Facial identity across the lifespan

Mila Mileva, Andrew W. Young, Rob Jenkins, & A. Mike Burton

Department of Psychology, University of York, UK

Correspondence to:

Mila Mileva

Department of Psychology

University of York

Heslington, York

YO10 5DD, UK

mila.mileva@york.ac.uk

Funding: This work was supported by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.323262 to A. Mike Burton.

Abstract

We can recognise people that we know across their lifespan. We see family members age, and we can recognise celebrities across long careers. How is this possible, despite the very large facial changes that occur as people get older? Here we analyse the statistical properties of faces as they age, sampling photos of the same people from their 20s to their 70s. Across a number of simulations, we observe that individuals' faces retain some idiosyncratic physical properties across the adult lifespan that can be used to support moderate levels of age-independent recognition. However, we found that models based exclusively on image-similarity only achieved limited success in recognising faces across age. In contrast, more robust recognition was achieved with the introduction of a minimal top-down familiarisation procedure. Such models can incorporate the within-person variability associated with a particular individual to show a surprisingly high level of generalisation, even across the lifespan. The analysis of this variability reveals a powerful statistical tool for understanding recognition, and demonstrates how visual representations may support operations typically thought to require conceptual properties.

Keywords: face perception; face recognition; age perception

Facial identity across the lifespan

1. Introduction: Recognising people at different points in their lives

Paul McCartney has been famous for nearly sixty years, at the time of writing. As a highly-photographed celebrity, his image appears over a huge range of viewing conditions. Some aspects of his appearance are transient from moment to moment (e.g. pose, expression, lighting), some involve deliberate changes in self-presentation (such as hairstyle and moustache) and some are much slower to change (e.g. effects of health and ageing). Despite this, many viewers familiar with Sir Paul can recognise him accurately and consistently across this extensive range (see Fig. 1). In this study, we focus on changes due to time: how can we recognise someone's face over a sixty-year period? Are multiple visual representations necessary ('young McCartney', 'middle-aged McCartney' etc.) or is there sufficient commonality across all his images to support recognition? Indeed, is this problem solvable for human viewers using only visual processes? Can we create a common representation capable of recognising all of the images in Fig. 1, or must we, of necessity, recognise photos from the top and bottom rows of Fig. 1 using independent visual representations, supported by a common network of semantic knowledge and experience? What would be the characteristics of a visual representation capable of incorporating this level of age-related variability?

Familiar face recognition has two components, 'telling people apart', i.e. discriminating one identity from another, and 'telling people together', i.e. cohering multiple images of the same person (Andrews, Jenkins, Cursiter, & Burton, 2015; Burton, 2013; Jenkins, White, Van Montfort, & Burton, 2011). In fact, people are generally very accurate at recognising familiar faces, and relatively unaffected by superficial image variability due to pose and viewing

conditions (Bruce, Henderson, Newman, & Burton, 2001; Burton, Wilson, Cowan, & Bruce, 1999). In contrast, viewers are much less accurate in identity tasks with unfamiliar faces, for example finding it hard to match different images of the same person (e.g. Bruce et al., 1999, 2001; Megreya & Burton, 2006, 2008), and are severely affected by superficial image changes (e.g. Bruce, 1982; Hill & Bruce, 1996; O'Toole, Edelman, & Bülthoff, 1998; Patterson & Baddeley, 1977).

Fig. 1. Images of Sir Paul McCartney across the three time periods used throughout this paper. Images on the top row were taken in the 1960s-70s, images in the middle row were

taken in the 1980s-90s and images in the bottom row were taken in the 2000s-10s. See image attributions in the Acknowledgements section.

We have argued that a key part of face learning, by which a face becomes familiar, is the acquisition of knowledge about how that particular face varies (Burton, 2013; Jenkins et al., 2011; Young & Burton, 2018). Examining the physical properties of multiple images of the same person reveals that individuals *vary in idiosyncratic ways* (Burton, Kramer, Ritchie, & Jenkins, 2016). In consequence, learning to recognise a familiar face must involve learning about the identity-specific ways in which that particular face can change in appearance across different encounters. However, this variability has so far been studied only over relatively limited timescales. Here we ask whether a mechanism for learning face representations over variability could cope with the large changes in appearance across a whole adult lifespan.

Of course, there are general properties of face transformations due to ageing, some of which are well-understood at a population level. For example, the effects of age on 3d face shape, pigmentation and texture can be captured, on average, by known physical manipulations (Burt & Perrett, 1995; Geng, Zhou, & Smith-Miles, 2007; Pittenger & Shaw, 1975). However, these general transformations are of limited use in capturing the ageing process for a particular individual. Instead, lifestyle choices (e.g. diet), biological factors (e.g. proneness to balding) and self-presentation choices (e.g. make-up, facial hair), mean that one cannot reliably predict how a particular person's appearance will change over time. Despite this, previous multidimensional scaling studies consistently identify age as one of the key dimensions for face recognition (Johnston et al., 1997; Shepherd, Ellis, & Davies, 1977) and there is converging evidence for significant impairments to recognition accuracy with the introduction of age as a factor in face recognition, face matching and live person

identification (Davis & Valentine, 2009; George & Hole, 1998; Megreya, Sandford, & Burton, 2013. However, all of these studies use relatively small age ranges (often across months) and thus tell us rather little about how we recognise identities across their whole lifespan.

From the perspective of the perceiver, there are several ways in which one might build a representation of a face over time. For a personally-familiar individual, one can make incremental updates on each new encounter, observing the ageing process as time passes. For a celebrity, e.g. a film star, one might track a person as they age, but nevertheless be exposed to frequent encounters with earlier manifestations of the face, e.g. through watching films from different time periods. For a young viewer, who has not aged alongside Paul McCartney, it may be that one encounters images of him in a relatively arbitrary sequence across his age. Any mechanism for representing a face over the lifespan must be able to incorporate all of these differing types of incremental development.

In fact, surprisingly little is known about face learning, and still less is known about long-term representational updating. Early studies showed standard effects of exposure on face learning, such that increasing number and duration of exposures leads to better subsequent recognition (MacLin, MacLin, & Malpass, 2001; Reynolds & Pezdek, 1992). More recently, it has become clear that the quality of exposure is a critical component in face learning: to develop a generalisable representation, capable of supporting recognition of novel encounters, a large *range* of experience is necessary (Dunn, Kemp, & White, 2018; Matthews & Mondloch, 2018; Murphy, Ipser, Gaigg, & Cook, 2015; Ritchie & Burton, 2017). The use of highly variable, naturally occurring ('ambient') images shows that exposure to variability between images of the same face supports generalisable learning that can lead to accurate

recognition of new (previously unseen) images of a particular face. Existing studies demonstrating this point use images with high variance in camera, lighting, expression and personal presentation (hairstyle etc). However, they do not incorporate the level of variability encountered when one recognises a person over a sixty year adult lifespan, as with our example of Paul McCartney.

1.1 Bottom-up and top-down contributions to recognition

Research in face identification generally focusses on the visual aspects of recognition: how can a visual input be mapped onto a stored representation? However, in real-world encounters, recognition is typically supported by many additional sources of information. As we talk to colleagues at work we have rather diverse evidence to support their identification, such as their voices, the content of the conversation and so forth. Furthermore, although the presentations of their faces will vary mid-conversation, through pose, expression or lighting change, we do not need to make a new visual match to check their identity (Young, 2018). In media exposure to famous people we also often have a great deal of top-down identification support, for example as we watch films whose cast we know or watch a newscaster introducing an item with 'In Washington today, the President said ….'.

What is the effect of top-down support for face recognition? We know, of course, that recognition of a well-known face can take place without it. A great body of psychological research relies on studies in which viewers see a sequence of familiar faces, and have to discriminate between these and unfamiliar faces. Despite the lack of any warning about who will appear in the experiment, this is generally not too challenging a task if the familiar faces are sufficiently well-known (typically media celebrities). Nevertheless, we also know that top-down support can bring about significant changes in face recognition. For example, many

studies show that familiarity decisions are faster when faces are preceded by those of related people, the phenomenon of semantic priming (Bruce & Valentine, 1986; Schweinberger, 1996; Young, Flude, Hellawell, & Ellis, 1994). There are also well-known effects of context in face recognition, such that faces of moderate familiarity are easier to recognise when they arise in an expected context than when they occur in unexpected places (e.g. Hanczakowski, Zawadzka, & Macken, 2015; Watkins, Ho, & Tulving, 1976; Young, Hay, & Ellis, 1985). While these effects are widely-acknowledged, the literature typically does not address the problem of how to disentangle the contributions of visual and non-visual processes to recognition.

In the studies below, we examine face recognition supported by bottom-up and top-down processes. Taking the difficult problem of recognising someone over their adult lifespan, we ask how much can be achieved by representations based on image-similarity alone, and how much (if anything) is added by incorporating a top-down component of guided recognition at the face learning stage. If faces retain some level of physical characteristic that is preserved across ageing, then physical analysis of multiple images may in itself deliver a robust representation, capable of generalisation to novel instances. On the other hand, if a contribution from top-down processes is needed for this task, then statistics based on image descriptions alone may not be enough to solve this difficult problem.

To operationalise bottom-up image-similarity and top-down guided recognition, we use established statistical descriptions of faces. We measure image-similarity in a space derived from principal components analysis (PCA) of faces. This is a very simple, linear description of the factors underlying variation in face pictures, and is common in psychological, neurophysiological and automatic recognition research (Chang & Tsao, 2017; O'Toole, Abdi,

Deffenbacher, & Valentin, 1993; Sirovich & Kirby, 1987; Turk & Pentland, 1991). To model guided recognition, we use linear-discriminant function analysis (LDA). This is a common top-down clustering method which captures the covariance structures within and between different stimuli classes. It extracts the most efficient features used to describe a class of stimuli (in our case – different images of the same person) which can also best serve to discriminate it from members of another class (different images of another person). This is done by minimising within-group differences (here, the differences between images of the same face) while maximising separation between groups (i.e. between different faces). Once we have used LDA to establish these optimal dimensions (or features), we can then test this model by using it to classify novel images of the same or new identities. This allows us to determine whether the identity clusters from LDA produce a covariance structure that could potentially be useful to human observers. The technique has been used in previous models of face perception (Bekios-Calfa, Buenaposada, & Baumela, 2011; Kramer, Young, & Burton, 2018), and provides a simple but effective way of establishing between-person differences.

With these statistical tools, we examine images taken across a very wide age period, as illustrated in Fig. 1. We use photos that are deliberately unstandardized to represent the types of pictures people recognise day to day; so-called ambient images (Burton, Jenkins, & Schweinberger, 2011; Sutherland et al., 2013). Across simulations, we ask: to what extent is it possible to 'recognise' novel photos of someone based solely on the physical similarity between images? We then investigate to what extent top-down processes help – i.e. when models are built to cohere disparate images of the same person (using a combination of PCA and LDA), does this help to recognise new pictures, beyond simple physical image similarity?

To summarise, our interest in age here is two-fold. First, since previous models have not typically incorporated this level of variability, we are interested to establish the limits of simple statistical classifications across this range. Second, we wish to establish whether there is a core characteristic of a face which transcends age: to what extent does learning a 20-year old face allow one to recognise the person at 60? The following simulations address these issues with the faces of celebrities for whom photos exist over this long period.

2. Method

2.1 Image Set

We collected over 200 images of each of 10 people spanning the period from 1960 to 2018. As we intended to compare variability within and between individuals, we took the decision not to introduce variability due to gender or ethnicity. The ten individuals were all white (Caucasian), male musicians, active from around the mid-1950s to the present time.

Images for each identity were collected for 3 separate time periods: 1960s-70s, 1980s-90s and 2000s-2010s. The set (referred to as the 'lifespan set' from now on) contained a total of 2100 images – 70 images for each identity in each time period. The images were selected from the online archive Getty Images which contains over 200 million assets. They varied greatly in lighting conditions, pose, hairstyle, emotional expression, facial hair and image quality. We used only images in which all the internal features were visible (i.e. within about +/- 30° from full face) and not obscured by clothing or accessories (see Fig. 1). For the purpose of computational analysis all images were represented in grayscale using a lossless image format (bitmap) and resized to 190 x 285 pixels.

