Cognition 249 (2024) 105792

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

A cognitive template for human face detection

Jonathan E. Prunty [c,*], Rob Jenkins [b], Rana Qarooni [b], Markus Bindemann [a]

[a] School of Psychology, University of Kent, Canterbury, UK
[b] Department of Psychology, University of York, York, UK
[c] Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK

ARTICLE INFO

Keywords:
Face
Detection
Cognitive
Templates
Averages

ABSTRACT

Faces are highly informative social stimuli, yet before any information can be accessed, the face must first be detected in the visual field. A detection template that serves this purpose must be able to accommodate the wide variety of face images we encounter, but how this generality could be achieved remains unknown. In this study, we investigate whether statistical averages of previously encountered faces can form the basis of a general face detection template. We provide converging evidence from a range of methods—human similarity judgements and PCA-based image analysis of face averages (Experiment 1–3), human detection behaviour for faces embedded in complex scenes (Experiment 4 and 5), and simulations with a template-matching algorithm (Experiment 6 and 7)—to examine the formation, stability and robustness of statistical image averages as cognitive templates for human face detection. We integrate these findings with existing knowledge of face identification, ensemble coding, and the development of face perception.

1. Introduction

The detection of faces in the visual environment is a necessary prerequisite to all other tasks with faces. Detection is made challenging because the faces we encounter in everyday life can vary considerably in appearance, for example in terms of sex, ethnicity, age and size. Despite this variability, human face detection is remarkably efficient, proceeding with speed and accuracy (Crouzet, Kirchner, & Thorpe, 2010; Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Hershler & Hochstein, 2005; Kelly, Duarte, Meary, Bindemann, & Pascalis, 2019; Qarooni, Prunty, Bindemann, & Jenkins, 2022; Rousselet, Macé, & Fabre-Thorpe, 2003). How this feat is accomplished by the visual system is unknown, though the process is assumed to involve mapping of sensory inputs onto stored mental representations that serve as 'templates' for detection (Bindemann & Lewis, 2013; Lewis & Edmonds, 2005; Prunty, Qarooni, Jenkins, & Bindemann, 2023). One way in which human detection could achieve such efficiency would be to match a range of inputs to a *single* cognitive template (many-to-one mapping). Here we investigate whether such a template could be formed by computing an average of previously encountered faces.

Reliance on a single template for detection of all faces may seem implausible. If the tuning of the template is too narrow, faces in the environment will be too often missed. If the tuning is too broad, non-face objects will trigger too many false alarms. Yet statistical summaries, such as averages, may be appropriate to the task. By selectively emphasising features that are consistent across category members (e.g., different faces), a statistical summary offers a way to capture highly variable information within a single description (Cohen, Dennett, & Kanwisher, 2016; Jenkins & Burton, 2011). Averaging multiple photographs of one individual, for instance, can provide a reliable representation of that group's shared characteristic: the person's identity (Burton, Jenkins, Hancock, & White, 2005; Burton, Jenkins, & Schweinberger, 2011; Jenkins & Burton, 2008, 2011). Averaging images of multiple identities can produce a face image that portrays the group's most dominant shared trait – attractiveness, age, or trustworthiness, for instance (Benson & Perrett, 1993; Perrett, May, & Yoshikawa, 1994; Sutherland et al., 2013). By the same logic, an average image from a sufficiently large and varied pool of faces should form an image that is representative of faces *in general*, making it an ideal template for face detection.

This possibility has not yet been tested. However, it has long been known that typical or average faces are classified faster *as faces* than distinctive faces (Valentine & Bruce, 1986). Moreover, in theorising in this domain, face norms are often assumed to be established based on all encountered faces (Leopold, Bondar, & Giese, 2006; Rhodes, Brennan, & Carey, 1987; Valentine, 1991), which resonates with the notion of a

* Corresponding author at: School of Psychology, Keynes Colleges, University of Kent, Canterbury CT2 7NP, UK.
E-mail address: jep84@cam.ac.uk (J.E. Prunty).

https://doi.org/10.1016/j.cognition.2024.105792
Received 25 May 2023; Received in revised form 5 February 2024; Accepted 8 April 2024
Available online 18 May 2024

detection template that is formed from a large and varied pool of faces. The possibility of an average template for face detection also seems tenable in light of recent work on ensemble perception, which provides evidence that the visual system uses statistical summaries to efficiently represent sets of objects, including faces (Cohen et al., 2016; Whitney & Yamanashi Leib, 2018). For example, observers can rapidly extract average summaries from sets of faces across a number of different dimensions including expression (Elias, Dyer, & Sweeny, 2017; Haberman & Whitney, 2007, 2009; Ying & Xu, 2017), gender (Haberman & Whitney, 2007), viewpoint (Sweeny & Whitney, 2014), and attractiveness (Ying, Burns, Choo, & Xu, 2017; Ying, Burns, Lin, & Xu, 2019). Average identity can also be encoded, irrespective of whether faces are familiar (Neumann, Schweinberger, & Burton, 2013) or unfamiliar (de Fockert & Wolfenstein, 2009), or whether faces are presented simultaneously (de Fockert & Wolfenstein, 2009; Neumann et al., 2013) or sequentially (Neumann et al., 2013; Yamanashi Leib, Fischer, Liu, Whitney, & Robertson, 2014). Humans appear to extract average representations from the faces they encounter across a wide range of conditions.

The study of averages of multiple images of the *same* face provides some insight into why a similar cognitive process, that averages across multiple identities, might also be beneficial for detection (Kramer, Ritchie, & Burton, 2015). Averaging across face images of the same person can form a stable identity representation (Burton et al., 2005, 2011; Jenkins & Burton, 2011) that can outperform individual photos on recognition tasks (Burton et al., 2005; Jenkins & Burton, 2008; Robertson, Kramer, & Burton, 2015). A stable identity representation is formed by removing image-specific properties that are uncorrelated with identity – such as lighting, pose or hairstyle – leaving only information consistently represented across images (i.e., information related to identity, see Jenkins & Burton, 2011). Stable cognitive representations of an individual, formed through exposure to multiple instances of the same face, could thus provide a model of face learning and underpin the recognition advantage that exists for familiar faces (Burton, Wilson, Cowan, & Bruce, 1999; Ritchie et al., 2015). In a similar manner, exposure to multiple instances of *different* faces could form a stable, generic face representation, that would be less reliant on exemplar-specific properties such as identity, and could provide an explanation for detection expertise.

The current series of experiments will investigate the viability of multiple-identity average representations as generic face detection templates. Experiments 1 to 3 examine how many faces must be combined before a multiple-identity average stabilises, so that its appearance is not meaningfully changed when further images are incorporated. Experiments 4 and 5 then compare detection performance for stable face averages and their constituent exemplars when these are embedded in naturalistic scenes. This provides a direct test for the average template hypothesis. Similar to the recognition advantage conferred by single-identity averages over individual photographs (Burton et al., 2005; Jenkins & Burton, 2008), we predicted a *detection* advantage for average images of multiple identities relative to exemplars. Finally, Experiments 6 and 7 use a template-matching algorithm to simulate the formation of the cognitive detection template by averaging images of multiple identities.

2. Experiment 1

Previous work shows that stable within-identity averages for face identification can be formed surprisingly quickly (Burton et al., 2005; Jenkins & Burton, 2011). For example, an average based on as few as 20 images can retain its apparent identity even when contaminated with other identities (Jenkins & Burton, 2011; Jenkins, Burton, & White, 2006). However, a key difference between detection and recognition is that multiple identities can span more variability in appearance than a single identity. Considering factors such as identity, sex, age or race, the formation of a sufficiently broad and stable average detection template might require exposure to a very large number of faces.

In this experiment, we systematically manipulate both the number of face identities contributing to an average, and the variability of these faces, to estimate the point of stability. For this purpose, we constructed average images from small or large sets of faces, drawn from demographic categories that varied by age, sex and race. To assess how quickly these averages stabilised, participants decided whether pairs of averages look like the same person or different people. The underlying logic is that averages constructed from random samples of images should converge, as indexed by the proportion of 'same person' responses, as the number of contributing images increases.

2.1. Methods

2.1.1. Participants

All methodological procedures, sample size and planned analyses were preregistered on the Open Science Framework (OSF; see osf.io/z58ay) and the data for these experiments are also available on OSF (osf.io/k9dnm). Forty paid participants (26 females, 14 males; Age $M = 30.83$, $SD = 10.60$) were recruited via Prolific in 2020 (Prolific, 2021). Samples of this size (or smaller) have proven sufficient in previous experiments, in both the face-matching (e.g., Fysh, 2018; Young, Hay, McWeeny, Flude, & Ellis, 1985) and categorical perception literature (e.g., Calder, Young, Perrett, Etcoff, & Rowland, 1996; Peng et al., 2010). All participants were in the age range 18–60, had English as their first language, and self-reported normal or corrected-to-normal vision. In this and all subsequent experiments, informed consent was obtained from participants prior to data collection.

2.1.2. Design and stimuli

To assess the stability of cross-identity averages, we constructed average face images that varied according to the number of images they contained (Number of Constituent Images; NOCI) and their demographic variability. Ambient face images were sourced from an online face generator (thispersondoesnotexist.com), a generative adversarial network (GAN) trained on 70,000 face images (Karras, Aila, Laine, & Lehtinen, 2017; Karras, Laine, & Aila, 2019). Four hundred and twenty full-face images were selected based on three categorical dimensions, reflecting perceived age (young, old adults), race (Asian, Black, White) and gender (male, female). Combining age (2), race (3), and gender (2) dimensions resulted in 12 face categories, as shown in Fig. 1.

To validate these assignments, we selected 20 faces at random from each of the 12 face categories (240 faces in total) and asked 90 independent observers to classify the intermixed faces according to their perceived age ($N = 30$), race ($N = 30$), or gender ($N = 30$). We found high concordance between observers' classifications and our category assignments for age ($M = 88.04\%$, $SD = 10.58\%$), race ($M = 97.61\%$, $SD = 1.83\%$), and gender ($M = 91.64\%$, $SD = 5.32\%$; see Table S1).

As documented elsewhere, face databases can reflect the demographic distribution of online images rather than the demographic distribution of the global population (Cavazos, Phillips, Castillo, & O'Toole, 2020). Consistent with this observation, faces generated by thispersondoesnotexist.com skewed towards young White appearance. We used 120 images of young female and male White faces to create single-category conditions. In addition, we gathered 30 images of both old female and male White faces, 20 images of both young Black and Asian faces, and 10 images of old, Black, and Asian faces for use in the 2, 4, 6 and 12-category average conditions (see Table A1).

All face images were cropped and resized to 380 × 570 pixels. Image landmarking and image averaging was carried out using InterFace (Kramer, Jenkins, & Burton, 2017). In this software tool, face shape is stored as a set of *xy* coordinates. Calculating the average x and y values for each coordinate across images produces the average shape. Face texture is stored as a matrix of RGB pixel intensities. Calculating the average RGB values for each pixel across images produces the average

Fig. 1. Examples of the 12 categories of computer-generated face images utilised for the construction of face averages, varying by sex (female, male), age (old, young), and race (Black, Asian, White). These categories were employed to construct averages for different heterogeneity conditions, which varied by the number of contributing demographic categories from 1 to 12. The categories are illustrated in Fig. 2 and comprised of averages from a single category of images (1F = White young female; 1M = White young male), two categories (2C = White young male and female), four categories (4C = White young and old x male and female), six categories (6C = Black, Asian, White x young male and female), and 12 categories (12C = Black, Asian and White x young and old x male and female).

texture. The final average face representation is produced by morphing the average texture to the average shape.

