Establishing a Database for Studying Human Face Photograph Memory

Wilma Alice Bainbridge* (wilma@mit.edu)
Department of Brain and Cognitive Sciences, MIT, 77 Massachusetts Avenue
Cambridge, MA 02139 USA

Phillip Isola* (phillipi@mit.edu)
Department of Brain and Cognitive Sciences, MIT, 77 Massachusetts Avenue
Cambridge, MA 02139 USA

Idan Blank (iblank@mit.edu)
Department of Brain and Cognitive Sciences, MIT, 77 Massachusetts Avenue
Cambridge, MA 02139 USA

Aude Oliva (oliva@csail.mit.edu)
Computer Science and Artificial Intelligence Lab (CSAIL), MIT, 77 Massachusetts Avenue
Cambridge, MA 02139 USA

Abstract

Contemporary visual environments bombard us with hundreds of face images every day, and this places a non-trivial demand on long-term memory. However, little is known about what makes certain faces remain in our memories, while others are quickly forgotten. To establish a basis for face memorability exploration, we assembled a database of 8,690 face photographs from online sources, spanning diverse face and image characteristics. Workers on Amazon's Mechanical Turk were asked to identify repetitions within a stream of these stimuli. Variations in image memorability (hit rates, false alarm rates, and their interactions) were reliable across participants, suggesting that face images may have different intrinsic levels of memorability. We discuss future directions in using this database to quantify face photograph memorability, as well as potential scientific and commercial applications.

Keywords: face recognition; image memorability; face photograph memory database

Introduction

Every day, we encounter an overwhelming number of photographs and images of people's faces. Many inter-personal interactions are mediated by such images: we view people's Facebook profile pictures; memorize photographs of our students; browse personals on dating websites; skim through pictures attached to job applications; and encounter countless face images published on advertisements on billboards, in magazines, and online. As social creatures, we remember many of these faces.

Large-scale visual memory experiments have shown that people have a remarkable ability to remember which specific image they saw even after seeing thousands of pictures depicting objects, scenes or events (Konkle, Brady, Alvarez, & Oliva, 2010a; Standing, 1973). Importantly, these studies have shown that we do not just remember the gist of a picture, but we are able to recognize which precise image we saw and some of its visual details (Brady, Konkle, Alvarez, & Oliva, 2008; Konkle, Brady, Alvarez, & Oliva, 2010b). In addition to remembering particular images as icons, we also have the intuition that not all images are remembered equally. While the reasons why some images are remembered are varied, recent works have found that

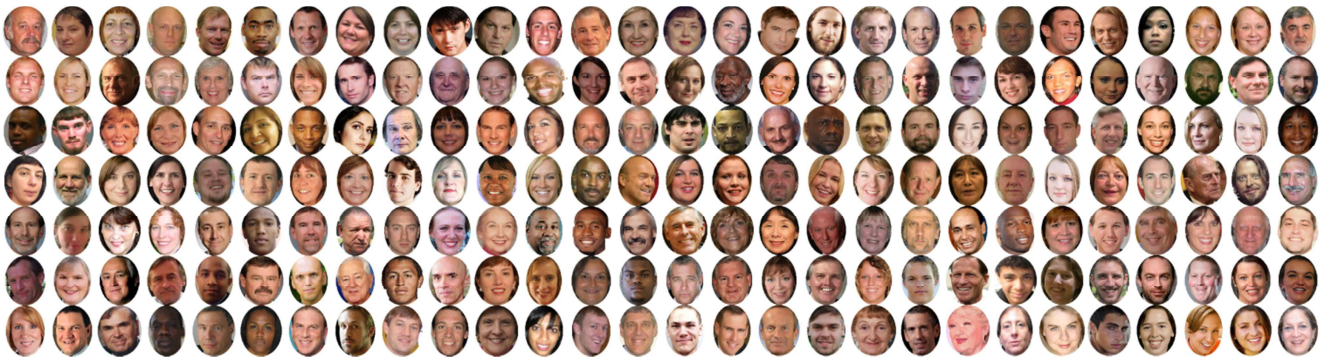

Figure 1: An example set of 196 random images from the face photo database used for this study.

Figure 2: An illustration of the behavioral procedure. Participants were required to identify repeats amongst a stream of face photos.

images containing people with visible faces are highly memorable (Isola, Parikh, Torralba, & Oliva, 2011a; Isola, Xiao, Torralba, & Oliva, 2011b).

