, and this trend is consistent across many of the algorithms that were tested. This is likely due to the fact that there is larger separability between the clusters of bona fide and morph scores when multiple images of the subject(s) are retrieved at rank 1 and 2, as visualized in Figure 11.

For the most accurate 1:N algorithm under an unconsolidated gallery scenario (sensetime-007), at a false detection rate of 0.001, its rank 1 and 2 score pairs generated a morph classifier where morph miss rate is 0.171 and 0.166 when either one or both subjects exist in the gallery, respectively. This means the classifier successfully detected around 83% of morphs when the threshold is set to generate a false detection 1 out of every 1 000 bona fide searches. Relaxing the threshold where false detection rate is 0.01 (one false detection out of every 100 bona fide searches), the trained sensetime-007 morph classifier would successfully detect morphs approximately 98% of the time.

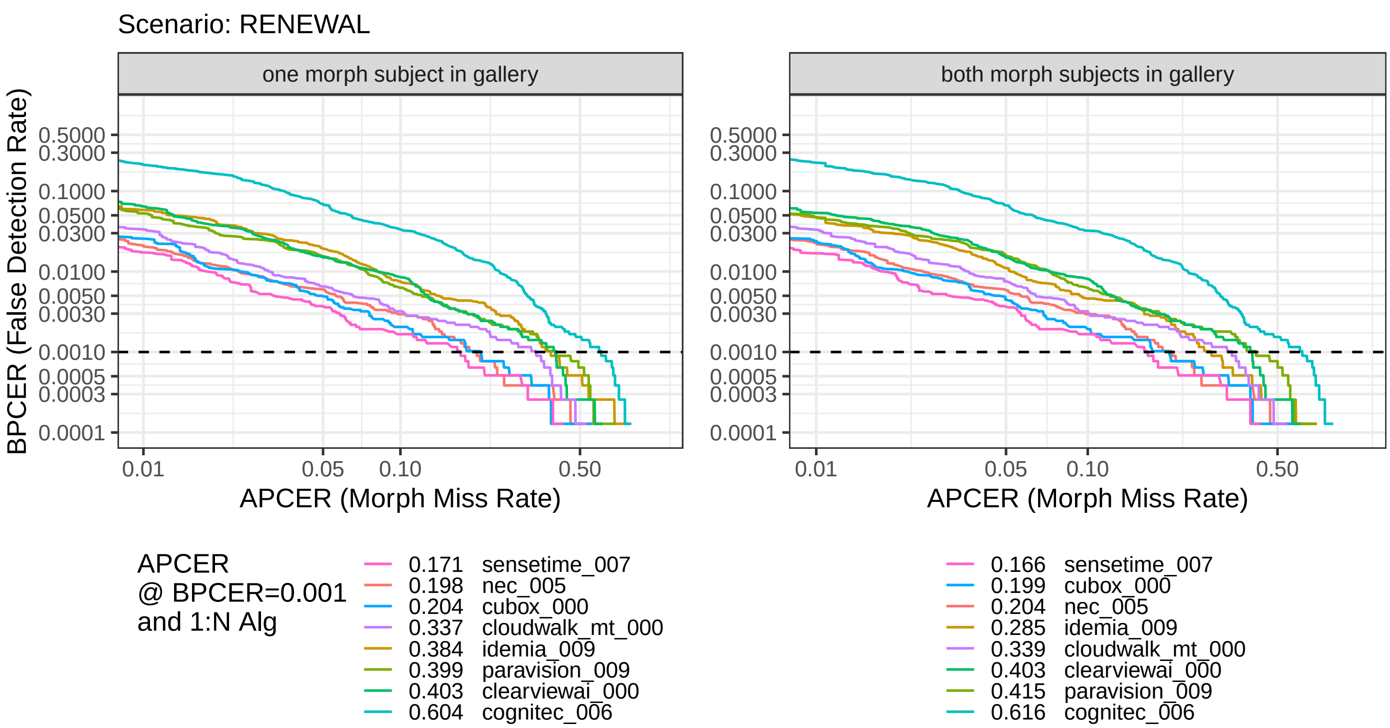


Figure 12: Accuracy of morph classifiers trained on different 1:N algorithm rank 1 + 2 scores for morphed searches (one or both subjects in gallery) and mated bona fide searches against a ***unconsolidated*** *gallery.*

7 New Enrollment/Application for ID credential

In a scenario where a person is applying for a new ID credential under an identity that presumably does not exist in the database, the expected outcome of a 1:N search of the application photo against the gallery would be the retrieval of very low similarity scores at rank 1 and 2 indicating that no existing matching identity was found. This behavior would be expected regardless of whether the photo is a bona fide or a morph, as illustrated in Figure 13.

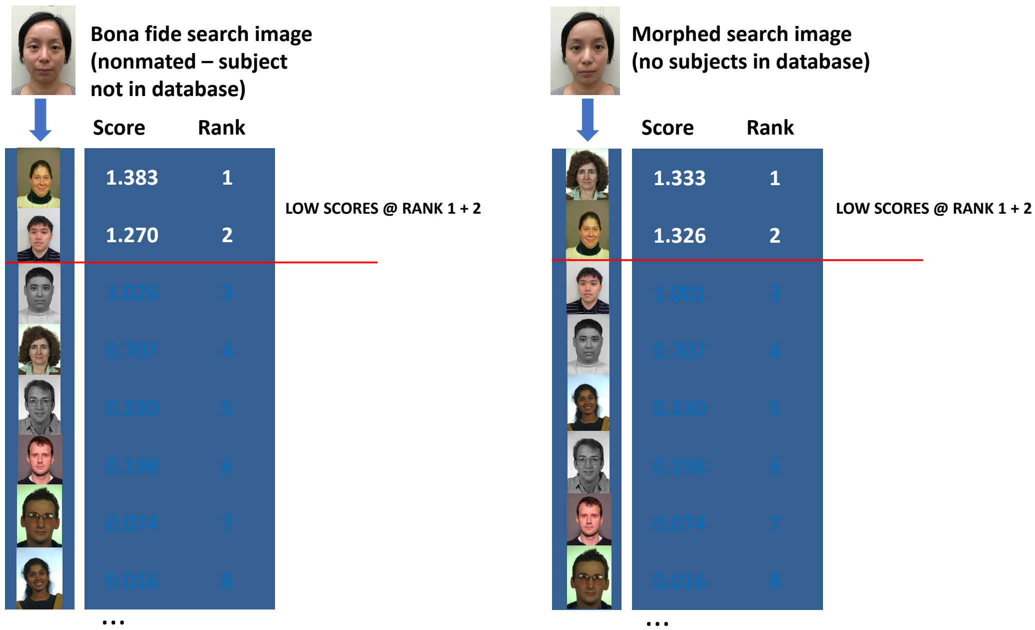


Figure 13: Notional examples of similarity scores that are returned on a candidate list for a nonmated search against a gallery where no images of the subject(s) exist in the gallery. Examples show possible candidate list outcomes based on the type of probe (bona fide or morph). The images in this figure are of NIST employees + subjects from the public NIST FERET Database [14].

7.1 Experiment

Following the same experimental procedures from Section 6.2, the same set of 21 393 morphs and 42 786 bona fides were searched against a gallery where no prior enrollment of the subject(s) exists.

7.2 Analysis of Rank 1 and Rank 2 Scores

Figure 14 shows the distribution of rank 1 and rank 2 scores when bona fide and morph probes are searched against a consolidated gallery where no prior photo of the subject(s) exists. Regardless of whether the probe is a bona fide or a morph, searches against a gallery where the applicant's identity does not exist generate very low scores at rank 1 and 2. Figure 14 presents score distributions for one algorithm. Results for a number of other 1:N algorithms we conducted this analysis on are presented in the Appendix.