Using this method, 288 average faces were created from 12 different NOCI (2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 30, 40). Averages were generated from a selection of 120 distinct identities, and consisted of the face region only, against a black background. We included six conditions (see Table A1), which varied according to the number of face categories represented within each image pool. For example, two single-category conditions were included that contained a pool of 120 faces drawn from the same face category (e.g., either all young White females or young White males), while the twelve-category condition contained a pool of 120 faces: 10 drawn from each of the twelve face categories. The other four conditions captured two (gender), four (age and gender), and six categories (race and gender), respectively. Fig. 2 shows examples from each face category.

To create stimulus displays for the experiment, the average images were paired as follows. For each NOCI, twelve face pairs were generated by shuffling the image pool and dividing this into halves (60 identities each). The left-side average in a face pair was then created from the first half, and the right-side average from the second half, using the number of identities determined by the NOCI. In this manner, 144 face pairs (12 for each of the 12 NOCI) were constructed from a total of 288 face averages (corresponding to two averages for each of the 144 pairs). This process was repeated for each of the six face category conditions (1F, 1M, 2C, 4C, 6C, 12C), giving a total of 864 trials. Therefore, each pair was comprised of two average images generated from an equal number of randomly selected constituent images (NOCI), which were drawn from the same composition of categories (e.g., young White females) but from identities that did not overlap.

Fig. 2. Example average faces in Experiment 1. Columns depict averages constructed from different number of constituent images (NOCI), increasing from left to right, from averages created from 2 to 40 NOCI. Rows depict different heterogeneity conditions, with the number of demographic categories contributing to each average increasing top to bottom (see Table A1). 1F = female single-category; 1M = male single-category; 2C = two categories (gender); 4C = four categories (age and gender); 6C = six categories (gender and race); 12C = twelve categories (age, gender, and race).

2.1.3. Procedure

Each trial began with a 1-s fixation cross, followed by a pair of face averages, which remained onscreen until a response was registered. Averages were displayed at 50% size (190 × 285 pixels), 120 pixels to the left and right of the screen centre. Participants were instructed to press 'S' if they perceived the face images as depicting the same identity or 'D' for different identities. Trials from all six conditions were presented in a fully randomised order, with the option of a break every 144 trials. The experiment was conducted online using Inquisit 6.1 (Inquisit (6.1), 2020). To monitor data quality, we added 36 extra trials as attention checks. In these extra trials, the face averages were presented upside-down. Participants were instructed to press the space bar whenever they saw inverted faces instead of making the usual identity judgement. All participants met our inclusion criterion of 30 or more correct attention checks (M = 35.23, SD = 1.12).

2.1.4. Transparency and openness

All methodological procedures and planned analyses for Experiments 1–3 were preregistered on the Open Science Framework (OSF; see osf.io/z58ay). All data and materials have been made publicly available on OSF (see osf.io/k9dnm). The code used to analyse these data are available from the corresponding author upon reasonable request.

2.2. Results

To assess the impact of NOCI and face category on observers' perceived similarity of face pairs, we calculated the proportion of trials classified as depicting the same person, separately for each condition. As can be seen from Fig. 3a, the proportion of same-identity responses increased with NOCI and differed between category conditions. As NOCI increased from 2 to 20, 'same' responses in all conditions increased from below 20% to above 60% then remained high. For all NOCI, 'same' responses were higher for homogenous conditions (constructed from fewer face categories) than for heterogenous conditions (more categories) – by a margin of up to 30%.

A 6 (Category: 1F, 1M, 2C, 4C, 6C, 12C) × 12 (NOCI: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 30, 40) repeated-measures ANOVA revealed a main effect of NOCI, $F(11,429) = 335.08$, $p < .001$, $\eta_p^2 = 0.90$, with more 'same' responses for high NOCI, a main effect of Category, $F(5,195) = 106.84$, $p < .001$, $\eta_p^2 = 0.73$, with more 'same' responses for homogenous sets, and an interaction between these factors, $F(55,2145) = 6.91$, $p < .001$, $\eta_p^2 = 0.15$.

Fig. 3b shows a breakdown of this interaction, summarising pairwise comparisons for adjacent NOCI in each variability condition (Holm-Bonferroni corrected; full statistical analyses are presented in supplementary materials). Beyond NOCI 10, significant differences between adjacent NOCI were rare, particularly for homogeneous conditions. By NOCI 20, 'same' responses had reached asymptote, such that adding further images had little effect on performance. These observations suggest that cross-identity averages had reached perceptual stability with as few as 10–20 images.

Fig. 3c shows a similar breakdown for adjacent variability conditions in each NOCI group. In general, averages constructed from homogenous faces were more likely to elicit 'same' responses than averages constructed from heterogenous faces. Such differences persisted at the highest NOCI. For example, 'same' judgements for the 12C condition were lower than in the 1F, 1M, and 2C conditions even at NOCI 40. Thus, cross-identity averages did not converge, even when they had reached perceptual stability. Averages based on heterogenous faces were less likely to be seen as the same person.

To quantify the trend of 'same' judgements across NOCI, we fitted polynomials of degree 0 through 4 (constant, linear, quadratic, cubic, quartic) to 'same' responses across NOCI, collapsed across condition. We then compared the fit of these models using stepwise least-squares regression. As can be seen in Table 1, adding linear and quadratic terms substantially improved fit. There was also a small reduction in the residual sum of squares after adding a third-order polynomial term. This indicates that the increase in perceptual similarity with increasing NOCI was nonlinear, with a steep rise in participants' 'same' judgements across early NOCI.

2.3. Discussion

As the number of faces contributing to each average increased from 2 to 20, so did participants' tendency to perceive the averages as the same person. From 20 to 40, however, similarity judgements plateaued. This nonlinear pattern was evident from our curve-fitting analysis. Identity decisions were also influenced by the number of demographic categories contributing to the averages. For most NOCI, averages composed of demographically varied faces were less likely to be seen as the same person than averages composed of homogenous faces. Yet even 12-category averages stabilised by NOCI 20, such that incorporating further images produced no further change in responses.

3. Experiment 2

Experiment 1 investigated the *perceived* similarity of cross-identity averages. Participants' 'same' judgements of pairs of averages increased with NOCI number, before stabilising by NOCI 20. Here we investigated whether these increases in similarity judgements might be related to the *physical* similarity of test images, by conducting an image analysis of the experimental materials. Principal Components Analysis

Table 1
NOCI Model Fit Parameters for Experiment 1.

Model	RSS	AIC	R2	F	p
Intercept	0.663	3.31	–	–	–
Linear	0.044	−27.24	0.934	140.68	<.001
Quadratic	0.009	−43.91	0.986	33.66	<.001
Cubic	0.004	−53.10	0.995	12.30	.008
Quartic	0.004	−51.27	0.995	0.11	.75

Note: RSS = residual sum of squares, AIC = Akaike information criterion.

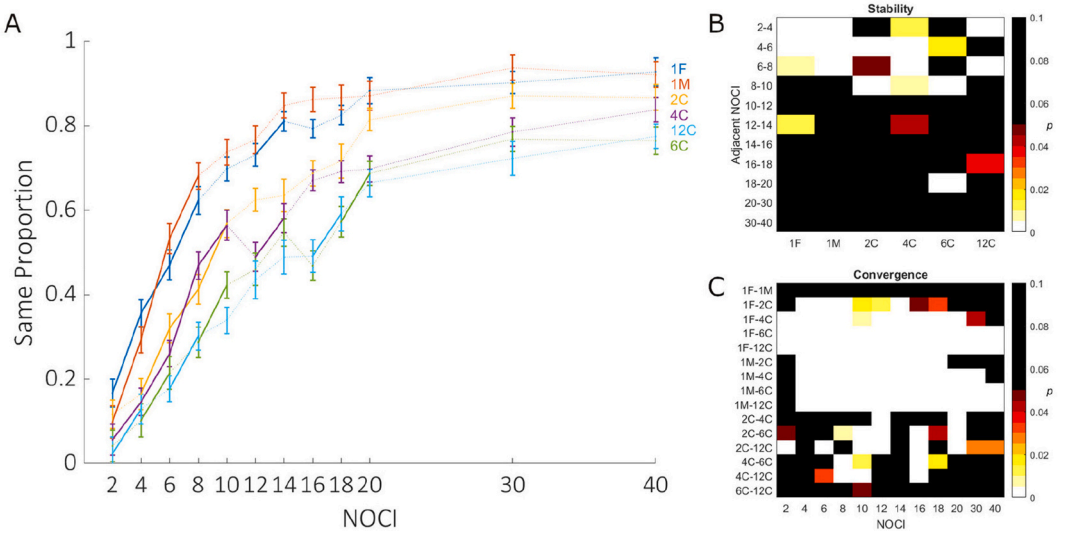

Fig. 3. Average proportion of 'same' decisions across images constructed from two to 40 NOCI for all six conditions (1F, 1M, 2C, 4C, 6C, and 12C) in Experiment 1 (A). Error bars represent within-subjects variability via 95% Cousineau-Morey confidence intervals (see Baguley, 2012), and solid lines represent significant adjacent NOCI comparisons (Holm-Bonferroni corrected). Two-tailed *p*-values for adjacent NOCI (B; 'Stability') and condition (C; 'Convergence') comparisons are presented as heatmaps (degrees of freedom = 39, black denotes all non-significant comparisons where $p \geq .05$). There were 396 set-size comparisons, and 180 condition comparisons in total.

(PCA) was used to measure the variability of average images, with the expectation that image variability would decrease as more faces contribute to each average.

3.1. Methods

To determine the physical similarity of averages, we randomly generated 12 averages per NOCI from the pool of face images used in the participant-based analysis (to model trial number). PCA was then performed on each set of 12 averages (see Burton et al., 2005). To measure variance across averages, the standard deviation for texture and shape eigenvectors was computed separately for each NOCI, repeating this process 32 times with newly generated averages (to model participant number). To arrive at a single *variance score* for each NOCI, the mean of the standard deviations was calculated across iterations. Separate variance values for texture and shape were converted to z-scores before averaging.

3.2. Results

Statistical analysis of the images mirrored the behavioural data of Experiment 1, with variance score substituting for proportion of 'same' judgements in the perceptual task. As Fig. 4a shows, variance score declined sharply over initial increases in NOCI, then gradually levelled off. Analogous to the behavioural analysis, variance scores were higher in the most heterogenous condition than in the least heterogenous conditions, though the graded effect of heterogeneity was less clear.

A 6 (Category) × 12 (NOCI) repeated-measures ANOVA revealed a main effect of NOCI, $F(11, 341) = 3637.23$, $p < .001$, $\eta_p^2 = 0.97$, with less variability at higher NOCI, a main effect of Category, $F(5, 155) = 99.24$, $p < .001$, $\eta_p^2 = 0.76$, with less variability for homogenous sets, and an interaction between these two factors, $F(55, 1705) = 3.11$, $p < .001$, $\eta_p^2 = 0.09$. Pairwise comparisons for adjacent NOCI (Fig. 4b) and adjacent variability conditions (Fig. 4c) indicate substantial stability among cross-identity averages (non-significant differences between adjacent pairs) with as few as 10–20 images.

3.3. Discussion

The PCA analysis converges with participants' similarity judgements in Experiment 1, with a sharp decline in variability over initial increases in NOCI, followed by a more gradual reduction across higher NOCI. This analysis suggests that perceived similarity relates to differences in physical variability across average images. Interestingly, the PCA analysis also indicated greater physical similarity across demographic category conditions than might be predicted by participants' identity decisions alone. In Experiment 1, averages with heterogenous source images were perceived as less similar than averages with homogenous source images, even at the highest NOCI. Experiment 3 examines why category conditions did not converge in the behavioural data.

