Article:

Neumann, Markus F., Schweinberger, Stefan R. and Burton, A. Mike orcid.org/0000-0002-2035-2084 (2013) Viewers extract mean and individual identity from sets of famous faces. Cognition. pp. 56-63. ISSN 0010-0277

https://doi.org/10.1016/j.cognition.2013.03.006

Running Title: average set identity

Viewers extract mean and individual identity from sets of famous faces

Markus F. Neumann[1], Stefan R. Schweinberger[1,2], & A. Mike Burton[3]

[1]Department of General Psychology, University of Jena, Germany

[2]DFG Research Unit Person Perception, University of Jena, Germany

[3]University of Aberdeen, UK

Correspondence to: Markus F. Neumann, School of Psychology, The University of Western Australia; 35 Stirling Highway, Crawley, WA 6009, Australia; Phone +61 8 6488 113; Fax +61 8 6488 1006

E-mail: markus.neumann@uwa.edu.au

Abstract

When viewers are shown sets of similar objects (for example circles), they may extract summary information (e.g., average size) while retaining almost no information about the individual items. A similar observation can be made when using sets of unfamiliar faces: Viewers tend to merge identity or expression information from the set exemplars into a single abstract representation, the set average. Here, across four experiments, sets of well-known, famous faces were presented. In response to a subsequent probe, viewers recognized the individual faces very accurately. However, they also reported having seen a merged 'average' of these faces. These findings suggest abstraction of set characteristics even in circumstances which favour individuation of the items. Moreover, the present data suggest that, although seemingly incompatible, exemplar and average representations co-exist for sets consisting of famous faces. This result suggests that representations are simultaneously formed at multiple levels of abstraction.

Keywords: set representation, ensemble coding, face, identity, averaging

Introduction

"Set representations" have recently attracted increasing research interest. When seeing groups of perceptually similar objects, information such as size, or motion, may be coded via summary statistics in terms of a mean value across exemplars (Albrecht & Scholl, 2010; Chong & Treisman, 2003). Whenever observers can capitalize on redundancy of information – a common observation in structured sets – they can compress this information into a single representation such as the set average (Alvarez, 2011). In a seminal investigation, Ariely (2001) investigated size representations from sets containing differently sized circles. Critically, participants tended to identify a test circle as having been presented when it had a similar size to the mean of the whole set, even when such an item had not been present. Moreover, participants were near chance when they had to choose which of two circles had been presented. Taken together, these findings suggest that i) mean size information was computed and retained for the set and ii) size information of individual set members was unavailable. There are different potential explanations for weak exemplar representations. First, encoding of precise exemplar representations may not routinely occur, or may simply contain too much noise, perhaps due to the lack of focal attention to set exemplars. Alternatively, an individual representation may initially be computed but may then be discarded extremely fast.

Recently, statistical representations have been demonstrated for sets of perceptually complex stimuli, such as faces. When asked to compare the emotional intensity of a single image with that of a set (up to 16 face photographs varying in emotional intensity), participants performed highly accurately (Haberman & Whitney, 2007, 2009). Performance was actually comparable to a control "exemplar" condition, in which participants compared an image with a homogeneous set with constant emotional intensity. Beyond extraction of mean emotion (and gender, see Haberman & Whitney, 2007) information from sets of faces, a similar mechanism may compute the mean *identity* from sets of *unfamiliar* faces. In one recent study (de Fockert & Wolfenstein, 2009), participants initially saw sets containing photographs of 4 unfamiliar faces from different individuals. In a "match" condition, a subsequent single image could either be an exemplar image from the previous set, or an average morph created from the four set images. Strikingly, the set averages (which had never been seen) received more 'present' responses than the (seen) exemplars. The authors concluded that averaging identity information might serve as the "default mode" for generating mental representations from groups of faces.

Given that facial representations should serve person recognition, this is a surprising finding, since mean identity representations should actually *prevent* identification of a specific person in a group. It is relatively straightforward to understand how superficial averaging of abstract shapes might take place in the visual system, but much harder to account for averaging over such high-level characteristics as someone's identity. For this reason, it is important to note that the authors used unfamiliar faces. Crucially, unfamiliar face recognition is strongly image-dependent and sensitive to superficial picture similarity (Bruce et al., 1999), and is thus based on very different mechanisms than familiar face recognition. For example, viewers are very good at matching different images of a familiar person, but very poor at matching unfamiliar faces (Bruce, Henderson, Newman, & Burton, 2001; Burton, Bruce, & Hancock, 1999;

Kemp, Towell, & Pike, 1997; Clutterbuck & Johnston, 2004). This discrepancy suggests a qualitative difference in perception of familiar and unfamiliar face identities (Hancock, Bruce, & Burton, 2000), which may also have consequences for the interpretation of the identity set averaging data. Accordingly, increased percentages of “present” responses to matching averages in the study of de Fockert and Wolfenstein (2009) could reflect *image* averaging across similar pictures, rather than *identity* averaging. If viewers are failing to differentiate between the unfamiliar people shown to them, they might plausibly construct a set average combining these images. So, while this study certainly demonstrates set averaging for a class of high-level stimuli (faces), we argue that evidence for *identity* set averaging would be much more compelling if it could also be shown to exist for familiar faces sets.

