Individual differences in face-looking behavior generalize from the lab to the world

Matthew F. Peterson
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Jing Lin
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Ian Zaun
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Nancy Kanwisher
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Recent laboratory studies have found large, stable individual differences in the location people first fixate when identifying faces, ranging from the brows to the mouth. Importantly, this variation is strongly associated with differences in fixation-specific identification performance such that individuals' recognition ability is maximized when looking at their preferred location (Mehoudar, Arizpe, Baker, & Yovel, 2014; Peterson & Eckstein, 2013). This finding suggests that face representations are retinotopic and individuals enact gaze strategies that optimize identification, yet the extent to which this behavior reflects real-world gaze behavior is unknown. Here, we used mobile eye trackers to test whether individual differences in face gaze generalize from lab to real-world vision. In-lab fixations were measured with a speeded face identification task, while real-world behavior was measured as subjects freely walked around the Massachusetts Institute of Technology campus. We found a strong correlation between the patterns of individual differences in face gaze in the lab and real-world settings. Our findings support the hypothesis that individuals optimize real-world face identification by consistently fixating the same location and thus strongly constraining the space of retinotopic input. The methods developed for this study entailed collecting a large set of high-definition, wide field-of-view natural videos from head-mounted cameras and the viewer's fixation position, allowing us to characterize subjects' actually experienced real-world retinotopic images. These images enable us to ask how vision is optimized not just for the statistics of the "natural images" found in web databases, but of the truly natural, retinotopic images that have landed on actual human retinae during real-world experience.

Introduction

The crux of the problem of visual recognition is the ability to appreciate that an object is the same across the very different images it casts on the retina due to changes in position, size, lighting, and viewing angle, to name a few (DiCarlo & Cox, 2007). Recent work suggests that for the case of face recognition, position invariance is achieved in part by behavior rather than by computation: People fixate a consistent and stereotyped position on the face, thus minimizing variability in the retinal position of face images (Gurler, Doyle, Walker, Magnotti, & Beauchamp, 2015; Mehoudar, Arizpe, Baker, & Yovel, 2014; Peterson & Eckstein, 2012). In particular, robust individual differences are found in the precise location where people make their first saccade into the face, with a continuous distribution ranging from the brows to the mouth. These differences are robust over time, task, face familiarity, and variation in low-level properties such as

Citation: Peterson, M. F., Lin, J., Zaun, I., & Kanwisher, N. (2016). Individual differences in face-looking behavior generalize from the lab to the world. *Journal of Vision*, *16*(7):12, 1–18, doi:10.1167/16.7.12.

doi: 10.1167/16.7.12 Received December 15, 2015; published May 18, 2016 ISSN 1534-7362

color, size, and contrast (Gurler et al., 2015; Mehoudar et al., 2014; Or, Peterson, & Eckstein, 2015; Peterson & Eckstein, 2012, 2013). Most importantly, face recognition performance drops by nearly 20% when faces are presented at another subject's preferred looking position if it differs from one's own (Or et al., 2015; Peterson & Eckstein, 2013). This work suggests that the representations that underlie face recognition are retinotopically specific, with position invariance largely attained not by cortical computations (Riesenhuber & Poggio, 1999; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007) but by looking behavior. However, all of this work has been conducted in laboratory settings, with eye movements monitored as subjects performed tightly controlled tasks in which photographs of faces are presented at a fixed distance while head and body movements are restricted by a chinrest.

The lab testing situation differs from real-world face viewing in a number of respects, yet few studies have investigated real-world gaze on faces in nonclinical populations (Einhäuser et al., 2009; Macdonald & Tatler, 2013, 2015). In the lab, visual stimulation is limited to a centrally presented computer screen, whereas real-world faces generate a wide array of retinal images of unpredictable sizes and positions anywhere in the visual field. In the world, unlike the lab, retinal stimulation is determined not only by eye movements, but also by head direction and body orientation. Further, real-world vision is dynamic and interactive, with goals shifting moment to moment, rather than fixed by task instructions. Perhaps most importantly, in the real world the face we are looking at is often looking back at us, engendering a social context associated with tasks, signals, actions, and behavioral consequences that are distinct from the lab. Given the dramatic differences between these conditions, it is important to know whether the consistent individual differences in face-looking behavior documented in previous lab studies are also found in everyday real-world vision. Here we asked this question by measuring each subject's preferred face-fixation position in the lab with the same methods used previously, and then by sending them off for a walk around the Massachusetts Institute of Technology (MIT) campus while wearing a mobile eye tracker. This design enabled us to monitor where individuals fixated on faces that came into view during naturalistic real-world vision. If position invariance for face recognition is indeed solved in large part by looking behavior (rather than computation), then individual differences in preferred face-fixation positions measured in lab should generalize to real-world behavior. Failure to find this result would suggest that the prior results reflect a special case, and would cast doubt on the hypothesis that position invariance in face recognition is solved by eye movements. A failure to generalize would also call into question the extent to which face recognition behavior measured in the lab should be applied to our understanding of how the brain processes faces during normal operation.