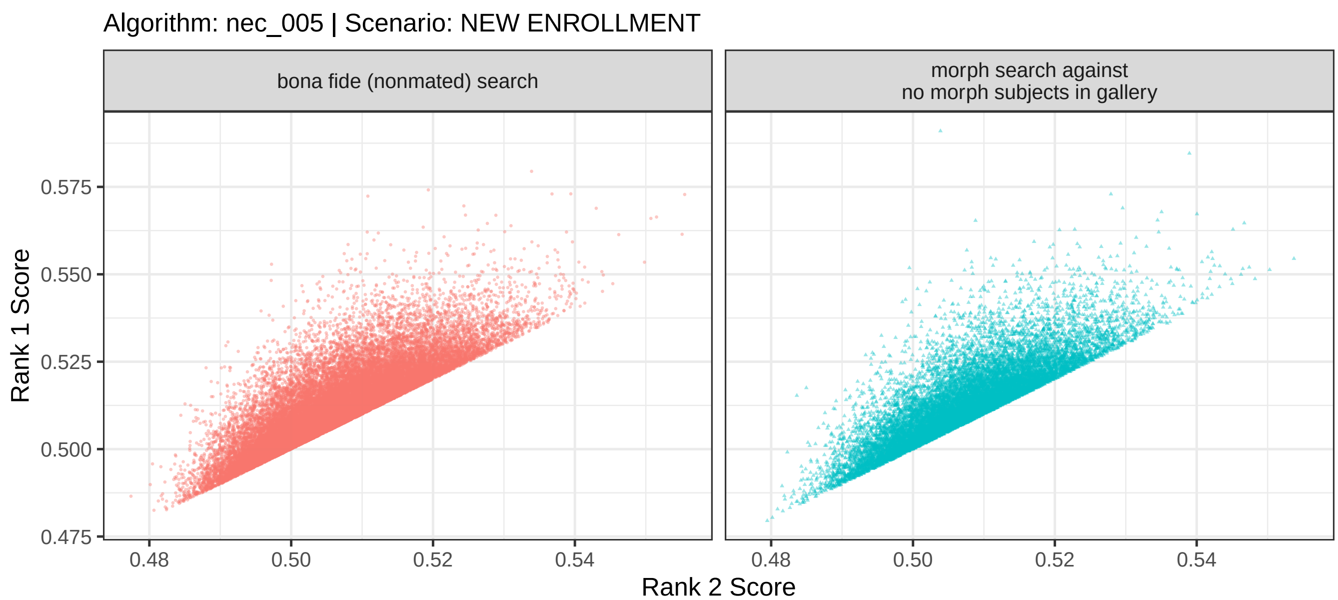


Figure 14: The position of each point in the scatter plot represents the native rank 1 and 2 similarity scores for bona fide or morph probes searched against a ***consolidated*** *gallery when no prior photos of the subjects exist in the gallery. All galleries contain 1.6 million unique people.*

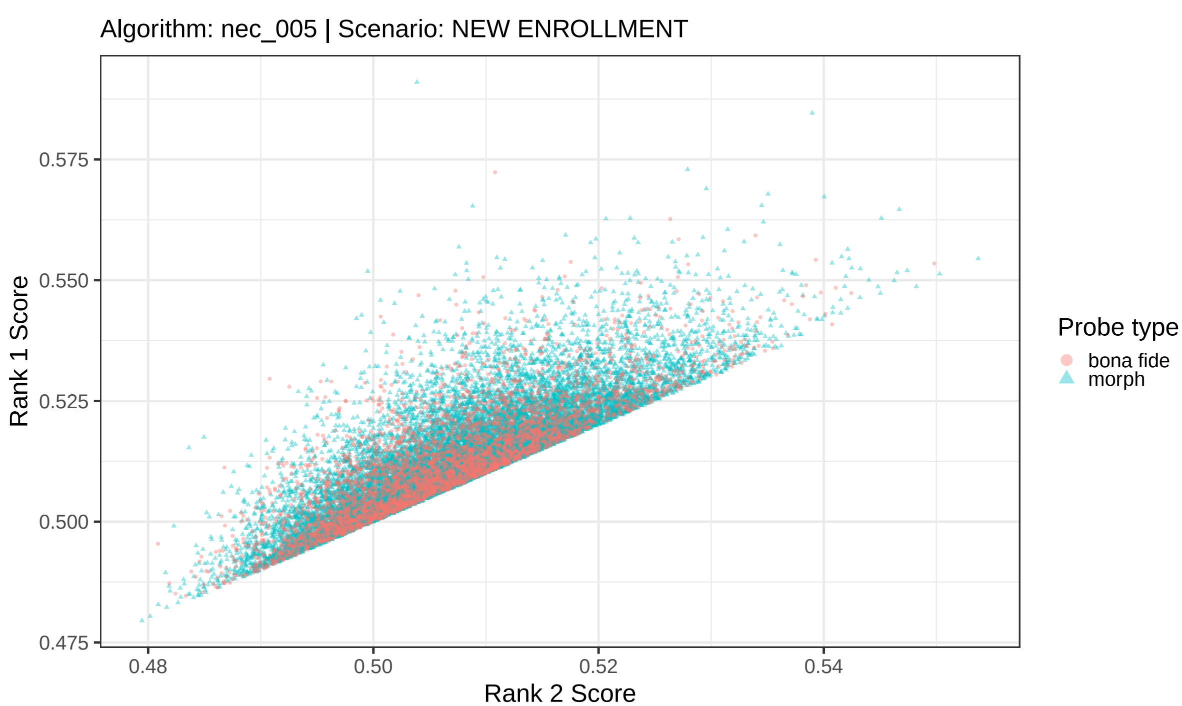


Figure 15: This plot combines the rank 1 and 2 scores from the bona fide and morph search scenarios against an ***consolidated*** *gallery from Figure 14.*

7.3 Morph Detection Performance

Using the rank 1 and 2 scores generated from the scenarios in Figure 14, we follow the same training and testing methodology from Section 6.4. Rank 1 and 2 score pairs and partitions generated by the same 1:N algorithms were used to train and test morph classifiers under the scenario where bona fides and morphs were searched against a gallery that did not contain the subject(s) . Following the same 5-fold cross validation training and testing methodology, a random sample of 80% of the data (55 622 score pairs) is used for training, and the remaining 20% (12 836 score pairs) is used for testing, repeated over five iterations. The 64 180 morph prediction scores (42 806 bona fide and 21 374 morph) across the five iterations of testing are used to measure classification accuracy.

From the results shown in Figure 16, we observe that the approach of using rank 1 and 2 score pairs for doing morph detection is **not** effective in a new enrollment scenario where the subjects in the morph have not been previously encountered. The significant overlap between the bona fide and morph score distributions under this scenario, as visualized in Figures 14 and 15, makes it difficult for the morph classifier to differentiate between a legitimate photo versus a morph.