2.2 General Procedure

In the following simulations, we compared multidimensional models derived from: (i) image statistics (using PCA) and (ii) identity clustering (using PCA+LDA). In each case, a training set of images was used to build a space, while a different subset was retained for later test. After training, test images were projected into these PCA or PCA+LDA spaces, and proximity to training images was measured using simple Euclidean distance. Test images were regarded as correctly classified if the nearest training image in the space was a photo of the same identity.

For PCA, we followed the general procedure adopted in previous studies (Burton et al., 2016; Kramer et al., 2018; Kramer, Young, Day, & Burton, 2017; see Smith, 2002 for a detailed tutorial) by using normalised face images, computed by aligning 82 fiducial points on each image (corners of the eyes, tip of the nose, etc. – see Burton et al., 2016 for details). The location of these fiducial points was determined using a standard semi-automatic process requiring five manually-aligned landmarks (see Kramer et al., 2017 for details). All images were then shape-standardised by morphing them to the average shape of the entire face set (Burton, Miller, Bruce, Hancock, & Henderson, 2001; Craw, 1995). PCA was applied to these shape-normalised textures (see Kramer, Jenkins, & Burton, 2017 for appropriate software). Only the first N components explaining 95% of the variance in the image pixel intensity information were retained for subsequent analysis (N is reported for each specific model below). LDA was applied to the image projections in PCA space, such that images of the same person were clustered together, minimising intra-class and maximising inter-class differences. As with PCA, we retained derived dimensions which accounted for 95% of the

input variance. A more detailed description of the LDA procedure can be found in Kramer et al., (2017; 2018; see also Tharwat et al., 2017 for a comprehensive tutorial on LDA).

3. Simulations

3.1 Simulation 1: Recognition in image-based and identity-clustered spaces

In this first simulation, we examine face classification of images across the whole age range, building spaces based on photos sampled from the entire set. Our aim is to establish whether face classification is possible in the context of such large within-person variability, and whether the addition of top-down identity-clustering will enhance classification accuracy.

3.1.1 Simulation 1a: Learning time-invariant identities

We derived spaces from PCA and from PCA+LDA, as described above. From the lifespan set (total of 2100 images) we created training sets (1800 images) comprising 180 images per identity (60 images in each time period) and test sets (300 images) comprising 30 images per identity (10 images in each time period). This procedure was repeated 7 times with counterbalancing, such that any particular image was included 6 times in training sets and once in a test set. So, in each model, images in the test set were always novel compared to those in the corresponding training set.

The procedure for construction and test of PCA and PCA+LDA models is summarised in Fig. 2. In these simulations, we used a PCA space built on 178 components, the number needed to account for 95% of the image variance. Test items were projected into this space, and the closest training item was established using Euclidean proximity. For the PCA+LDA classification, we applied LDA to the PCA projections of the images in the training set, then

projected the test images in the newly-created LDA space, and, again, computed the Euclidean distance between each test and training image. For both the PCA Only and PCA+LDA models we calculated the Euclidean distance between each image in the test and training sets and then identified the closest image from the training set to each of the images in the test set. In each case, model accuracy was estimated as the average classification success (% of correct responses) across all 7 iterations. As there are 10 identities represented here, chance accuracy is 10%.

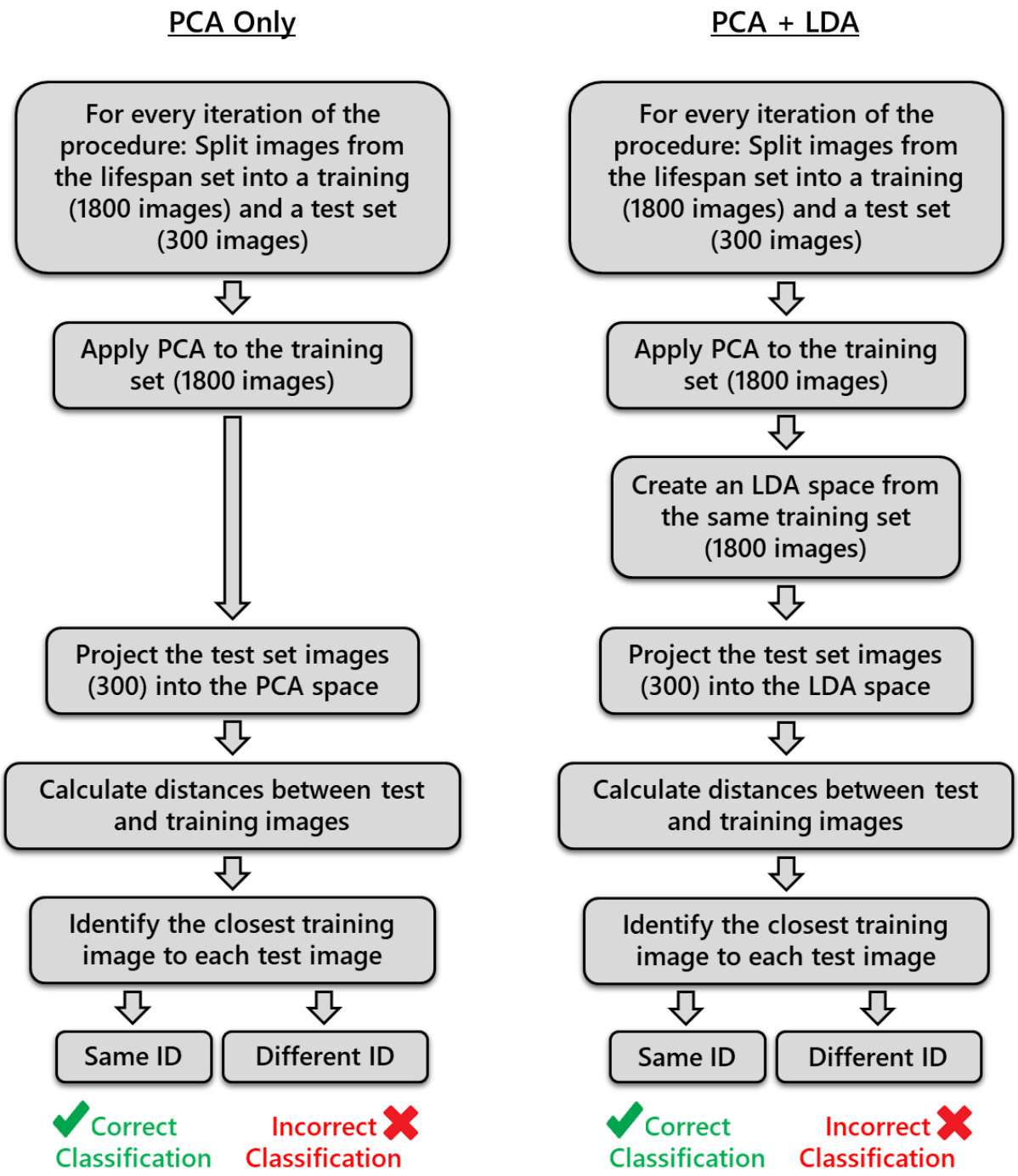

Fig. 2. Classification procedure for the PCA (left) and PCA+LDA (right) models in Simulation 1. First, we split the images from the lifespan set into seven different training (1800 images) and test (300 images) sets and we use the same sets for both the PCA Only and the PCA+LDA models. Both sets contain images from all three time periods. For every iteration of the procedure in both models, we then perform PCA on the training set. For the

PCA Only model, we project the images from the test set into the newly created PCA space, then for each image in the test set, we calculate its distance to every image in the training set and determine the identity of the closest image. For the PCA+LDA model, we perform LDA on the training set and project all images from the test set into the newly created LDA space. Again, we calculate the distance between each test and each training image and determine the identity of the closest image.

3.1.1.1 Simulation 1a: Results and Discussion

Fig. 3 shows the average classification accuracy for novel test images from the PCA-only and PCA+LDA models against chance performance. Both PCA and PCA+LDA perform considerably above chance levels, with the identity-clustered spaces created by LDA providing a substantial boost to performance. (Details of individual runs are given in Supplementary Materials, Table S1; each iteration shows the same pattern.)

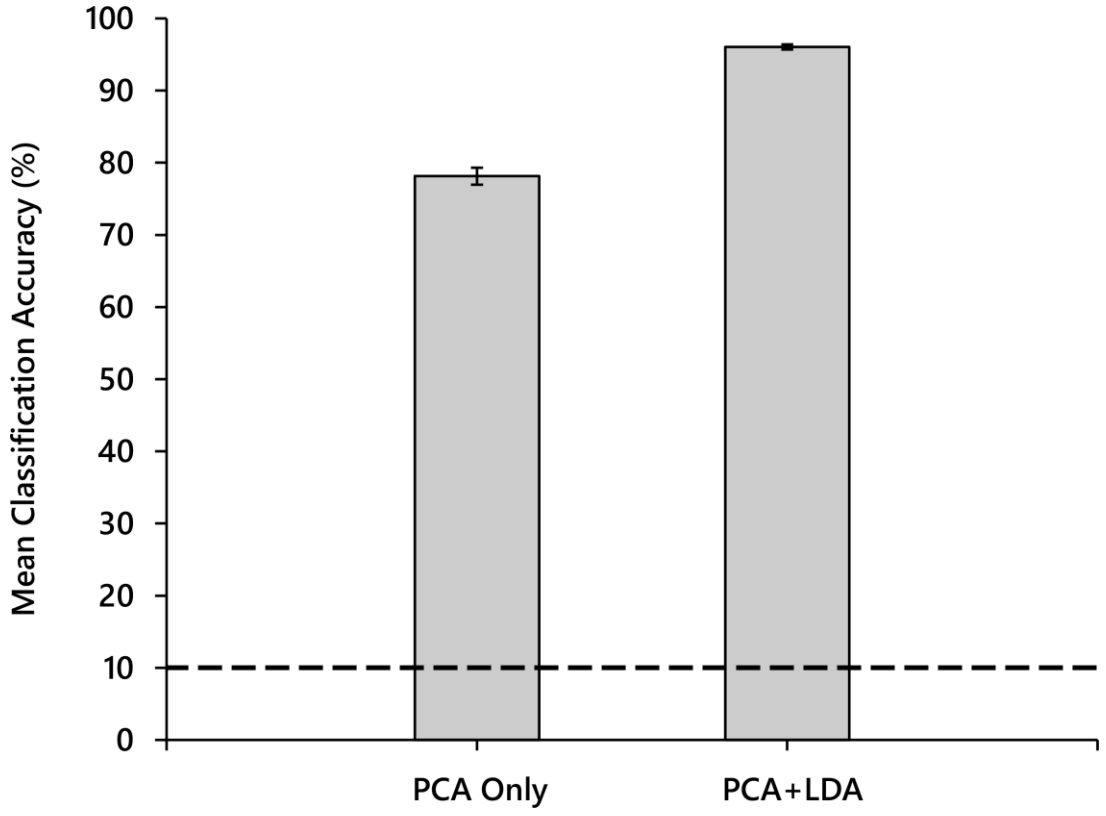

Fig. 3. Mean classification accuracy for novel test images for both the PCA and PCA+LDA approaches. Dashed line represents chance performance. Error bars show standard error.

The performance of the PCA-only model is perhaps surprising. Using simple image-descriptions, the model performs rather well (around 80% accuracy) when identifying people over a very large time period. So, despite images showing people ranging from their 20s to their 70s, and despite all the people sharing similar demographics (white male musicians) there is enough commonality between superficial image-descriptions to support a reasonably good level of classification accuracy, well above chance. This suggests that there may be something in common between images of Paul McCartney (say) across his entire lifespan, which can be used for recognition. Of course, the 'training' and test images cover the same time periods, and so this result may rely on within-period similarity (we test generalisability between periods in simulations below). However, in this setting, the addition of top-down information, in the form of identity-clustering, gives a considerable added advantage, taking the PCA+LDA model to near perfect performance. In other words, a simple image-recognition method can benefit considerably from incorporating, at learning, the information that particular photos (for example, Paul McCartney at different ages) all belong to the same person.

