Categorization training changes the visual representation of face identity

Fabian A. Soto¹

Published online: 31 May 2019
© The Psychonomic Society, Inc. 2019

Abstract

Previous research suggests that learning to categorize faces along a new dimension changes the perceptual representation of that dimension, but little is known about how the representation of specific face identities changes after such category learning. Here, we trained participants to categorize faces that varied along two morphing dimensions. One dimension was relevant to the categorization task and the other was irrelevant. We used reverse correlation to estimate the internal templates used to identify the two faces at the extremes of the relevant dimension, both before and after training, and at two different levels of the irrelevant dimension. Categorization training changed the internal templates used for face identification, even though identification and categorization tasks impose different demands on the observers. After categorization training, the internal templates became more invariant across changes in the irrelevant dimension. These results suggest that the representation of face identity can be modified by categorization experience.

Keywords Face identity · Face encoding · Categorization · Perceptual learning

Previous research suggests that learning to categorize faces along a new dimension changes the perceptual representation of that dimension. Such studies have used stimuli varying along artificially created morphed dimensions. A morphed face dimension is created by selecting two unfamiliar “parent” faces (Fig. 1a, Step 1), and morphing from one parent to the other in several steps (Step 2). For example, the first level could include 0% of one face and 100% of the other, the second level 25% of one face and 75% of the other, the third level 50% of each face, and so on. To obtain a two-dimensional morph space, each of the levels in the first dimension is morphed with each of the levels in the second dimension, with each face contributing 50% to the combination (Step 3). This results in a space of faces that are varying combinations of four parent faces, as shown in Fig. 1b, where each dot

represents the coordinates of a face in the morph space. Such morphed dimensions are known to be integral (Soto & Ashby, 2015); that is, different faces in the space (e.g., each point in Fig. 1b) are not perceived as being composed of two dimensions, but rather as unique identities. Points in the space that are close to each other are perceived as being the same identity (e.g., the central circle of points in Fig. 1b), but enough change in *any* direction produces a change in perceived identity (e.g., any of the outer circles of points in Fig. 1b). Integrality is a desirable property, as these stimuli are used to determine what conditions lead to a change from this holistic representation to a dimensional representation. Goldstone and Steyvers (2001) found that training in a categorization task in which one morphed dimension was relevant and another was irrelevant (i.e., according to the red category bound in Fig. 1b) produced such *dimension differentiation*, as suggested by transfer of performance to a new task in which either the relevant or the irrelevant dimension was replaced by a completely new morphed dimension. Transfer was also observed when the categorization rule was rotated by 90 degrees (i.e., the relevant dimension became irrelevant and vice versa).

Other results suggest that dimension differentiation is neither a general learning phenomenon nor a stimulus-specific (i.e., similarity-based) effect, but rather a dimension-specific effect (Goldstone & Steyvers, 2001).

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13414-019-01765-w>) contains supplementary material, which is available to authorized users.

✉ Fabian A. Soto
fasoto@fiu.edu

¹ Department of Psychology, Florida International University, 11200 SW 8th St, AHC4 460, Miami, FL 33199, USA

Fig. 1 Tasks and stimuli used in this study. **a** Procedure used to create a two-dimensional space of morphed faces used in the categorization task. **b** Representation of the categorization task and the estimated templates in the morphing space. *Points* represent stimuli obtained from a specific combination of levels for each dimension. The *red dotted line* represents the category boundary used for training, so that stimuli on each side of the boundary were assigned to a different category. The *numbered green lines* represent the estimated templates, with the *green dot* representing the approximate position of the base image, and the *arrow* represents the direction of the target identity.

Performance drops more when both dimensions are replaced instead of only one (not a general learning phenomenon) and when the categorization rule is rotated 45 degrees rather than 90 degrees (not a similarity-based effect, as in the former case stimulus assignment to categories is more similar to the original task). Finally, we know some of the conditions that produce this form of learning. Categorization feedback is sufficient to produce the effect, as learning occurs with circular arrangements that do not suggest any privileged direction in the morphing space (see Fig. 1b). Learning of new dimensions is blocked by already-available dimensions to describe the categorization rule (e.g., mouth size or eyes shape). Finally, the stimulus space must be created through factorial combination of morphing sequences, as other ways of building the space produce no learning (Folstein, Gauthier, & Palmeri, 2012).