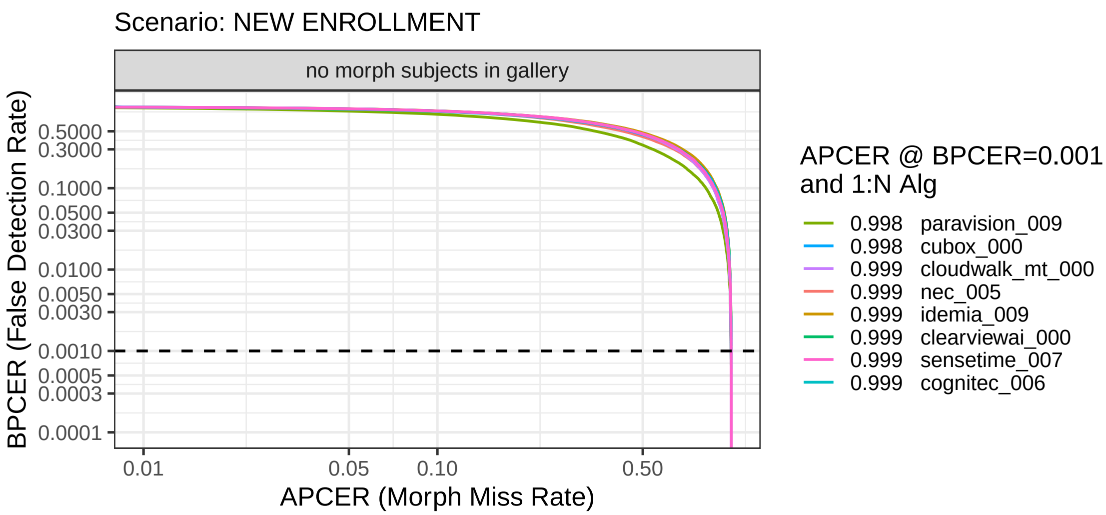


Figure 16: Accuracy of morph classifiers trained on different 1:N algorithm rank 1 + 2 scores for morphed or bona fide searches against a ***consolidated*** *gallery when no prior photos of the subjects exist in the gallery.*

8 Discussion

8.1 Utility of a 1:N Algorithm for Morph Detection in ID Renewal Processes

Based on the outcomes of the experiments discussed in this report, 1:N face recognition systems may have utility in detection of morphed photos in operational pipelines, with particularly promising results under an ID renewal scenario. One potential advantage of using this 1:N approach is that many ID issuance agencies (e.g., passport offices) will already have a 1:N face recognition system within their operational pipeline, so there is opportunity to reuse existing infrastructure in lieu of procuring a dedicated morph detection capability.

At a high level, the following methodology for leveraging a 1:N face recognition algorithm for morph detection is proposed:

- Generate a set of morphs with relevant imagery, ideally operational data from the operational system (see Section 3.2).

- Run searches with both morph and bona fide probes against the operational gallery (which contains existing photo(s) of the search subject(s)).
- Conduct analysis of how your 1:N system behaves on morph vs. bona fide searches (our approach proposes analyzing rank 1 and 2 score pairs).
- Identify potential thresholds for flagging suspicious activity based on trends observed between morph vs. bona fide score pairs. This might be accomplished through visual inspection, training a morph classifier with the score pairs, or some other approach.
- Continuously and iteratively adjust thresholds/policies based on actual morph detection outcomes with human review and additional supporting information.

There will be errors associated with automated morph detection, and in our assessment of similarity scores as a potential way to classify morphs, it is possible for some legitimate mated bona fide searches to generate low similarity scores. The goal may be to establish thresholds such that false detection rates are at acceptable levels and even if morph detection rates are low, it would still yield gains in operations compared to not having any morph detection capability at all. We do not conceive of automated morph detection as being a lights-out operation, and there will almost always be human review, investigation, and remediation of suspiciously flagged images. The investigation process may not be trivial and may involve asking the applicant to conduct facilitated in-person photo recollection or some other process to reduce the opportunity for image manipulation.

8.2 Future Work

This report presents a methodology and initial quantitative results on the use of automated one-to-many search algorithms as a mechanism to potentially detect the presence of morphs. There are factors that may have an impact on the morph detection outcomes reported in our study. NIST plans on releasing a series of updates to this report, iteratively assessing different aspects relevant to face morphing and their potential influence on morph detection performance using one-to-many face recognition.

Appendix A Individual Algorithm Results

ID Renewal Scenario: The following plots show 1:N algorithm rank 1 and 2 similarity scores for bona fide or morph probes searched against a **consolidated** gallery, simulating possible scenarios for ID renewal:

- The application photo is a legitimate bona fide photo of the applicant, and a prior photo of the applicant exists in the database
- The application photo is a morphed photo of the applicant and a ”hidden identity” and
  - A prior photo of the applicant exists in the database
  - A prior photo of both the applicant and the ”hidden identity” exist in the database

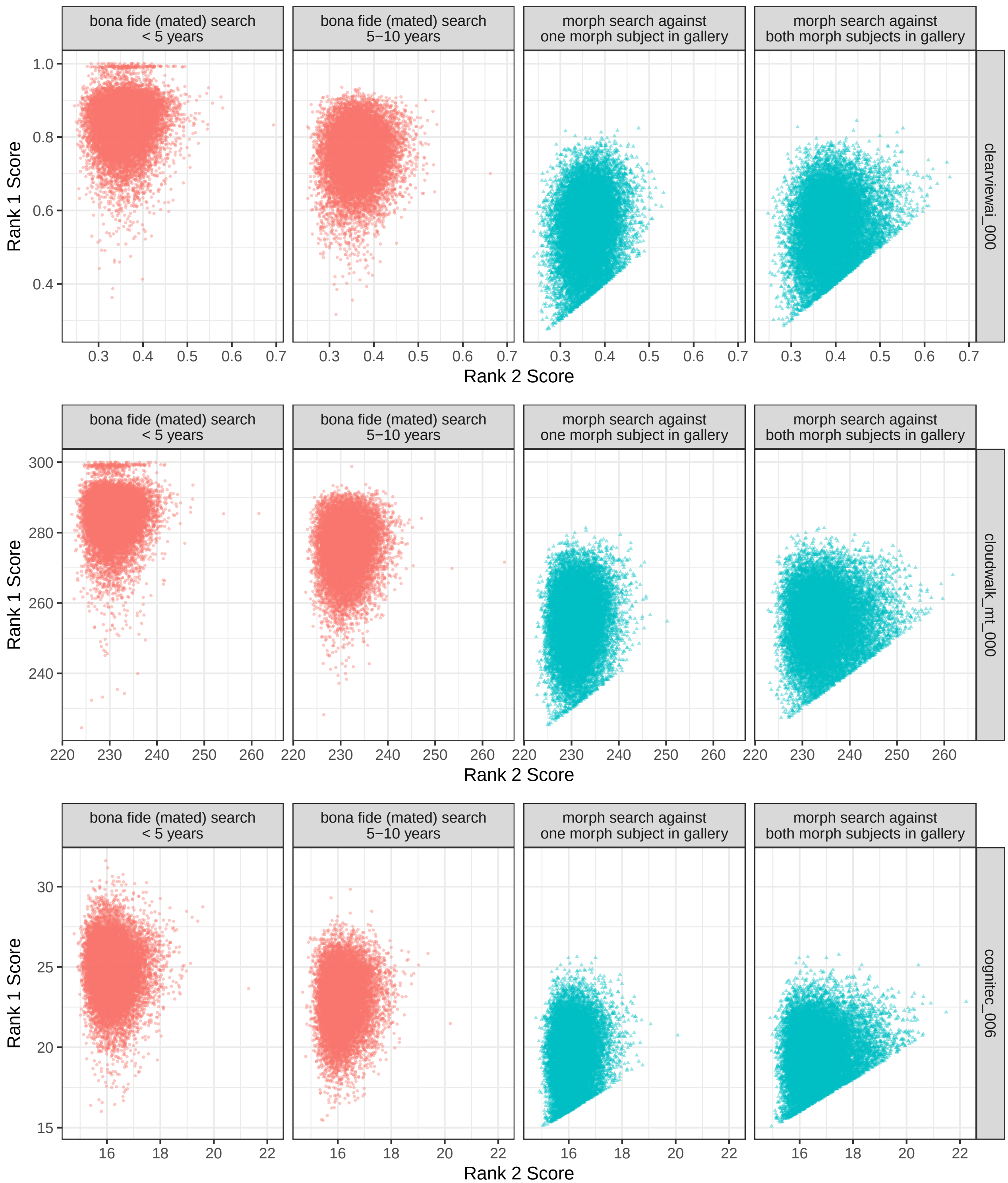


Figure 17: The scatter plots show native 1:N algorithm rank 1 and 2 similarity scores for bona fide or morph probes searched against a ***consolidated*** *gallery that contains a prior photo of the subject(s). For bona fides, scores are broken out by the time elapsed (< 5 years or between 5 - 10 years) between the probe and gallery photo. For morphs, scores are shown for galleries that include either one subject or both subjects that went into the morph. All galleries contain 1.6 million unique people.*