3.1.2 Simulation 1b: Learning time-specific identities

The simulation above seems to suggest that fairly simple statistical approaches to face recognition are robust against extensive within-person variability, incorporating decades of change. However, we would also like to establish whether there is any advantage to be gained by forming more specific representations for recognition. For example, would it be

efficient to store several representations of Paul McCartney, such that a Beatles-era Paul would be recognised somewhat independently of a modern-era Paul? In order to do this, we repeated the simulation above, but this time treating each time-period independently. So, we now seek to make a 30-way classification, treating each of three time periods of the 10 identities as representing a 'different person'. Exactly the same training and test image sets were used as in the previous simulation.

3.1.2.1 Simulation 1b: Results and Discussion

Fig. 4 shows the average classification accuracy for novel test images for both the PCA and PCA+LDA models. As with the 10-way classification, we see that both PCA and PCA+LDA perform substantially above chance. Overall performance is lower than in Simulation 1a, consistent with the fact that there are now 30 rather than 10 classification options. Nevertheless, there is a high level of classification accuracy, and this is boosted considerably by identity-clustering.

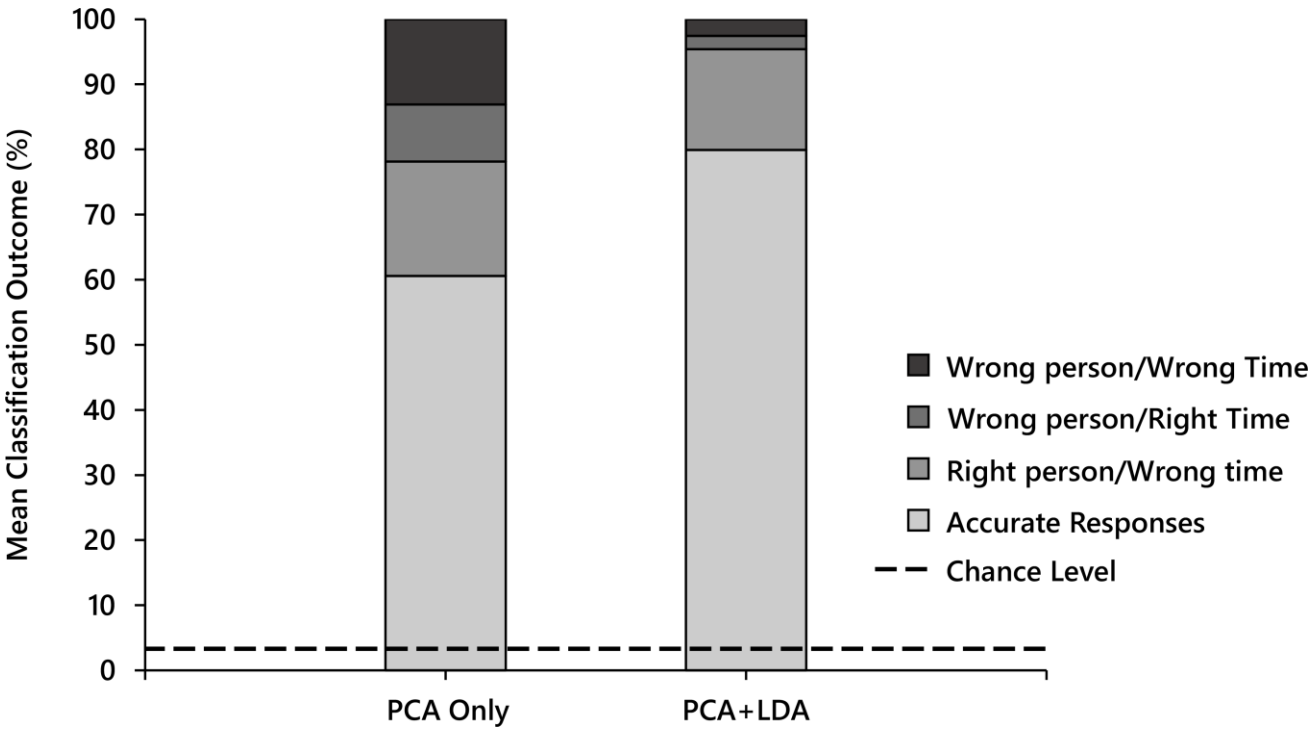

Fig. 4. Mean classification accuracy and errors for both the PCA and PCA+LDA approach using a 30-way classification. Dashed line represents chance performance.

The analysis of errors in this simulation is also instructive. Imagine a test image Paul McCartney from the 1960s which is not correctly recognised. Is it more likely that the error will arise because somebody else's photo from the 1960s is chosen instead, or because another photo of Paul McCartney, but from a later period, is chosen? Either is plausible here. Simulation 1a suggests quite high commonality across images of the same person over time, and so we might predict errors in this situation to reflect 'Right person/Wrong time'. On the other hand, we might expect different people of the same era to look more alike than the same person in different eras. For example, many of these musicians had fashionably copious hair in the 1960s, whereas several of them had no hair at all in later photos. It should also be noted that pictures taken in the 1960s have some low-level characteristics in common, which differ from modern photos, due to the substantial changes in camera technology across this time period. If recognition is highly influenced by these characteristics, then we might expect a predominance of 'Wrong person/Right time' errors.

Fig. 4 shows that in the PCA+LDA model, almost all errors are 'Right person/Wrong time', retaining the identity of the test image. This is interesting, because the clustering in each case was only made within a single time period (i.e. there were 30 categories derived from learning items). Nevertheless, incorporating the within-person variability in a single time period appears to be enough to derive identity representations which can be useful across the entire adult lifespan – leading to almost perfect recognition of identity. In contrast, errors deriving from the PCA model, based solely on image properties, are more evenly distributed across the different options. This pattern of results was the same in all seven iterations of the

procedure (see Table S2 in the Supplementary Materials for further information). Here we see a suggestion of a very useful property of within-person clustering. It appears that mapping different images onto the same representation, even within a restricted time period, results in a robust representation which can be recruited outside that time period – but this requires more than the simple image-similarity offered by PCA. Top-down information during encoding, as used in LDA, results in a representation that is highly robust with respect to identity.

3.2 Simulation 2: Recognition in the context of a large prior exposure to faces

Our first simulation demonstrates that while access to physical variability alone is sufficient to support a degree of accurate classification across time, the addition of top-down identity information during learning increases performance substantially. However, this simulation, like many previous face recognition models, fails to incorporate our vast background knowledge of faces and the way they change with time. In order to address this, we next compare PCA and PCA+LDA models in spaces derived from a much larger database of faces, incorporating many more individuals than those we are explicitly testing. These extra faces vary in all the ways typical of ambient face images, including age. We hypothesise that abstraction over a background variability which includes 'ambient' age ranges may support the development of age-independent representations of the faces in our training and test sets.

3.2.1 Simulation 2: Background set images

For the next simulations, we collected a large set of 6100 images, which we label the 'background set'. The set comprised 434 identities (243 males) who were mostly celebrities such as actors, musicians, comedians and TV presenters. Critically, it contained no identities from the lifespan set (used in Simulation 1). To mimic the wide range of exposure to faces of

differing levels of familiarity in everyday life, the background set included multiple images for 263 identities while the remaining 171 identities were represented with a single image only (cf. Kramer et al., 2018). For people represented by multiple images, these ranged from 2 to 170 photos per person, with a mean of 23. All images were 'ambient' and varied in pose, lighting conditions, emotional expressions, age and image quality.

Fig. 5. Examples of the types of images used in the background set. The first and second pair of images on the bottom row depict the same identities. They illustrate sampling images of the same identity across time. See image attributions in the Acknowledgements section.

As we are interested in face recognition across time, the background set included images of people aged between 18 and 88 years. In order to ensure a broad coverage of ages, we included at least 50 identities in each of six pre-specified age groups (20-29, 30-39, 40-49, 50-59, 60-69 and 70-79). Each age group included an average of 10 images per identity. We also tried to achieve broad coverage of image differences across time by including photos of similar-age identities from different decades. For example, the set contained images of both the English actor Daniel Radcliffe who reached 29 in 2008 and the English singer-songwriter Marc Bolan who reached 29 in 1976. Furthermore, for identities where this was possible (e.g., Tom Jones, Cher, etc.), we collected images spanning different decades, similar to those shown in Fig. 1. The majority of identities were White (Caucasian), although other ethnicities were also represented. Images were resized to 190 x 285 pixels and represented in grayscale, in a lossless image format (bitmap). The images in this set were landmarked in exactly the same way as these in the lifespan set. Fig. 5 shows examples of the types of images used in the background set.

3.2.2 Simulation 2a: Learning time-invariant identities within an established face space

PCA was first performed on the background set (6100 images). We followed the same procedure as in the first simulation – PCA was performed on shape-normalised images and we used the first 258 PCs which explained 95% of the variance. The images from the lifespan set were then projected into this PCA space. We derived classification ('recognition') data from the same 7 training and test sets as used in Simulation 1. We next ran LDA on the PCA signatures of the training set and projected images from the test set in this newly-created LDA face space. The identity of the test images was classified by calculating the distance between each test and each training image and determining the

identity of the closest image. Fig. 6 shows a simplified flow chart of the procedure. Again, the model had 10 identities to choose from for classification, meaning that chance level was 10%.

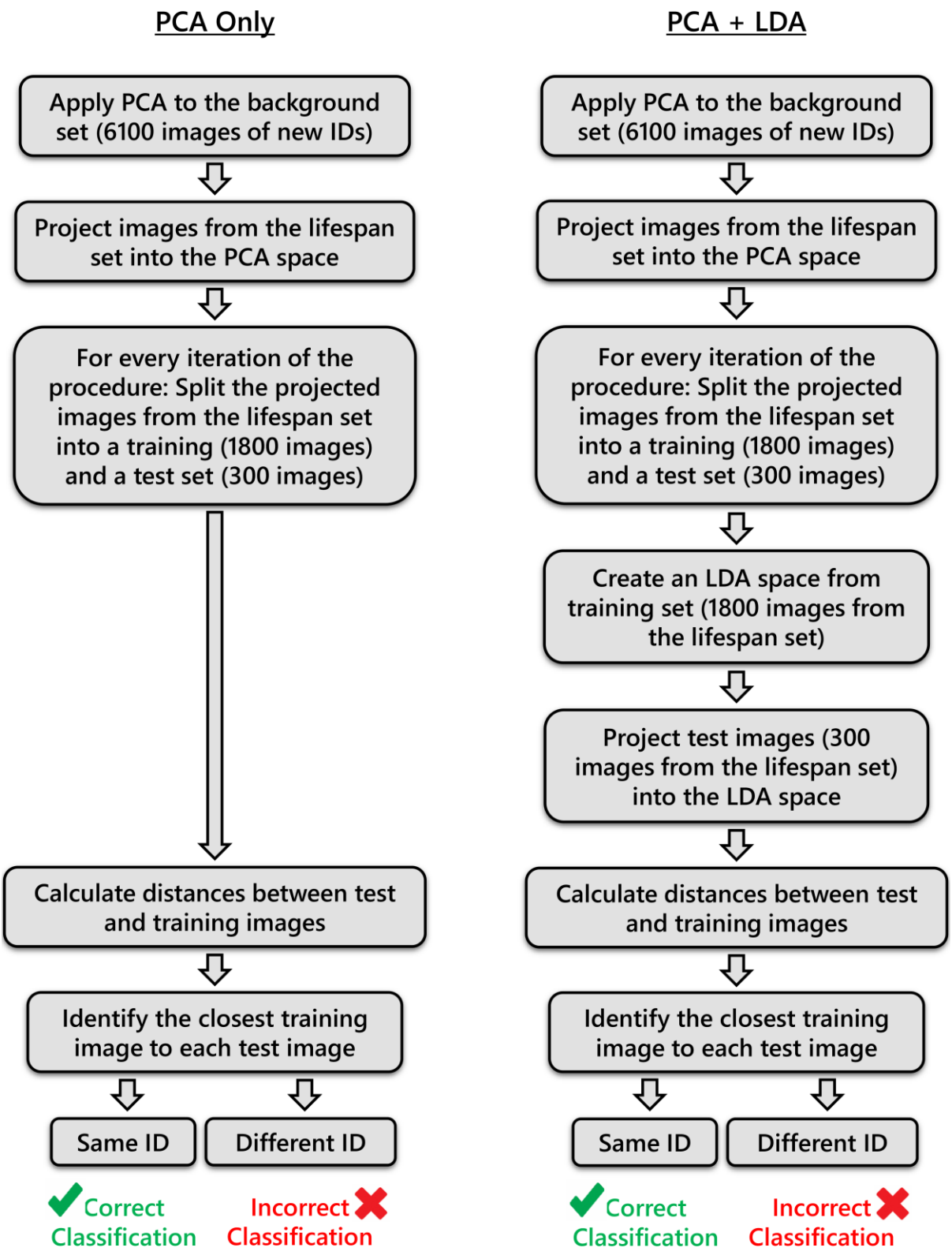

Fig. 6. Simplified representation of our classification procedure for both the PCA (left) and the PCA+LDA (right) approach in Simulations 2a and 2b. First, we perform PCA on all