However, little is known about the contents of this form of learning (Rescorla, 1988), and in particular how the visual representation of specific face identities changes after category learning. This is an important open issue, as the influence of categorization on face identification might

help explaining the effects of social categorization (e.g., by race or sex) on face perception (Hugenberg, Young, Sacco, & Bernstein, 2011) and more generally is informative about the influence of high-level cognition on perception (Collins & Olson, 2014). Note that a categorization task such as that used by Goldstone and Steyvers (2001) (see Fig. 1b) involves assigning stimuli that vary along *identity-changing* dimensions to the same category. This is similar to categorization by race or sex, albeit with much lower within-category variability. On the other hand, this is fundamentally different from learning of new identities (e.g., Jenkins, White, Van Montfort, & Burton, 2011; Ritchie & Burton 2017), which involves images varying along *identity-preserving* dimensions such as pose, expression, illumination, age, etc. While categorization and identification might share learning mechanisms (Soto & Wasserman, 2010), they require extracting different forms of information from faces.

How can one study changes in visual representations, which cannot be directly measured? A common strategy in psychophysics is assuming that observers use a relatively

simple representation and designing experiments to estimate those representations. For example, a popular model assumes that people detect a target signal by matching an internal template containing the signal and the input stimulus (Murray, 2011). The green box in Fig. 1c depicts such a simple model. In this example, the model's task is to detect a face identity. The observer matches the presented image and template using a dot product; the resulting value, which is perturbed by perceptual noise, represents the perceptual evidence that the presented image contains the target identity. The observer compares this random variable against a decision bound to make a behavioral decision (e.g., “yes” or “no”).

The observer's internal template can be estimated using a reverse correlation procedure, in which participants are presented with a base image combined with random noise images (see left of Fig. 1c) and asked whether they see the target identity (*Observer Decision* in Fig. 1c). Different noise patterns are presented in each trial; the observer reports seeing the identity in some of them, but not in others. Averaging the noise images identified as the target identity produces an estimate of the template used to solve the task (see right of Fig. 1c). The estimated templates, which are usually called *classification images* (Murray, 2011), highlight areas of the face that provide task-relevant information. Thus, different templates are useful for different tasks. For example, the templates used to determine emotional expression, gender, and identity are all different (e.g., Mangini & Biederman 2004; Schyns, Bonnar, & Gosselin, 2002). That is, face categorization and identification rely on different sources of information.

Here, we used reverse correlation to estimate templates used for identification of the parents of the category-relevant dimension (A and B in Fig. 1). We estimated such visual templates before and after categorization training, in the context of a task requiring participants to identify a single individual (either parent A or B) against 12 foil individuals (none of them parent faces). Templates were estimated using two different base images, represented by the two green dots in Fig. 1b. These base images carried no information about the relevant dimension, but they did carry strong information about the irrelevant dimension. This allowed us to determine whether irrelevant face information influenced the templates used for identification. Four identification templates (for A and B, estimated at two levels of the irrelevant dimension) were estimated before and after categorization training, to evaluate the effect of such training. The templates are represented by the four arrows in Fig. 1b, which are pointing from the base image on top of the category boundary toward the target identities. Note that these templates are defined in the space of image pixels rather than in the morphing space, and that they are estimated in the context of an identification task rather than

the categorization task shown in Fig. 1b. Thus, templates for two opposite identities (1–2 and 3–4) do not necessarily point in opposite directions, which would suggest a unified representation of the relevant dimension. Also, templates for the same identity (1–3 and 2–4) do not necessarily point in parallel directions, which would suggest a representation of the relevant dimension that is separable (i.e., invariant) from the irrelevant dimension. However, we hypothesized that categorization training would increase the degree to which both of those properties were present in the estimated internal templates.