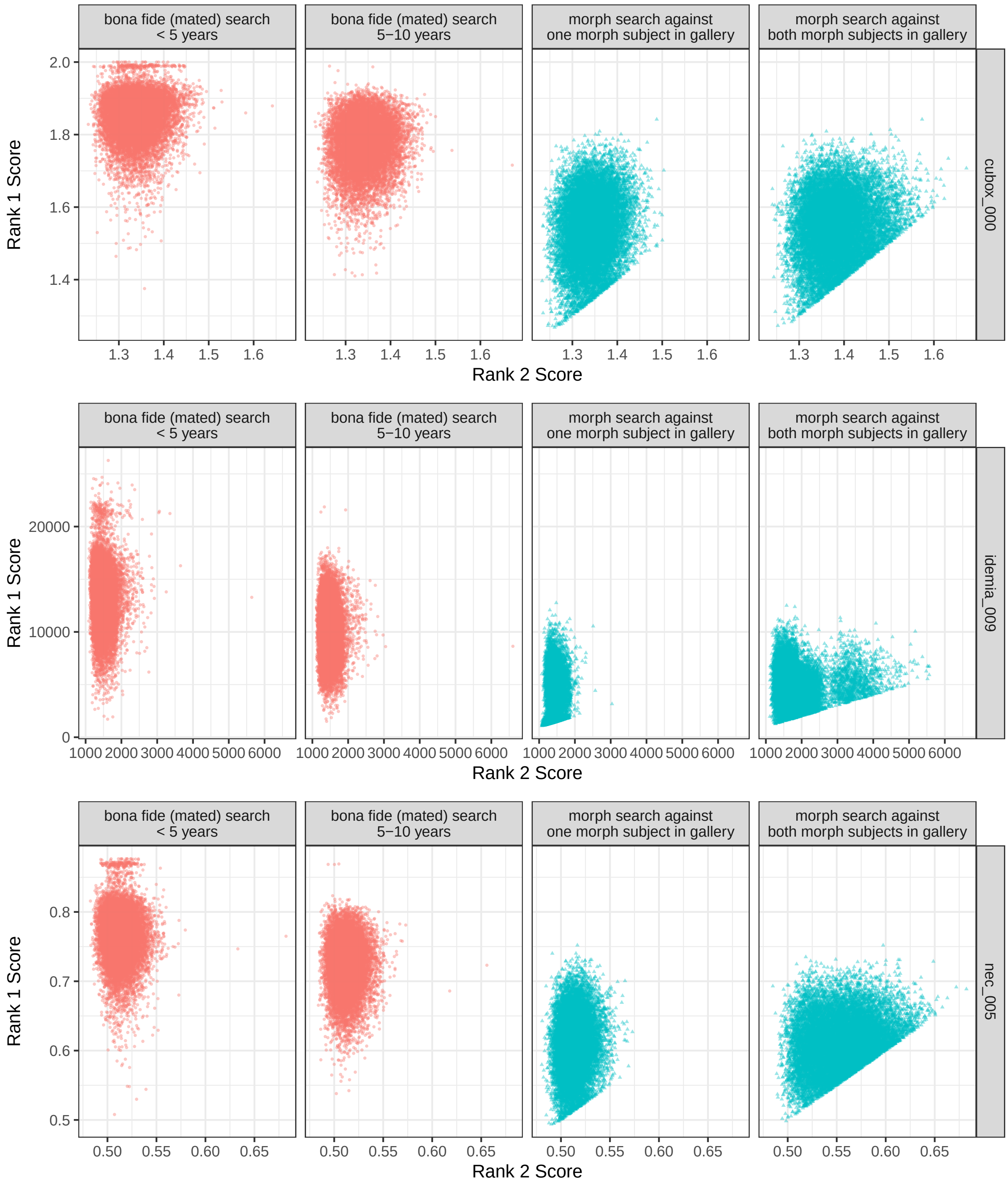


Figure 18: The scatter plots show native 1:N algorithm rank 1 and 2 similarity scores for bona fide or morph probes searched against a ***consolidated*** *gallery that contains a prior photo of the subject(s). For bona fides, scores are broken out by the time elapsed (< 5 years or between 5 - 10 years) between the probe and gallery photo. For morphs, scores are shown for galleries that include either one subject or both subjects that went into the morph. All galleries contain 1.6 million unique people.*

