Population Variation and Equity
Volume 1 · Number 1 · 2026

A Small Average Change, Different Identity-Level Outcomes: Two Simple Face Representations

Adam Johnston[1] · Eric Fisher[2] · Paul Lewis[3]

[1] Switzerland · [2] Ireland · [3] Finland

Original publication: November 3, 2026 | Article 667

Abstract

Centering and normalizing face images changed which queries were identified correctly without producing a clear average benefit. In one seeded Olivetti gallery split, raw-pixel nearest-gallery accuracy was 87.5 percent and normalized-image accuracy was 85.5 percent. Six queries became correct and ten became incorrect. The paired difference was -2.0 percentage points, with an identity-clustered interval of -6.05 to 2.05 points. Four identities had positive net gains, six had losses, and thirty had zero net change. Gallery-only brightness and image-spread associations remained uncertain. This small closed-set comparison illustrates why aggregate identity accuracy needs paired outcomes and subject-level context, while providing no benchmark of contemporary face embeddings or subject-preserving generation.

Closed-set nearest-gallery identification on a public historical grayscale image collection, with no learned recognition network or generated video.

A Small Average Change, Different Identity-Level Outcomes: Two Simple Face Representations

Closed-set nearest-gallery identification on a public historical grayscale image collection, with no learned recognition network or generated video.

Abstract

Centering and normalizing face images changed which queries were identified correctly without producing a clear average benefit. In one seeded Olivetti gallery split, raw-pixel nearest-gallery accuracy was 87.5 percent and normalized-image accuracy was 85.5 percent. Six queries became correct and ten became incorrect. The paired difference was -2.0 percentage points, with an identity-clustered interval of -6.05 to 2.05 points. Four identities had positive net gains, six had losses, and thirty had zero net change. Gallery-only brightness and image-spread associations remained uncertain. This small closed-set comparison illustrates why aggregate identity accuracy needs paired outcomes and subject-level context, while providing no benchmark of contemporary face embeddings or subject-preserving generation.

1. An average representation gain can hide who changes

A representation can leave most predictions unchanged, improve a few, and damage a few others. Comparing only its final accuracy with a reference score conceals that paired pattern. This is especially consequential in a small identity collection, where each subject has only a handful of query photographs. A percentage-point difference at the collection level may arise from very few changed decisions, and apparent subgroup effects can be much less stable than their numerical precision suggests.

The research question is whether per-image centering and normalization benefit identities similarly in a simple nearest-gallery identification task. We first estimate the overall paired correctness difference. We then describe identity-specific net differences and ask whether gallery-only image spread and brightness are associated with those differences. The primary directional proposition favors a positive average normalization gain. The property associations are exploratory descriptions, without a claim that modifying brightness or gallery variability would cause the estimated change.

Subject-preserving generation, as studied in Meng et al., 2026, makes identity measurement an important practical issue. Our experiment does not generate images or video, recover identity from a synthesized sequence, or execute that method. Instead, it uses ordinal labels in an existing face-image collection to examine a basic representation choice. The distinction is substantive: a closed-set identification baseline on historical frontal photographs cannot certify preservation of a person's identity in a new generative setting.

2. A small public archive with a constrained photographic setting

The data resource in AT&T Laboratories Cambridge, 2026 contains four hundred grayscale images of forty identities, ten images per identity. We obtain the Sam Roweis MATLAB version referenced by the scikit-learn maintainer loader and verify that its SHA256 exactly matches the loader's published checksum. The matrix has 4,096 image features per photograph, corresponding to sixty-four by sixty-four pixels. We follow the documented row grouping for labels and transpose the spatial image layout to match that maintained version.

Credit belongs to AT&T Laboratories Cambridge. The collection consists of historical photographs acquired under a constrained, largely frontal setting with some variation in expression, lighting, and facial

details. The identification work represented by Samaria and Harter, 1994 provides the original research context. We do not treat the archive as a random sample of current users, all imaging devices, or unconstrained video frames. Its small size and regular background are central limitations, even when it gives a convenient inspectable baseline.