Methods

Participants

Twenty students at Florida International University (13 females and seven males; ages between 22 and 29 years, with median = 22) participated in exchange for a monetary compensation (US\$10/h). The study was approved by an Institutional Review Board and written informed consent was obtained from all participants.

Stimuli

A detailed description of the stimuli can be found in the [Supplementary Material \(SM\)](#). Stimuli were created from 16 computer-generated face images. Stimuli used in the categorization task were generated using a factorial morphing procedure (Goldstone & Steyvers, 2001; Folstein et al., 2012; Soto & Ashby, 2015), illustrated in Fig. 1a. Each dimension was obtained from a pair of parent faces (Step 1) morphed in a sequence of 19 steps (Step 2). A two-dimensional space was generated by factorially combining each of the faces in each dimension with each of the faces in the other dimension (Step 3). These two-dimensional morphs included 50% from each of the one-dimensional morphs. The specific stimuli presented to participants in the categorization task are represented by the black points in Fig. 1b. The base images presented during the reverse correlation task are represented by the green points in Fig. 1b. An additional 12 faces were included during identification training in the reverse correlation task. All images in the reverse correlation task were combined with noise masks, created using a weighted sum of sinusoids as proposed by Mangini and Biederman (2004).

Procedure

Participants completed four sessions of a reverse correlation task and three sessions of categorization training. A detailed description of the tasks is included in the *SM*.

The experiment started with two 90-min sessions of the reverse correlation task. A session involved identification of one parent of the relevant dimension (A or B; see Fig. 1b); each parent was identified in a different session. Participants were instructed to identify which of two images was more similar to the parent face. They were warned that the two faces would be very similar in most trials. Participants completed ten blocks, each comprising 36 identification training trials followed by 100 reverse correlation trials, for a total of 360 identification training trials and 1000 reverse correlation trials. Due to a coding error early in the experiment, two participants completed six rather than ten blocks of the task.

Two different faces were presented during each identification training trial, each combined with a different noise pattern. One image contained the target face (parent A or B) and the other included one of 12 foils. The participant's task was to identify which of the two images contained the target, by pressing the "left arrow" or "right arrow" keys. A choice was followed by "CORRECT" in blue font color or "INCORRECT" in red font color. Each distractor face was repeated three times within a block. The order of trials and the left–right position of each face within a trial were random.

During a reverse correlation trial, participants were presented with a choice between the same base image combined with a pattern of noise and its reversed pattern (i.e., noise $\times -1$). The base images (green points in Fig. 1b) contained ambiguous information about the category-relevant dimension (50% from each parent face). Different base images were used in different blocks, and they contained strong information about the category-irrelevant dimension (i.e., 100% from one parent face and 0% from the other).

After completion of the two initial reverse correlation sessions, participants completed three 1-h sessions of the categorization task depicted in Fig. 1b, where the black dots represent the morphed stimuli presented and the red dotted line represents their separation into two different categories. A single face was presented in each trial, and participants reported whether it belonged to category A or B through a key press, which was always followed by feedback about the correctness of the response. Each session consisted of nine blocks of 72 trials each, for a total of 648 trials. Each stimulus (36 per category) was presented once in a block, with the order randomized within the block.

After categorization training, all participants completed two new reverse correlation sessions, identical to those completed before categorization training.

Data analysis

See *SM* for details about inclusion/exclusion criteria, data analysis software, and template estimation and analysis.

The main analysis was performed on data from the reverse correlation task. A rejection criterion of $\alpha = .05$ was used.

Individual templates and anti-templates were estimated using the R package *rcicr* (Mangini & Biederman, 2004; Dotsch & Todorov, 2012). We estimated four templates (for faces A and B at 2 levels of the irrelevant dimension) per participant and their corresponding anti-templates, both before and after training. Templates are estimated by averaging noise related to selected (template) or not selected (anti-template) noise patterns. The final templates are grayscale images with a single intensity value at each pixel.