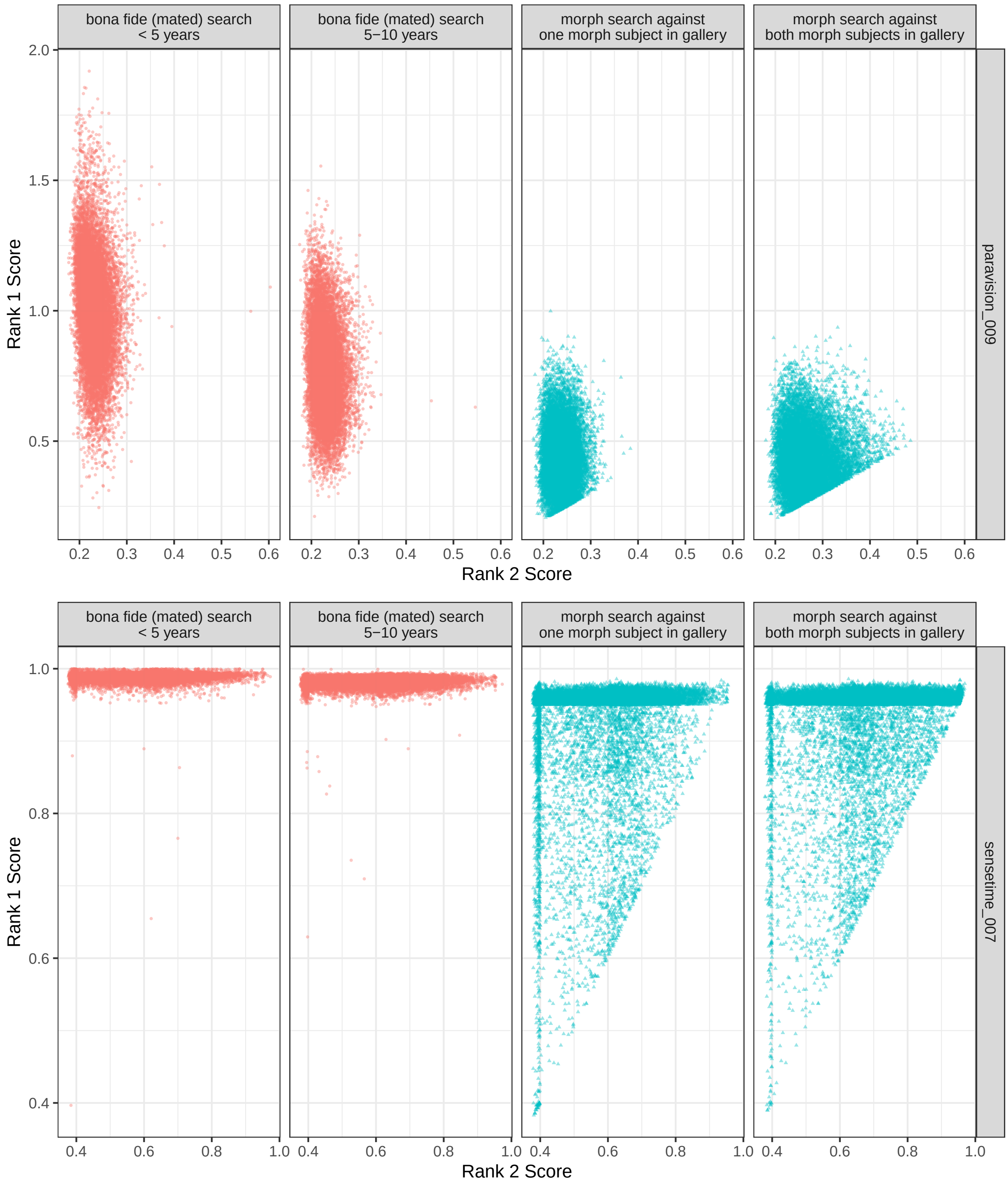


*Figure 19: The scatter plots show native 1:N algorithm rank 1 and 2 similarity scores for bona fide or morph probes searched against a **consolidated** gallery that contains a prior photo of the subject(s). For bona fides, scores are broken out by the time elapsed (< 5 years or between 5 - 10 years) between the probe and gallery photo. For morphs, scores are shown for galleries that include either one subject or both subjects that went into the morph. All galleries contain 1.6 million unique people.*

ID Renewal Scenario: The following plots show 1:N algorithm rank 1 and 2 similarity scores for bona fide or morph probes searched against an **unconsolidated** gallery that contains multiple (two or more) prior photos of the subject(s), simulating possible scenarios for ID renewal:

- The application photo is a legitimate bona fide photo of the applicant, and multiple prior photos of the applicant exist in the database
- The application photo is a morphed photo of the applicant and a ”hidden identity” and
  - Multiple prior photos of the applicant exist in the database
  - Multiple prior photos of both the applicant and the ”hidden identity” exist in the database

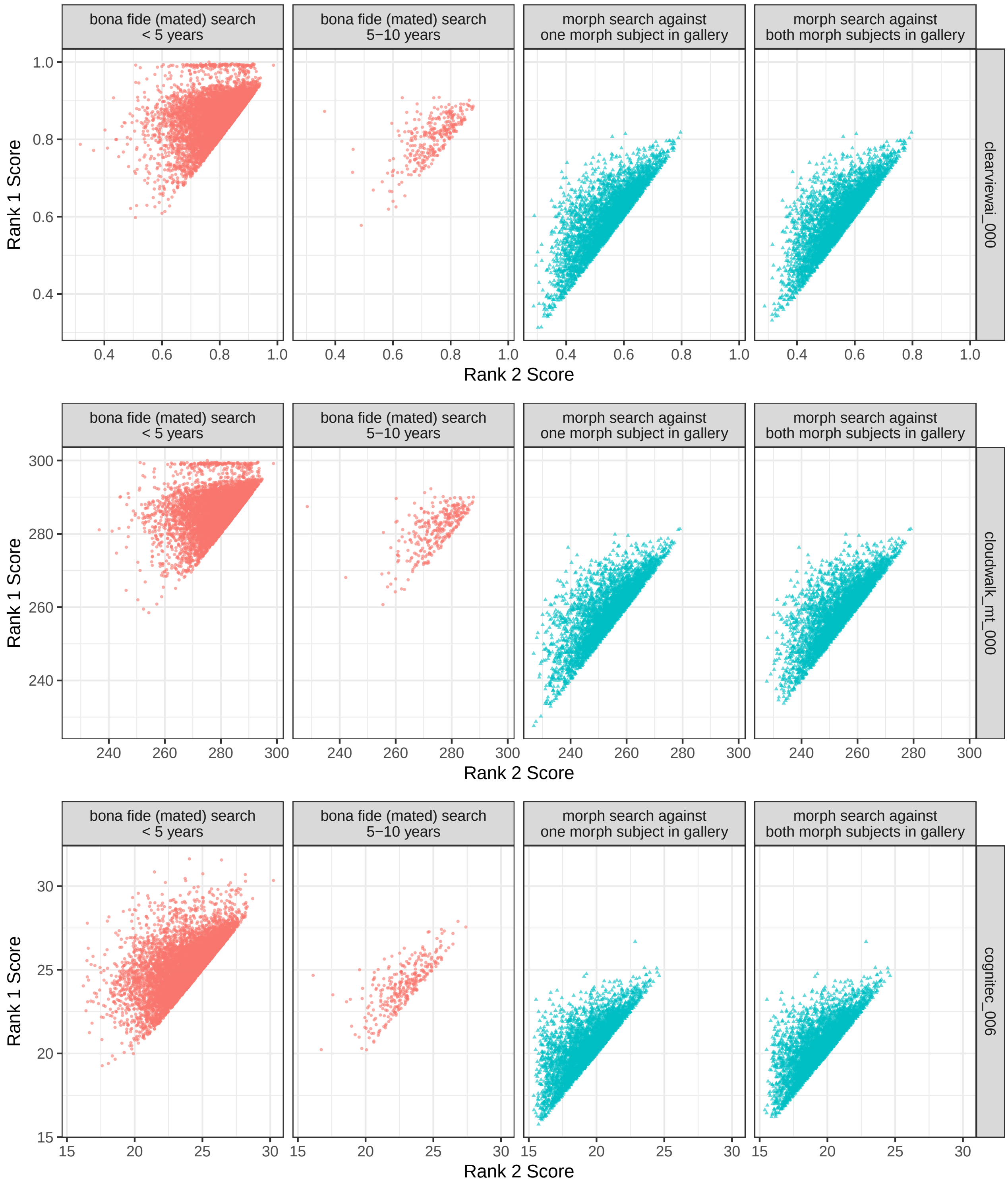


*Figure 20: The scatter plots show native 1:N algorithm rank 1 and 2 similarity scores for bona fide or morph probes searched against an **unconsolidated** gallery that contains multiple (two or more) prior photos of the subject(s). For bona fides, scores are broken out by the time elapsed (< 5 years or between 5 - 10 years) between the probe and gallery photos. For morphs, scores are shown for galleries that include either one subject or both subjects that went into the morph. All galleries contain 1.6 million unique people.*