Another important characteristic of previous studies examining set averaging for faces was relatively small image variability within sets. For instance, set averaging for facial expressions was generally investigated by assembling sets from a single identity, using slightly different emotional intensities from a morph continuum between two veridical expressions (Haberman, Harp, & Whitney, 2009; Haberman & Whitney, 2007, 2010). One study on set *identity* averaging actually involved 4 true set photographs, but had sets deliberately arranged to comprise similar identities (de Fockert & Wolfenstein, 2009). Therefore, low recognition rates for set exemplars may have originated from participants being unable to differentiate between exemplars at encoding. It is important to see if the use of more naturally diverse sets could increase exemplar memory, and whether this would in turn affect the quality and strength of set representations.

In sum, previous studies have investigated set averaging using face sets that varied little on either identity or image properties. In the present study, we tested facial identity averaging by using diverse pictures from highly familiar identities, for which participants have rich pre-existing mental representations. We further encouraged identity processing for half of the participants by instructing them to indicate whether a specific *person* had been seen in a set of faces, while the other half indicated whether a specific *image* had occurred. We expected that set averaging would be strongly reduced or absent for highly familiar faces, and that performance would reflect accurate representation of exemplars instead; Since viewers know these identities, and faces in the set were quite diverse, there appears to be no advantage in averaging across them.

Material and methods

The present article includes 4 experiments that share the following aspects. Stimuli were 240 original faces collected from various internet sources, 10 each from 24 well-known celebrities (12 German and 12 International). Sixty gender-homogeneous sets were created from these photographs, each containing 4 images of different identities. Images contributing to a set were chosen to be roughly similar with respect to head angle and gaze direction. Five sets from 12 different identity combinations were assembled. Note that as a result of obtaining the images from the internet, image variation within the sets was large. All images were taken under entirely non-standardized conditions, causing considerable variation on image parameters such as lighting. Additional *set averages* were created for each of the 60 sets by morphing across the respective 4 set images. Image size was 247 x 387 pixels, all images were presented grey-scaled and fitted in an oval mask, excluding most of the hair.

Set displays contained 4 images randomly assigned to 4 specified positions on the screen (cf. Fig. 1), and were presented for 1500 ms. Immediately following the set display (ISI = 0), probe images were displayed for 500 ms, in smaller size than the set images (200 x 300 pixels). Participants used both index fingers to indicate via button press (“f” and “j” on a standard German keyboard) whether or not the probe image had been present in the previous S1 set. Probe images were: i) a set exemplar (i.e., an image from the previous set); ii) a new exemplar of one of the 4 identities of the previous set; iii) a new exemplar of a different familiar identity; iv) the average of the 4 set images; v) the average of 4 different images of the set identities; or vi) the average of 4 images of different familiar identities.

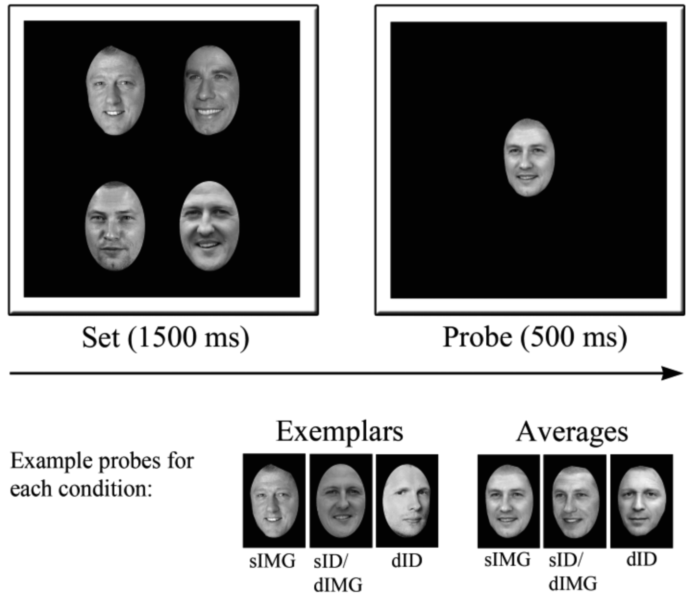

Fig. 1: Example of a set, followed by a probe (sIMG average). Sets were presented simultaneously in Experiments 1-3, and sequentially in Experiment 4. Celebrities in the example set depict (top left to bottom right): Bill Clinton, John Travolta, Till Schweiger (German actor), and Michael Schumacher (German race car driver). Examples for all probe conditions of this set are given below.

In each of these six conditions, 60 trials were presented, with 10 trials per condition in each of 6 experimental blocks. Response button assignment for "present" and "absent" was counterbalanced across participants. A blank screen for 2200 ms allowed for a total response window of 2700 ms.