Two similarity measures were obtained from each participant's set of templates and anti-templates, using the Pearson correlation of corresponding pixel intensity values in the two images, within an oval-shaped face area mask that excluded non-facial features. The first similarity measure was computed between templates for two different identities (the two parent faces of the relevant dimension, A and B) estimated using the same base image (i.e., same level of the irrelevant dimension). Using the numeration from Fig. 1b, template 1 was correlated with anti-template 2, and template 3 was correlated with anti-template 4. Higher similarity here would indicate a stronger dimensional representation, with the templates for identities A and B pointing in opposite directions. The second similarity measure was computed between templates for the same identity (i.e., same parent face of the relevant dimension), but estimated using different base images (i.e., different values in the irrelevant dimension). Using the numeration from Fig. 1b, template 1 was correlated with template 3, and template 2 was correlated with template 4. Higher similarity here would indicate representations of the identities that were more invariant to (or separable from) changes in the irrelevant dimension.

Both similarity measures were analyzed with a 2 (learning stage: before vs. after categorization training) $\times$ 2 (dimension level) within-subjects ANOVA. Partial eta-squared (η_p^2), along with a 90% confidence interval on this estimate, was calculated as a measure of effect size.

Average templates were computed for each combination of level of the relevant dimension, level of the irrelevant dimension, and pre- versus post-categorization-training conditions. *Separability images* were obtained by subtracting two average templates for identification of the same face (A or B), estimated at different levels of the irrelevant dimension: template 1 minus 3, and template 2 minus 4. The two separability images were obtained both pre- and post-categorization-training. Separability images were tested for significance using a two-tailed cluster test from random field theory (Chauvin, Worsley, Schyns, Arguin, & Gosselin, 2005).

Results

After obtaining classification images, we obtained two measures summarizing changes in these templates that could be produced by categorization training. First, the templates of the two parents of the relevant dimension (A and B) could become negatively correlated, as it has been observed with natural face categories (Dotsch & Todorov, 2012), indicating a stronger dimensional representation in which such extremes are encoded as opposite values. Categorization training had an effect in this direction (Fig. 2a), but it was not significant, $F(1, 15) = 3.03$, $p > .1$, $\eta_p^2 = .168$ [.000, .410]. This result could be due to low statistical power, but our data does not allow us to reach a conclusion regarding this point (see power analysis in *SM*). The effect of irrelevant level, $F(1, 15) = 1.32$, $p > .1$, $\eta_p^2 = .080$ [.000, .317] and the interaction, $F(1, 15) = .5$, $p > .1$, $\eta_p^2 = .032$ [.000, .244], were also not significant. Second, the templates for the same identity at different levels of the irrelevant dimension could become more similar (i.e., correlated), which would explain why categorization training increases invariance of the relevant dimension (i.e., perceptual separability; see Soto & Ashby 2015). Our results show a significant effect of training in this direction (Fig. 2b), $F(1, 15) = 5.95$, $p = .028$, $\eta_p^2 = .284$ [.019, .509]. There was no effect of relevant level, $F(1, 15) = 1.044$, $p > .1$, $\eta_p^2 = .065$ [.000, .296], or the interaction, $F(1, 15) = .098$, $p > .1$, $\eta_p^2 = .006$ [.000, .161]. An exploratory analysis (for details, see *SM*) revealed that identification training was unlikely to have contributed to these results.

In sum, the main result of our study is that categorization training influences the internal templates used for face identification, which become more invariant to changes in face information that is irrelevant for categorization. Importantly, previous research has shown that identification and categorization tasks impose different demands on an observer (Schyns et al., 2002). In an ideal observer analysis, we corroborated that the two tasks used here demanded sampling from different areas of the image to be solved optimally (for details, see *SM*).

To explore more precisely what changes in internal templates produced such effect, we computed average templates (Fig. 2c) and subtracted templates for identification of the same face (A or B), estimated at different levels of the irrelevant dimension: template 1 minus 3, and template 2 minus 4 (see Fig. 1b). Next, we obtained significant clusters from these difference images (Chauvin et al., 2005), representing face areas used differently across changes in the irrelevant dimension. Figure 2d displays significant clusters before (red) and after (blue) categorization training. Comparing these clusters allows us to determine what changes in the internal templates underlie the increase in invariance observed earlier.