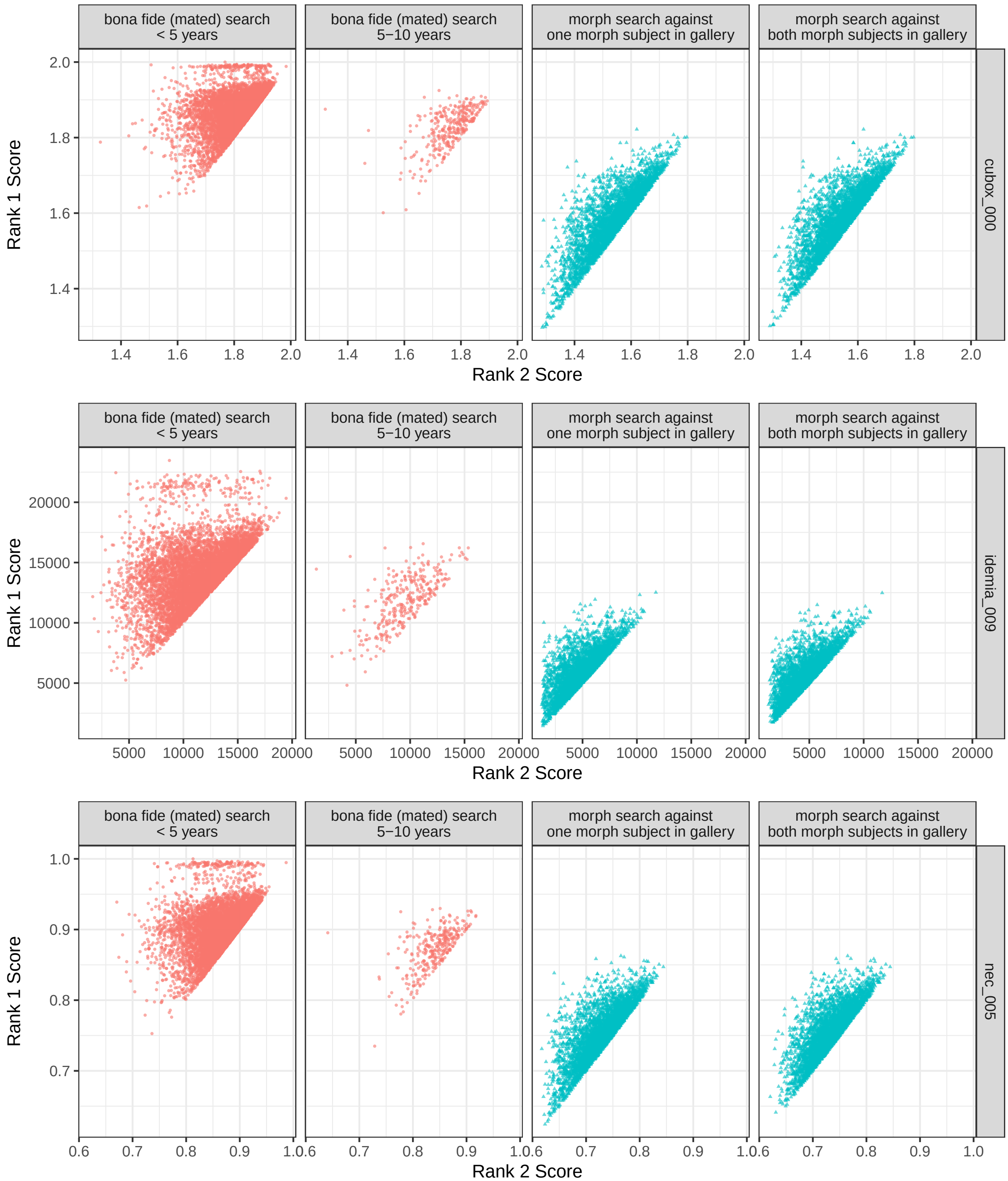


*Figure 21: The scatter plots show native 1:N algorithm rank 1 and 2 similarity scores for bona fide or morph probes searched against an **unconsolidated** gallery that contains multiple (two or more) prior photos of the subject(s). For bona fides, scores are broken out by the time elapsed (< 5 years or between 5 - 10 years) between the probe and gallery photos. For morphs, scores are shown for galleries that include either one subject or both subjects that went into the morph. All galleries contain 1.6 million unique people.*

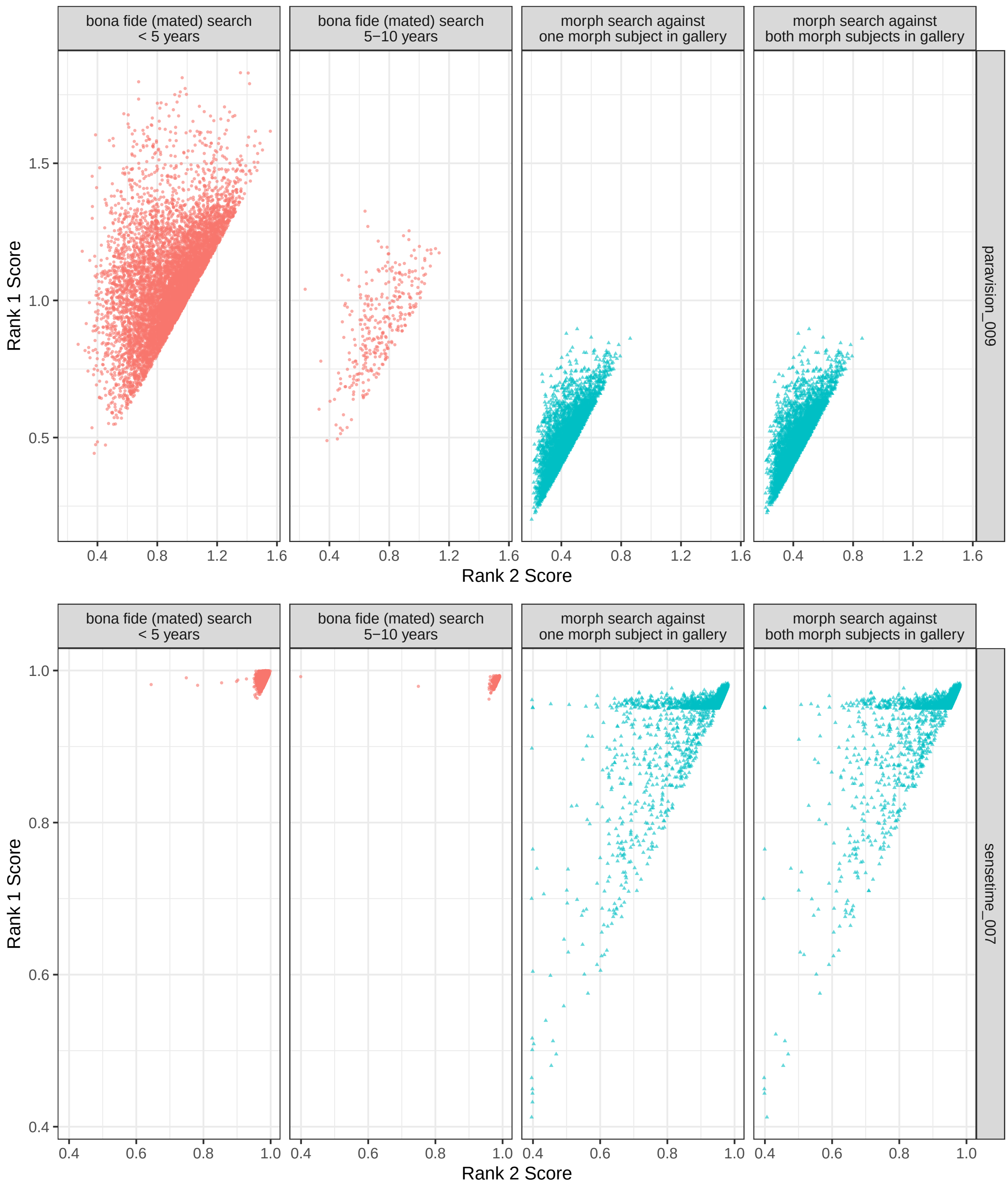


Figure 22: The scatter plots show native 1:N algorithm rank 1 and 2 similarity scores for bona fide or morph probes searched against an ***unconsolidated*** *gallery that contains multiple (two or more) prior photos of the subject(s). For bona fides, scores are broken out by the time elapsed (< 5 years or between 5 - 10 years) between the probe and gallery photos. For morphs, scores are shown for galleries that include either one subject or both subjects that went into the morph. All galleries contain 1.6 million unique people.*