4. Experiment 3

In Experiment 1, participants encountered all face category conditions intermixed. Considering that Experiment 2 found few differences in image variability between category conditions, here we ask whether intermixed presentation might have reduced convergence. When participants judge each pair of images in the context of all experimental conditions, their judgements can be informed by other conditions. For example, averages from the 12C condition might look more dissimilar when seen among averages from homogenous conditions (e.g., 1M, 1F, 2C, 4C) than when seen among other 12C averages. Such context effects are commonly reported for other perceptual judgements (Marshall, Lazar, Krakauer, & Sharma, 1998; Schneider & Parker, 1990). Here we tested for similar effects for face averages, by presenting category conditions in a blocked design. If intermixed presentation reduces convergence, then such convergence should now be stronger.

4.1. Methods

All methodological procedures, sample size, and planned analyses were preregistered on the Open Science Framework (OSF; see osf.io/z58ay), and the data for these experiments are also available on OSF (osf.io/k9dnm). Participants were recruited online via Prolific in 2020, using the same inclusion criteria as in Experiment 1. One participant who failed the attention check (20/36) was replaced, resulting in a sample size of 40 (24 females, 16 males; Age $M = 28.88$, $SD = 8.06$). This sample size has proven sufficient in Experiment 1 and previous studies in the face-matching (e.g., Fysh, 2018; Young et al., 1985) and categorical perception literature (e.g., Calder et al., 1996; Peng et al., 2010). The stimuli and procedure were the same as for Experiment 1, except that the six face category conditions (1M, 1F, 2C, 4C, 6C, 12C) were now

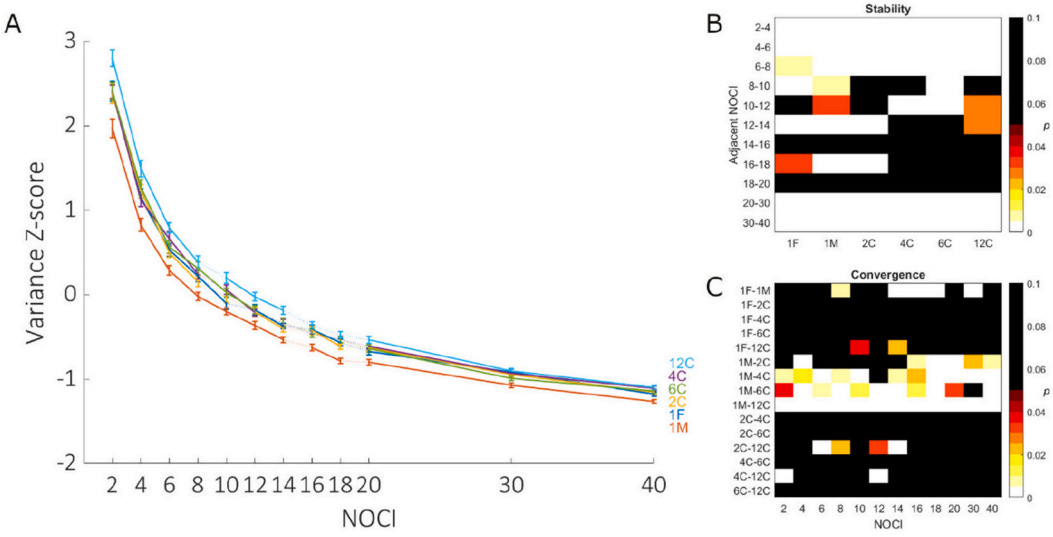

Fig. 4. Image variance scores across NOCI for all six conditions in Experiment 2 (A). Error bars represent within-subjects variability via 95% Cousineau-Morey confidence intervals (see Baguley, 2012), and solid lines represent significant adjacent NOCI comparisons (Holm-Bonferroni corrected). Two-tailed *p*-values for adjacent NOCI (B; 'Stability') and condition (C; 'Convergence') comparisons are presented as heatmaps (degrees of freedom = 31, black denotes all comparisons where $p \geq .05$).

presented in separate blocks rather than being intermixed. The order of trials within each block was randomised, and block order was counterbalanced across participants.

4.2. Results

As in previous experiments, the data were analysed with a 6 (Category) x 12 (NOCI) repeated-measures ANOVA, which revealed main effects of Category, $F(5,195) = 10.27$, $p < .001$, $\eta_p^2 = 0.21$, and NOCI, $F(11,429) = 250.89$, $p < .001$, $\eta_p^2 = 0.87$, and an interaction, $F(11,2145) = 4.996$, $p < .001$, $\eta_p^2 = 0.11$. A graphical summary of the statistical comparisons for NOCI and face category is provided in Fig. 5a and b. As can be seen from these data, all face categories became increasingly stable with higher NOCI, with only sporadic differences between NOCI 20 and 30 (the 12C condition) and 30 to 40 (1F). This indicates that stability of face averages was reached generally at these higher NOCI.

Fig. 5c shows the differences between variability conditions, with more homogeneous conditions receiving a greater proportion of 'same' responses for several comparisons up to NOCI 18. Yet by NOCI 20, only the 1M and 4C condition differed, indicating general convergence between all face category conditions from this NOCI onward.

As in Experiment 1, polynomials were also fitted to the mean perceptual similarity scores, collapsed across conditions (see Table 2). This analysis showed that only linear and quadratic terms substantially improved fit, indicating that similarity judgements of face pairs increased rapidly at small NOCI and then levelled off (see Fig. 5).

4.3. Discussion

As in Experiment 1, 'same identity' responses increased with higher NOCI, generally stabilising by NOCI 20. In contrast to that experiment, the different face categories also converged by NOCI 20, in the sense that significant differences between homogenous and heterogeneous sets were now few. This new result more closely reflects the physical variability of average images (see Experiment 2) and underscores the core finding from the preceding experiments: perceptually stable and robust cross-identity face averages can be formed from a surprisingly small number of faces. This observation holds for averages constructed from a single demographic category and for averages constructed from multiple demographic categories.

Table 2
NOCI Model Fit Parameters for Experiment 4.

Model	*RSS*	*AIC*	*R2*	*F*	*p*
Intercept	0.678	3.58	–	–	–
Linear	0.012	−42.64	0.982	546.17	<.001
Quadratic	0.004	−53.23	0.994	16.68	.003
Cubic	0.003	−55.14	0.996	3.09	.117
Quartic	0.002	−56.69	0.997	2.40	.165

Note: RSS = residual sum of squares, AIC = Akaike information criterion.

5. Experiment 4

Given the variability of ambient faces, an effective average detection template could conceivably require a large sample of faces to form. For example, it is estimated that humans know about 5000 faces (Jenkins, Dowsett, & Burton, 2018) and maintain social networks of around 150 individuals (Dunbar, 1993; Dunbar, Arnaboldi, Conti, & Passarella, 2015). In contrast to these estimates, Experiments 1 to 3 found that stabilising a face detection template could require just 10 to 20 examples, depending on demographic heterogeneity. This finding suggests that a general face detection template could form quickly, after exposure to a surprisingly small number of faces (Kramer, Young, Day, & Burton, 2017). This theory is consistent with the observation that infants, who necessarily have limited experience of faces (Sugden, Mohamed-Ali, & Moulson, 2014), are able to rapidly and reliably detect faces embedded in complex natural scenes (Di Giorgio et al., 2012; Kelly et al., 2019; Prunty, Jackson, Keemink Jolie, & Kelly, 2020).

Experiment 4 directly compares detection performance for average versus exemplar faces embedded in naturalistic scenes. To focus specifically on faces, we present these stimuli without body cues to remove any potential influence of the body on detection (Bindemann, Scheepers, Ferguson, & Burton, 2010). Although face averages are not stimuli that are encountered naturally, we reasoned that an average face in the scene should match the putative face template better than an exemplar face in the scene. A closer match should be evident from faster detection times and higher response accuracy for averages than exemplar faces.

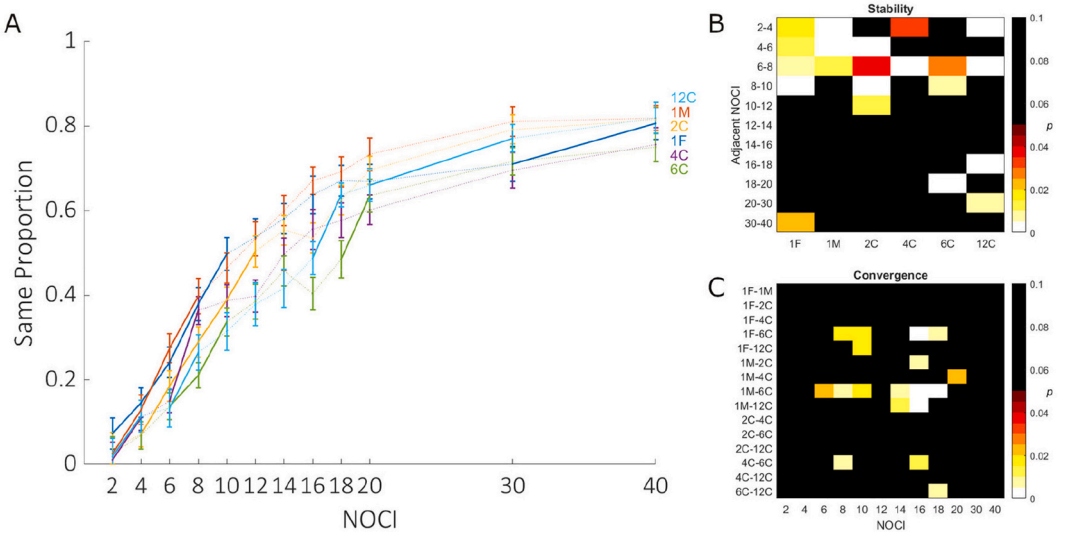

Fig. 5. Average proportion of 'same' decisions across NOCI for all six conditions in Experiment 3 (A). Error bars represent within-subjects variability via 95% Cousineau-Morey confidence intervals (see Baguley, 2012), and solid lines represent significant adjacent NOCI comparisons (Holm-Bonferroni corrected). Two-tailed *p*-values for adjacent NOCI (B; 'Stability') and condition (C; 'Convergence') comparisons are presented as heatmaps (degrees of freedom = 39, black denotes all non-significant comparisons where $p \geq .05$). There were 396 set-size comparisons, and 180 condition comparisons in total.

5.1. Methods

5.1.1. Participants

The data for this experiment is available on OSF (osf.io/bf8yr). Due to the inclusion of sex as a factor, sample size was increased from Experiment 4 to 80 participants comprising of 40 females and 40 males (Age $M = 31.79$, $SD = 8.97$). These were recruited online using Prolific in 2021, 75 identified as White and five as Asian. Participants were pre-screened using the same eligibility criteria as Experiment 1. To ensure the collection of high-quality data online, we included a screen calibration procedure to ensure detection stimuli were displayed at a standard size for all participants. We also included 12 additional trials to measure participants' attention, and set an inclusion threshold of 75% (9 out of 12) correct responses. An additional 36 participants were excluded prior to analysis for failing attention checks ($N = 8$), or for failing to correctly complete screen calibration ($N = 28$).