images from the background set which models our previous experience with faces and the way they age. We then represent all images from the lifespan set into the context of the PCA space created from the background set. All images from the lifespan set are then split into seven different training (1800 images) and test (300 images) sets and we use the same sets for both the PCA Only and the PCA+LDA models. Both sets contain images from all three time periods. For the PCA Only model, we use the projections of images from the lifespan set into the PCA space created from the background set. For each image in the test set, we calculate its distance to every image in the training set and determine the identity of the closest image. For the PCA+LDA model, we perform LDA on the training set (still, represented as PCA projections of images from the lifespan set into the PCA space, created from the background set) and project all images from the test set into the newly created LDA space. Again, we calculate the distance between each test and each training image and determine the identity of the closest image.

3.2.2.1 Simulation 2a: Results and Discussion

Fig. 7 shows the average classification accuracy for both the PCA only and the PCA+LDA models against chance performance. As in Simulation 1, both models perform considerably above chance levels, with identity-clustered spaces providing an extra boost to performance. This was true in all seven iterations of the procedure (see Table S1 in the Supplementary Materials for further information). Once again, it is perhaps surprising that simple PCA projections, based entirely on image similarity, are able to capture substantial identity information across a lifetime of images. This demonstrates that the PCA model in *Simulation 1a* does not derive its high accuracy levels simply by having encoded variance specific to the 10 identities tested. In the present simulation, the face space is built from a much wider set of identities, and does not include the test people at all. In fact, PCA performance in Simulation

2a slightly exceeds that in Simulation 1a, suggesting that encoding extra variability from prior experience with a range of faces helps to some degree in image-level identity-classification.

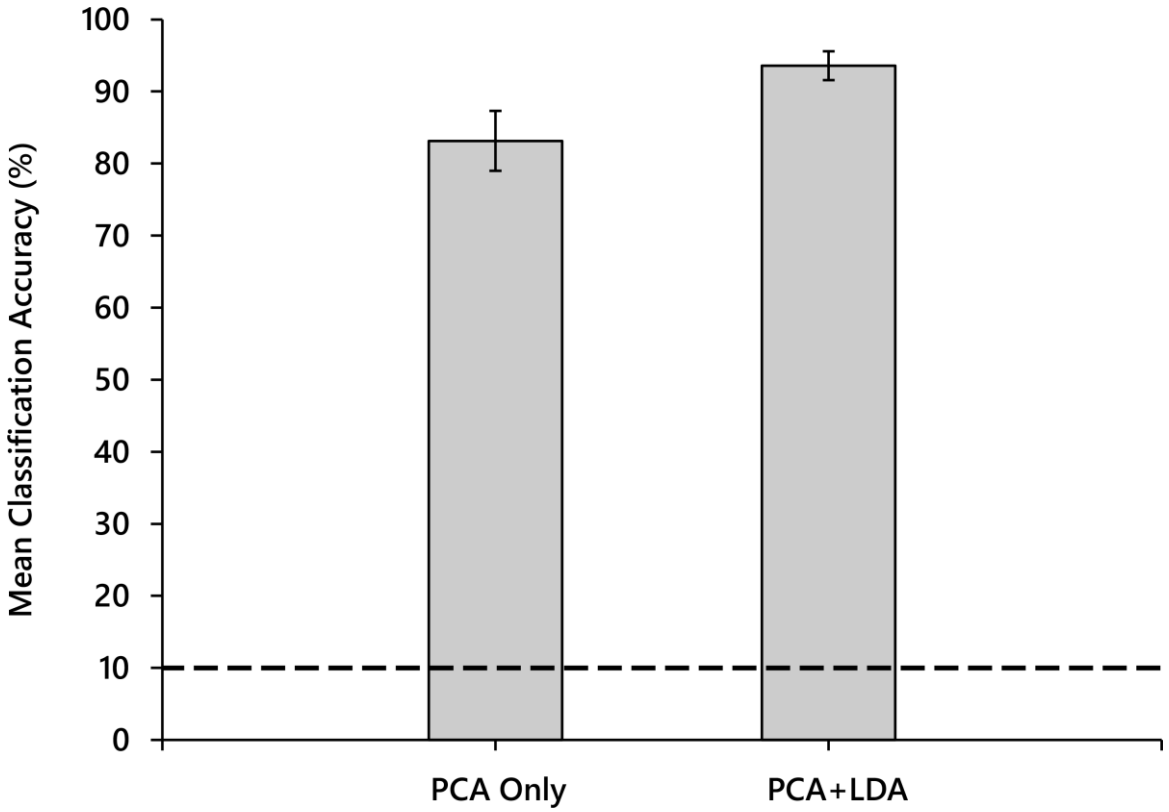

Fig. 7. Mean classification accuracy for novel test images for both the PCA and PCA+LDA approach with a large PCA background set. Dashed line represents chance performance. Error bars show standard error.

For the PCA+LDA model, we once again see near-ceiling performance. The top-down identity-clustering information provides the statistical model with enough information to identify people across their adult lifespan, even using PCA dimensions derived from an entirely different set.

3.2.3 Simulation 2b: Learning time specific identities within an established face space

As in Simulation 1, we also examined the effect of establishing time-specific representations within the space derived from the background image set. Is there any observable advantage in developing three representations per test identity – one for each time period? As previously, we repeated the analysis above, but this time seeking a 30-way classification. Fig. 8 shows mean accuracy for both PCA and PCA+LDA models.

Once again, we see a high degree of accuracy in both models, with LDA adding a substantial boost to accuracy in classification. As with Simulation 1, the large majority of errors in the PCA+LDA model are 'Right person/Wrong time'. This suggests that the clustering algorithm is preserving important identity information which can, to quite a large extent, be used across different age periods. This is less clear in the PCA model, again suggesting that top down information about the range of a person's possible appearance, even within a single age period, is useful for identification across the whole lifespan. More detailed information about the classification outcomes in all seven iterations of the procedure can be found in Table S3 in the Supplementary Materials.

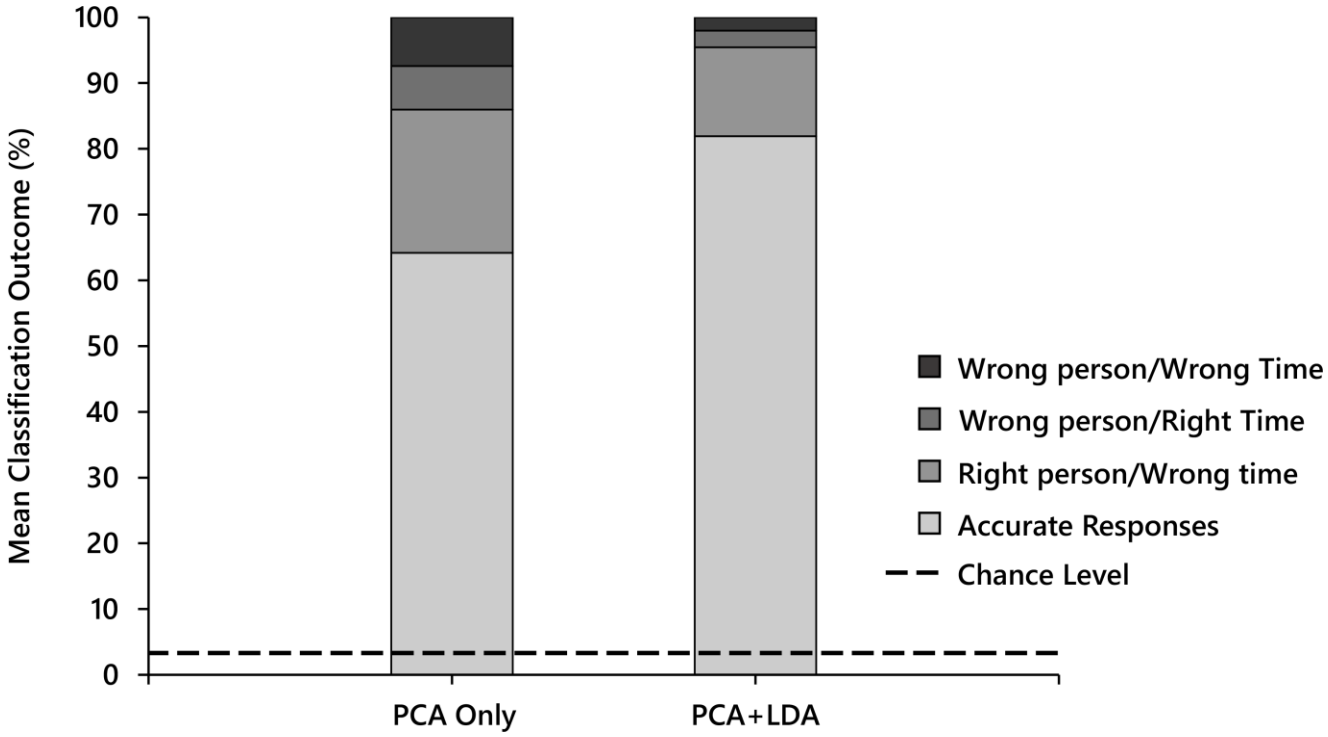

Fig. 8. Mean classification accuracy and errors for both the PCA and PCA+LDA approach with a large PCA base set and a 30-way classification. Dashed line represents chance performance.

3.2.4 Simulation 2c: Increasing the number of classification identities

In all of our simulations so far, we have used a simple classification task across just ten identities. Despite the fact that all of the identities were male musicians active during the same period of time, it is still possible that we have overestimated classification accuracy by limiting the number of possible classification identities. In order to address this issue and make our simulations more representative of real life recognition decisions, we included all identities from the background set (434 identities) in the classification for both the PCA Only and the PCA+LDA models, making our task a 1-in-444 choice (434 identities from the background set + 10 identities from the lifespan set).

The procedure of the present simulation was similar to that of Simulation 2a: First, we applied PCA to all images in the background set, thus generating a simulation of previous experience with faces. We then split the lifespan set into seven different training and test sets. For the PCA Only model, we calculated the distance in PCA space between each image in the test set (300 images) and each image in the training and the background set (total of 7900 images, 444 identities) and finally determined the identity of the closest image. For the PCA+LDA model, we created an LDA space using each training set together with the entire background set, and projected all images from the test set into this newly created LDA space. Finally, we estimated the distance in LDA space between each image in the test set and each image in the training and the background sets (total of 7900 images, 444 identities) and identified the closest image. Unlike our previous simulations, the models had 444 identities to choose from for classification, meaning that chance level was 0.2%.