Note first that some clusters that are significant before categorization training disappear or show a reduction in size after training. Clusters that do not disappear show a reduction in their extension (volume reduced by 62.52% in eye cluster and 60.11% in mouth cluster of Fig. 2d-left, and by 59.42% in eye cluster of Fig. 2d-right), but not in their location (91.47%, 30.14% and 39.75% overlap of the post-training cluster with the corresponding pre-training cluster). The fact that some significant clusters are similarly located before and after categorization training suggest that such training did not radically change sampling of facial information. This is confirmed by looking at the classification images in Fig. 2c, and by a positive correlation between corresponding templates (e.g., template 1) before and after categorization training (mean correlations for templates 1–4: .19, .17, .18, and .18; all $p < .001$ uncorrected).

The observed reduction in number and size of significant clusters helps explaining the main result shown in Fig. 2b. Importantly, these clusters are all on or near internal face features (eyes and mouth), which are areas that are known to provide information for identification (Butler, Blais, Gosselin, Bub, & Fiset, 2010; Schyns et al. 2002). On the other hand, some significant clusters appear only after categorization training, and they tend to be located around external face features (forehead, jaw line, and cheeks), which people tend to ignore in face identification (Butler et al., 2010; Schyns et al., 2002). This suggests that perhaps the main effect of categorization training was to increase template separability (i.e., similarity of templates across changes in the irrelevant dimension) around the internal face features, but to decrease such separability around external face features. An exploratory analysis revealed that not to be the case, as a similar increase in similarity was observed in internal and external face areas (for details, see *SM*).

In conclusion, our study suggests that categorization training changes the internal templates used for face identification. The most important change involved an increase in the invariance of these templates to changes in the irrelevant dimension (Fig. 2b), which explains previously observed increments in perceptual separability (Soto & Ashby, 2015). While such increase in invariance was observed across the face, the effect seemed to be particularly strong around internal face features, where significant violations of separability either disappeared or decreased in extent.

Discussion

Categorization training changed the internal templates used for face identification, even though identification and categorization tasks impose different demands on an observer (Schyns et al., 2002). After categorization

Fig. 2 Experimental results. **a** Mean similarity between the internal template of a face at one extreme of the relevant dimension and the anti-template of the face at the other extreme of the relevant dimension; higher values represent a stronger dimensional representation. **b** Mean similarity between the internal template for the same face (one extreme of the relevant dimension) at different levels of the irrelevant dimension; higher values represent an increase in invariance across

changes in the irrelevant dimension. **c** Average templates estimated in this study; different colors represent whether more (*red*) or less (*blue*) luminance would produce identification of the target face (*green* is neutral). **d** Facial areas represented differently across changes in the irrelevant dimension, both before (Pre, in *red*) and after (Post, in *blue*) categorization training. *Error bars* represent standard errors of the mean

training, the internal templates used to identify a face became more invariant across changes in the irrelevant dimension, which relates to previously observed increments in dimensional separability (Soto & Ashby, 2015). Such increase in invariance was observed across the face, but the effect seemed particularly strong around internal face features, which are known to provide important information for identification (Butler et al., 2010; Schyns et al., 2002).

However, categorization training did not dramatically change the face information sampled for identification. Many features of the internal templates remained relatively unchanged after categorization training (Fig. 2c), and some areas of the face represented differently across changes in the irrelevant dimension were similarly located before and after training (Fig. 2d). One interesting possibility is that the identity representations already-existing before categorization training were used as a “scaffold” for the development of new representations after categorization training.

Our results suggest that the representation of individual face identities can be modified by categorization experience, opening the possibility that naturalistic categorization tasks (e.g., by race, sex, age, etc.) have a similar effect. The influence of categorization experience on internal face templates might thus explain the multiple effects of social categorization on face perception (Hugenberg et al., 2011) or why internal templates of emotional expression vary with cultural experience (Jack, Caldara, & Schyns, 2012). An important caveat is that naturalistic tasks involve much larger variation in identity than what we can obtain using the morphing procedure common to studies of dimensional differentiation. More research is needed to understand to what extent the two paradigms are comparable.