5.1.2. Design and stimuli

To assess whether humans demonstrate a detection advantage for average over exemplar faces, we presented participants with 288 scenes: 144 contained a face, 144 did not contain a face. Faces varied according to their perceived sex (male or female) and age (young or old), and the number of identities they represented: either a single identity (i.e., an exemplar face) or an average of 40 identities. Together, these formed eight face categories (see Fig. 6). Previous work has distinguished between the detection of faces in the visual periphery, and the classification of faces at fixation (Bindemann & Lewis, 2013), and has encouraged the use of ecologically-valid scene contexts over artificially constructed search grids (Kelly et al., 2019). Accordingly, faces were embedded in complex natural scenes and face location was distributed evenly across all scene regions.

To accomplish this, two hundred computer-generated faces were selected in total, comprising of 50 faces for each of the four demographic categories. All faces were front facing and shared the same perceived ethnicity (i.e., White). Eighteen faces were randomly selected from each demographic category to be used as exemplars. For each of the 18 exemplar images per category, an equivalent average face was constructed by morphing 40 randomly selected faces from that category using InterFace. This formed an image set of 144 face items, with 18 average and 18 exemplar faces per demographic category.

Scene images were sourced from online image repositories (e.g., Unsplash.com, Pexels.com) that provide freely usable images (CC0 license). Fifty scene images were collected for each of six scene categories, reflecting child-centred scenes (e.g., classrooms, playgrounds), garages, homes, offices, restaurants, and shops (for examples, see Fig. 7). These scene images were edited to remove any existing faces or people, and cropped to a standard size (2500 × 1500 pixels). Twenty-four scenes from each of the six scene categories were randomly selected to contain a face (144 in total), the remaining 156 scenes were used in face absent (144) and attention-check (12) trials. To define face locations, the face present scenes were divided into an invisible 4 × 3 grid of 12 regions (500 × 500 pixels each). Locations were counterbalanced across participants by rotating around four region groupings such that each face category had an equal chance of appearing in each scene region. All face items were then cropped to remove the image background, and embedded in the scenes (see Fig. 7).

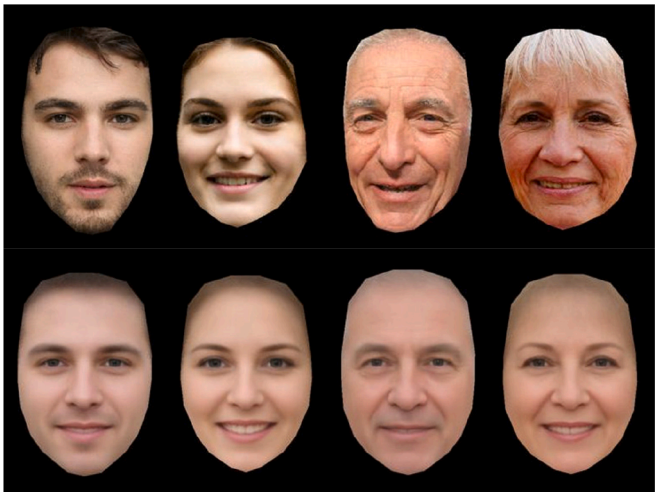

Fig. 6. Examples of the eight face categories of Experiment 4, which varied according to age: young (left) and old (right), sex: male (columns 1 and 3) and female (columns 2 and 4), and type: exemplars (top) and averages (bottom).

5.1.3. Procedure

Online participants first completed a screen calibration procedure in which they were asked to adjust the length of an on-screen line to a standard size (85.6 mm, the length of a credit card). Using this ratio, scene images (144 face present, 144 face absent) were displayed at a standard 21.00 × 15.75 cm size for all participants, regardless of their screen size or resolution. Within those scenes, faces measured 2.00 × 3.00 cm. The stimuli remained on display until a response was registered, and participants were instructed to press 'P' (present) for scenes with faces, and 'A' (absent) if no face was present. After the response, an inter-stimulus interval of one second was applied before the appearance of the next scene image. The 12 attention check trials, consisting of inverted face-absent scenes, were presented at pseudo-random intervals. For these attention-check trials, participants were instructed to press 'Spacebar'. Experimental trials were presented in a fully randomised order, with the option of a break every 72 trials.

5.1.4. Transparency and openness

All data and materials for Experiments 4–5 have been made publicly available on OSF (osf.io/bf8yr). The code used to analyse these data are available from the corresponding author upon reasonable request. The design and analysis plan for these studies were not preregistered.

5.2. Results

To investigate detection performance, we measured participants' accuracy, as indicated by the proportion of correctly classified trials, and response speed, by calculating median response times (RTs) for correct trials. Firstly, we compared face-present and face-absent trials. Consistent with visual search logic (e.g., Eckstein, 2011), accuracy was lower for face-present than face-absent trials ($M_{FP} = 96.35\%$, $M_{FA} = 97.76\%$), $t(79) = 3.73$, $p < .001$, as participants were more likely to miss faces that were present than to find faces in scenes were there were none. Participants also responded faster on face-present than face-absent trials ($M_{FP} = 652$ ms, $M_{FA} = 1044$ ms), $t(79) = 10.60$, $p < .001$, as finding a face effectively terminated their search.

The data of main interest were responses to the different face categories of face-present trials. These data are summarized in Table 3. We compared participants' detection performance for average and exemplar faces with 2 (face type: exemplars, averages) × 2 (face sex: male, female) × 2 (face age: young, old) repeated-measures ANOVAs for accuracy and RTs. For accuracy, this analysis revealed a main effect of face type, $F(1,79) = 10.07$, $p = .002$, $\eta_p^2 = 0.113$, whereby average faces were detected with greater accuracy relative to exemplars ($M_{Av} = 96.86\%$, $M_{Ex} = 95.85\%$). In addition, main effects of face age, $F(1,79) = 9.18$, $p = .003$, $\eta_p^2 = 0.10$, and face sex were found, $F(1,79) = 5.48$, $p = .022$, $\eta_p^2 = 0.07$, and an interaction between the two factors, $F(1,79) = 14.17$, $p < .001$, $\eta_p^2 = 0.15$ (all other interactions, $Fs < 2.0$, $ps > 0.16$). Pairwise comparisons (Holm-Bonferroni corrected) showed that young faces were detected more accurately than old faces, and this effect was specific to male, $t(159) = 4.18$, $p < .001$, but not female faces, $t(159) = 0.77$, $p = .445$ (see Table 3).

Fig. 7. Examples of scenes used in Experiment 4. The large image depicts a young male average face embedded within a 'home' scene. The small images depict examples of all eight face categories in a selection of scenes including home, office, garage, shop, child-centred and restaurant scenes.

Table 3
Mean Detection Response Time (RT) and Detection Accuracy (Acc) for Exemplars and Average faces in Experiment 4.

	Face Type						Difference (E–A)	
	Exemplars			Averages				
Face Category	RT	CI	Acc	RT	CI	Acc	RT	Acc
Combined	663	4	95.85	650	4	96.86	13	−1.01
Young Male	661	9	96.60	643	9	97.85	18	−1.25
Young Female	658	12	96.81	651	11	96.39	7	0.42
Old Male	669	13	93.33	657	10	95.97	12	−2.64
Old Female	666	10	96.67	650	10	97.22	16	−0.55

Note. Within-subjects variability for RTs (CI) are represented via 95% Cousineau-Morey confidence intervals (see Baguley, 2012). RTs are measured in milliseconds, Accuracy in percentages.

The same analysis for RTs also found a main effect of face type, as average faces were detected faster than exemplars, (M_{Av} = 650 ms M_{Ex} = 663 ms), $F(1,79) = 8.90$, $p = .004$, $\eta_p^2 = 0.10$. For RTs, main effects of face sex, $F(1,79) = 0.09$, $p = .766$, $\eta_p^2 < 0.01$, and face age were not found, $F(1,79) = 2.68$, $p = .105$, $\eta_p^2 = 0.03$, and there were no interactions between factors, $Fs < 0.5$, $ps > 0.41$.

Taken together these data show a clear detection advantage for face averages over exemplars, both in terms of the accuracy with which faces are found, and the speed with which they are detected.

Because the detection advantage for average faces was numerically small in accuracy (~1%) and RTs (~13 ms), we investigated whether this could be driven by the repetition of average faces over the course of the experiment compared with the exemplars, which were more varied (and hence less repetitive) in appearance. To examine this, we compared the detection speed of averages and exemplars across time, by calculating RTs on a trial-by-trial basis over the course of the experiment (see Fig. 8). We then correlated detection speed for averages and exemplars with trial order (from 1 to 72, with 72 trials in each of these conditions, collapsed across face sex and age). This analysis showed that response times for average and exemplar faces decreased over the course of the experiment (Fig. 8, left panel), but a detection speed advantage for average over exemplar faces was consistent over time (Fig. 8, right panel), $r(70) = 0.10$, $p = .398$ (accuracy: $r(70) = 0.05$, $p = .667$). The detection advantage of averages over exemplars is therefore unlikely to be driven by stimulus repetition.

It is also possible that the detection advantage for averages was driven by low-level visual properties such as visual saliency (Itti & Koch,

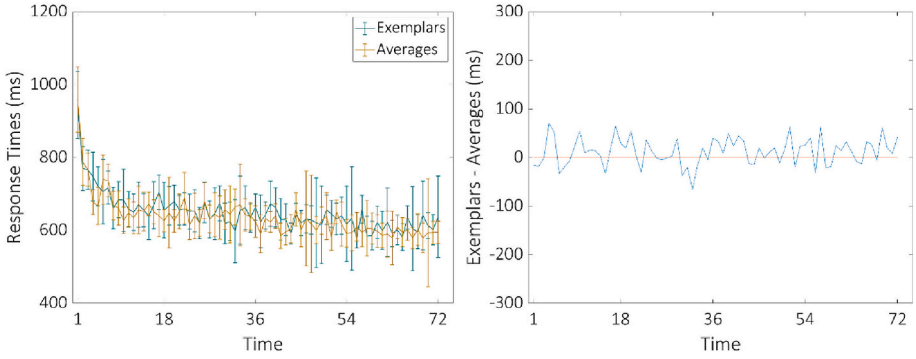

Fig. 8. Mean differences in RTs between exemplar and average faces across time in Experiment 4. Both the mean RTs (left) and the difference between RTs (right) for exemplar and average faces are plotted across time. Error bars represent within-subjects variability via 95% Cousineau-Morey confidence intervals (see Baguley, 2012).

2001; Itti, Koch, & Niebur, 1998). For example, it could be that the relative uniformity of these stimuli against a cluttered scene background renders them better detection targets than more variable exemplar faces. To test this alternative explanation, we compared the visual saliency of average and exemplar faces embedded in scenes, and then examined the impact of saliency differences on detection performance.

We first generated saliency maps for all face-present scenes and their face-absent equivalents, using a software package provided by the 'OpenCV' computer vision library in python (Itseez, 2015; Montabone & Soto, 2010). We then focused our analysis on a 180 × 200 pixel region of the saliency map surrounding the location of the face, and computed the difference in average pixel intensity between the corresponding face-absent and face-present scenes. In this manner, we produced a dataset ($N = 288$) that captured the change in regional saliency following the addition of a face stimulus in the average and exemplar conditions.