Despite the fact that the memorability of face photos is of both psychological and commercial significance, it is not clear how findings illuminating scene and object memorability will generalize to face images. First, memorability has been shown to be heavily influenced by the distinctiveness of stimuli (Konkle et al., 2010a, 2010b). Compared to scenes and objects, faces are a relatively homogeneous category and have low variation in visual features. However, faces could be coded with rich sub-categorical structure (e.g., gender, race, age, emotional expression, dominance level, attractiveness) that may render their representations more distinguishable in memory. Second, evidence suggest that faces are processed by specialized cognitive (Duchaine, Yovel, Butterworth, & Nakayama, 2006; Robbins & McKone, 2007) and neural (Kanwisher & Yovel, 2006) mechanisms (c.f., McKone, Crookes, & Kanwisher, 2009). For these reasons, face memorability deserves special attention.

In this study, we establish a large-scale face photograph database on which we have quantified performance on a repetition detection task. We examined inter-image variability, and its reliability, on this task. Specifically, we analyzed two memory-related behavioral measures – hit rate and false alarm rate – which we term "memorability scores".

Methodology

We conducted a large-scale experiment that used photos from a database of diverse faces, run on 337 participants on Amazon's Mechanical Turk. The following section describes the assembly of the database and the experiment run on Mechanical Turk.

Face Photo Database Generation

We assembled a diverse database of 10,000 photos of faces. First, we generated a list of approximately 25,000 first and last name pairs from a database of names from the United States census (Kleimo, 2011), using parameters for a balance of both genders and names of high commonality. Use of the US census allowed us to collect names from a diverse range of ethnic backgrounds, representing the general gender, racial, and age distribution of the United States adult population. However, because the first and last names were generated randomly, they did not necessarily represent specific people from the US population. Example names included "Wilma Reno," "Phillip Robichaux," "Lori Blank," and "Arlene Olivarez".

Each of the 25,000 names was used as a search query, and, for each query, approximately 10 photos were automatically downloaded from Google Images. Our Google Image Search parameters included that all photos be at least 400×300 pixels, full-color, and of faces. The experimenters went through the set of photos and deleted those that were low-quality, depicted children, were obscured by other objects, included accessories such as hats and glasses, or had unusual makeup. The database was filtered down to over 10,000 photos of faces that were diverse over a wide range of ages, genders, races, and attractiveness levels. Faces had both eyes visible and open and, in general, expressions tended to range from neutral to smiling. Five experimenters then went through the set and deleted recognizable celebrities for the purposes of this study, bringing the set used for this experiment to a final size of 8,690 photos. We expect that only a small percent of our database should be celebrity photos that were not identified through our initial screening. The stimuli for the experiment were then generated by placing ovals around the faces to frame them and to diminish the influence of irrelevant background features in the photo. All photos were resized to a standard of 256 pixels in height with variable width to preserve aspect ratio. Figure 1 shows a collection of example photos from the database.

The Behavioral Experiment

Face memory performance was measured through a behavioral study called the "Face Memory Game" run on Amazon's Mechanical Turk. Mechanical Turk is a tool belonging to Amazon.com's Web Services that allows researchers to crowdsource tasks and experiments for monetary compensation to a large Internet population. Mechanical Turk served as an ideal environment for this study, allowing us to obtain memory scores for thousands of images.

The methodology for this game is based off the methodology from a previous image memorability study conducted with scenes (Isola et al., 2011b; see Figure 2). The task was structured into a series of 30 levels, each taking about 4.8 minutes and consisting of 120 photos. Although labeled "levels" to give a sense of progress to the participant, the levels did not differ from each other in

difficulty or stimulus type. For each level, the participant saw a constant stream of stimuli, each displayed for 1 second and then followed by a 1.4 second fixation point before the next stimulus was presented. Stimulus presentation order was different for each participant. Participants had to press the key 'r' (for "repeat") whenever the current stimulus was the same photo as one they had seen before (sometimes across levels). When they responded correctly to a repeat, a green cross appeared as feedback. When participants missed a repeated photo or pressed 'r' for a novel photo, a gray X appeared to indicate an error. The game was first preceded by a short qualification and training round of 30 photos. Between levels, participants were given a brief break of up to five minutes and were presented with their correct response score for that level. After 30 levels of the game were over, the game ended. However, participants could choose to end the game at any time, and their data was used up to that point.