Experiments were run in two versions, varying in task requirements. Version a) required participants to indicate whether a particular *image* had been a set member, whereas version b) required participants to match *identity* (i.e., whether a person had been a set member). Participants in version a) were explicitly informed that a different image for one of the set identities could occur as a probe stimulus and were instructed to respond "absent" in this case. Overall, 84 young adult participants (mean age = 22.01, *SD* = 3.38; 19 male) were tested and received monetary compensation or course credit. Participants gave written informed consent and reported normal or corrected-to-normal visual acuity. Experiments 1a and 1b each comprised 18 participants, and all remaining experiments (2a, 2b, 3a, 3b, 4a, 4b) comprised 8 participants each.

Experiment 1 - main study

Method

Experiments 1a and 1b followed the procedure laid out above, differing only in the response required by participants (image-present, or person-present). These and subsequent experiments followed a 2 (Probe Type) x 3 (Match Type) design. Probe types were either exemplars (i.e., original images), or set averages. Match Type referred to the relation of the probe face to the set images in that it involved either one, or an average of all i) image(s) from the set (sIMG), ii) different image(s) from the same set identities (sID/dIMG), or iii) image(s) of different identities (dID).

Prior to the experiment proper, participants were given 24 practice trials, and provided with trial-by-trial feedback on accuracy. Note that the correct answer to average probes is always "absent". In order to prevent participants from learning this association, averages were not presented in the practice phase. In order to assess familiarity of the identities used, new pictures of the 24 celebrities were shown following the main procedure in Experiment 1b. Participants were presented images consecutively in the middle of the screen for an unlimited duration, and for each face they indicated by button press whether or not they were familiar with the person. For a "familiar" response, participants were additionally asked to indicate the name, or if they were unable to do so, some identifying semantic information for that person (i.e., occupation, nationality).

Results

Face familiarity task

Written responses from one participant were unavailable. Overall, recognition performance was high, and nearly all the celebrities used in the Experiment could be spontaneously named. On average, celebrities were successfully identified by unique semantic information or name in 92.4% of the cases.

"Present" responses to probe faces

Figure 2 (row 1) shows the proportion of present responses for each of the probe types in Experiment 1 for the image matching (left) and the identity matching group (right). First, and as expected, participants performed very accurately on probe exemplars. Proportions of "present" responses during both tasks were clearly largest for sIMG conditions, indicating good matching performances when a probe image was identical to one of the set images. New identities in dID conditions received few "present" responses overall, i.e., false positives were rare. During image matching, new images from one of the set identities (sID/dIMG) were rejected quite accurately, but less well than dID images. During identity matching, "present" hits to sID/dIMG images were frequent, though reduced compared to present responses to identical images (sIMG).