This analysis revealed that the resultant change in visual saliency within the target regions was different after the addition of average and exemplar faces, $t(287) = 10.82$, $p < .001$. This was characterised by larger increases in saliency for exemplars ($M = 1.24$) compared with averages, where regional saliency was lower relative to the unaltered scene background ($M = -0.16$). This demonstrates that exemplar faces, with their wider range of idiosyncratic features (e.g., differences in lighting, hairstyle and so forth) form better detection targets than average faces based on low-level visual properties alone. It also suggests that the detection advantage for average faces cannot be explained by differences in visual saliency, which should favour exemplar faces instead. Indeed, correlations between average-exemplar differences in visual saliency and detection performance showed that a reduced saliency advantage for exemplars was related to an increased detection advantage for averages, Accuracy: $r(286) = 0.16$, $p = .005$; RTs: $r(286) = -0.10$, $p = .102$. This emphasises that better performance for average faces was present in spite of, not because of, differences in low-level visual saliency.

5.3. Discussion

Experiments 1 to 3 found that the concept of an average detection template was plausible, in principle. In Experiment 4, we tested this idea by contrasting detection performance for average and exemplar faces embedded in complex natural scenes. We found support for this hypothesis as observers detected average faces with both greater speed and accuracy compared with exemplar faces. The detection advantage for averages was numerically small but consistent over the course of the experiment. Therefore, this advantage cannot be attributed to a simple repetition effect, whereby observers were more likely to detect faces in a homogenous category, such as the averages, compared with the visually more varied exemplars. It also cannot be attributed to low-level visual properties such as visual saliency, as exemplars differed more from the surrounding scene context than averages.

6. Experiment 5

Experiment 4 found that averages were detected more quickly and more accurately than exemplar faces. In Experiment 5 our aim was to replicate this average advantage and test its robustness. In natural settings, detection operates at a range of viewing distances. We therefore examined whether the detection advantage for averages persists when the faces in scenes are reduced in size.

6.1. Methods

The data for this experiment is available on OSF (osf.io/bf8yr). Eighty participants (40 females; Age $M = 30.78$, $SD = 9.44$) were recruited online using Prolific in 2021. Sixty-six participants identified as White, twelve as Asian, and two as Black. An additional 35 participants were removed prior to analysis for failing attention checks ($N = 15$), or for failing to correctly complete screen calibration ($N = 20$).

The experiment design and stimuli were the same as Experiment 4, except for the size of the faces embedded in scenes. In the previous experiment, faces were displayed at a size of 2.00 × 3.00 cm, within scenes that measured 21.00 × 15.75 cm. Faces in this experiment were half that size (1.00 × 1.50 cm), but the size of the scenes did not change (see Fig. 9). The locations of the faces were kept the same, and all other aspects were identical.

6.2. Results

Participants' mean detection performance is summarized in Table 4. These data show that reducing face size increased task difficulty, as participants were slower and less accurate compared with the previous experiment. Again, we found higher accuracy for face absent trials ($M_{FP} = 89.31\%$, $M_{FA} = 98.98\%$), $t(79) = 8.79$, $p < .001$, and faster responses for face present trials ($M_{FP} = 1029$ ms, $M_{FA} = 2335$ ms), $t(79) = 12.00$, $p < .001$. We then analysed participants' detection performance using two 2 (face type) × 2 (face sex) × 2 (face age) ANOVAs for accuracy and RTs.

For accuracy, this analysis revealed a main effect of face type, $F(1,79) = 10.09$, $p = .002$, $\eta_p^2 = 0.113$, whereby average faces were detected with greater accuracy relative to exemplars ($M_{Av} = 90.05\%$, $M_{Ex} = 88.56\%$). A main effect of face age was also found $F(1,79) = 6.84$, $p = .011$, $\eta_p^2 = 0.080$, reflecting improved detection of young ($M = 90.00\%$) relative to old faces ($M = 88.61\%$). A main effect of face sex, $F(1,79) = 0.11$, $p = .743$, $\eta_p^2 = 0.001$, and interactions between factors were not observed, $Fs < 2.1$, $ps > 0.15$.

In contrast, RTs did not show a main effect of face type, $F(1,79) = 1.44$, $p = .234$, $\eta_p^2 = 0.018$, though there was a main effect of face age, $F(1,79) = 4.11$, $p = .046$, $\eta_p^2 = 0.049$, due to the faster detection of young faces. The main effect of face sex was not significant, $F(1,79) = 3.20$, $p = .078$, and no interactions between any of the factors were found, $Fs < 3.0$, $ps > 0.08$.

Overall, these data therefore show that younger faces were detected faster and more accurately than older faces. Of main interest, an advantage for average faces over exemplars was also found, but this was only observed in detection accuracy but not detection speed.

6.3. Discussion

In Experiment 5, we investigated whether the average advantage in face detection would persist for faces presented at a smaller size. The average advantage was statistically significant in accuracy but not in reaction times, possibly due to slower search times overall. The presence of an average advantage in accuracy across both contexts—rapid search for large faces and slower search for small faces—is striking, given that accuracy for exemplars was already high. This advantage can be explained by a closer match between seen target and stored template.

In addition, an age effect was also observed in Experiment 4 and 5, whereby older faces were detected more slowly and less accurately. This appears to be consistent with an average template account for face detection. Older faces appear to be more distinctive than younger-appearing faces that are closer to the average (Deffenbacher, Vetter, Johanson, & O'Toole, 1998). These effects are more pronounced in young and middle-aged perceivers (such as the participants in the current experiments) than in older perceivers (Ebner et al., 2018). In addition, distinctive faces require more time to be classified as faces than typical faces (Valentine & Bruce, 1986). Therefore, the reduced detection of older faces in Experiments 4 and 5 might reflect their distinctiveness - and therefore distance - from an average detection template. Alternatively, these results might also reflect an own-age bias (Anastasi & Rhodes, 2005; Rhodes & Anastasi, 2012) whereby the participants here (with an average age of ~30) were impaired in the detection of faces that were more distal from their own. As these age biases also reflect observers' experience with faces, either of these explanations is consistent with a stored detection template that is based on face

Fig. 9. An illustration of the decrease in face size for Experiment 5. One trial image (young, male average face) from Experiment 4 is shown on the left, and its Experiment 5 equivalent is shown on the right.

Table 4
Mean Detection Response Time (RT) and Detection Accuracy (Acc) for Exemplars and Average faces in Experiment 5.

	Face Type						Difference (E–A)	
	Exemplars			Averages				
Face Category	RT	CI	Acc	RT	CI	Acc	RT	Acc
Combined	1057	11	88.56	1044	11	90.05	13	−1.49
Young Male	1077	30	89.51	1019	28	91.39	58	−1.88
Young Female	1037	27	89.10	1027	28	90.00	10	−0.90
Old Male	1079	29	87.01	1062	32	88.96	17	−1.95
Old Female	1036	27	88.61	1069	27	89.86	−33	−1.25

Note. Within-subjects variability for RTs (CI) are represented via 95% Cousineau-Morey confidence intervals (see Baguley, 2012). RTs are measured in milliseconds, Accuracy in percentages.

averages, and in which faces that are more different from the average are less likely to be detected.

7. Experiment 6

Thus far, we have established the feasibility of an average template for face detection (Experiments 1–3), and have provided the first behavioural evidence in support of this notion (Experiments 4–5). In the preceding experiments, we used participants' detection performance to infer whether the cognitive template underlying detection better resembles a single-identity exemplar face or a multiple-identity average of faces. Although we found a detection advantage for averages over exemplars, it is difficult to separate visual processing in this task from other aspects of cognition (e.g., selective attention, top-down expectations). Without a clean separation, the observed detection advantage could be explained by other factors besides the correspondence between average faces and human detection templates. In Experiments 6 and 7, we will address possible alternative explanations by attempting to reproduce our behavioural findings using a template-matching algorithm. Simulating human detection using straightforward image matching will determine whether target-template similarity *alone* is sufficient to produce the pattern of findings reported, or whether other aspects of human perception are required.

In Experiment 6, we assess the detection performance of averages formed from different numbers of constituent images (NOCI) using the template-matching algorithm. The algorithm functions by comparing the visual similarity of image regions to a template in order to locate the region that provides the best match. This algorithm can therefore be used to simulate detection, as it is able to identify the region of a scene that is most similar to an average face template – if that region contains a face, it can be said to have been 'detected'. Moreover, we can simulate template formation by varying the NOCI within each average face image and noting when detection performance stabilises. Based on the human performance data, we would expect stable average templates to outperform exemplar templates, and that simulated detection accuracy will increase with template NOCI, before plateauing once stability is reached (see Experiments 1 to 3). Conversely, we would expect an exemplar template to be the best match for its own identity target, and that accuracy for this *specific* identity will be reduced with increasing NOCI, as the template representation broadens to include additional face identities (Burton et al., 2005). We would also expect the demographics of constituent face images to affect detection performance. For instance, a narrowly-tuned template that only includes examples from a single demographic group should show better detection performance for faces from the same demographic category, compared with those from a different demographic category (Prunty, Jackson, Keemink Jolie, & Kelly, 2020; Prunty, Qarooni, Jenkins, & Bindemann, 2023; Stein, End, & Sterzer, 2014).

7.1. Methods

7.1.1. Simulating face detection

To simulate human face detection, we used a template-matching algorithm (Kroon, 2021) to determine the location of a face within a larger scene image. The algorithm has been used previously to simulate detection (Prunty, Qarooni, Jenkins, & Bindemann, 2023), and uses the Sum of Squared Differences (SSD; see Di Stefano & Mattoccia, 2003) and Normalised Cross Correlation (NCC; see Kaso, 2018) to identify the region of an image most similar to a template. The algorithm thus 'detects' exemplar faces embedded in scenes using average face images as detection templates. For each scene image, if the algorithm was able to correctly identify the location of the face, we recorded a 'hit', otherwise we recorded a 'miss'. To reiterate, this algorithm was not selected because its mechanism might be a good model of human face detection per se. Instead, we are interested in whether detection is mediated by the visual similarity of targets to average face templates. Our focus is not the algorithm, but the representations it compares (Jenkins & Burton, 2008).

To examine the efficacy of detection templates constructed from faces that are either of the same or of a different category to the exemplars embedded in scenes, we first defined two demographic categories: Category A, which corresponded to young Black female faces, and Category B, which corresponded to young White male faces. When applied to template and scene exemplars, these categories formed four experimental conditions: AA (Black female templates, Black female exemplars), AB (Black female templates, White male exemplars), BA (White male templates, Black female exemplars) and BB (White male

templates, White male exemplars). These comparisons are analogous to two observer groups from different non-diverse societies, detecting faces from their own and the other group's society.

Stimuli for each condition were generated from a set of 300 scenes and a set of 240 exemplar faces (120 per demographic category). For each of the 20 iterations ('participants'), eight face templates (NOCI: 1, 2, 4, 6, 8, 10, 20, 40) and a unique set of detection stimuli ('trials') were generated. To create each stimulus set, 20 exemplars (143 × 214 pixels) from the relevant face category were embedded within 240 randomly selected scenes (40 per scene category; see Fig. 7 and Experiment 4 for further details). Each exemplar appeared once within each of the 12 scene locations. Within matching conditions (AA, BB), one of these exemplars was selected to be the 'specific' target. Face templates were formed by averaging the specific target face with other remaining exemplar faces according to the template's designated NOCI, such that templates with a NOCI of 1 were identical to the specific target face, and a NOCI 40 template would be an average of the specific target face and 39 other face identities. Within mismatching conditions (AB, BA), a random exemplar from the template face category was designated as the specific target. Templates were 143 × 214 pixels, the same size as the faces in scenes. To minimise the influence of the template image background on the matching process, each template image was cropped to the face outline, and the remaining background was filled with gaussian noise.

During the simulation, the template-matching algorithm compared each of the eight NOCI templates to the 240 scene stimuli, and the proportion of hits and misses for the specific target face and the remaining 'generic target' faces (i.e., all other exemplars) were recorded. This process was repeated with different sets of randomly selected target faces and scenes for 20 iterations.