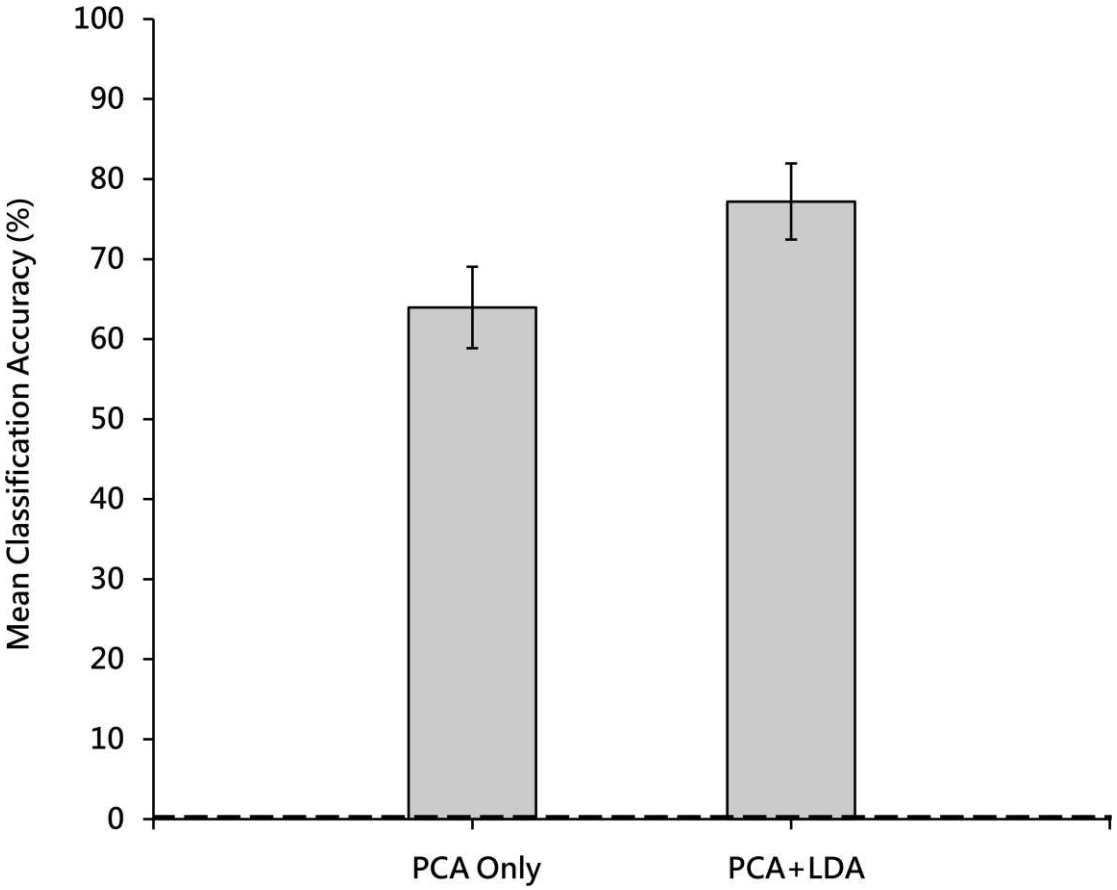

Fig. 9. Mean classification accuracy for novel test images for both the PCA and PCA+LDA approach with the identities from the large background set included in the simulation classification to create a 1-in-444 chance correct choice rate. Dashed line represents chance performance. Error bars show standard error.

Fig. 9 shows the average classification accuracy for both the PCA only and the PCA+LDA models against chance performance. Compared to Simulation 2a, we see a reduction in accuracy for both the PCA Only and the PCA+LDA models following the increased sample of possible identities for classification. Nevertheless, both models still perform considerably above chance levels, with identity-clustered spaces providing an extra boost to performance. This was true in all seven iterations of the procedure (see Table S1 in the Supplementary Materials for further information). The results from this simulation show that our previous models show higher absolute levels of accuracy than models with more response options.

However, they also replicate and strengthen our key finding that top-down information about identity provides a clear advantage for classification across the lifespan.

3.3 Simulation 3: Classifying identity across time

The results from the simulations above imply that there is substantial idiosyncratic, time-invariant information in faces which can be used for identification by simple image similarity (PCA models) and, even more efficiently, by identity clustering statistics (PCA+LDA). However, in each of these simulations, we trained models on images spanning all time periods together. Therefore, in our final section, we ask how well we can recognise a particular identity when we have been familiar with the way they look in only one of these time periods. Moreover, as we have only presented simulation data so far, it is also important to consider human recognition behaviour. This will help us establish the extent of cross-age face recognition difficulties in human observers as well as provide a valuable benchmark for the performance of our models.

Behavioural data were collected from a sample of 34 students at the University of York (23 female, Mean Age = 20.4, Age Range = 18-32). All participants had normal or corrected-to-normal vision, were unfamiliar with the identities included in the lifespan set and earned course credit for their participation. The study was approved by the ethics committee at the University of York. We used a face matching task in order to assess face recognition within and across all three time periods (1960s-70s, 1980s-90s and 2000s-2010s). For this task, participants were presented with two images on a computer screen and were asked to indicate whether they depicted the same person (match trials) or two different people (mismatch trials). We used 240 images from the lifespan set in order to create 120 trials (12 per ID, half match trials). For each ID, participants completed 6 trials with no age gap (match and

mismatch pairs for each time period), 4 trials with a 20-year gap between images (60s vs 80s and 80s vs 00s) and 2 trials with a 40-year gap (60s vs 80s). For mismatch trials identities were paired based on visual similarity.

Fig. 10 shows the mean matching accuracy across all conditions. A 2x3 within subjects ANOVA with factors: age gap (no gap vs 20-year gap vs 40-year gap) and trial type (match vs mismatch) revealed a significant main effect of age gap (F (2, 66) = 27.55, $p < .001$, η_p^2 = .45). There was no significant main effect of trial type (F (1, 33) < 1, $p < .05$, η_p^2 = .01) nor a significant interaction between age gap and trial type (F (2, 66) < 1, $p < .05$, η_p^2 = .02). Follow-up Tukey HSD tests showed that face matching accuracy with no age gap (M = 0.80, SD = 0.13, 95% CI [0.75, 0.84]) was significantly higher than matching accuracy with a 20-year age gap (M = 0.69, SD = 0.15, 95% CI [0.63, 0.74]) and a 40-year gap (M = 0.66, SD = 0.19, 95% CI [0.59, 0.73]) between the two images in the face pair. There was no difference in accuracy with a 20- and a 40-year age gap. One-sample t-tests (two-tailed) against chance (50%) showed that accuracy in each condition was significantly above chance (t_{min} = 4.24, all ps < .001). These results illustrate the difficulties experienced by unfamiliar viewers to recognise identities across time when they only have image-based information available to them.

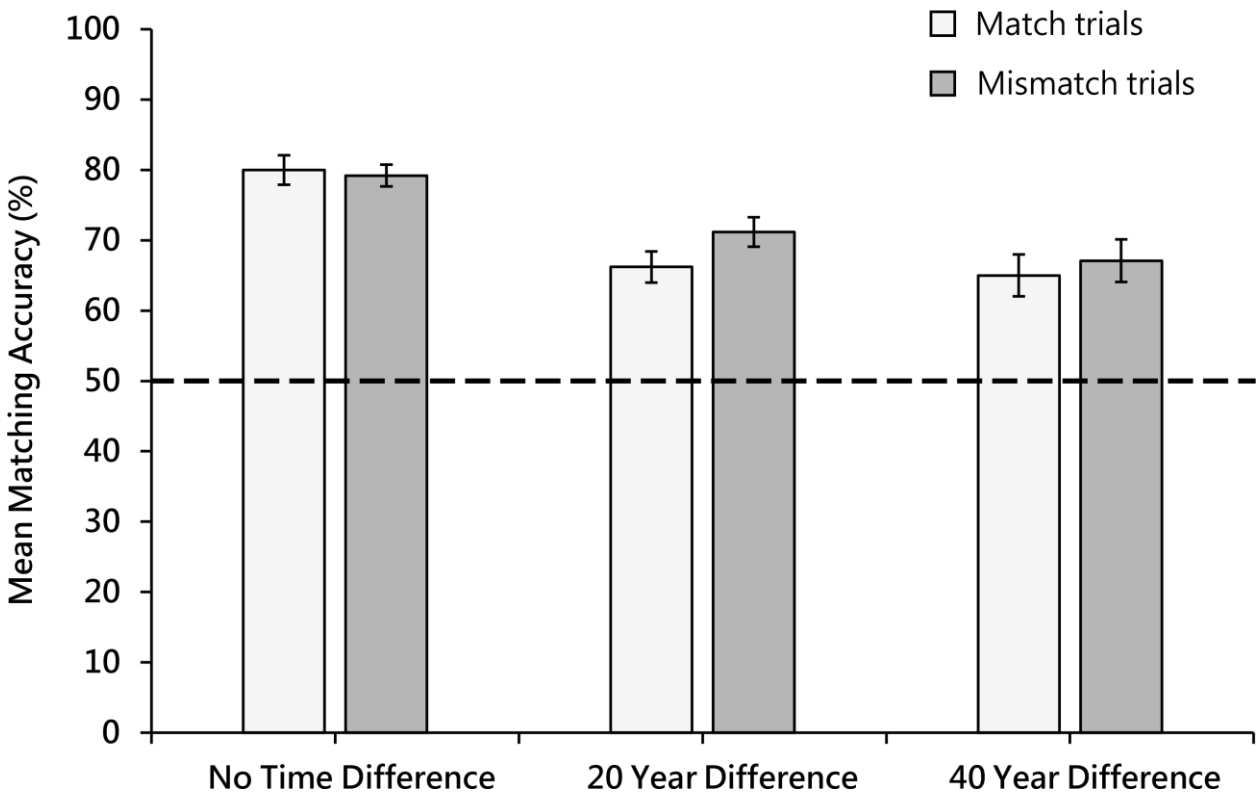

Fig. 10. Mean face matching performance across three age gap conditions (no age gap, 20-year age gap and 40-year age gap). Dashed line represents chance performance. Error bars show within-subjects standard error (Cousineau, 2005).

Next, we return to our simulations, this time training on single time periods and testing identification inside and outside the trained age-range. For example, does knowledge of someone's appearance in their twenties help recognise them in their sixties, and *vice versa*? The human data suggests a sharp drop-off in accuracy as the time-gap between images increases. Thus, the only difference between Simulation 2 and the present simulation is that here we train on images from a single time period and we classify new images from the same or from a different time. As with previous simulations, we compare classification accuracy using PCA only (based on physical properties of the images) and from PCA+LDA to determine the benefits of top-down clustering information.

3.3.1 Simulation 3: Procedure

First, all images from the lifespan set (2100 images) were projected into the PCA space created from the background image set (6100 images). Then, we split the lifespan set according to time period – images from the 1960s/70s, 1980s/90s and 2000s/10s. There were 700 images in each of these categories, 70 per identity. As with the previous simulations, we created 7 different training sets (600 images each, 60 per identity) and test sets (100 images each, 10 per identity). Sets were counterbalanced so that each image was included in a training set 6 times and once in a test set.

In order to calculate classification performance for novel test images from PCA, we computed the distance between each image in the test set and each image in the training set. For PCA+LDA, we ran an LDA on the images from the training set, then projected the novel images from the test set into the newly-created LDA space and calculated the distance between each test and each training item. We used the first 258 PCs for the base PCA (which explained 95% of the variance) and retained only the first LDA dimensions (up to 9) which explained 95% of the 'discriminability' of the overall LDA model.