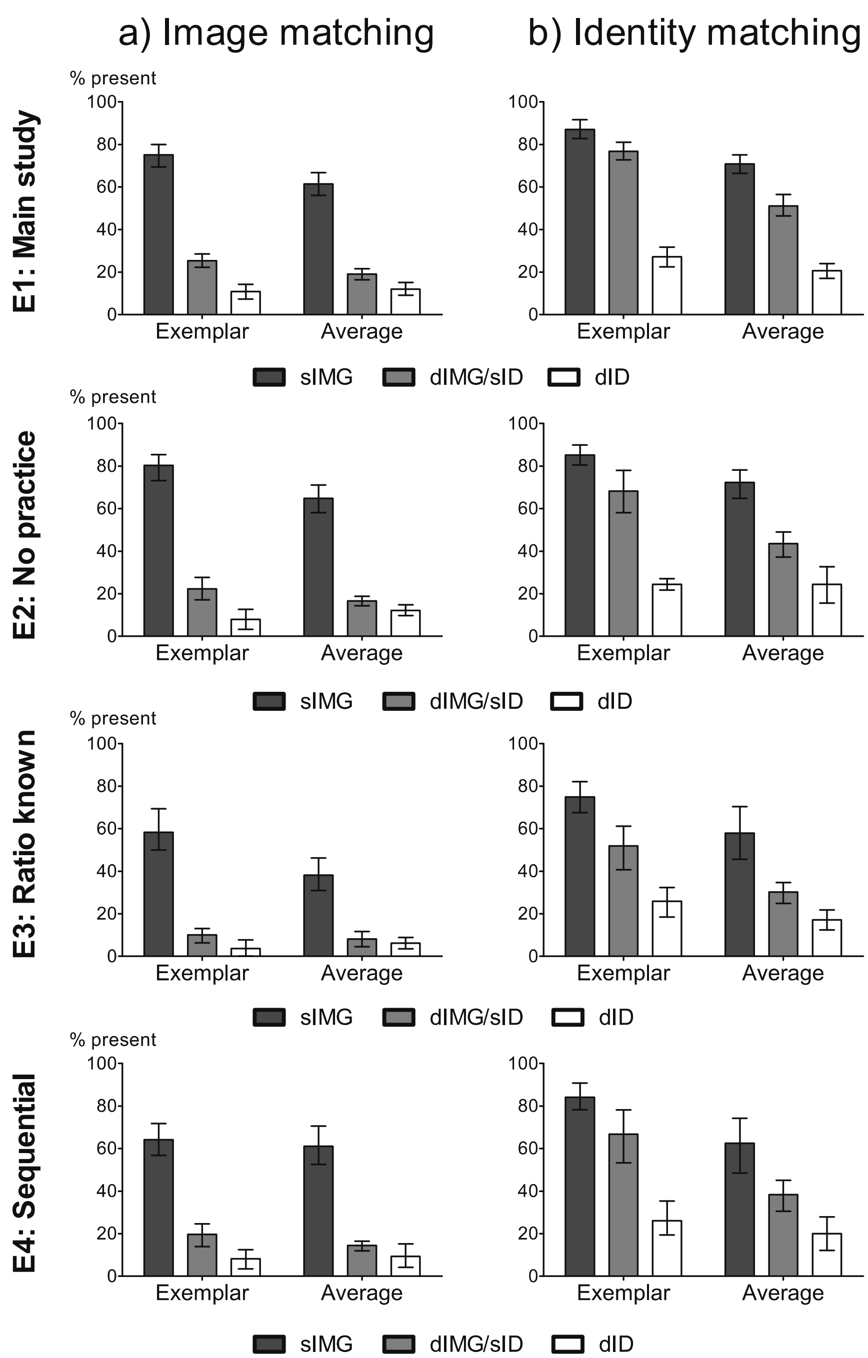

Fig. 2: Percentage "present" responses to probe images in all 4 Experiments. Left column: image matching; Right column: identity matching. Error bars represent 95% CI based on normalized data (see Cousineau, 2005). *N* = 18 in Experiments 1a and 1b, *N* = 8 in all control Experiments 2a,b; 3a,b; and 4a,b, respectively.

Unexpectedly, a strikingly similar pattern was elicited by set average images. sIMG averages elicited remarkably large proportions present responses, indicating that participants erroneously identified the set average as an actual set member. This was not because averages per se tended to elicit responses (e.g., due to inflated typicality), since averages of different identities (dID) were reliably rejected. During image matching, averages across 4 new images from the set identities (sID/dIMG) were rejected - correctly - almost as accurately as averages from new identities (dID). By contrast, a much larger and intermediate level of incorrect present responses occurred during identity matching (incorrect, because an average *never* represented an identity from a set).

Statistical analyses were performed by entering data from both experiment versions to separate 2 by 3 ANOVAs. These revealed reliable main effects of Match Type and Probe Type (all $F > 8$, all $p < .011$, all $\eta^2_P > .320$), and significant interactions of both factors, $F(2, 34) = 9.55$, $p < 0.01$, $\eta^2_P = .360$; $F(2, 34) = 17.50$, $p < 0.01$, $\eta^2_P = .507$, for Experiments 1a and 1b, respectively. Accordingly, present responses were most frequent for sIMG, intermediate for sID/dIMG, and infrequent for dID conditions. In addition, while present responses occurred overall more often for exemplar than average probes, the amount of the difference varied with Match Type, and was rather small (Exp. 1b) or absent (Exp. 1a) in dID conditions. Critically, even when averages were analysed separately, Match Type was still highly significant (both $F > 170$, $p < .001$, $\eta^2_P > .810$), as were all pairwise contrasts between sIMG and sID/dIMG, and between sID/dIMG and dID conditions in both tasks (all $t(17) > 3.81$, all $p < .002$). Importantly, this confirms that averages were more often selected not only when created from the identical set images (sIMG), but also when created from different images of the set identities (sID/dIMG), compared to averages from new identities (dID). More detailed descriptions of all 2 by 3 ANOVAs and follow-up paired comparisons are detailed in Table 1 in the appendix.

In a second step, we examined differences between task conditions by including Task as between-subjects factor in an ANOVA on combined data from Experiments 1a and 1b. The 3-way interaction of Task by Probe Type by Match Type, $F(2, 68) = 6.88$, $p < .01$, $\eta^2_P = .360$ was significant, indicating differences in patterns elicited during image and identity matching, respectively. Fig.2 suggests that a main source for this interaction were large differences in sID/dIMG exemplar conditions in both tasks. This was unsurprising, because a “present” response had been the correct answer during identity matching, but the incorrect response during image matching.

Of greater theoretical interest were differences in present responses elicited by average probes across the two tasks. ANOVA on data from average probes with Match Type and Task revealed a significant interaction, $F(2,68) = 20.34$, $p < .001$, $\eta^2_P = .374$). Independent sample t-tests carried out on corresponding Match Type conditions between the two tasks indicated comparable proportions present responses for sIMG averages in image and identity matching, $t(34) = 1.670$, $p = .104$, and slightly more present responses during identity than image matching for dID conditions $t(34) = 2.619$, $p = .014$. Most importantly, sID/dIMG present responses differed substantially between task conditions, $t(34) = 6.460$, $p < .001$, with more present responses given in the identity than in the image matching task. Thus, sID/dIMG averages were not easily mistaken as a *set image*, but were frequently mistaken as a *person* occurring in a set.

Control Experiments 2-4

Considering that set averaging was typically observed in combination with impaired exemplar memory, the finding from Experiment 1 is particularly challenging, because it suggests that viewers are extracting identity-average information from a set, while simultaneously representing individual exemplar information. Moreover, while it seems reasonable to suppose that viewers might code a set of circles using summary statistics, or even a set of unknown faces, there seems no reason why one should extract an average of, for instance, Bill Clinton and John Travolta. In the following control experiments, we tested for a number of possible alternative explanations for this effect.