Our results might also inform studies about learning of new faces across changes in identity-preserving variables (e.g., Jenkins et al. 2011; Ritchie & Burton 2017). Although the present study involved images varying along identity-changing morphed dimensions, in both cases the task is technically one of categorization, in which multiple images are assigned the same label (i.e., either an individual or a group of individuals) and the participant must learn to extract the label-predicting information while ignoring the label-irrelevant information in faces. There is reason to believe that a common learning mechanism might underlie such learning across category levels (Soto & Wasserman, 2010). An important question for future research is whether the representation of face identities learned from multiple identity-preserving images (rather than from a single image, as in the present study) is affected by categorization training. Thus, the current study is only a first step toward understanding the influence of categorization on identity representation. Future research will require richer stimulus sets and estimated representations (e.g., by assuming more complex observer models).

In sum, dimensional differentiation through categorization may have important consequences for face perception in naturalistic circumstances. The present study contributes to our understanding of the contents of this form of perceptual learning (Rescorla, 1988), by showing how it changes the visual representation of individual faces.

Acknowledgements S.D.G. This work was supported in part by an NIMH Grant to F.A.S. (1R21MH112013-01A1). The author thanks Jefferson Salan and Claudia Wong for their help collecting data for this study.

Open Practices Statement The data and materials for all experiments are available at <https://osf.io/dy23v/> and none of the experiments was preregistered. The work described here conformed to Standard 8 of the American Psychological Association’s Ethical principles of Psychologists and Code of Conduct (<http://www.apa.org/ethics/code/index.aspx>).

References

- Butler, S., Blais, C., Gosselin, F., Bub, D., & Fiset, D. (2010). Recognizing famous people. *Attention, Perception, and Psychophysics*, 72(6), 1444–1449. [10.3758/APP.72.6.1444](https://doi.org/10.3758/APP.72.6.1444)
- Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., & Gosselin, F. (2005). Accurate statistical tests for smooth classification images. *Journal of Vision*, 5(9), 1.
- Collins, J. A., & Olson, I. (2014). Knowledge is power: How conceptual knowledge transforms visual cognition. *Psychonomic Bulletin and Review*, 21(4), 843–860. <https://doi.org/10.3758/s13423-013-0564-3>
- Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychology and Personality Science*, 3(5), 562–571.
- Folstein, J. R., Gauthier, I., & Palmeri, T. (2012). Not all morph spaces stretch alike: How category learning affects object discrimination. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 38(4), 807–802.
- Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, 130(1), 116.
- Hugenberg, K., Young, S. G., Sacco, D. F., & Bernstein, M. J. (2011). Social categorization influences face perception and face memory. In A. J. Calder, G. Rhodes, M. H. Johnson, & J. V. Haxby (Eds.) *The Handbook of Face Perception*, (pp. 245–261). Oxford University Press.
- Jack, R. E., Caldara, R., & Schyns, P. (2012). Internal representations reveal cultural diversity in expectations of facial expressions of emotion. *Journal of Experimental Psychology: General*, 141(1), 19–25.
- Jenkins, R., White, D., Van Montfort, X., & Burton, A. (2011). Variability in photos of the same face. *Cognition*, 121(3), 313–323.
- Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, 28(2), 209–226.
- Murray, R. (2011). Classification images: A review. *Journal of Vision*, 11(5), 2. <https://doi.org/10.1167/11.5.2>
- Rescorla, R. (1988). Pavlovian conditioning: It’s not what you think it is. *American Psychologist*, 43(3), 151–160.

- Ritchie, K. L., & Burton, A. (2017). Learning faces from variability. *The Quarterly Journal of Experimental Psychology*, 70(5), 897–905.
- Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological Science*, 13(5), 402–409.
- Soto, F. A., & Ashby, F. (2015). Categorization training increases the perceptual separability of novel dimensions. *Cognition*, 139, 105–129. <https://doi.org/10.1016/J.journalcognition.2015.02.006>
- Soto, F. A., & Wasserman, E. (2010). Missing the forest for the trees: Object discrimination learning blocks categorization learning. *Psychological Science*, 21(10), 1510–1517. <https://doi.org/10.1177/0956797610382125>
- Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.