From the face stimulus database, 2222 photos were randomly selected as target photos, while the remaining 6468 photos were used as filler photos. Repetitions of photos in the task happened with both target and filler photos. The memory performance measures are based off the results from the target photos, where repetitions were spaced 91-109 photos apart. The repetition with the filler photos acted as a "vigilance task" to test the reliability of participants, with repetitions spaced 1-7 photos apart. The filler photos were also used as spacing between the target photos, and some had no repetitions. Neither target photos nor filler photos had more than one repetition across the entire study.

A total of 337 Mechanical Turk workers participated in the game, and 90% of the data came from 168 workers. The average worker played over 8 levels. We limited the game to only Mechanical Turk workers in the US, so that the workers' demographics would approximately match the demographics of the faces used as stimuli. Workers were paid $0.40 per level, or approximately $5 an hour. Workers were screened in several ways throughout the study to ensure they were attentive to the task. First, only workers with at least a 95% Mechanical Turk approval rate were allowed to participate in the study. During the study, if a participant's error rate for false alarms exceeded 50% for the last 30 photos, or if their hit rate for vigilance task repeats fell below 50% for the last 10 photos, then the data from that level were discarded and the participant received a flag. Rejection criteria were reset for each level. If the participant received three flags, they were blocked from continuing in the experiment. Otherwise, participants could restart the game as many times as they liked, until they had completed 30 levels. When restarting the game, unseen photos were always selected as stimuli.

Results

We collected an average of 30.4 hit rate (HR) scores per photo and 35.4 false alarm rate (FAR) scores per photo. The

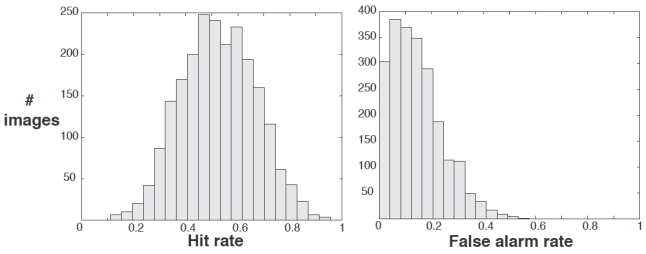

Figure 3: Hit rate and false alarm rate histograms over the target photos in our experiment.

average HR was 53.6% (*SD*=14.3%), and the average FAR was 14.5% (*SD*=9.9%). The distributions of these memorability scores followed simple unimodal forms (Figure 3).

Is Memory Performance on Some Images Reliably Different than on Other Images?

To evaluate the reliability of our measurements, we split our participant pool into two independent halves, and quantified how well memorability scores measured on the first half of the participants matched memorability scores measured on the second half of the participants. Averaging over 25 random split-half trials, we calculated a Spearman's rank correlation ρ of 0.44 between HRs on the two halves and a ρ of 0.48 on FARs. The strength of these correlations demonstrates that we have characterized real differences between photos.

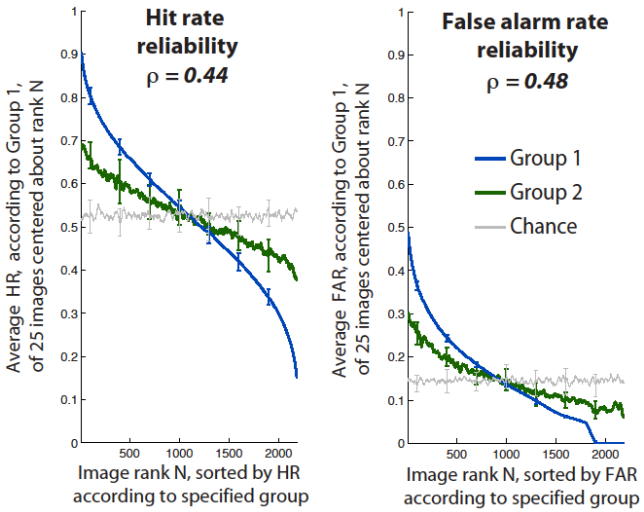

Figure 4: Data split-half reliability. Photos are ordered on the x-axis by the HR (left) and FAR (right) of a random half of the participants, and are plotted against these measures on the same half (blue line) or the remaining half (green line) of participants. Chance reliability is shown by randomly ordering the photos on the x-axis (gray line). Plots are averaged across 25 such random splits of the participant pool.