Method

Experiments 2-4 were identical to Experiment 1 except as follows. Experiment 2 did not include practice trials. During practice in Experiment 1, the ratio of correct "present" responses was larger than in the actual experiment, such that one might be concerned that participants developed exaggerated expectations about the required ratio of present responses. To exclude this possibility, practice trials were omitted in Experiment 2 and all further experiments. In Experiment 3, participants were additionally informed, correctly, that present responses were required in 16.6 % (Exp. 3a), or 33.3 % (Exp. 3b) of the trials. In Experiment 4, set images were presented sequentially rather than at the same time (order: top left, top right, bottom left, bottom right). Each image was shown for 375 ms, such that total presentation duration was equivalent to Experiments 1-3 (i.e., 1500 ms).

Results

Control Experiments 2-4 yielded results completely consistent with Experiment 1 (cf. Fig. 2, rows 2-4). Most importantly, performance in sIMG conditions was in each case quite accurate for exemplars, and very inaccurate for averages, with large proportions present responses to both sIMG exemplars and, only slightly reduced, to sIMG averages.

Separate 2 by 3 ANOVAs for each experiment corroborated the pattern of Experiment 1. Again, more present responses were given to exemplars than to averages (except for Experiment 3a, where the main effect of Probe Type only approached significance, $p = .076$, and in Experiment 4a, $p = .334$). Main effects of Match Type indicated more present responses to sIMG vs. sID/dIMG conditions, and to sID/dIMG vs. dID conditions throughout. Probe Type interacted with Match Type in all experiments except for Experiment 3b and 4a. Further descriptions of 2 by 3 ANOVAs for all control experiments are detailed in Table 1 in the appendix.

Experiments 2 and 3 controlled for possible expectation effects in Experiment 1a regarding the correct proportion present responses. Such expectations could either originate from practice trials, or from a more general expertise with psychological experimentation methods. However, Experiment 2 replicated all key results of Experiment 1 in virtually identical form, despite excluding practice trials (cf. Fig. 2). Similarly, informing participants about the correct ratio of present trials in Experiment 3 did not differentially affect responses to set averages, although it led to an overall decrease in present responses, indicating that this information successfully induced a more conservative response criterion. We conducted additional ANOVA on combined data from Experiments 2 and 3, and included "Ratio Information" (Experiment 2: not informed, Experiment 3: informed) as an additional between-subjects factor. No significant 4-way interaction was found, $F < 1$, and no other interaction including Ratio Information, all $p > .05$, except for an interaction of Match Type by Ratio Information, $F(2, 56) = 8.09$, $p = .002$, $\eta^2_P = .224$. The latter interaction simply reflects the fact that informing participants about correct ratio led to a greater reduction of present responses in sIMG matching (18.3%) conditions (critically, both for exemplars and averages), and less reduction in the other two conditions (sID/dIMG = 12.6%; dID = 4.0%), in which present responses were already less frequent. Importantly, Experiment 3 provides no evidence that the ratio of present responses might explain the remarkably large proportions of present responses to "matching" set averages.

Experiment 4 addressed a different possibility. Specifically, when presented simultaneously, set images could have been processed to a different extent (e.g., with a focus on the top two faces, and only brief inspection of the bottom faces). By presenting the set faces sequentially for the same amount of time, participants are encouraged to process all faces equivalently. Note that simultaneous presentation is not essential for statistical processing (Chong & Treisman, 2005b; Haberman & Whitney, 2009). In the ANOVA on combined data from Experiments 4a and 4b, the 3-way interaction only approached significance, $F(2, 28) = 2.75$, $p = .086$, $\eta^2_P = .164$, possibly due to relatively low power. However, interactions of Task by Probe Type, $F(2, 28) = 4.66$, $p = .049$, $\eta^2_P = .250$, and Task by Match Type, $F(2, 28) = 8.75$, $p = .002$, $\eta^2_P = .385$, were revealed. Overall, the pattern of results strikingly resembles the previous findings. Most importantly, sequential presentation caused no selective reduction in present responses to sIMG set averages compared to Experiment 2. If anything, sIMG *exemplar* detection was slightly compromised during image matching in Experiment 4a: Exemplars received comparable proportions present responses as averages, and neither the main effect of Probe Type, $F(1, 7) = 1.08$, $p = .334$, $\eta^2_P = .134$, nor the interaction of Probe Type and Match Type $F(1, 7) = 2.62$, $p = .111$, $\eta^2_P = .273$ were significant.

General Discussion

We examined set averaging for identity information in face sets. In contrast to previous work, sets in the present study involved both familiar faces, and large image variability. Compared to earlier work, we used an extended experimental procedure by including both an image-change condition (sID/dIMG) and an additional task (identity matching) to promote identity processing of sets exemplars. Across four experiments, we consistently received two key results that extend the current knowledge regarding set representations for complex stimuli, and that can be summarized as follows.