The forty labels identify groups of photographs belonging to the same released subject. They do not provide demographic categories, and we do not infer such attributes from appearance or label numbering. The demographic evaluation motivation in Kärkkäinen and Joo, 2019 therefore cannot be implemented by relabeling these forty identities as demographic groups. The subgroup unit in this article is simply the released identity. Public mirror access and file integrity also do not establish unrestricted rights to redistribute the photographs; the analysis preserves provenance and creator credit.

3. One gallery division, with every identity represented

Using random seed 20261107, we permute the ten source-row indices within each identity. The first five form its gallery and the remaining five its query set. Across identities there are two hundred gallery images and two hundred query images. No source row appears in both sets. Every query identity has five gallery examples, so the task is closed-set identification: assign each query to one of the forty identities already represented in the gallery.

This division is not a split between seen and unseen people. It evaluates new photographs of identities available to the lookup rule. Nor is it a chronological split; source image indices do not supply a validated capture-time ordering. A new seed could place different expressions or lighting conditions in the gallery and alter the comparison. We retain the full realized assignment and do not search repeated seeds for the largest normalization gain. Uncertainty from repeatedly drawing galleries is outside the single-split interval.

The unconstrained photographic context of Huang et al., 2008 and the broader pose variation represented by Cao et al., 2017 illustrate settings that differ from this archive. Neither dataset contributes images to the executed comparison. Cross-context generalization would require actual additional observations and an evaluation design compatible with their labels. It cannot be inferred by citing a larger benchmark or assuming that a simple preprocessing result has the same sign under different pose, cropping, or acquisition conditions.

4. Two distances, both using the same gallery photographs

The reference representation is the flattened grayscale vector. We subtract the archive-wide minimum intensity and divide by its resulting maximum, matching the maintained source convention while retaining float64 arithmetic. These common intensity operations use the complete archive, but they are a single shared positive affine transformation. Such a transformation preserves raw Euclidean nearest-neighbor rankings. It does not estimate an identity-specific classifier from the query labels or provide a tuned test-image feature transformation.

The second representation subtracts each image's own mean and divides the centered vector by its Euclidean norm. All selected image norms are positive. We apply the same rule independently to gallery and query photographs. Each query is assigned the identity of its closest gallery photograph under squared Euclidean distance in the selected representation. Gallery order supplies deterministic tie resolution. No dimension reduction, neural embedding, augmentation, learned threshold, or fitted class centroid is introduced.

The projection-based history in Turk and Pentland, 1991 and Belhumeur et al., 1997 helps locate this intentionally elementary baseline. We do not compute either projection. Likewise learned recognition approaches in Schroff et al., 2015, Deng et al., 2018, Wang et al., 2018, Kim et al., 2022, and Meng et al., 2021 use representations whose behavior cannot be inferred from per-image pixel normalization. Our two rules provide a controlled comparison of untrained transformations. Their numerical accuracies should not be advertised as a competitive ranking against modern recognition systems.

5. The diagram makes the identity task explicit

Centering removes an image's average intensity, and unit normalization removes a common positive contrast scale from its centered vector. These are mathematical properties of the representation. Whether they help a nearest-gallery identity decision is an empirical question because absolute intensity and contrast can carry both nuisance variation and useful similarity information. We do not assume that removing them must preserve identity-relevant evidence, and no synthetic lighting change is imposed on the queries.