New Enrollment/ID Application Scenario: The following plots show 1:N algorithm rank 1 and 2 similarity scores for bona fide or morph probes searched against a **consolidated** gallery, simulating possible scenarios of a new application for an ID credential:

- The application photo is a legitimate bona fide photo of the applicant and no prior photo(s) of the applicant exist in the database
- The application photo is a morphed photo of the applicant and a ”hidden identity” and no prior photo(s) of the contributing subjects exist in the database

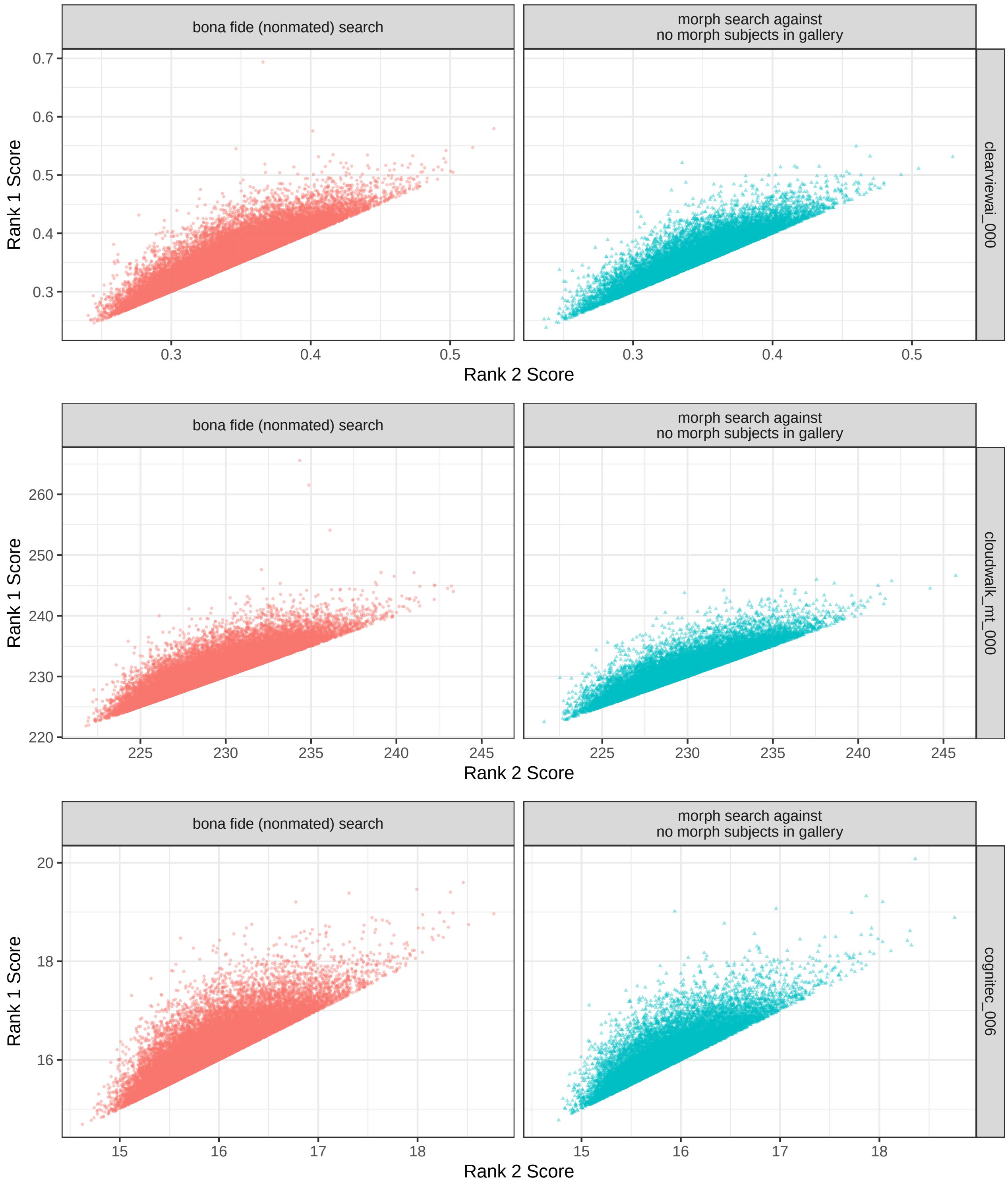


Figure 23: The position of each point in the scatter plot represents the native rank 1 and 2 similarity scores for bona fide or morph probes searched against a consolidated gallery when no prior photos of the subjects exist in the gallery. All galleries contain 1.6 million unique people.

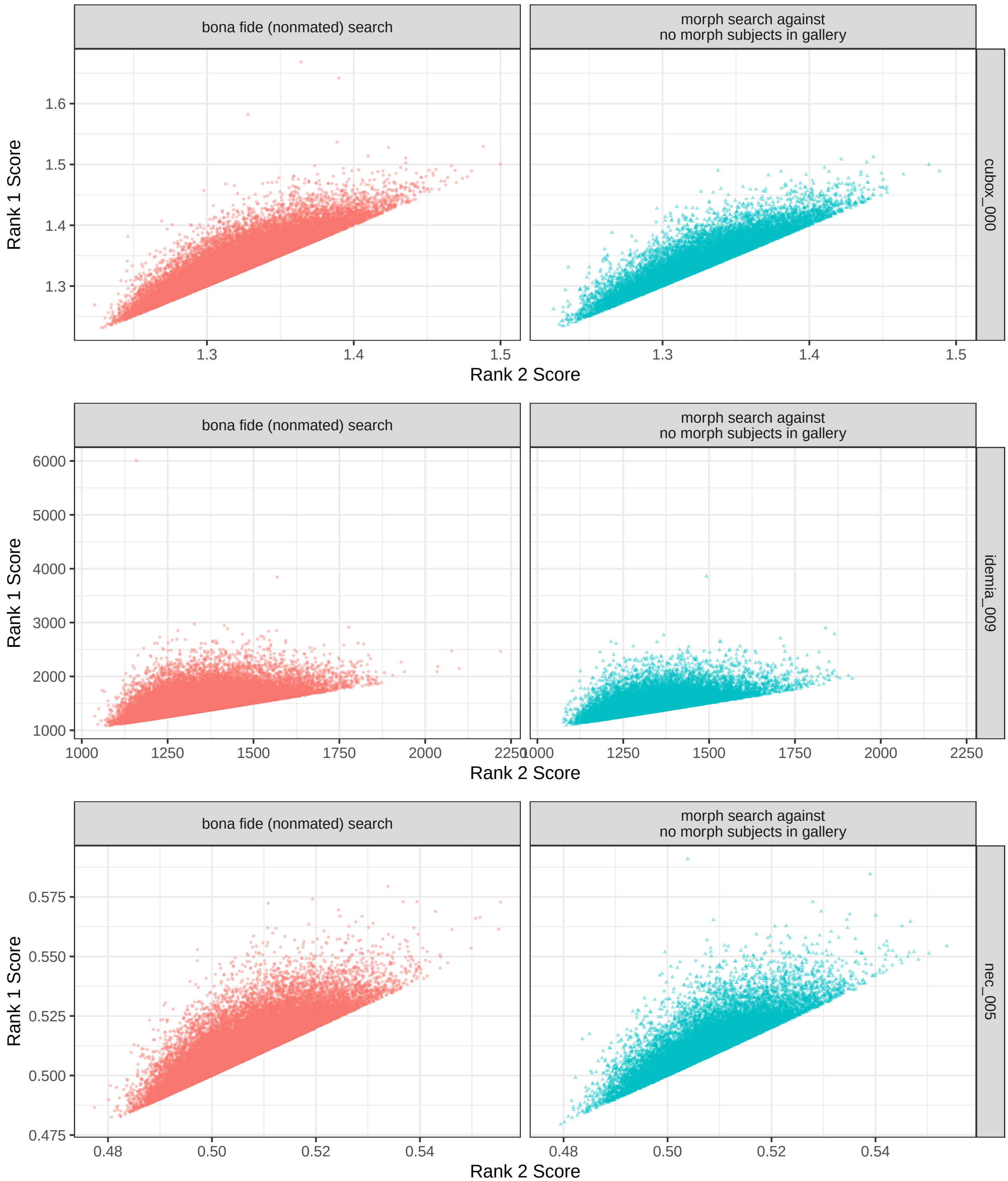


Figure 24: The position of each point in the scatter plot represents the native rank 1 and 2 similarity scores for bona fide or morph probes searched against a consolidated gallery when no prior photos of the subjects exist in the gallery. All galleries contain 1.6 million unique people.