First, and as predicted, the use of familiar faces in briefly presented sets produces good memory for set exemplars. Second, and surprisingly, viewers nevertheless show clear and consistent evidence for averaging identity information in faces, even across highly familiar set exemplars. Three control studies ruled out alternative explanations based on participants' expectations, or a potential selective processing of a subgroup of set items. We will first discuss these novel findings in the context of our specific approach to create variable sets from familiar faces, and then relate these results to the concepts of set averaging and individual face recognition more generally.

Previous studies had used low image variability within sets. Set images were either taken from standardized databases and set identities were chosen to resemble each other (de Fockert & Wolfenstein, 2009), or – more commonly – sets comprised perceptually similar levels from a morph continuum (e.g. happy to neutral expression, see Haberman & Whitney, 2007, 2009). One reason why participants in previous studies were almost unable to recall individual set exemplars may have been simply because when presented in the set, they all looked alike. By contrast, sets in the present study employed images from different internet sources, and therefore varied more naturally on various dimensions including lighting, viewing angle, head posture, and expression. We expected that set exemplars would consequently be easier to discriminate and that this would lead to improved exemplar memory, which was the pattern we observed in the present study. However, we also assumed that increased exemplar memory would coincide with little if any evidence for set average representations. This assumption was based on our understanding of set averages as an efficient process to capture the essential information from a set in situations where accurate encoding of the set constituents is impossible, for instance by short presentations of crowded displays. Such an idea seemed intuitively plausible and was supported by many previous studies using both simple and complex stimulus material (for a recent review, cf. Alvarez, 2011).

Here we observed a strikingly different pattern: Despite the expected good performance in exemplar memory, set averaging was remarkably robust. In actual fact, present response rates for sIMG averages of about 60% in the present study were even higher when compared to a analogous condition of a different study, where unfamiliar faces had been used (approximately 40%, de Fockert & Wolfenstein, 2009). Accordingly, set averaging of facial identity appears robust to substantial image variability within sets, at least for familiar faces.

Importantly, the use of familiar faces enabled us to address alternative low level explanations for this identity set averaging effect, which previous work could not completely rule out. Specifically, it was unclear whether participants generated average *identity*, or rather average *image* representations from sets. Here, we tested separate groups of participants either with an image matching task as in previous work (e.g., de Fockert & Wolfenstein, 2009), or with an identity matching task. Such a task should have promoted identity processing for the set exemplars, and participants could not simply rely on matching certain low-level aspects of an image due to the potential image change in sID/dIMG conditions.

Critically, we found clear evidence for set averaging in the identity matching group. This suggests that the abstraction of identity information into a summary statistic is not simply a low-level stimulus-driven process, but includes averaging of actual identity information from several faces. This argument receives further support when taking into account the results from sID/dIMG conditions, where participants of the identity matching group often misinterpreted an average across 4 different identities as an actual person from the previous set, even though the probe average involved different images of these identities! Note that this was not a result of inaccurate person memory due to the rather short presentation duration: Identity recognition for exemplars was generally accurate even across the image change in the present experiments: Participants in the identity matching group very accurately accepted sID/dIMG exemplars, while the very same sID/dIMG exemplars were rejected – again very accurately – by participants from the image matching group.

We had expected that both using familiar faces and more variable images would increase exemplar recognition, but reduce or abolish set averaging. By contrast, while accurate exemplar recognition was indeed observed, set averaging for facial identity was also robust. This is remarkable since compelling evidence for set averaging was previously associated with absent or noisy memory for instances, irrespective of stimulus type (Ariely, 2001; Chong & Treisman, 2005b; Alvarez & Oliva, 2008; Haberman & Whitney, 2007, 2009). Accordingly, average set processing has been thought of as an effective and efficient method to extract only the most important information from a complex visual scene (Alvarez, 2011). Supporting this idea, it has been shown that abstractive representations are more precise under distributed than under focused attention (Chong & Treisman, 2005a), and summary coding of high-level information can proceed even in the near absence of attention (Alvarez & Oliva, 2009). In fact, set averaging seems to be so efficient that it can be performed almost as accurately as coding of a single exemplar (Chong & Treisman, 2003). This research suggests that precise exemplar and set average representations are incompatible to the extent that only one representation is extracted at a time, according to task needs. Most research on set averaging employs settings in which it is difficult for viewers to extract precise exemplar representations for their experience. Sets were usually quite crowded or perceptually very similar. Here, task conditions (distinct, familiar faces) allowed forming of precise exemplar representations, accompanied with strong set average representations. To our knowledge, this is the first demonstration of robust simultaneous exemplar and average representations.

In our experiments, “present” responses for exemplars exceeded those for sIMG averages, a pattern that contradicts the commonly described preponderance of average over exemplar representations. This is clearly not reflecting weak average representations in the present study, but rather a consequence of increased recognition of familiar face exemplars (approximately 80%, compared to 30-35% for unfamiliar faces in de Fockert & Wolfenstein, 2009). Our data demonstrate that robust set average representations can co-exist with precise exemplar representations.