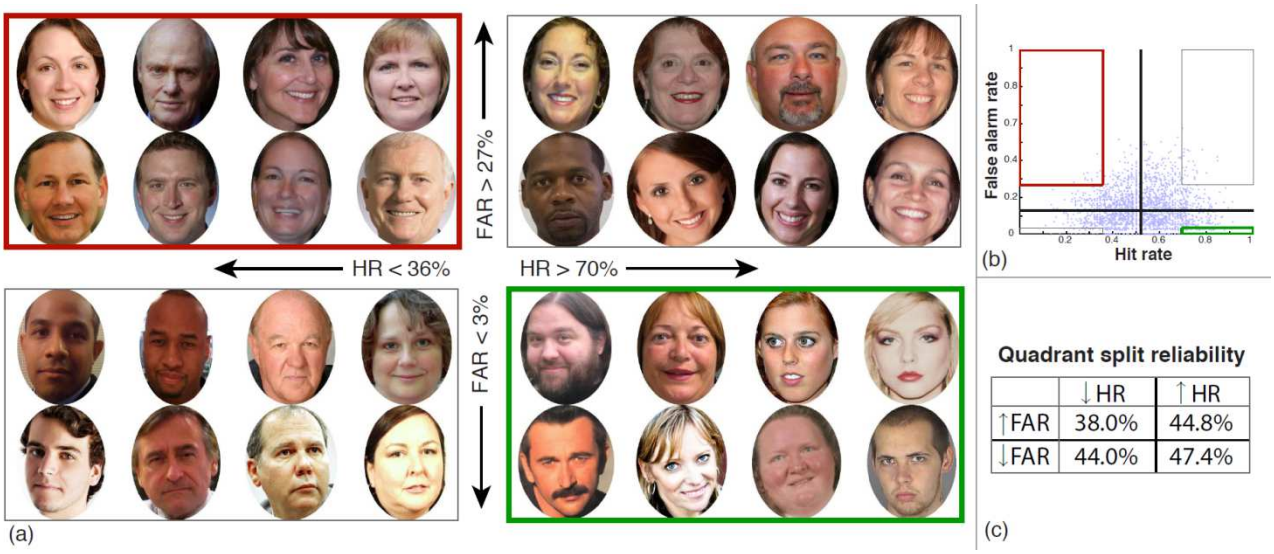

Figure 5: (a) Sample images of four performance profiles. The image set was broken into a 5×5 grid of HR quintiles crossed with FAR quintiles. Each quadrant shows a random sample of the photos at each of the four corners of this distribution (highest/lowest HR/FAR). The set outlined in green can be characterized as more memorable than the set outlined in red since the green set has both a higher HR and a lower FAR than the red set. (b) A scatterplot showing HR versus FAR. Rectangles indicate the same corners of the quintile grid as in (a). The black lines split the distribution along the median HR and FAR creating four performance profiles. (c) Reliability computed as percent overlap of HR/FAR profile assignments of photos between two halves of the participants (averaged across 25 random splits of the participant pool). Profiles correspond to the quadrants defined by the black lines in (b).

Are these differences large enough to be interesting? We examined the reliability of the size of the memorability differences as follows. We sorted photos by their scores given by the first half of the participants and plotted this against memorability scores according to the second half of the participants (Figure 4). For clarity, we convolve the resulting function with a length-25 box filter. This shows that, for example, if a repeat is correctly detected 80% of the time by one half of the participants, we can expect the other half of the participants to correctly detect this repeat around 66% of the time, corroborating that this photo is truly memorable. At the other end of the spectrum, if a repeat is only detected 30% of the time by one half of the participants, the other half will tend to detect it only 42% of the time – this photo is consistently forgotten. It thus appears that there really is sizable variation in face photo memorability.

Thus, our data show enough variation and enough reliability that it should be possible to use these data to model detailed aspects of photo memorability in later work (c.f., Isola et al., 2011a, 2011b). Individual differences and random variability in the sequence of photos each participant viewed add noise to these estimations; nonetheless, this level of reliability suggests that information intrinsic to the photos plays a key role in determining which photos are remembered.