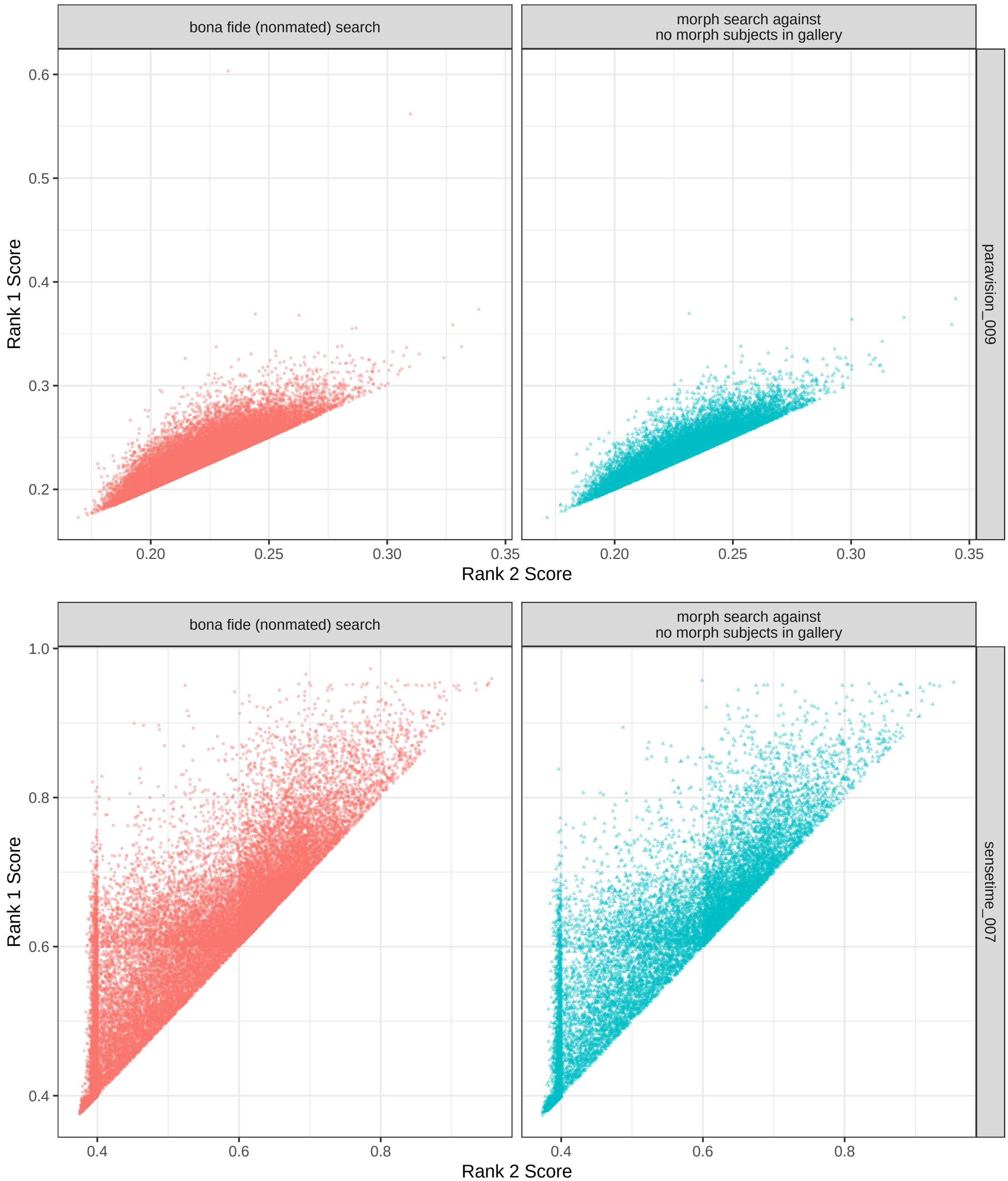


Figure 25: The position of each point in the scatter plot represents the native rank 1 and 2 similarity scores for bona fide or morph probes searched against a consolidated gallery when no prior photos of the subjects exist in the gallery. All galleries contain 1.6 million unique people.

References

[1] M. Ferrara, A. Franco, and D. Maltoni. The magic passport. In *IEEE International Joint Conference on Biometrics*, pages 1–7, Sep. 2014.

[2] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. *On the Effects of Image Alterations on Face Recognition Accuracy*, pages 195–222. Springer International Publishing, Cham, 2016.

[3] David J. Robertson, Robin S. S. Kramer, and A. Mike Burton. Fraudulent id using face morphs: Experiments on human and automatic recognition. *PLOS ONE*, 12(3):1–12, 03 2017.

[4] Robin S. S. Kramer, Michael O. Mireku, Tessa R. Flack, and Kay L. Ritchie. Face morphing attacks: Investigating detection with humans and computers. *Cognitive Research: Principles and Implications*, 4(1):28, 2019.

[5] Face Recognition Vendor Test (FRVT) MORPH Webpage. https://pages.nist.gov/frvt/reports/morph/frvt_morph_report.pdf.

[6] Ulrich Scherhag, Andreas Nautsch, Christian Rathgeb, Marta Gomez-Barrero, Raymond Veldhuis, Luuk Spreeuwers, Maikel Schils, Davide Maltoni, Patrick Grother, Sebastien Marcel, Ralph Breithaupt, R. Raghavendra, and Christoph Busch. Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting. pages 1–7, 09 2017.

[7] The CentOS Project. https://www.centos.org.

[8] M. Ferrara, A. Franco, and D. Maltoni. Face demorphing. *IEEE Transactions on Information Forensics and Security*, 13(4):1008–1017, April 2018.

[9] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. *On the Effects of Image Alterations on Face Recognition Accuracy*, pages 195–222. Springer International Publishing, Cham, 2016.

[10] M. Ferrara, A. Franco, and D. Maltoni. Decoupling texture blending and shape warping in face morphing. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7, 2019.

[11] Mei Ngan, Patrick Grother, Kayee Hanaoka, and Jason Kuo. Face Recognition Vendor Test (FRVT) Part 4: MORPH - Performance of Automated Face Morph Detection, NISTIR 8292 Draft Supplement, April 2022. https://pages.nist.gov/frvt/reports/morph/frvt_morph_report.pdf.

[12] JTC 1/SC 37. International organization for standardization: Information technology biometric presentation attack detection part 3: Testing and reporting. In *ISO/IEC 30107-3*, 2017.

[13] E. J. Berg. *Heaviside's operational calculus as applied to engineering and physics*. Electrical engineering texts. McGraw-Hill book company, inc., 1936.

[14] P. Jonathon Phillips, Hyeonjoon Moon, Syed Rizvi, and Patrick Rauss. The feret evaluation methodology for face-recognition algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22:1090 – 1104, 10 2000.