Given that ensemble coding is supposed to foster efficient extraction of information, as suggested by previous studies, a simultaneous extraction of exemplar and set average representations does not appear to be particularly efficient. The extent to which exemplar and average representations may draw upon identical or distinct resources is a matter of current debate. Of particular interest, a recent study suggested that hierarchical representations in working memory may simultaneously be formed on multiple levels of abstraction (Brady & Alvarez, 2011). In this study, participants remembered the size of an individual circle at clearly above-chance precision, but size judgements were consistently biased towards the average size in the set. Accordingly, items in working memory could be represented via a combination of set ensemble statistics and individual exemplar information, with statistical representations increasing accuracy in situations of inaccurate exemplar memory. Data from the present experiments are in line with the general idea of a hierarchical representation system.

In the experiments reported here, there seems no obvious advantage to be gained from constructing a representation that merges the individuals. For example, when interacting with groups, there is no communicative advantage to forming a single visual representation of all faces. A tentative suggestion is that set averaging could serve compensatory purposes. For instance, while impaired at recognizing individuals, participants with developmental prosopagnosia nevertheless showed preserved identity and expression set averaging for unfamiliar faces (Leib et al., 2012). Additionally, face recognition performance did not correlate with set averaging performance in that study, suggesting that both tap into distinct processes. While this is an important finding, it remains unclear how set averaging could compensate for poor *individual* face recognition. Further research is needed to clarify the relation between the different coding mechanisms (individual exemplars versus set averages) and their respective relevance for typical and impaired identity processing of both unfamiliar and familiar faces.

It remains to be seen whether the accurate simultaneous computation of exemplar and average representations – which were expected to be incompatible – is a feature of categories beyond faces. These have made a good starting-point, because it is simple to manipulate familiarity without affecting stimulus structure, and because there are well-understood technical mechanisms for combining different images. However, even within the class of faces, a thorough understanding of this phenomenon will require further investigation into the role of encoding time to test efficiency of set representations, set characteristics (e.g., male vs. female, own-race vs. other-race) and other operational variables.

Acknowledgements

The authors would like to thank Kathrin Rauscher, Stefanie Luttmann, and Michaela Kessler for their assistance with data collection, and Stefanie Broncel, Kristin Gottschlich, Marlena Itz, and Jan Rehbein for helping with stimulus preparation and data collection in Experiment 1a. This work is supported by a young researchers grant awarded by the University of Jena to MFN.

References

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually Averaging in a Continuous Visual World: Extracting Statistical Summary Representations Over Time. *Psychological Science, 21*(4), 560-567.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122-131.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19*(4), 392-398.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences of the United States of America, 106*(18), 7345-7350.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157-162.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items. *Psychological Science, 22*(3), 384-392.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology-Applied, 5*(4), 339-360.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology-Applied, 7*(3), 207-218.

Burton, A. M., Bruce, V., & Hancock, P. J. B. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science, 23*(1), 1-31.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*(4), 393-404.

Chong, S. C., & Treisman, A. (2005a). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics, 67*(1), 1-13.

Chong, S. C., & Treisman, A. (2005b). Statistical processing: computing the average size in perceptual groups. *Vision Research, 45*(7), 891-900.

Clutterbuck, R., & Johnston, R. A. (2004). Matching as an index of face familiarity. *Visual Cognition, 11*(7), 857-869.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorial in Quantitative Methods for Psychology, 1*(1), 42-45.

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology, 62*(9), 1716-1722.

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision, 9*(11), -.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*(17), R751-R753.

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.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention Perception & Psychophysics, 72*(7), 1825-1838.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences, 4*(9), 330-337.

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology, 11*(3), 211-222.

Leib, A. Y., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012). Crowd perception in prosopagnosia. *Neuropsychologia, 50*(7), 1698-1707.