False Memories versus True Memories

Our data allow us to look at both false memories and true memories. False memories may arise in response to highly typical faces, because they resemble many other faces (Vokey & Read, 1992). True memories should relate to specific encodings of the photos seen in our experiment. Can we separate these two signals in our data? If a photo receives both a high hit rate and a high false alarm rate, it may be highly memorable, but it also may just be a face that always feels familiar, regardless of whether or not it has been previously seen. A stronger case for high memorability can be made when we find photos that have high hit rates and low false alarm rates – what is termed a "mirror effect" (Glanzer & Adams, 1985, 1990). If one photo consistently has both a higher hit rate and a lower false alarm rate than another photo, then we can confidently say that the first photo evoked a stronger true memory than the second.

To isolate truly memorable photos, we split our photo set about the median HR and then again about the median FAR, producing four performance profiles (high/low HR/FAR) (see Figure 5). Are some photos consistently assigned to the high-HR/low-FAR profile, whereas others are consistently assigned to the low-HR/high-FAR profile? If so, we can say the former photos are more memorable than the latter. We tested this level of consistency by splitting our photos into profiles according to one random half of the participants and comparing these assignments to those given by the other half of the participants. Averaging over 25 such trials, the

two halves of the subjects agree 47% of the time on assignments to the high-HR/low-FAR profile (chance level would be 25%). Interestingly, we see similar levels of agreement in each of the remaining quadrants, as reported in Figure 5c.

These quadrants may reflect different types of photos with respect to memory: some photos may be distinctive and strongly remembered; some may be prototypical and produce both strong memories and many false alarms; others may evoke many false memories while, interestingly, generating relatively few true memories (low-HR/high-FAR); and still others may simply be ignored all together (low-HR/low-FAR).

Discussion

This study has established a database for the exploration of face photograph memory, and shows that memorability of face photographs can be reliably measured. We found an average hit rate of 53.6% across the target face photos, compared to a false alarm rate of only 14.5%. In contrast, Isola et al. (2011b) used the same experimental protocol and found an average hit rate of 67.5% and false alarm rate of 10.7% for scene photo memory capacity. Do these numbers for face photos indicate that we are worse at remembering faces than scenes? Or, is the face photo performance high, considering that faces vary at the exemplar level (i.e., all belong to the same basic-level category), while the scenes used by Isola, et al. (2011b) vary at the categorical level (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976)? It is difficult to compare across separate studies and participant pools – for example, Isola et al. (2011b) recruited international participants, while the current study limited participants to the United States. It will also be essential to find a way to quantify the differences between face and scene photos in order to meaningfully compare memorability between the two different groups of stimuli.

A second interesting question to explore is what attributes lead to the separation of photos into the four performance profiles we identified based on hit rate and false alarm rate (Figure 5). Previous research has suggested that more distinct faces have high hit rates and low false-alarm rates in an old/new task (Deffenbacher, Johanson, Vetter, & O'Toole, 2000; Light, Kayra-Stuart, & Hollander, 1979). In contrast, both hit and false-alarm rates are high for typical faces, due to the effect of "context-free familiarity", a sense of familiarity not related to a specific previous encounter with a face (Vokey & Read, 1992). The other two profiles we explored may also have interesting qualifying characteristics to examine that were not explicitly addressed in the past literature.

Beyond distinctiveness and typicality, we advocate the exploration of several other attributes and their correlations with memorability. Previous research has noted that memorability of a face, both perceived and actual, may differ based on viewer characteristics, such as race (Chiroro & Valentine, 1995; Meissner, Brigham, & Butz, 2005) or recent experience with other face images (Lewis & Johnston, 1997); however, the current study shows surprising reliability across subjects of diverse backgrounds, viewing a widespread distribution of photos. This suggests there are similarities across participants in how they represent different photos in memory. One next important step will be to examine how the demographic characteristics of the participant (e.g., race, gender, and age) may or may not predict the memorability of face photos with matching or non-matching characteristics. Other properties to examine in the context of memorability include perceived memorability (do people actually remember what they think they will remember?), attractiveness, and eye contact. While the current work focuses on memory for photos of faces, future work will also explore memory for face identity across different photos of the same person.