To test the generalisability of representations over time, we ran learning/test set combinations of three types: (i) training and test sets from the same time period; (ii) learning and test sets 20 years apart; and (iii) learning and test sets 40 year apart. We created separate models for all possible combinations of learning/test sets to generate identification data. So, for example, separate models were created to identify 60s/70s images from the 2000s/10s learning set, and *vice versa*. Classification accuracy was averaged across 7 different iterations for the within time simulations and across 49 iterations (7 training sets x 7 test sets) for the between time simulations. Fig. 11 shows a simplified flow chart of the procedure.

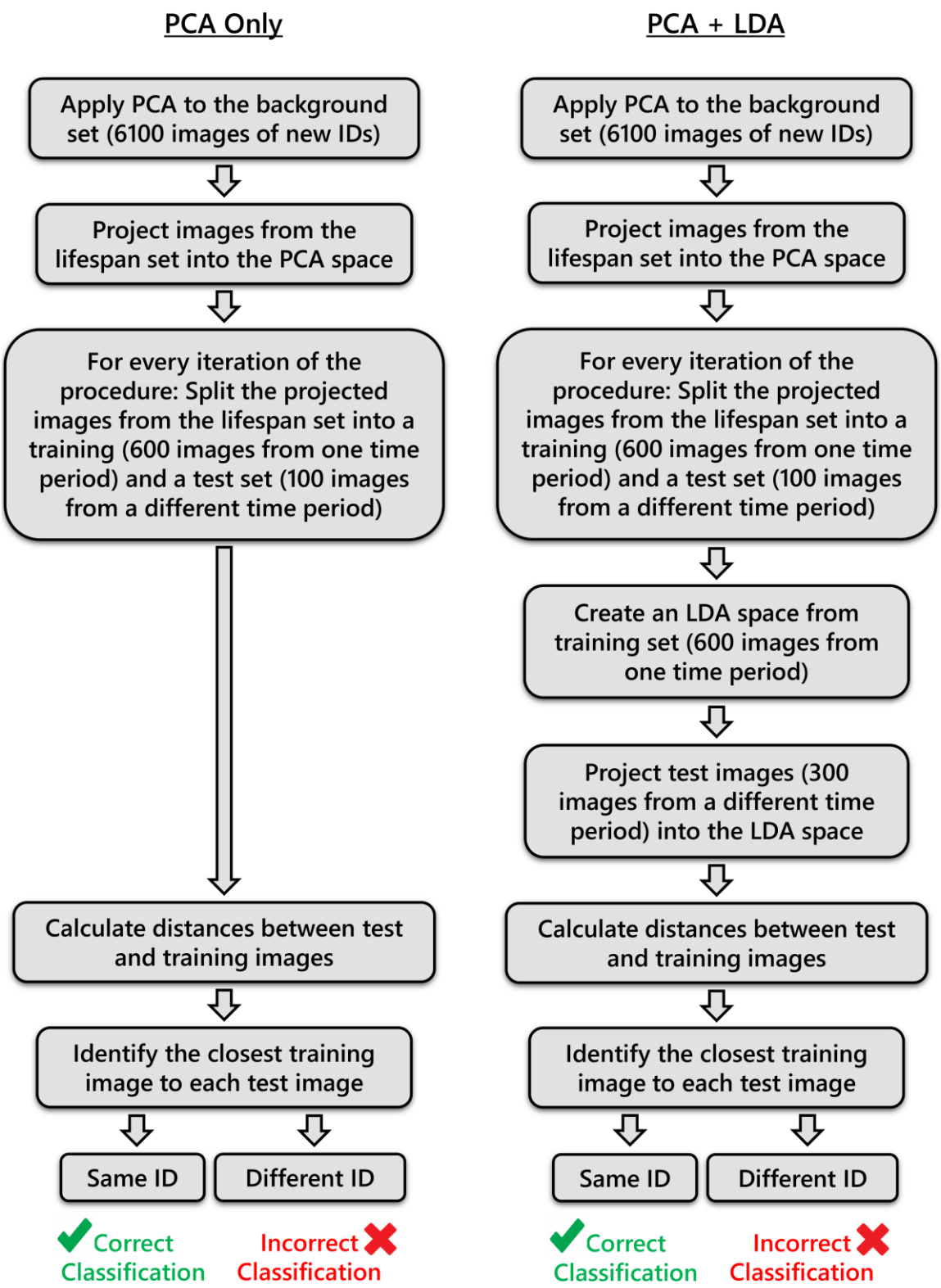

Fig. 11. Simplified representation of our classification procedure for both the PCA (left) and the PCA+LDA (right) approach in Simulation 3. First, we perform PCA on all images from

the background set which models our previous experience with faces and the way they age. We then represent all images from the lifespan set into the context of the PCA space created from the background set. All images from the lifespan set are then split into seven different training (600 images) and test (100 images) sets and we use the same sets for both the PCA Only and the PCA+LDA models. The training sets always contained images from a single time period only and the test sets contained images from the same (for within time period classification) or a different time period (for between time period classification). For the PCA Only model, we use the projections of images from the lifespan set into the PCA space, created from the background set. For each image in the test set, we calculate its distance to every image in the training set and determine the identity of the closest image. For the PCA+LDA model, we perform LDA on the training set (still, represented as PCA projections of images from the lifespan set into the PCA space created from the background set) and project all images from the test set into the newly created LDA space. Again, we calculate the distance between each test and each training image and determine the identity of the closest image.

3.3.2 Simulation 3: Results and Discussion

Fig. 12 shows mean classification accuracy for the three types of simulations using the PCA and the PCA+LDA approach. Here, for the first time, we observe a profound difference between image-similarity models based on PCA only, and models based on image-clustering (PCA+LDA). First, we note good performance for identification within a given time-period. This pattern is consistent with Simulations 1 and 2 above: when we use training images from the same period as test images, a simple image-similarity technique is reasonably accurate, though there is still a marked advantage to be derived from identity clustering. However, when training and test images come from differing periods (i.e. when learning 20 year olds

and testing on 40 year olds) the two statistical models diverge markedly. Here, the clustering approach continues to provide well above chance levels of accuracy but the image-based PCA approach fails catastrophically. We saw the same pattern of results in all seven iterations of the procedure (see Table S4 in the Supplementary Materials for further details).

These are interesting results. We observed in our previous simulations that it is possible to derive a representation (i.e. a functional face space) based on images sampled across a whole adult lifespan, and that such a representation could be used for quite accurate identification. However, in the present case, we observe that simple image-similarity breaks down as a useful representation when tested outside the range of learning. So, while image-similarity can support recognition within a time period, it cannot do so across periods. However, the introduction of top-down information using LDA is enough to maintain continued high levels of recognition. Of course, as the time between training and test images increases, performance falls, as we would expect for a harder task. Nevertheless, clustering together multiple images of a person *within a time period* forms a representation which is useful enough to generalise beyond it. This implies that the identification of *person-specific* variability leads to representations which are much more robust than those based only on physical similarity.

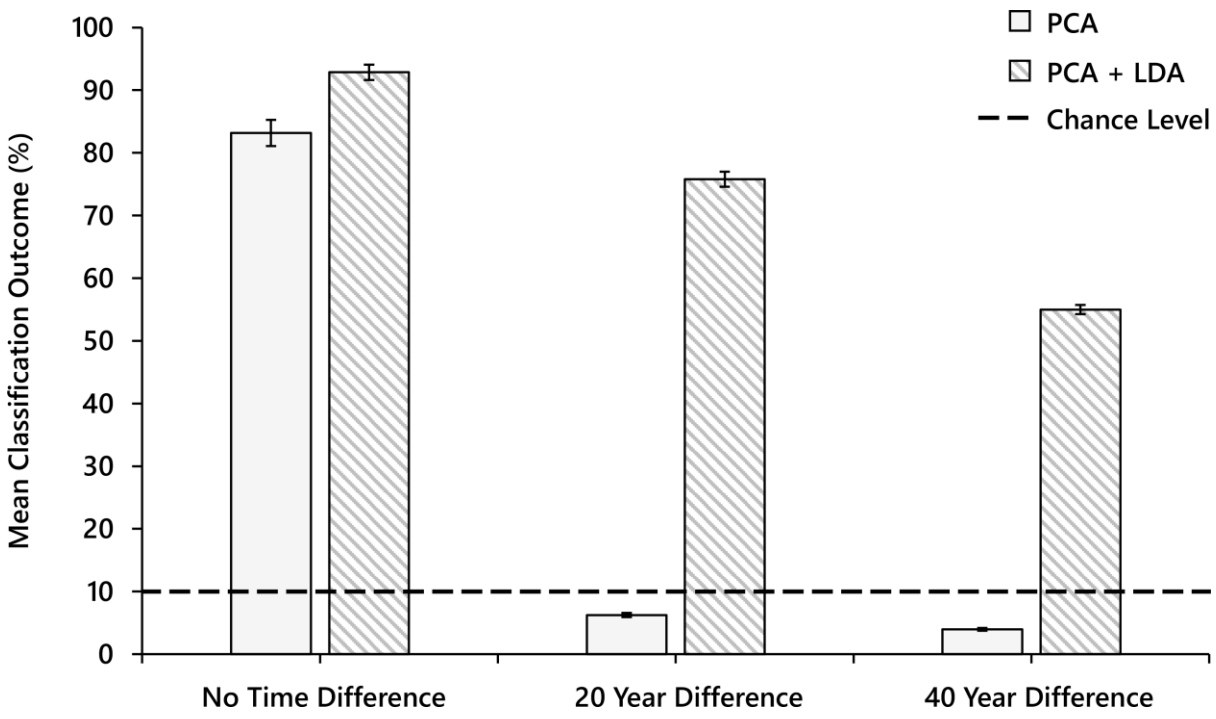

Fig. 12. Mean classification accuracy with no time difference, 20-year difference and 40-year difference between images in the training and test sets for the PCA and PCA+LDA procedures. Dashed line represents chance performance. Error bars show standard error.

Representation of idiosyncratic variability has been shown to be key to familiar face recognition, though previously only within limited time periods (Burton et al., 2016). While we have now shown generalisation over a much larger range, we can also ask whether the trajectory of different people's aging is itself idiosyncratic? It seems intuitively the case that some people's appearance changes more over the aging process than others. As we noted in the introduction, there are some common effects due to skeletal and skin ageing, but these do not seem consistently to account for many of the decade-to-decade changes we observe in the people we know.

Fig. 13 shows data from the identity-clustered classification (PCA+LDA; Fig. 12) broken down for each of the ten identities, and ordered by mean overall performance. The figure shows large individual differences. While all people are well-recognised within-time period, some (to the left of Fig. 13) are much more difficult to recognise across time periods than others (those to the right). The images below the plot show the identities that were easiest and most difficult to classify). This is clear evidence of idiosyncrasy in ageing: some people change more over time than others (Brian Wilson representing an example of the former, and Frankie Avalon the latter). This is represented by the efficiency with which a clustering algorithm can form a generalisable representation when exposed to one time period only. While Fig. 13 shows significant individual differences in the ways our chosen identities have aged, it is important to acknowledge that these identities were celebrities, who may be more motivated than the general populus to preserve their youthful appearance for as long as possible. It is, therefore, possible that our data underestimate the magnitude of these individual differences in the general population and overestimate the overall identity recognition accuracy with images spanning the entire adult lifespan.