7.1.2. Transparency and openness

All data for Experiments 6–7 have been made publicly available on OSF (osf.io/bf8yr). The code used to conduct these simulations and analyse these data are available from the corresponding author upon reasonable request. The design and analysis plan for these studies were not preregistered.

7.2. Results

The template-matching algorithm's detection performance within each condition (AA, BB, AB, BA) is displayed in Fig. 10. The algorithm's accuracy scores were analysed separately for each condition using 2 (target type: specific and generic) x 8 (NOCI: 1, 2, 4, 6, 8, 10, 20, 40) repeated-measures ANOVAs. For condition AA, ANOVA revealed a main effect of target type, $F(1,19) = 16.09$, $p < .001$, $\eta_p^2 = 0.46$, reflecting greater accuracy for specific ($M = 79.69\%$) compared with generic target faces ($M = 61.80\%$). There was no main effect of NOCI, $F(7,133) = 0.33$, $p = .940$, $\eta_p^2 = 0.02$, but there was an interaction between the two factors, $F(7,133) = 28.71$, $p < .001$, $\eta_p^2 = 0.60$. Pairwise comparisons (Holm-Bonferroni corrected) between target type across NOCI revealed greater accuracy for specific over generic target faces for NOCI 1 to 4, $ts > 3.5$, $ps < 0.02$, but not NOCI 6 to 40, $ts < 2.7$, $ps > 0.07$. Conducting comparisons between adjacent NOCI separately for each target type revealed that for specific targets, only NOCI 1 and 2 differed, $t(19) = 4.09$, $p = .021$, all other $ts < 2.0$, $ps > 0.9$, but for generic targets, accuracy increased across NOCI 1 to 4, $ts > 3.7$, $ps < 0.04$, and from NOCI 6 to 8, $t(19) = 3.84$, $p = .032$, but not at any other adjacent NOCI, all other $ts < 3.4$, $ps > 0.08$.

Conducting the same analysis for the other matching condition, condition BB, we again found a main effect of target type, $F(1,19) = 32.52$, $p < .001$, $\eta_p^2 = 0.63$, reflecting greater accuracy for specific ($M = 82.03\%$) compared with generic target faces (62.89%), and no main

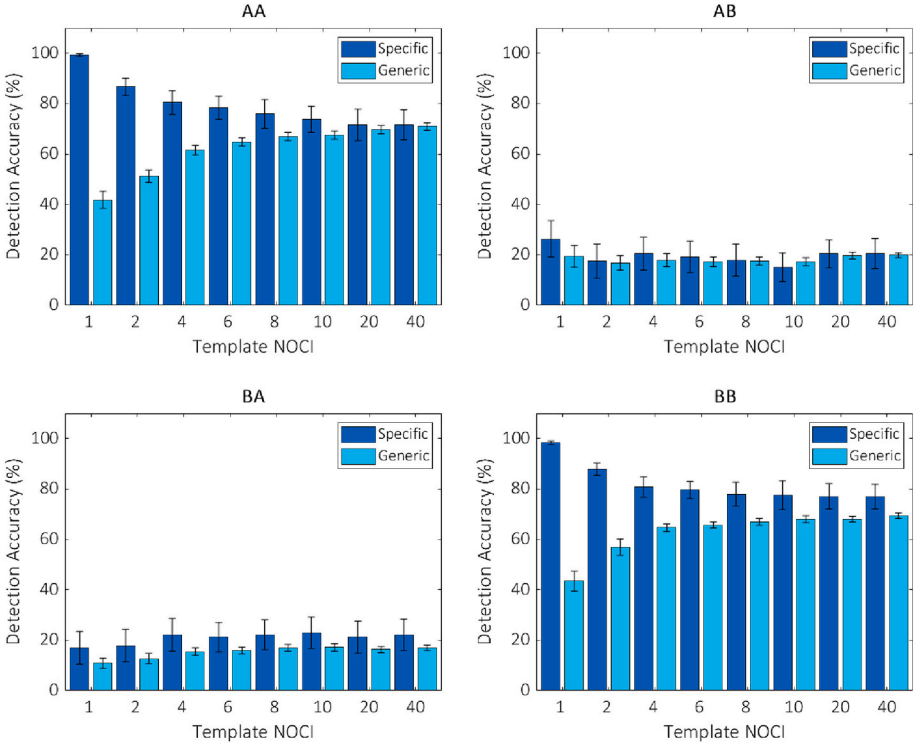

Fig. 10. Mean simulated detection accuracy for specific and generic (i.e., all other) target faces across NOCI. Simulation results are plotted separately for each condition: AA (Black female templates, Black female exemplars), AB (Black female templates, White male exemplars), BA (White male templates, Black female exemplars) and BB (White male templates, White male exemplars). Error bars represent within-subjects variability via 95% Cousineau-Morey confidence intervals (Baguley, 2012), whereby 'participants' reflects 20 iterations of different sets of randomly selected target faces and scenes for each condition (see Method of Experiment 6).

effect of NOCI, $F(7,133) = 0.46$, $p = .863$, $\eta_p^2 = 0.02$. An interaction between the two factors was also found, $F(7,133) = 18.09$, $p < .001$, $\eta_p^2 = 0.49$, with corrected comparisons between target type across NOCI revealing greater accuracy for specific over generic targets for NOCI 1 to 6, $ts > 3.4$, $ps < 0.02$, but not NOCI 8 to 40, $ts < 2.1$, $ps > 0.20$. For comparisons between adjacent NOCI, only the transition from NOCI 1 to 2 was significant within each target type, reflecting an accuracy decrease for specific targets, $t(19) = 4.47$, $p = .012$, and an accuracy increase for generic targets, $t(19) = 4.96$, $p = .004$, all other $ts < 2.6$, $ps > 0.6$.

For the mismatching conditions, the ANOVA for condition AB revealed no main effects of target type, $F(1,19) = 0.09$, $p = .762$, $\eta_p^2 = 0.01$, or NOCI, $F(7,133) = 0.74$, $p = .635$, $\eta_p^2 = 0.04$, and no interaction between the two factors, $F(7,133) = 0.82$, $p = .573$, $\eta_p^2 = 0.04$. For condition BA, ANOVA revealed a main effect of NOCI, $F(7,133) = 3.03$, $p = .006$, $\eta_p^2 = 0.14$, reflecting slight increases in accuracy across the lowest NOCI, though no corrected comparisons between adjacent NOCI reached significance, $ts < 2.1$, $ps > 0.9$. However, ANOVA found no main effect of face type, $F(1,19) = 0.86$, $p = .365$, $\eta_p^2 = 0.04$, and no interaction between factors, $F(7,133) = 0.12$, $p = .997$, $\eta_p^2 < 0.01$.

7.3. Discussion

In Experiment 6, we simulated the formation of the human detection template using a template-matching algorithm. To do this, we constructed average representations to act as detection templates that varied according to their number of constituent face images. Faces were drawn from a single demographic category and thus represented a narrowly tuned template. The results of the simulation indicate that the general detection performance of templates increases rapidly after incorporating additional identities, before plateauing at approximately NOCI 6 to 8. Conversely, we observed an equally rapid decrease in detection performance for specific target faces, as detection templates incorporated an increasing number of additional images. While this demonstrates that a single-identity face image is its own best template, this best-case scenario is also a laboratory artefact as perfect *image* matches between template and target do not occur under naturalistic viewing conditions. Instead, it is an average of multiple face identities that formed the most effective *generic* detection template here. Importantly, this pattern was present only when detecting faces from a matching demographic category. For mismatching categories, the algorithm's detection performance was poor for specific and generic target faces across all NOCI. This suggests that a narrowly tuned template may lead to decrements in our ability to detect faces from other demographic categories (Prunty, Qarooni, Jenkins, & Bindemann, 2023).

8. Experiment 7

Experiments 1 to 3 demonstrated that the number of constituent face identities required to produce a stable template representation was comparatively small (approximately 10 to 20 faces). In a parallel of these experiments, the simulation in Experiment 6 found that the detection performance of average templates also stabilised after incorporating just a small number of constituent images (approximately 6 to 8 faces). In Experiment 7, we will attempt to replicate the findings from Experiments 4 and 5 with further simulations using the same template-matching algorithm as in Experiment 6. That is, we will directly assess the performance of exemplar and average images. In contrast to the template-matching simulations of Experiment 6, which employed average templates to detect exemplar targets, exemplar and average images will therefore serve both as detection templates and as target stimuli embedded in scenes in Experiment 7. In Experiments 4 and 5, human participants detected averages more quickly and more accurately than exemplars. If this finding is driven by the similarity of average faces to the human detection template, then we would expect to find the same pattern using similarity-based image matching in Experiment 7. Specifically, we would expect average template images to show superior detection performance for averages compared with exemplar faces in scenes. In addition, we would also expect average templates to outperform exemplar templates when used to locate exemplar faces in scenes.

8.1. Methods

In this experiment, the same template-matching algorithm used in Experiment 6 was employed here to simulate behavioural Experiments 4 and 5. For this purpose, the 144 face images from the behavioural experiments were utilised as detection templates, and consisted of 40-identity averages and single-identity exemplars (72 each). Template faces were the same size as the faces in scenes, with backgrounds cropped to the face outline and filled with gaussian noise. Scene stimuli were identical to the face-present scenes used in Experiments 4 and 5. Each template was then compared to each scene stimulus in turn, and the ratio of hits and misses for each template was recorded. The entire process was carried out firstly for large faces (Experiment 7a, using stimuli from Experiment 4), then for small faces (Experiment 7b, using stimuli from Experiment 5). Face location and demographics were also counterbalanced across template 'participants', as in the behavioural experiments.

8.2. Results

To analyse the template-matching algorithm's accuracy for detecting exemplar and average faces in scenes, we conducted a 2 (face type: exemplars, averages) × 2 (template type: exemplars, averages) mixed ANOVA for Experiment 7a (large faces) and Experiment 7b (small faces). The results for both sub-experiments are depicted in Fig. 11. For Experiment 7a, ANOVA revealed a main effect of face type, $F(1,142) = 1256.61$, $p < .001$, $\eta_p^2 = 0.90$, reflecting greater accuracy for average faces in scenes ($M = 81.54\%$) relative to single-identity exemplars ($M = 62.04\%$). There was also a main effect of template type, $F(1,142) = 61.33$, $p < .001$, $\eta_p^2 = 0.30$, as average templates ($M = 80.85\%$) were more accurate compared with exemplar templates ($M = 62.73\%$). An interaction of the two factors was also found, $F(1,142) = 11.57$, $p < .001$, $\eta_p^2 = 0.08$. Corrected comparisons (Holm-Bonferroni) indicated differences between face templates were significant for both averages, $t(71) = 6.58$, $p < .001$, and exemplars, $t(71) = 9.39$, $p < .001$, in scenes. Likewise, differences between face type were significant for both average, $t(71) = 47.49$, $p < .001$, and exemplar templates, $t(71) = 20.64$, $p < .001$, though the effect of face type within average templates was stronger (average templates $d = 11.27$, exemplar templates $d = 4.90$).

