Partitioning natural face image variability emphasises within-identity over between-identity representation for understanding accurate recognition

Supplementary Materials

David White*, Tanya Wayne & Victor Varela

*Corresponding author: david.white@unsw.edu.au

Supporting methodology for General Methods: List of stimulus identities

US celebrities: Al Pacino, Bill Clinton, Brad Pitt, Cameron Diaz, Gwyneth Paltrow, Harrison Ford, Jack Nicholson, Jennifer Aniston, Jennifer Lopez, John Travolta, Julia Roberts, Leonardo DiCaprio, Madonna, Marilyn Monroe, Meg Ryan, Princess Diana, Russell Crowe, Sarah Jessica Parker, Sylvester Stallone, Tom Cruise.

Dutch celebrities: Andre van Duin, Bridget Maasland, Caroline Tensen, Chantal Janzen, Danny de Munk, Frans Bauer, Froukje de Both, Gerard Joling, Gordon, Henkjan Smit, Irene Moors, Jamai, Jantje Smit, Jim Bakkum, Karin Bloemen, Katja, Marco Borsato, Patty Brard, Tatjana Simic, Wendy van Dijk.

Supporting methodology for Experiment 4: Super-recognizers screening test accuracy

Accuracy data in Table S1 show substantial and stable superiority of super-recognisers relative to normative test performance, across repeated tests of face identification with diverse test conditions and procedures. The two experimental groups show roughly equivalent levels of accuracy on all tests.

In addition to the main screening tests reported in the main text, super-recognisers had also previously completed the Models Face Matching Test (Robertson et al., 2016) and a fingerprint matching test (from Thompson & Tangen, 2014), which were presented in the same pairwise matching format as the Glasgow Face Matching Test. These tests were not used to select participants and so provided a way to verify the ability of our super-recognisers in perceptual discrimination tasks using faces (models test) and a novel stimulus class (fingerprints). Although both groups show numerically higher accuracy in fingerprint

matching relative to normative performance, this was within 1SD of the normative mean for both groups, whereas mean accuracy on the Models Face Matching Test was more than 1.5SD above the mean for both groups, confirming that at least part of the expertise of our super-recogniser group in perceptual discrimination was specific to face images.

	UNSW Face Test	*GFMT*	*CFMT+*	*Models*	*Fingerprint*
Between-identity variation SR raters (n = 23)	76.4 (4.8)	100 (0)	94.3 (3.4)	91.6 (4.7)	82.3 (10.2)
Within-identity variation SR raters (n = 19)	75.4 (3.9)	100 (0)	93.4 (3.7)	90.6 (5.2)	86.6 (7.5)
Normative test scores	58.9 (5.8)	81.3 (9.7)	69.6 (11.5)	73.6 (10.9)	77.0 (9.0)

Table S1. Mean scores on tests used to identify super-recognisers in Experiment 4 (standard deviations in parenthesis). Normative test scores for face identification tests are from previously published work: UNSW Face Test (Dunn et al., 2020; n = 302), Glasgow Face Matching Test (GFMT, Burton et al., 2010; n = 194); Cambridge Face Matching Test extended version (CFMT+, Russell et al., 2009; n = 25); Models Face Matching Test (Robertson et al., 2016; n = 54); Fingerprint Matching Test (Towler et al., submitted; n = 584). Details of procedure for selecting super-recognisers is provided in the main text.

Supporting methodological details for section 'Computational study: Image similarity in low-level statistics and DCNN'

First, we measured low-level visual properties of image statistics using a GIST descriptor (Oliva & Torralba, 2001) that is designed to extract global properties of images. For each image, a vector of 512 values was obtained by passing the image through a series of Gabor filters across eight orientations and four spatial frequencies, and windowing the filtered images along a 4 × 4 grid. Each vector represents the image in terms of the spatial frequencies and orientations present at different positions across the image. We then computed cosine distances between the vectors representing each image in a pair. All cosine

scores were then normalized and converted to similarity scores (i.e. 1 minus distance) for consistency with human rating data. Similarity scores for the GIST descriptor analysis are shown in Figure 5 of the main manuscript.

Second, we measured image statistics at progressively higher levels of abstraction: from raw pixel information in an image, through to highly abstracted features that are optimised to recognise the identity of faces, using an open source Deep Convolutional Neural Network (DCNN) face recognition algorithm (VGGFace: Parkhi, Vedaldi, & Zisserman 2015). Because the DCNN was not trained on any of the celebrity faces in our stimulus set, we were able to measure representation of between- and within-identity variation in a model that was unfamiliar with all the faces used in our experiments.