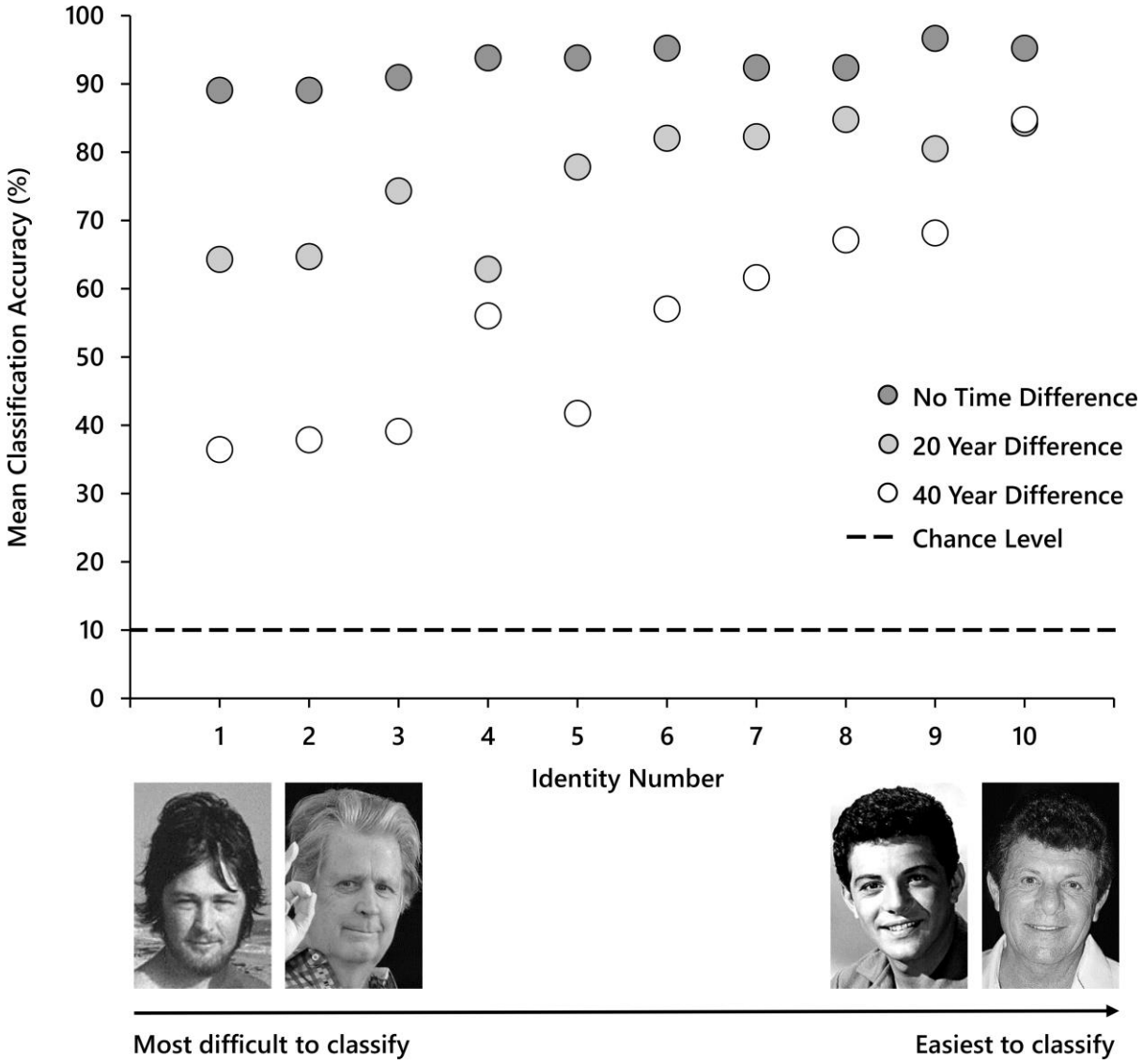

Fig. 13. Mean classification accuracy within and between time periods separately for each identity. The identities are ordered by the mean accuracy across all three conditions. Dashed line represents chance performance. The images below show the person most difficult to recognise across time periods (left pair – Brian Wilson) and the person who was easiest (right pair – Frankie Avalon). See image attributions in the Acknowledgements section.

4. General Discussion

Previous research on facial ageing has focused on generic transformations which occur over the lifespan. It is well-established that viewers are sensitive to these and can, for example, make judgements of comparative age relatively well over a wide range of face stimuli. Here, we address a different question: when we recognise a familiar person over a whole adult lifespan, what type of representation is necessary? In the simulations above, we ask what information is available in face images, and how this might be used by viewers to recognise people, as modelled by fairly simple statistical procedures. These simulations suggest the following:

1. Each person's face contains some physical commonality across the whole adult lifespan, and this can be used for recognition. It is therefore not necessary to store several different representations of Paul McCartney.
2. Top-down support during learning helps when forming robust generalisable representations for face recognition. Knowing that an image is an instance of Paul McCartney enhances the formation of a generalisable representation.
3. If one has learned a face over a limited time period, then top-down support has an even greater benefit for the formation of an age-independent representation.

Running through these simulations is a consistent advantage for incorporating top-down information into statistical representations for face recognition. In Simulations 1 and 2, we see that physical 'image-matching' delivers well-above chance levels of performance, even for images sampled across a lifetime. However, in order to achieve near-perfect performance (as is commonly observed for highly familiar faces in behavioural tasks with human

observers), it is necessary to incorporate knowledge about which images should be clustered together. When we make the task very hard (Simulation 3), by learning images from one time period (e.g. Brian Wilson in the 1960s) and testing in another (Brian Wilson in the 1980s), image-matching using PCA Only breaks down altogether, implying that its above-chance performance in Simulations 1 and 2 is probably due to the availability of images from the same time period in the training set. More importantly, we show that identity recognition across the lifespan is only observed in these simulations when top-down information is incorporated into the representation (using PCA+LDA).

It has been known for many years that there are key behavioural markers of familiarity in face recognition (Bruce, 1986; Bruce & Young, 1986; Johnston & Edmonds, 2009). When tasks are equated, such as asking viewers to match two facial images, there is a large performance advantage for familiar faces (Bruce et al., 2001; Burton et al., 1999; Clutterbuck & Johnston, 2002, 2004, 2005). We have argued elsewhere that this arises because processing unfamiliar faces is dominated by image-level descriptions (Hancock, Bruce, & Burton, 2000; Longmore, Liu & Young, 2008. Megreya & Burton, 2006; Young & Burton, 2018). Accuracy depends on the extent to which two images of an unfamiliar face can be matched on physical similarity. In contrast, recognition of familiar people relies on a representation that incorporates their own idiosyncratic variability (Burton, 2013; Burton et al., 2016; Jenkins & Burton, 2011; Jenkins et al., 2011). Here, we have shown that being able to incorporate this idiosyncratic variability can benefit recognition across transformations in appearance due to ageing.

An important question concerns the potential involvement of the vast and complex semantic, experiential and emotional aspects of our responses to a familiar person (Bobes, Lage

Castellanos, Quiñones, García, & Valdes-Sosa, 2013; Gobbini & Haxby, 2007;Wiese et al, 2019); in particular, whether these play a central role in creating the profound differences between recognition of familiar and unfamiliar faces. Accessing semantic, episodic and emotional memories forms a critical aspect of our everyday experience with familiar people, but we have shown here that, from the standpoint of visual recognition, these properties are not essential to differences in recognisability between familiar and unfamiliar faces. Instead, manipulations based on clustering visual images are sufficient to support generalisable recognition. While this intervention does involve some degree of top-down influence and reshaping of the underlying perceptual space, this operates on purely visual representations.

The simulations above suggest that a key component of familiarity is captured by simply clustering together multiple, highly varying, images of the people to be recognised. But how could this clustering happen in daily life? We propose that our everyday encounters support recognition through multiple sources. We can gather many 'instances' of a person's face throughout a single interaction, but also through multimodal recognition (voices, names) and contextual information. This range of support for recognition allows one to add to one's visual representation, by knowing which inputs to group together: if I am confident from multiple sources that this is Paul McCartney, then I can update my visual representation of him with this encounter. Of course, representation of the people we know well is not simply visual; we form highly complex associations for the people we know, including their personal information, their relationship with us, and emotional associations. It seems highly likely that the people we can recognise over long time periods are those for whom we have rich and detailed knowledge. Using LDA as a simple top-down intervention to cluster together different instances of a face does not come close to capturing this complexity. However, our intention in using it is to point out that that recognition in our daily lives is multi-faceted, and

there are very many instances in which we might plausibly believe that a photo we are viewing is Paul McCartney. It seems highly likely that this multimodal support assists in the development of complex visual representations, by proving information about which face images should be incorporated into the representation of particular person.

We propose that the distinction between bottom-up/top-down processing goes some way to understanding the fundamental distinction between unfamiliar and familiar face recognition. Using the particularly challenging case of recognition over the adult lifespan, we can observe the effects of incorporating variability into representations for recognition. Representations based on physical similarity can get one so far, but the addition of top-down processes significantly enhance the development of robust performance.

We now return to the initial problem of recognising someone over their entire adult life. These simulations do suggest that people's faces have some key visual characteristics which they preserve over life. However, in order to use these effectively, they must be sampled over the entire lifespan, as in Simulations 1 and 2. In contrast, Simulation 3 demonstrated that a set of simple physical descriptions incorporated in a single time period cannot be used to recognise someone at a different age, at least using our simple image-similarity model based on PCA. Generalisation beyond the era of exposure, on the other hand, is greatly enhanced by the formation of clustered representations, in other words the acquisition of familiarity with a person. The clear prediction from Fig. 12, is that it will be difficult to match two images of unfamiliar faces across eras, as is borne-out by our behavioural data. However, a viewer familiar with someone at a particular era will have some prospect of recognising that person in a later era – though accuracy will decrease with the time gap. This

seems to be consistent with what is known about long-term recognition of familiar and unfamiliar faces (Megreya et al., 2013; Özbek & Bindemann, 2011).

Of course, the simple image statistical procedures we have used here are intended as illustrative only of the nature of underlying constraints that must be faced for recognition. We are not trying to develop a practical automated face recognition system. In fact, recent developments in the use of deep convolutional neural networks (DCNNs), have demonstrated that with enough computational power and a large enough training set, it is now possible to perform astonishingly accurate face recognition automatically (e.g. Ranjan, Sankaranarayanan, Castillo & Chellappa, 2017; Taigman, Yang, Ranzato & Wolf, 2014). The latest networks can match faces more accurately than human viewers unfamiliar with the faces, out-performing even trained experts in the field (Phillips et al., 2018). However, in this paper we are concerned only with what is known about human viewers: we clearly recognise those known to us across the lifespan (e.g. Paul McCartney), and the normal difficulties of matching unfamiliar faces are exacerbated by large age-differences. The impressive recent developments in automated recognition are interesting, but do not provide an explanation of human performance.

We should also make clear that we have no commitment either to PCA or LDA as capturing the mechanics of the brain itself; what they do very effectively is to allow us to analyse the nature of the computational problems that the brain has to solve. Both techniques are used extensively in the literature on face processing (e.g. Bekios-Calfa et al., 2011; Chang & Tsao, 2017; Kramer et al., 2018) but we use them here mainly to capture the differences in statistical procedures between bottom-up and top-down analyses. However, despite their simplicity and the challenging stimuli employed here (large datasets with multiple samples of

multiple people in unconstrained photos) we are able to discern general principles for learning faces. The nature of faces is to change, within encounters, between encounters and over time. At the core of our approach is to treat this variability as data, not noise, an idea introduced into modern theories of face processing by Bruce (1994), but not operationalised until more recently. We hope to have demonstrated that this approach can help make headway towards solving the difficult problem posed by the pictures in Fig. 1.