The future possibility of quantifying "memorability" of a face lends itself to many useful applications in both the field of psychology and mainstream society. For instance, Todorov (2011) identified features in faces linked to different subjective judgments of those faces, such as attractiveness and trustworthiness. These were used to build computer models that generated faces varying along these featural dimensions. A score of memorability could similarly be added to the feature set of a face, and thus be used to rate, manipulate, and generate face images. For animated films, animators could create cartoon characters with different levels of memorability (c.f., Gooch, Reinhard, & Gooch, 2004), such as a highly memorable protagonist surrounded by forgettable extras. Makeup artists could use software that would identify where to apply makeup to make celebrities memorable for a photoshoot. Algorithms could automatically identify the most memorable face photographs out of an album to use in textbooks, magazines, or even social network profiles.

Conclusion

This study serves as an initial, empirical look at a new large, diverse database of face photos and the average rate and reliability of memorability measurements across this database. When viewing a stream of hundreds, sometimes thousands, of novel face photos, participants in our experiment were able to accurately identify repeats about half the time they appeared, while making relatively few false alarms. This suggests that participants were holding in memory detailed representations of hundreds of face photos even though each photo was presented with just a single one-second view. In addition, we found that photos of faces vary substantially in memorability; these reliable differences indicate the importance of memorability for understanding how we process face images. This research opens the door to future investigation in various fields, from cognitive psychology to cognitive neuroscience to computer vision, as to what makes some face images or facial features more memorable than others.

Acknowledgments

We would like to thank Marc Howard for helpful discussions and advice. This work is partly funded by a NSF grant (1016862) and a Google research award to A.O. W.A.B. is funded by the Leventhal Graduate Fellowship, P.I is funded by an NSF graduate research fellowship, and I.A.B. is funded by the Henry E. Singleton Fund.

The face photograph database will be publicly available on the author's website.

References

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, 105*(38), 14325-14329.

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology, 48*(4), 879-894.

Deffenbacher, K. A., Johanson, J., Vetter, T., & O'Toole, A. J. (2000). The face typicality-recognizability relationship: encoding or retrieval locus? *Memory & Cognition, 28*(7), 1173-1182.

Duchaine, B. C., Yovel, G., Butterworth, E. J., & Nakayama, K. (2006). Prosopagnosia as an impairment to face-specific mechanisms: elimination of the alternative hypotheses in a developmental case. *Cognitive Neuropsychology, 23*(5), 714-747.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13*(1), 8-20.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 5-16.

Gooch, B., Reinhard, E., & Gooch, A. (2004). Human facial illustrations: creation and psychophysical evaluation. *ACM Transactions on Graphics (TOG), 23*(1), 27-44.

Isola, P., Parikh, D., Torralba, A., & Oliva, A. (2011a). *Understanding the intrinsic memorability of images*. Paper presented at the 25th Conference on Neural Information Processing Systems (NIPS), Granada, Spain.

Isola, P., Xiao, J. X., Torralba, A., & Oliva, A. (2011b). What makes an image memorable? *24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 145-152.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences, 361*(1476), 2109-2128.

Kleimo, A. (2011). The Random Name Generator, from http://www.kleimo.com/random/name.cfm

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010a). Scene Memory Is More Detailed Than You Think: The Role of Categories in Visual Long-Term Memory. *Psychological Science, 21*(11), 1551-1556.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010b). Conceptual Distinctiveness Supports Detailed Visual Long-Term Memory for Real-World Objects. *Journal of Experimental Psychology-General, 139*(3), 558-578.

Lewis, M. B., & Johnston, R. A. (1997). Familiarity, target set and false positives in face recognition. *European Journal of Cognitive Psychology, 9*(4), 437-459.

Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory, 5*(3), 212-228.

McKone, E., Crookes, K., & Kanwisher, N. (2009). The cognitive and neural development of face recognition in humans. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences IV*. Cambridge, MA: MIT Press.

Meissner, C. A., Brigham, J. C., & Butz, D. A. (2005). Memory for own- and other-race faces: a dual-process approach. *Applied Cognitive Psychology, 19*(5), 545-567.

Robbins, R., & McKone, E. (2007). No face-like processing for objects-of-expertise in three behavioural tasks. *Cognition, 103*(1), 34-79.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382-439.

Standing, L. (1973). Learning 10,000 Pictures. *Quarterly Journal of Experimental Psychology, 25*, 207-222.

Todorov, A. (2011). Evaluating faces on social dimensions. In A. Todorov, S. T. Fiske & D. A. Prentice (Eds.), *Social Neuroscience: Toward Understanding the Underpinnings of the Social Mind*. New York, NY: Oxford University Press.

Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition, 20*(3), 291-302.