We computed similarity of all stimulus comparisons made by humans, separately from vectors of neuron activation at each layer of this network. All images in our dataset were first cropped and resized to 224x224 pixels automatically using a face detection algorithm (Multi-task Cascaded CNN, MTCNN: Zhang, Zhang, Li, & Qiao, 2016) so that they conformed to input conventions for the recognition algorithm. Similarity scores were computed as cosine distances between feature vectors at each layer (for convolutional layers with multiple feature vectors nested in cells, this was computed as the average distance across all the cells for that layer). All cosine scores were normalized and converted to similarity scores (i.e. 1 minus distance) for consistency with the human rating data. Average similarity scores for between and within-identity variation comparisons at each layer of the DCNN are shown in Figure 5 of the main manuscript.

Supplementary analysis of typicality ratings

In Footnote 3 of the main manuscript we describe an analysis that aimed to address a query raised by a reviewer. The reviewer queried whether identity averages created by our image averaging technique are representative of the central tendencies – 'centroids' – of person-specific subspaces in facespace. To address this query, we analysed the relationship between similarity ratings made between individual face images and the identity averages in Experiments 2 and 3, and typicality ratings to individual images that were made by a new set of participants.

We recruited 475 participants (296 female; M_{age} = 33.8, SD_{age} = 11.1) via Amazon Mechanical Turk and asked them to rate how typicality each individual image was of the persons appearance. Each participant rated typicality of one exemplar image for each of the 20 US celebrities identities using a Likert scale (1 = 'Very Atypical'; 9 = 'Very Typical'), giving an average of 24 typicality ratings for each of the 200 images in the database.

We then correlated the typicality ratings for each individual image with the average similarity rating between that image and the identity average, computed from all participant responses in Experiments 2 and 3. We found a high correlation between typicality and image-to-average similarity, $r(400) = 0.685$, $p < 0.001$, suggesting that image averages are relatively good approximations of the centroids of person-specific subspaces. Data from this analysis are shown in Figure S1 and full item data are provided in supporting data (White, Wayne & Varela, 2021).

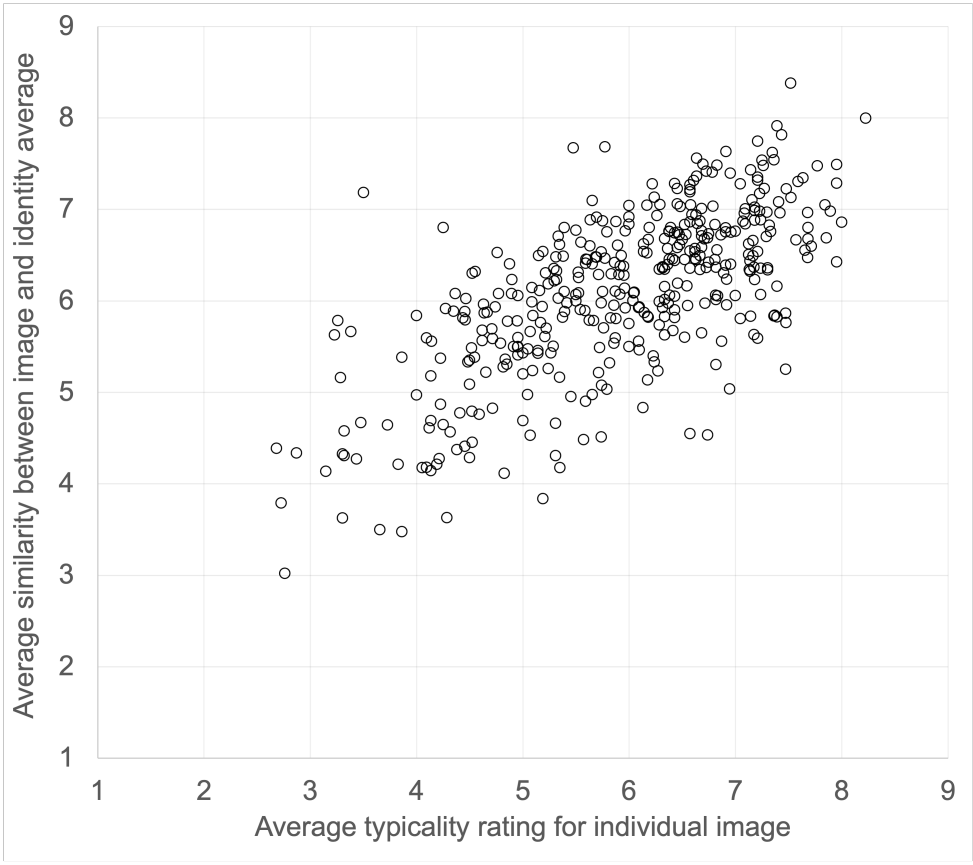

Figure S1. Correlation between typicality ratings made to 200 individual images of US celebrities, and average similarity rating between those images and their respective identity average (computed from all participant responses in Experiments 2 and 3).

References not included in main manuscript

Thompson, M. B., & Tangen, J. M. (2014). The nature of expertise in fingerprint matching: experts can do a lot with a little. *PloS one*, *9*(12), e114759.

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by metropolitan police super-recognisers. PloS one, 11(2), e0150036.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503.