Appendix

Exp.	*Effect*	*F-statistics*	*Effect size (partial η^2)*	*Description*		
E1a	ProbeType	$F(1,17) = 8.28$, $p = .010$	$\eta^2_p = .328$	Exemplars ($M = 37.0$ %) > Averages ($M = 30.8$ %)		
	MatchType	$F(2,34) = 303.84$, $p < .001$	$\eta^2_p = .947$	sIMG ($M = 68.2$ %) > sID/dIMG ($M = 22.2$ %) > dID ($M = 11.4$ %)		
	ProbeType x MatchType	$F(2,34) = 9.67$, $p = .003$	$\eta^2_p = .363$	$Exp_{sIMG} > Avg_{sIMG}$	$t(17) = 3.24$	$p = .005$
				$Exp_{sID/dIMG} > Avg_{sID/dIMG}$	$t(17) = 2.86$	$p = .011$
				$Exp_{d/ID} = Avg_{d/ID}$	$t(17) = -0.70$	$p = .493$
E1b	ProbeType	$F(1,17) = 37.16$, $p < .001$	$\eta^2_p = .686$	Exemplars ($M = 63.6$ %) > Averages ($M = 47.5$ %)		
	MatchType	$F(2,34) = 244.53$, $p < .001$	$\eta^2_p = .935$	sIMG ($M = 78.9$ %) > sID/dIMG ($M = 63.9$ %) > dID ($M = 23.9$ %)		
	ProbeType x MatchType	$F(2,34) = 17.51$, $p < .001$	$\eta^2_p = .507$	$Exp_{sIMG} > Avg_{sIMG}$	$t(17) = 4.88$	$p < .001$
				$Exp_{sID/dIMG} > Avg_{sID/dIMG}$	$t(17) = 6.31$	$p < .001$
				$Exp_{d/ID} > Avg_{d/ID}$	$t(17) = 3.26$	$p = .005$
E2a	ProbeType	$F(1,7) = 6.24$, $p = .041$	$\eta^2_p = .471$	Exemplars ($M = 36.8$ %) > Averages ($M = 31.1$ %)		
	MatchType	$F(2,14) = 224.48$, $p < .001$	$\eta^2_p = .970$	sIMG ($M = 72.5$ %) > sID/dIMG ($M = 19.4$ %) > dID ($M = 10.0$ %)		
	ProbeType x MatchType	$F(2,14) = 10.42$, $p = .006$	$\eta^2_p = .598$	$Exp_{sIMG} > Avg_{sIMG}$	$t(7) = 3.41$	$p = .011$
				$Exp_{sID/dIMG} = Avg_{sID/dIMG}$	$t(7) = 2.09$	$p = .075$
				$Exp_{d/ID} = Avg_{d/ID}$	$t(7) = -1.67$	$p = .134$
E2b	ProbeType	$F(1,7) = 7.11$, $p = .032$	$\eta^2_p = .504$	Exemplars ($M = 59.2$ %) > Averages ($M = 46.7$ %)		
	MatchType	$F(2,14) = 111.43$, $p < .001$	$\eta^2_p = .941$	sIMG ($M = 78.7$ %) > sID/dIMG ($M = 55.9$ %) > dID ($M = 24.3$ %)		
	ProbeType x MatchType	$F(2,14) = 18.26$, $p < .001$	$\eta^2_p = .723$	$Exp_{sIMG} > Avg_{sIMG}$	$t(7) = 2.77$	$p = .028$
				$Exp_{sID/dIMG} > Avg_{sID/dIMG}$	$t(7) = 3.70$	$p = .008$
				$Exp_{d/ID} = Avg_{d/ID}$	$t(7) < 0.01$	$p > .999$
E3a	ProbeType	$F(1,7) = 4.33$, $p = .076$	$\eta^2_p = .382$	Exemplars ($M = 24.0$ %) = Averages ($M = 17.5$ %)		
	MatchType	$F(2,14) = 109.16$, $p < .001$	$\eta^2_p = .940$	sIMG ($M = 48.2$ %) > sID/dIMG ($M = 9.1$ %) > dID ($M = 4.9$ %)		
	ProbeType x MatchType	$F(2,14) = 10.50$, $p = .010$	$\eta^2_p = .600$	$Exp_{sIMG} > Avg_{sIMG}$	$t(7) = 2.89$	$p = .023$
				$Exp_{sID/dIMG} = Avg_{sID/dIMG}$	$t(7) = 0.78$	$p = .460$
				$Exp_{d/ID} = Avg_{d/ID}$	$t(7) = -1.84$	$p = .108$
E3b	ProbeType	$F(1,7) = 6.94$, $p = .034$	$\eta^2_p = .498$	Exemplars ($M = 50.9$ %) > Averages ($M = 35.0$ %)		
	MatchType	$F(2,14) = 57.55$, $p < .001$	$\eta^2_p = .892$	sIMG ($M = 66.4$ %) > sID/dIMG ($M = 41.0$ %) > dID ($M = 21.5$ %)		
	ProbeType x MatchType	$F(2,14) = 3.14$, $p = .077$	$\eta^2_p = .310$			
E4a	ProbeType	$F(1,7) = 1.08$, $p = .334$	$\eta^2_p = .134$	Exemplars ($M = 30.7$ %) = Averages ($M = 28.3$ %)		
	MatchType	$F(1,7) = 78.58$, $p < .001$	$\eta^2_p = .918$	sIMG ($M = 68.2$ %) > sID/dIMG ($M = 22.2$ %) > dID ($M = 11.3$ %)		
	ProbeType x MatchType	$F(2,14) = 2.62$, $p = .111$	$\eta^2_p = .273$			
E4b	ProbeType	$F(1,7) = 6.77$, $p = .035$	$\eta^2_p = .492$	Exemplars ($M = 59.0$ %) = Averages ($M = 40.2$ %)		
	MatchType	$F(2,14) = 63.82$, $p < .001$	$\eta^2_p = .901$	sIMG ($M = 73.2$ %) > sID/dIMG ($M = 52.6$ %) > dID ($M = 23.0$ %)		
	ProbeType x MatchType	$F(2,14) = 6.42$, $p = .013$	$\eta^2_p = .478$	$Exp_{sIMG} = Avg_{sIMG}$	$t(7) = 2.33$	$p = .052$
				$Exp_{sID/dIMG} > Avg_{sID/dIMG}$	$t(7) = 3.17$	$p = .016$
				$Exp_{d/ID} = Avg_{d/ID}$	$t(7) = 1.12$	$p = .301$

Table S1: Results from all four Experiments' 2x3 ANOVAs and, where applicable, post-hoc comparisons.