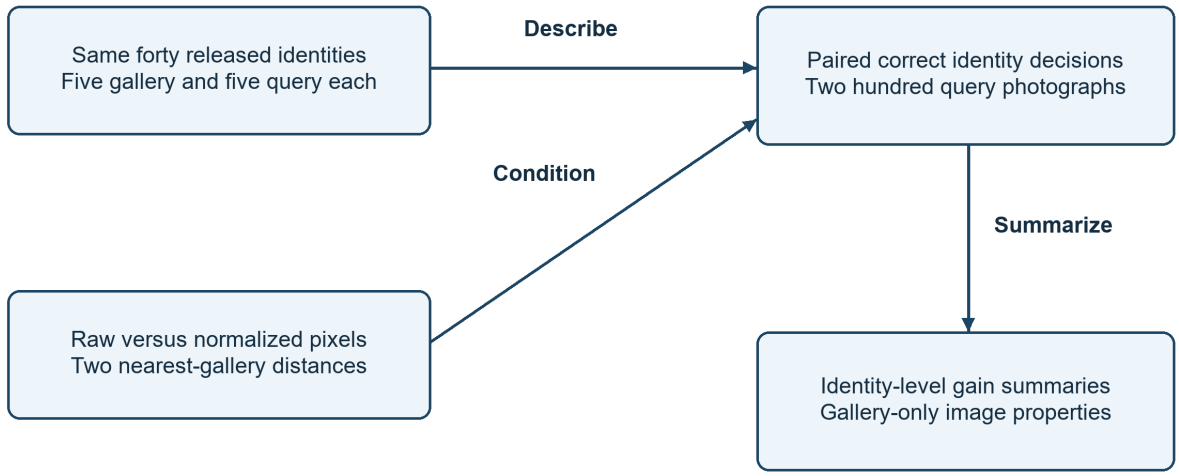


Figure 1. The same seeded gallery and query images enter two distance representations; paired correctness is summarized by identity before gallery-property associations are examined.

The figure ends with correctness and identity-level contrasts for a fixed photographic archive. Mitchell, 1997 supplies a separate historical face-image resource, while Tan and Triggs, 2010 studies recognition under difficult illumination; neither is substituted for the present cohort or preprocessing. Personalization in Guo et al., 2024 and video identity settings in Yuan et al., 2024 and He et al., 2024 introduce generative or temporal questions absent from this experiment. The retained diagram therefore describes static matching outcomes, without claiming generated-subject fidelity or measuring temporal consistency.

6. Paired correctness and gallery-only covariates

For each query we record whether the raw and normalized representations assign its released identity correctly. Subtracting raw correctness from normalized correctness yields a paired outcome taking minus one, zero, or one. An intercept-only ordinary least-squares model estimates its mean over two hundred queries. Covariance clusters the five queries belonging to each of forty identities, uses a finite-sample correction, and applies a t reference with thirty-nine degrees of freedom. The dependence reasoning in Cameron and Miller, 2015 supports making those shared identities explicit.

For the second model we aggregate the paired outcome to a mean for each identity. Gallery spread is the mean pixelwise squared difference across the ten unordered pairs among that identity's five gallery images. Gallery brightness is their mean intensity. Both predictors are computed without query photographs, then standardized across the forty galleries using population standard deviations. An ordinary least-squares regression relates identity mean gain to these two predictors; its design matrix has full column rank. We report HC3 covariance and normal-reference intervals in the tradition of White, 1980.

Table 1. Overall normalized-minus-raw correctness and exploratory identity-level associations with gallery properties.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	-0.0200 (0.0200)	[-0.0605, 0.0205]	200
M1: fit	R-squared = 0.0000	-	-	200
M2	Gallery spread, per SD	0.0343 (0.0201)	[-0.0050, 0.0737]	40
M2	Gallery brightness, per SD	0.0165 (0.0166)	[-0.0161, 0.0490]	40
M2: fit	R-squared = 0.1085	-	-	40

Model labels: M1 = Mean normalized-minus-raw identification gain; M2 = Subject-level gain versus gallery image properties.

The first model has two hundred rows and a mean-only R-squared of zero. The second has forty rows, two standardized predictors, and R-squared 0.108512. These fits answer different questions and use different covariance assumptions. A repeated-outcome model inspired by Laird and Ware, 1982 would be another modeling choice, not something already fitted. Pooling query rows while copying each gallery's covariates would also require careful dependence handling; we instead keep the property association visibly at its forty-identity level.

7. The paired table reveals how the accuracy changed

Raw-pixel accuracy is 0.875, or 175 correct queries. The normalized representation reaches 0.855, or 171 correct queries. Of the two hundred queries, 165 are correct under both rules and nineteen incorrect under both. Six switch from incorrect to correct, while ten switch from correct to incorrect. The total difference is therefore four fewer correct assignments. Reporting only the two marginal percentages would hide the sixteen query decisions on which the rules disagree.

The estimated mean gain is -0.020000, with clustered standard error 0.020000 and interval -0.060454 to 0.020454. In percentage-point language, normalization changes accuracy by -2.0 points, with an interval from approximately -6.05 to 2.05 points. The primary proposition of a positive gain receives no clear support in this realization. The interval also does not establish equivalence or prove that the true average

gain is exactly zero. A dedicated equivalence claim would need a justified margin and an appropriate design.