The algorithm showed a similar pattern of accuracy for small faces (Experiment 7b), though accuracy was reduced overall (see Fig. 11). For Experiment 7b, ANOVA uncovered main effects of face type, $F(1,142) = 1062.57$, $p < .001$, $\eta_p^2 = 0.88$, reflecting higher accuracy for averages ($M_{Av} = 70.74\%$, $M_{Ex} = 49.84\%$), and template type, $F(1,142) = 80.93$, $p < .001$, $\eta_p^2 = 0.36$, as average templates ($M = 71.05\%$) outperformed exemplar templates ($M = 49.53\%$). In Experiment 7b, face type and template type did not interact, $F(1,142) = 2.04$, $p = .155$, $\eta_p^2 = 0.01$. Corrected comparisons indicated that average templates outperformed exemplar templates in finding both average, $t(71) = 8.32$, $p < .001$, and exemplar faces in scenes, $t(71) = 10.56$, $p < .001$. In addition, average faces were found more frequently with both average templates, $t(71) = 32.10$, $p < .001$, and exemplar templates, $t(71) = 19.46$, $p < .001$.

8.3. Discussion

In this simulated experiment, we found that a standard template-matching algorithm was sufficient to replicate the findings from behavioural Experiments 4 and 5, in which average faces in scenes were detected more accurately than exemplars. In contrast to the behavioural

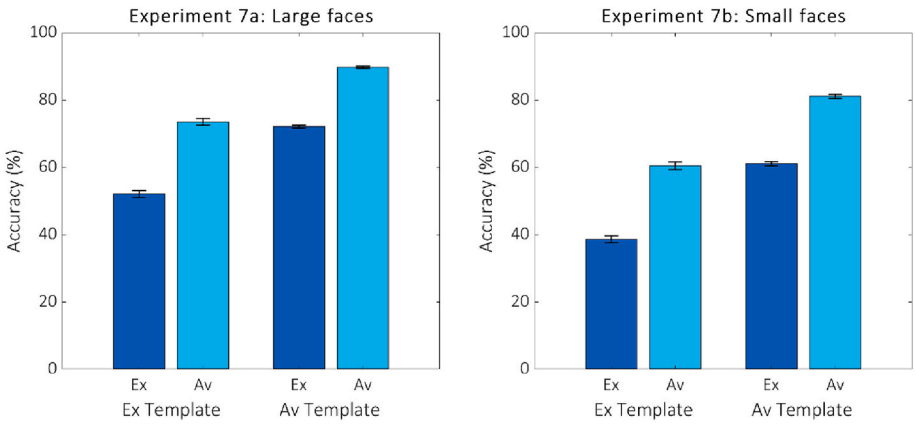

Fig. 11. Mean simulated detection accuracy for Exemplar (Ex) and Average (Av) faces in Experiment 7a and 7b. Simulation results are plotted separately for Exemplar and Average Templates. Error bars represent within-subjects variability via 95% Cousineau-Morey confidence intervals (Baguley, 2012), whereby 'participants' reflects the performance of 144 template faces (72 forty-identity average templates, 72 single-identity exemplar templates).

data from human participants, with this approach we were also able to directly model the performance of average and exemplar face images as templates for detection, by varying the template images used by the algorithm. Accordingly, this showed that average templates produce greater accuracy relative to exemplar templates, regardless of whether they are being matched with average or exemplar faces in scenes. The results of this simulation provide further evidence that a single representation in the form of a multiple-identity average could be utilised as an efficient and generalisable template for face detection.

9. General discussion

Humans can rapidly detect faces in scenes, yet how this is achieved is currently unknown. For detection to succeed, a face in the visual environment must be matched to a suitable mental representation. In this study, we have investigated whether this representation consists of an average of multiple faces. We have investigated this hypothesis across seven experiments, and using three converging approaches. First, in Experiments 1 to 3, we explored the feasibility of an average template, by estimating the number of different faces required for a perceptually stable average to form. Even for averages formed by drawing faces from multiple demographic categories, we found that no more than 20 different identities were required before pairs of averages were reliably perceived as a single identity. Second, in Experiments 4 and 5, we provided evidence for an average template in human observers by measuring their detection performance for single-identity exemplar faces and stable, multiple-identity average faces embedded in complex natural scenes. In these experiments we found observers showed improved detection performance for average faces, despite not encountering such faces in everyday life. Third, in Experiments 6 and 7, we directly modelled the detection performance of average face images using a template-matching algorithm. These simulations showed that the general detection performance of average templates rapidly improves as new faces are added, before reaching asymptote. Further, we found that averages provide a detection advantage over exemplars both as the targets in scenes, and as the template images used for detection.

One advantage of including these simulations alongside behavioural data is that they allow us to test the underlying assumptions of our experiments – namely that human detection performance reflects target-template correspondence, rather than low-level visual properties of averages (e.g., symmetry, uniformity, or smoothness) that may also capture attention. If such low-level properties can solely account for the average advantage, then averages produced from the same number of constituent faces should produce identical performance irrespective of composition. Yet, Experiments 4 and 5 also demonstrate performance differences between averages that were constructed from the same number of faces, but differed in terms of the faces' demographic groups (e.g., young vs. old).

Similarly, we also explored whether the detection advantage for averages is driven by low-level properties, by comparing differences in the visual saliency of average and exemplar faces when these were embedded in scenes (Experiment 4). This revealed that exemplar faces, with their wider variation in appearance, exhibited greater saliency and therefore suggests that the detection advantage for average faces cannot be explained by this factor. Finally, the question arises as to whether the average advantage can be attributed to a simple repetition or adaptation effect, whereby observers improved rapidly in the detection of these homogenous stimuli compared with the visually more varied exemplars. This was examined by comparing the detection of averages and exemplars over the course of Experiment 4. This showed that response times for average and exemplar faces decreased with repetition, but a detection speed advantage for average over exemplar faces was consistent over time. Together, these findings support the theory that an average of previously encountered faces could serve as the basis of a cognitive face detection template. We suggest that this advantage arises because averages capture the characteristics that are representative of faces *in general*, in a normal distribution of the population. This provides the best fit for the widest number of faces, making it an ideal template for face detection.

The role of average face representations in visual perception has been explored extensively in the field of face identification (Burton et al., 2005, 2011; Jenkins & Burton, 2011). For instance, averaging across multiple instances of a single identity has been shown to improve recognition performance in both automated (Jenkins & Burton, 2008) and human (Burton et al., 2005) observers. Such within-identity averages are robust representations for identification, and the process of refining a within-identity average has provided a model for identity learning (Kramer et al., 2015). In a similar manner, we suggest here that the formation of *cross-identity* averages could provide an explanation for the presence and acquisition of detection expertise. This raises the possibility that although face identification and face detection are clearly dissociable tasks that have generated distinct literatures, they might nonetheless be underpinned by a single cognitive mechanism that uses statistical averaging to summarise properties of faces. Recent evidence from the literature on ensemble perception further supports this idea, as it provides evidence that the visual system readily extracts such summaries from sets of faces – across multiple images of a single identity (Kramer et al., 2015; Neumann et al., 2013) and across individual images of multiple identities (de Fockert & Wolfenstein, 2009; Neumann et al., 2013).

The concept of an average template also generates testable predictions regarding the development of the human detection system. The

number of images required to form a stable general face template is surprisingly small when compared with the number of faces that we know (~5000; Jenkins et al., 2018), or maintain within our social networks (~150; Dunbar, 1993; Dunbar et al., 2015), and is instead similar to the number required to form a stable identity representation (~20, Jenkins & Burton, 2008, 2011). Detection expertise should therefore develop early in ontogeny, after only limited social exposure. Accordingly, recent work suggests that six-month-olds' ability to detect faces embedded in complex scenes is comparable to that of adult observers (Kelly et al., 2019; Prunty et al., 2020; Simpson, Maylott, Leonard, Lazo, & Jakobsen, 2019). Our investigation of the formation of an average template also suggests that, although infants may possess a general preference for face-like patterns at birth (Buiatti et al., 2019; Johnson, Dziurawiec, Ellis, & Morton, 1991; Reid et al., 2017), the development of detection expertise should require at least *some* exposure to faces. Consistent with this notion, monkeys deprived of exposure to faces do not develop face-selective cortical domains, show disrupted face orienting, and preferentially attend to the hands, not the faces, of familiar humans (Arcaro, Schade, Vincent, Ponce, & Livingstone, 2017; Sugita, 2008). Future research could test whether the specific visual diet of faces comprising the template impacts detection performance. For example, over-representation of a single demographic group during face learning could tune a detection template to that demographic group, resulting in more selective performance. Furthermore, the template of very young infants might be based predominantly on their primary caregiver, resulting in their preferential detection over unfamiliar faces.

In this study, we have provided the first evidence to suggest the human visual system uses statistical averaging to detect the presence of faces in the visual field. Our work has explored the properties (e.g., stability, robustness, formation) of average face representations and suggests that they convey a detection advantage over exemplars. This finding is counterintuitive in some respects, considering the detection system has developed to detect real faces (i.e., exemplars), not abstract representations of multiple faces. But although average faces are unlikely to be encountered naturally, previous work has demonstrated that the human visual system does routinely extract summary representations from sets of faces, including average representations of multiple identities (de Fockert & Wolfenstein, 2009; Kramer et al., 2015; Neumann et al., 2013). The possibility that such representations could form a cognitive template for detecting faces provides one explanation as to why extracting this information might benefit human vision.

Author note

All methodological procedures and planned analyses for the perceptual stability experiments (Experiments 1-3) were preregistered on the Open Science Framework (OSF; see osf.io/z58ay), but the face detection experiments (Experiments 4-5) and simulations (Experiments 6-7) were not preregistered. The data and materials for the perceptual stability experiments (osf.io/k9dnm) and face detection experiments (osf.io/bf8yr) have also been made publicly available on OSF. Simulated detection data have also been made publicly available (osf.io/bf8yr). The code and materials used to generate these data, as well as all code used to analyse data in this article, are available from the corresponding author upon reasonable request.

CRediT authorship contribution statement

Jonathan E. Prunty: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Rob Jenkins:** Conceptualization, Project administration, Supervision, Writing – review & editing, Funding acquisition. **Rana Qarooni:** Resources, Writing – review & editing. **Markus Bindemann:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Data availability

All experimental data are available on the Open Science Framework. Analysis code is available from the authors upon request.

Acknowledgements

This work was funded by a research grant from the Leverhulme Trust (RPG-2019-085) to Markus Bindemann and Rob Jenkins. Authors have no competing interests to declare.

Appendix A

Table A1
Design and stimuli for Experiment 1.

	Dimensions	Stimuli	No. of categories	No. of identities per category
Condition 1: '1F'	Single-category female	Young, White females	1	120
Condition 2: '1M'	Single-category male	Young, White males	1	120
Condition 3: '2C'	1 dimension: gender	Young, White males and females	2	60
Condition 4: '4C'	2 dimensions: gender and age	Young and old, White males and females	4	30
Condition 5: '6C'	2 dimensions: gender and race	Young, White, Black and Asian males and females	6	20
Condition 6: '12C'	3 dimensions: gender, age, race	Young and old, White, Black and Asian males and females	12	10

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2024.105792.