Acknowledgements

Image attributions for Fig. 1 - top row from left to right: Mikael J. Nordstrom [Public Domain], Pixabay [Public Domain – CC0], Michael Cooper [Public Domain]; middle row from left to right: Terry George [CC BY-NC-SA 2.0], Nationaal Archief [CC BY-SA 3.0 NL], Fresh on the net [CC BY 2.0]; bottom row from left to right: Oli Gill [CC BY-SA 2.0], Comediadeflowers [CC BY-SA 4.0], Angie Mills [CC BY-SA 3.0]

Image attributions for Fig. 5 - Movie Studio [Public domain], Debora Duarteb [CC BY-SA 4.0], Andrea Raffin [CC BY-SA 3.0], Eva Rinaldi [CC BY-SA 2.0], Glyn Lowe [CC BY 2.0], Walterlan Papetti [CC BY-SA 4.0], Eva Rinaldi [CC BY-SA 2.0], Nicolas Genin [CC BY-SA 2.0], Takahiro Kyono [CC BY 2.0], Gage Skidmore [CC BY-SA 3.0], Casablanca Records [Public domain], International Labour Organisation [CC BY-NC-ND 2.0], Ron Kroon [CC0 1.0], Hubert Burda Media [CC BY-NC-SA 2.0], State Farm [CC BY 2.0]

Image attributions for Fig. 10 – Capitol Records [Public domain], Takahiro Kyono [CC BY 2.0], Film Studio [Public domain], John Mathew Smith [CC BY-SA 2.0]

References

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology*, *68*(10), 2041–2050. https://doi.org/10.1080/17470218.2014.1003949

Bekios-Calfa, J., Buenaposada, J. M., & Baumela, L. (2011). Revisiting linear discriminant techniques in gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(4), 858–864. https://doi.org/10.1109/TPAMI.2010.208

Bobes, M. A., Lage Castellanos, A., Quiñones, I., García, L., & Valdes-Sosa, M. (2013). Timing and tuning for familiarity of cortical responses to faces. *PLoS ONE*, *8*(10), e76100. https://doi.org/10.1371/journal.pone.0076100

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, *73*(1), 105–116. https://doi.org/10.1111/j.2044-8295.1982.tb01795.x

Bruce, V. (1986). Influences of familiarity on the processing of faces. *Perception*, *15*(4), 387–397. https://doi.org/10.1068/p150387

Bruce, V. (1994). Stability from variation: the case of face recognition. The M.D. Vernon Memorial Lecture. *The Quarterly Journal of Experimental Psychology. A*, *47*(1), 5–28. https://doi.org/10.1080/14640749408401141

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*(4), 339–360. https://doi.org/10.1037/1076-898X.5.4.339

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental*

Psychology: Applied, *7*(3), 207–218. https://doi.org/10.1037//1076-898X.7.3.207

Bruce, V., & Valentine, T. (1986). Semantic priming of familiar faces. *The Quarterly Journal of Experimental Psychology Section A*, *38*(1), 125–150. https://doi.org/10.1080/14640748608401588

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305–327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Burt, D. M., & Perrett, D. I. (1995). Perception of age in adult Caucasian male faces: Computer graphic manipulation of shape and colour information. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *259*(1355), 137–143. https://doi.org/10.1098/rspb.1995.0021

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology*, *66*(8), 1467–1485. https://doi.org/10.1080/17470218.2013.800125

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology*, *102*(4), 943–958. https://doi.org/10.1111/j.2044-8295.2011.02039.x

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*(1), 202–223. https://doi.org/10.1111/cogs.12231

Burton, A. M., Miller, P., Bruce, V., Hancock, P. J. B., & Henderson, Z. (2001). Human and automatic face recognition: A comparison across image formats. *Vision Research*, *41*(24), 3185–3195. https://doi.org/10.1016/S0042-6989(01)00186-9

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, *10*(3), 243–248. https://doi.org/10.1111/1467-9280.00144

Chang, L., & Tsao, D. Y. (2017). The code for facial identity in the primate brain. *Cell*, *169*(6), 1013-1028.e14. https://doi.org/10.1016/j.cell.2017.05.011

Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception*, *31*(8), 985–994. https://doi.org/10.1068/p3335

Clutterbuck, R., & Johnston, R. A. (2004). Matching as an index of face familiarity. *Visual Cognition*, *11*(7), 857–869. https://doi.org/10.1080/13506280444000021

Clutterbuck, R., & Johnston, R. A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology*, *17*(1), 97–116. https://doi.org/10.1080/09541440340000439

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology, 1*(1), 42–45. doi: 10.20982/tqmp.01.1.p042

Craw, I. (1995). A manifold model of face and object recognition. In T. Valentine (Ed.), *Cognitive and computational aspects of face recognition* (pp. 201–221). https://doi.org/https://doi.org/10.4324/9780203428979

Davies, G. M., Ellis, H. D., & Shepherd, J. W. (Eds.). (1981). *Perceiving and remembering faces*. London: Academic Press.

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology, 23*(4), 482–505. https://doi.org/10.1002/acp.1490

Dunn, J. D., Kemp, R. I., & White, D. (2018). Search templates that incorporate within-face variation improve visual search for faces. *Cognitive Research: Principles and Implications*, *3*(1), 1–11. https://doi.org/10.1186/s41235-018-0128-1

Geng, X., Zhou, Z.H., & Smith-Miles, K. (2007). Automatic age estimation based on facial

aging patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(12), 2234–2240. https://doi.org/10.1109/TPAMI.2007.70733

George, P. A., & Hole, G. J. (1998). Recognising the ageing face: The role of age in face processing. *Perception, 27*(9), 1123–1134. https://doi.org/10.1068/p271123

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia*, *45*(1), 32–41. https://doi.org/10.1016/j.neuropsychologia.2006.04.015

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330–337. https://doi.org/10.1016/S1364-6613(00)01519-9

Hanczakowski, M., Zawadzka, K., & Macken, B. (2015). Continued effects of context reinstatement in recognition. *Memory and Cognition*, *43*(5), 788–797. https://doi.org/10.3758/s13421-014-0502-2

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(4), 986–1004. https://doi.org/10.1037/0096-1523.22.4.986

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1671–1683. https://doi.org/10.1098/rstb.2010.0379

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001

Johnston, R. J., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, *17*(5), 577–596. https://doi.org/10.1080/09658210902976969

Johnston, R. A., Milne, A. B., Williams, C., & Hosie, J. (1997). Do distinctive faces come

from outer space? An investigation of the status of a multidimensional face-space. *Visual Cognition, 4*(1), 59–67. https://doi.org/10.1080/713756748

Kramer, R. S. S., Jenkins, R., & Burton, A. M. (2017). InterFace: A software package for face image warping, averaging, and principal components analysis. *Behavior Research Methods*, *49*(6), 2002–2011. https://doi.org/10.3758/s13428-016-0837-7

Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition*, *172*, 46–58. https://doi.org/10.1016/j.cognition.2017.12.005

Kramer, R. S. S., Young, A. W., Day, M. G., & Burton, A. M. (2017). Robust social categorization emerges from learning the identities of very few faces. *Psychological Review*, *124*(2), 115–129. https://doi.org/10.1037/rev0000048

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. J*ournal of Experimental Psychology: Human Perception and Performance, 34,* 77-100. http://dx.doi.org/10.1037/0096-1523.34.1.77

MacLin, O. H., MacLin, M. K., & Malpass, R. S. (2001). Race, arousal, attention, exposure and delay: An examination of factors moderating face recognition. *Psychology, Public Policy, and Law*, *7*(1), 134–152. https://doi.org/10.1037/1076-8971.7.1.134

Matthews, C. M., & Mondloch, C. J. (2018). Improving identity matching of newly encountered faces: Effects of multi-image training. *Journal of Applied Research in Memory and Cognition, 7*(2), 280–290. https://doi.org/10.1016/j.jarmac.2017.10.005

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865–876. https://doi.org/https://doi.org/10.3758/BF03193433

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied*, *14*(4), 364–372. https://doi.org/10.1037/a0013464

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images raken on the same day or months apart: The limitations of photo ID. *Applied Cognitive Psychology*, *27*(6), 700–706. https://doi.org/10.1002/acp.2965

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 577–581. https://doi.org/10.1037/xhp0000049

O'Toole, A. J., Abdi, H., Deffenbacher, K. A., & Valentin, D. (1993). Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America A*, *10*(3), 405. https://doi.org/10.1364/JOSAA.10.000405

O'Toole, A. J., Edelman, S., & Bülthoff, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research*, *38*(15–16), 2351–2363. https://doi.org/10.1016/S0042-6989(98)00042-X

Özbek, M., & Bindemann, M. (2011). Exploring the time course of face matching: Temporal constraints impair unfamiliar face identification under temporally unconstrained viewing. *Vision Research*, *51*(19), 2145–2155. https://doi.org/10.1016/j.visres.2011.08.009

Patterson, K. E., & Baddeley, A. D. (1977). When face recognition fails. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(4), 406–417. https://doi.org/10.1037/0278-7393.3.4.406

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., … O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 201721355. https://doi.org/10.1073/pnas.1721355115

Pittenger, J. B., & Shaw, R. E. (1975). Aging faces as viscal-elastic events: Implications for a theory of nonrigid shape perception. *Journal of Experimental Psychology: Human*

Perception and Performance, *1*(4), 374–382. https://doi.org/10.1037//0096-1523.1.4.374

Ranjan, R., Sankaranarayanan, S., Castillo, C.D. & Chellappa, R. (2017) An all-in-one convo- lutional neural network for face analysis. *Proceedings of the 12th IEEE International Conference on Automatic Face Gesture Recognition Gesture Recognition*, pp 17–24. Available at https://ieeexplore.ieee.org/document/7961718/.

Reynolds, J. K., & Pezdek, K. (1992). Face recognition memory: The effects of exposure duration and encoding instruction. *Applied Cognitive Psychology*, *6*(4), 279–292. https://doi.org/10.1002/acp.2350060402

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology*, *70*(5), 897–905. https://doi.org/10.1080/17470218.2015.1136656

Schweinberger, S. R. (1996). How Gorbachev primed Yeltsin: Analyses of associative priming in person recognition by means of reaction times and event-related brain potentials. *Journal of Experimental Psychology: Learning Memory and Cognition*, *22*(6), 1383–1407. https://doi.org/10.1037/0278-7393.22.6.1383

Sirovich, L., & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A, 4*(3), 519–524. https://doi.org/10.1364/JOSAA.4.000519

Smith, L. I. (2002). A tutorial on Principle Components Analysis. Department of Computer Science, University of Otago, New Zealand. Retrieved from https://ourarchive.otago.ac.nz/bitstream/handle/10523/7534/OUCS-2002-12.pdf.

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118. https://doi.org/10.1016/j.cognition.2012.12.001

Taigman, Y., Yang, M., Ranzato, M., Wolf, L. (2014) Deepface: Closing the gap to human-level performance in face verification. *Proceedings ofthe IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Washington, DC), pp 1701– 1708.

Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications, 30*(2), 169–190. doi: 10.3233/AIC-170729

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86. https://doi.org/10.1162/jocn.1991.3.1.71

Watkins, M. J., Ho, E., & Tulving, E. (1976). Context effects in recognition memory for faces. *Journal of Verbal Learning and Verbal Behavior*, *15*(5), 505–517. https://doi.org/10.1016/0022-5371(76)90045-1

Wiese, H., Tüttenberg, S. C., Ingram, B. T., Chan, C. Y. X., Gurbuz, Z., Burton, A. M., & Young, A. W. (2019). A robust neural index of high face familiarity. *Psychological Science*, *30*(2), 261–272. https://doi.org/10.1177/0956797618813572

Young, A. W. (2018). Faces, people and the brain: the 45th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology, 71*, 569-594. https://doi.org/10.1177/1747021817740275

Young, A. W., & Burton, A. M. (2018). Are we face experts? *Trends in Cognitive Sciences*, *22*(2), 100–110. https://doi.org/10.1016/j.tics.2017.11.007

Young, A. W., Flude, B. M., Hellawell, D. J., & Ellis, A. W. (1994). The nature of semantic priming effects in the recognition of familiar people. *British Journal of Psychology*, *85*(3), 393–411. https://doi.org/10.1111/j.2044-8295.1994.tb02531.x

Young, A. W., Hay, D. C., & Ellis, A. W. (1985). The faces that launched a thousand slips: Everyday difficulties and errors in recognizing people. *British Journal of Psychology*, *76*(4), 495–523. https://doi.org/10.1111/j.2044-8295.1985.tb01972.x