Across identities, four have positive net gains, six have negative net gains, and thirty have zero net change. With five queries per identity, mean gains move in coarse increments. A zero net change can include offsetting improvements and losses, so those thirty identities should not all be described as having identical predictions. Their aggregate count is descriptive, not a collection of forty individually precise subgroup estimates. We do not rank individuals by supposedly inherent identification difficulty from such a small query set.

8. Gallery properties show uncertain associations

In the forty-identity model, a one-standard-deviation increase in gallery spread has coefficient 0.034343, with standard error 0.020067 and interval -0.004987 to 0.073673. Standardized gallery brightness has coefficient 0.016466, standard error 0.016592, and interval -0.016054 to 0.048986. Both intervals include zero. Their positive point estimates are compatible with a potential association, but this dataset does not locate either relationship precisely under the chosen approximation.

Gallery properties are measured from only five photographs. The measurement-error perspective in Carroll et al., 2006 matters because their numerical spread may depend strongly on which images entered that gallery. We have no repeated gallery measurement to separate a persistent identity property from split-specific photographic variation. The coefficients do not identify effects of increasing a person's brightness or adding variability to a gallery. Such an intervention would require specifying what actually changes in the images and evaluating it directly.

Influence analysis associated with Cook, 1977 would be useful when forty rows support a small regression, but we do not remove galleries with unusual properties or gains. Multiplicity principles in Benjamini and Hochberg, 1995 caution against expanding the feature list until an apparently decisive association appears. The two stated predictors remain exploratory, and their uncertainty is reported rather than reinterpreted as evidence of no possible heterogeneity. The interpretation follows Wasserstein and Lazar, 2016 in keeping effect sizes, sample units, and design limits visible.

9. Correct identification is narrower than identity preservation

Calibration ideas in Guo et al., 2017 concern whether a score supports an appropriate confidence interpretation. Our nearest-gallery rules return an identity label; we fit no probability calibrator and report no calibrated confidence. There is no reject option for unfamiliar people, no verification threshold, and no estimate of an open-set false-match rate. A closed-set accuracy of 87.5 percent should not be carried into a setting where the correct identity may be absent from the candidate gallery.

Validation concerns in Roberts et al., 2017 and model-selection bias in Varma and Simon, 2006 also constrain further use of these numbers. If the realized gallery split were used to choose a normalization rule, a subsequent independent evaluation would be needed for the selected rule. Neither the two hundred query images nor the forty identity labels establish performance on a new dataset. The single realized partition is a transparent small experiment, not an exhaustively validated identity representation or an account of protected-group fairness.

Generated video would add questions about changing pose, temporal correspondence, non-face appearance, and consistency across frames. The present observations contain none of those outputs. A future subject-preservation evaluation could use verified identity labels and a range of suitable recognition representations while checking whether its conclusions depend on the selected proxy. This study supports

the need for that care through a simple paired example; it does not determine which modern identity metric should serve as ground truth.

10. Pair the outcome before explaining a subgroup pattern

The completed comparison shows a small uncertain average loss from this normalization rule, with both beneficial and harmful query switches. Identity-level net outcomes differ, but gallery spread and brightness do not provide precise explanatory associations in the forty available summaries. The empirical conclusion remains tied to the two simple distances, the seeded within-identity split, and the constrained historical image collection. Those boundaries preserve the usefulness of a modest result without inflating it into a general recognition claim.

Documentation under Gebru et al., 2021 records source identity, checksum verification, row-label convention, random seed, exact gallery assignments, representation transformations, and the absence of demographic annotations. Preservation according to Wilkinson et al., 2016 includes per-query predictions and gallery-only summaries, with AT&T creator credit and source-specific redistribution status retained. These records allow the accuracy differences to be reproduced without displaying the photographs as generated evidence. Further scientific and domain review is needed before using the result to justify any broader identity-evaluation procedure.