References

Anastasi, J. S., & Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review, 12*(6), 1043–1047. https://doi.org/10.3758/BF03206441

Arcaro, M. J., Schade, P. F., Vincent, J. L., Ponce, C. R., & Livingstone, M. S. (2017). Seeing faces is necessary for face-domain formation. *Nature Neuroscience, 20*(10), 1404–1412. https://doi.org/10.1038/nn.4635

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods, 44*(1), 158–175. https://doi.org/10.3758/s13428-013-0347-9

Benson, P. J., & Perrett, D. I. (1993). Extracting prototypical facial images from exemplars. *Perception, 22*(3), 257–262. https://doi.org/10.1068/p220257

Bindemann, M., & Lewis, M. B. (2013). Face detection differs from categorization: Evidence from visual search in natural scenes. *Psychonomic Bulletin and Review, 20*(6), 1140–1145. https://doi.org/10.3758/s13423-013-0445-9

Bindemann, M., Scheepers, C., Ferguson, H. J., & Burton, A. M. (2010). Face, body, and center of gravity mediate person detection in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 36*(6), 1477–1485. https://doi.org/10.1037/a0019057

Buiatti, M., Di Giorgio, E., Piazza, M., Polloni, C., Menna, G., Taddei, F., Baldo, E., & Vallortigara, G. (2019). Cortical route for facelike pattern processing in human newborns. *Proceedings of the National Academy of Sciences, 116*(10), 4625–4630. https://doi.org/10.1073/pnas.1812419116

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*(3), 256–284. https://doi.org/10.1016/j.cogpsych.2005.06.003

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., 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

Calder, A. J., Young, A. W., Perrett, D. I., Etcoff, N. L., & Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition, 3*(2), 81–118. https://doi.org/10.1080/713756735

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2020). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science, 3*(1), 101–111. https://doi.org/10.1109/TBIOM.2020.3027269

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the bandwidth of perceptual experience? *Trends in Cognitive Sciences, 20*(5), 324–335. https://doi.org/10.1016/j.tics.2016.03.006

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision, 10*(4), 16. https://doi.org/10.1167/10.4.16

Deffenbacher, K. A., Vetter, T., Johanson, J., & O'Toole, A. J. (1998). Facial aging, attractiveness, and distinctiveness. *Perception, 27*(10), 1233–1243. https://doi.org/10.1068/p271233

Di Giorgio, E., Turati, C., Altoè, G., Simion, F., Di, E., Turati, C., Altoè, G., & Simion, F. (2012). Face detection in complex visual displays: An eye-tracking study with 3- and 6-month-old infants and adults. *Journal of Experimental Child Psychology, 113*(1), 66–77. https://doi.org/10.1016/j.jecp.2012.04.012

Di Stefano, L., & Mattoccia, S. (2003). Fast template matching using bounded partial correlation. *Machine Vision and Applications, 13*(4), 213–221. https://doi.org/10.1007/s00138-002-0070-5

Dunbar, R. I. M. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences, 16*(4), 681–694. https://doi.org/10.1017/S0140525X00032325

Dunbar, R. I. M., Arnaboldi, V., Conti, M., & Passarella, A. (2015). The structure of online social networks mirrors those in the offline world. *Social Networks, 43*, 39–47. https://doi.org/10.1016/j.socnet.2015.04.005

Ebner, N. C., Luedicke, J., Voelkle, M. C., Riediger, M., Lin, T., & Lindenberger, U. (2018). An adult developmental approach to perceived facial attractiveness and distinctiveness. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00561

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision, 11*(5), 14. https://doi.org/10.1167/11.5.14

Elias, E., Dyer, M., & Sweeny, T. D. (2017). Ensemble perception of dynamic emotional groups. *Psychological Science, 28*(2), 193–203. https://doi.org/10.1177/0956797616678188

Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception, 37*(4), 571–583. https://doi.org/10.1068/p5705

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology, 62*(9), 1716–1722. https://doi.org/10.1080/17470210902811249

Fysh, M. C. (2018). Individual differences in the detection, matching and memory of faces. *Cognitive Research: Principles and Implications, 3*(1), 20. https://doi.org/10.1186/s41235-018-0111-x

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*(17), R751–R753. https://doi.org/10.1016/j.cub.2007.06.039

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance, 35*(3), 718–734. https://doi.org/10.1037/a0013899

Hershler, O., & Hochstein, S. (2005). At first sight: A high-level pop out effect for faces. *Vision Research, 45*(13), 1707–1724. https://doi.org/10.1016/j.visres.2004.12.021

Inquisit (6.1). (2020). *Millisecond Software*.

Itseez. (2015). Open Source Computer Vision Library. https://github.com/itseez/opencv.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194–203. https://doi.org/10.1038/35058500

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259. https://doi.org/10.1109/34.730558

Jenkins, R., & Burton, A. M. (2008). 100% accuracy in automatic face recognition. *Science, 319*(5862), 435. https://doi.org/10.1126/science.1149656

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., Burton, A. M., & White, D. (2006). Face recognition from unconstrained images: Progress with prototypes. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)* (pp. 25–30). https://doi.org/10.1109/FGR.2006.45

Jenkins, R., Dowsett, A. J., & Burton, A. M. (2018). How many faces do people know? *Proceedings of the Royal Society B, 285*(1888), 20181319. https://doi.org/10.1098/rspb.2018.1319

Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition, 40*(1–2), 1–19. https://doi.org/10.1016/0010-0277(91)90045-6

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *ArXiv Preprint* (ArXiv:1710.10196v3).

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4401–4410*. https://doi.org/10.1109/CVPR.2019.00453

Kaso, A. (2018). Computation of the normalized cross-correlation by fast Fourier transform. *PLoS One, 13*(9), Article e0203434. https://doi.org/10.1371/journal.pone.0203434

Kelly, D. J., Duarte, S., Meary, D., Bindemann, M., & Pascalis, O. (2019). Infants rapidly detect human faces in complex naturalistic visual scenes. *Developmental Science, e12829*. https://doi.org/10.1111/desc.12829

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., Ritchie, K. L., & Burton, A. M. (2015). Viewers extract the mean from images of the same person: A route to face learning. *Journal of Vision, 15*(4), 1. https://doi.org/10.1167/15.4.1

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

Kroon, D.-J. (2021). *Fast/robust template matching*. MATLAB Central File Exchange.

Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature, 442*(7102), 572–575. https://doi.org/10.1038/nature04951

Lewis, M. B., & Edmonds, A. (2005). Searching for faces in scrambled scenes. *Visual Cognition, 12*(7), 1309–1336. https://doi.org/10.1080/13506280444000535

Marshall, R. S., Lazar, R. M., Krakauer, J. W., & Sharma, R. (1998). Stimulus context in hemineglect. *Brain, 121*(10), 2003–2010. https://doi.org/10.1093/brain/121.10.2003

Montabone, S., & Soto, A. (2010). Human detection using a mobile platform and novel features derived from a visual saliency mechanism. *Image and Vision Computing, 28*(3), 391–402. https://doi.org/10.1016/j.imavis.2009.06.006

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition, 128*(1), 56–63. https://doi.org/10.1016/j.cognition.2013.03.006

Peng, G., Zheng, H.-Y., Gong, T., Yang, R.-X., Kong, J.-P., & Wang, W. S.-Y. (2010). The influence of language experience on categorical perception of pitch contours. *Journal of Phonetics, 38*(4), 616–624. https://doi.org/10.1016/j.wocn.2010.09.003

Perrett, D. I., May, K. A., & Yoshikawa, S. (1994). Facial shape and judgements of female attractiveness. *Nature, 368*(6468), 239–242. https://doi.org/10.1038/368239a0

Prolific. (2021). *Prolific*.

Prunty, J. E., Jackson, K. C., Keemink Jolie, R., & Kelly, D. J. (2020). Caucasian infants' attentional orienting to own-and other-race faces. *Brain Sciences, 10*(1), 53. https://doi.org/10.3390/brainsci10010053

Prunty, J. E., Qarooni, R., Jenkins, R., & Bindemann, M. (2023). Ingroup and outgroup differences in face detection. *British Journal of Psychology, 114(S1), 94-111*. https://doi.org/10.1111/bjop.12588

Qarooni, R., Prunty, J. E., Bindemann, M., & Jenkins, R. (2022). Capacity limits in face detection. *Cognition, 228, 150227*. https://doi.org/10.1016/j.cognition.2022.105227

Reid, V. M., Dunn, K., Young, R. J., Amu, J., Donovan, T., & Reissland, N. (2017). The human fetus preferentially engages with face-like visual stimuli. *Current Biology, 27*(12), 1825–1828.e3. https://doi.org/10.1016/j.cub.2017.05.044

Rhodes, G., Brennan, S., & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology, 19*(4), 473–497. https://doi.org/10.1016/0010-0285(87)90016-8

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. *Psychological Bulletin, 138*(1), 146–174. https://doi.org/10.1037/a0025750

Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M., White, D., & Burton, A. M. (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. *Cognition, 141*, 161–169. https://doi.org/10.1016/j.cognition.2015.05.002

Robertson, D. J., Kramer, R. S. S., & Burton, A. M. (2015). Face averages enhance user recognition for smartphone security. *PLoS One, 10*(3), Article e0119460. https://doi.org/10.1371/journal.pone.0119460

Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision, 3*(6), 5. https://doi.org/10.1167/3.6.5

Schneider, B., & Parker, S. (1990). Does stimulus context affect loudness or only loudness judgments? *Perception & Psychophysics, 48*(5), 409–418. https://doi.org/10.3758/BF03211584

Simpson, E. A., Maylott, S. E., Leonard, K., Lazo, R. J., & Jakobsen, K. V. (2019). Face detection in infants and adults: Effects of orientation and color. *Journal of Experimental Child Psychology, 186*, 17–32. https://doi.org/10.1016/j.jecp.2019.05.001

Stein, T., End, A., & Sterzer, P. (2014). Own-race and own-age biases facilitate visual awareness of faces under interocular suppression. *Frontiers in Human Neuroscience, 8*, 582. https://doi.org/10.3389/fnhum.2014.00582

Sugden, N. A., Mohamed-Ali, M. I., & Moulson, M. C. (2014). I spy with my little eye: Typical, daily exposure to faces documented from a first-person infant perspective. *Developmental Psychobiology, 56*(2), 249–261. https://doi.org/10.1002/dev.21183

Sugita, Y. (2008). Face perception in monkeys reared with no exposure to faces. *Proceedings of the National Academy of Sciences, 105*(1), 394–398. https://doi.org/10.1073/pnas.0706079105

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & 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

Sweeny, T. D., & Whitney, D. (2014). Perceiving crowd attention: Ensemble perception of a crowd's gaze. *Psychological Science, 25*(10), 1903–1913. https://doi.org/10.1177/0956797614544510

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A, 43*(2), 161–204. https://doi.org/10.1080/14640749108400966

Valentine, T., & Bruce, V. (1986). The effects of distinctiveness in Recognising and classifying faces. *Perception, 15*(5), 525–535. https://doi.org/10.1068/p150525

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology, 69*, 105–129. https://doi.org/10.1146/annurev-psych-010416-044232

Yamanashi Leib, A., Fischer, J., Liu, Y., Whitney, D., & Robertson, L. (2014). Ensemble crowd perception: A viewpoint invariant mechanism to represent average crowd identity. *Journal of Vision, 14*(8), 26. https://doi.org/10.1167/14.8.26

Ying, H., Burns, E., Choo, A., & Xu, H. (2017). Ensemble representation of facial attractiveness adaptation by rapid serial visual presentation. *Journal of Vision, 17*(10), 840. https://doi.org/10.1167/17.10.840

Ying, H., Burns, E., Lin, X., & Xu, H. (2019). Ensemble statistics shape face adaptation and the cheerleader effect. *Journal of Experimental Psychology: General, 148*(3), 421–436. https://doi.org/10.1037/xge0000564

Ying, H., & Xu, H. (2017). Adaptation reveals that facial expression averaging occurs during rapid serial presentation. *Journal of Vision, 17*(1), 15. https://doi.org/10.1167/17.1.15

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception, 14*(6), 737–746. https://doi.org/10.1068/p140737