References

1. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.
2. Deng, Jiankang, et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition" *arXiv preprint arXiv:1801.07698*, 2018. https://arxiv.org/abs/1801.07698.
3. Guo, Zinan, et al. "PuLID: Pure and Lightning ID Customization via Contrastive Alignment." *arXiv preprint arXiv:2404.16022*, 2024. https://arxiv.org/abs/2404.16022.
4. Laird, Nan M. and James H. Ware. “Random-Effects Models for Longitudinal Data.” *Biometrics*, vol. 38, no. 4, 1982, pp. 963. https://doi.org/10.2307/2529876.
5. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.
6. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.
7. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.
8. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.
9. Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
10. Belhumeur, P.N.; Hespanha, J.P.; Kriegman, D.J.. (1997). Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/34.598228.

11. He, Xuanhua; Liu, Quande; Qian, Shengju; Wang, Xin; Hu, Tao; Cao, Ke; Yan, Keyu; Zhang, Jie. (2024). ID-Animator: Zero-Shot Identity-Preserving Human Video Generation. arXiv:2404.15275. https://arxiv.org/abs/2404.15275.

12. AT&T Laboratories Cambridge. (2026). Modified Olivetti faces: Sam Roweis MATLAB version [Data archive; mirror accessed 7 November 2026]. File identity and row-label layout verified against the scikit-learn maintainer loader. https://ndownloader.figshare.com/files/5976027.

13. Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

14. Huang, Gary B., Marwan Mattar, Tamara Berg, and Erik Learned-Miller. "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments." *ECCV Workshop on Faces in Real-Life Images*, 2008. https://cs.umass.edu/~elm/papers/Huang_eccv2008-lfw.pdf.

15. Meng, Qiang, et al. "MagFace: A Universal Representation for Face Recognition and Quality Assessment." *arXiv preprint arXiv:2103.06627*, 2021. https://arxiv.org/abs/2103.06627.

16. Wang, Hao, et al. "CosFace: Large Margin Cosine Loss for Deep Face Recognition" *arXiv preprint arXiv:1801.09414*, 2018. https://arxiv.org/abs/1801.09414.

17. Kim, Minchul, et al. "AdaFace: Quality Adaptive Margin for Face Recognition." *arXiv preprint arXiv:2204.00964*, 2022. https://arxiv.org/abs/2204.00964.

18. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

19. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

20. Kärkkäinen, Kimmo and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age" *arXiv preprint arXiv:1908.04913*, 2019. https://arxiv.org/abs/1908.04913.

21. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

22. Samaria, F.S.; Harter, A.C.. (1994). Parameterisation of a stochastic model for human face identification. Proceedings of 1994 IEEE Workshop on Applications of Computer Vision. https://doi.org/10.1109/acv.1994.341300.

23. Yuan, Shenghai; Huang, Jinfa; He, Xianyi; Ge, Yunyuan; Shi, Yujun; Chen, Liuhan; Luo, Jiebo; Yuan, Li. (2024). Identity-Preserving Text-to-Video Generation by Frequency Decomposition. arXiv:2411.17440. https://arxiv.org/abs/2411.17440.

24. Turk, M.A.; Pentland, A.P.. (1991). Face recognition using eigenfaces. Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.1991.139758.

25. Mitchell, Tom. (1997). CMU Face Images [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5JC79.

26. Xiaoyang Tan; Triggs, Bill. (2010). Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions. IEEE Transactions on Image Processing. https://doi.org/10.1109/tip.2010.2042645.

27. Meng, Zijie; Liu, Jiwen; Liu, Yufei; Tong, Chengzhuo; Liu, Xiaoqiang; Zhang, Yuanxing; Xu, Yulong; Wan, Pengfei. (2026). ARGUS: Stacked Multi-View Identity Mosaic Injection for Subject-Preserving Video Generation. arXiv:2606.11670. https://arxiv.org/abs/2606.11670.

28. Cao, Qiong, et al. "VGGFace2: A dataset for recognising faces across pose and age" *arXiv preprint arXiv:1710.08092*, 2017. https://arxiv.org/abs/1710.08092.

29. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

30. Schroff, Florian, et al. "FaceNet: A Unified Embedding for Face Recognition and Clustering" *arXiv preprint arXiv:1503.03832*, 2015. https://arxiv.org/abs/1503.03832.