Beyond answering whether face-fixation behavior observed in the lab generalizes to the world, the present study will enable us to make a first foray into a broader research program of characterizing what might be called "retinal image statistics" (RIS). Most prior studies of natural image statistics use photographs from the web that likely represent a biased sample of the images people actually see in everyday life. First, these photos reflect situations in which someone used a camera to select and frame a small portion of the visual world at a specific moment. The criteria for the photographer's selection likely differ from the criteria viewers use to select saccade targets. Second, most photographs are thrown away, and the ones that survive and get posted on the web are a nonrandom sample, less likely to be marred by the occlusions, blur, bad lighting, or other factors that reduce the intelligibility or attractiveness of the image but are common in real-world contexts. Third, and perhaps most importantly, images on the web do not come with information about where viewers were fixating. Fixation position matters enormously, because acuity declines sharply from the fovea toward the periphery, meaning that only a few degrees of the world around fixation are seen with high resolution. For all these reasons the standard web-photo–based analyses of natural image statistics do not represent an unbiased sample of the visual information that reaches the brain. Because our mobile eye tracking study records both the image seen by the subject, and the subject's eye position on that image, our study provides a collection of experienced images with the fixation point on each, a necessary first step in a broader study of the statistics of experienced natural retinal images.

Methods

The study was run in two stages. In Stage I, participants identified celebrity faces presented on a computer screen while their eye movements were monitored. Each subject was categorized into one of three groups according to where they tended to fixate on the faces. A subset of these subjects from each group were later recalled to participate in Stage II, in which they wore a mobile eye tracker to monitor their gaze while they walked around natural environments.

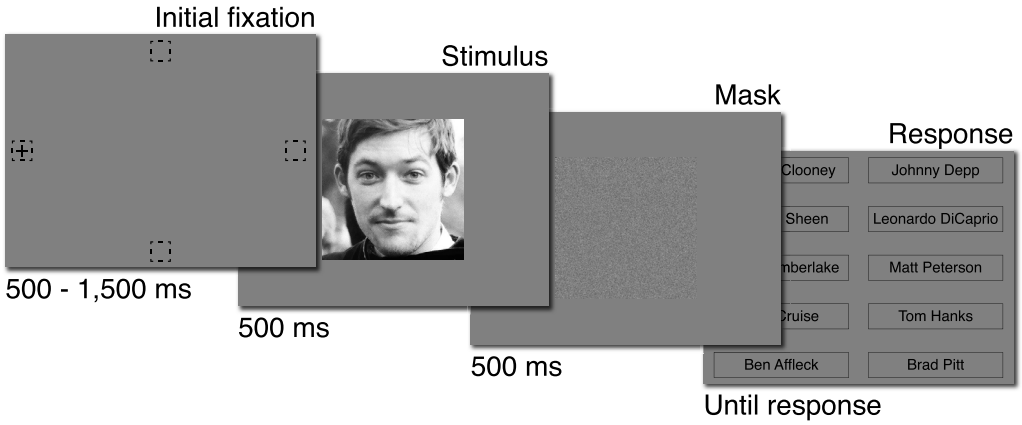

Figure 1. In-lab famous face identification paradigm (Stage I). Picture of the author for illustrative purposes only; in the actual study, subjects were shown pictures of real celebrities.

Methods (Stage I: In lab)

Participants

Seventy participants were recruited using flyers and departmental subject lists (40 MIT students and 30 from the Cambridge community; 48 female; M age = 28.0 years, min = 18 years, max = 62 years). Subjects received $20 for participation, gave informed consent, and had normal or corrected-to-normal vision. The study was approved by the MIT Committee on the Use of Humans as Experimental Subjects.

Eye tracking

The right eye of each participant was tracked using an SR Research EyeLink 1000 Desktop Mount sampling at 1000 Hz (SR Research Ltd., Ottawa, Ontario, Canada). A nine-point calibration and validation were run at the beginning of the session and after every 40 trials with a mean error of no more than 0.5° visual angle. Saccades were classified as events where eye velocity was greater than 22°/s and eye acceleration exceeded 4000°/s^2.

Stimuli and display

Stimuli were 160 frontal view images of 80 well-known Caucasian celebrities (e.g., Tom Cruise, Jennifer Lawrence) acquired using Google image search (two different images per celebrity, 40 male and 40 female). Images were converted to gray scale, rotated to an upright orientation, scaled so that the center of the eyes and center of the mouth were in the same position and separated by 6.0°, cropped from the top of the head to the chin (463 × 463 pixels or 16.9°) and contrast energy normalized. All stimuli were presented on a 17-in. CRT monitor with a resolution of 1024 × 768 pixels and refresh rate of 85 Hz. Subjects sat 50 cm from the monitor, with each pixel subtending 0.036°.

Procedure

Participants saw each of the 160 images in random order. Following the procedure used in our earlier studies (Peterson & Eckstein, 2012, 2013), a trial began with a fixation cross located 10° from the center of the monitor at either the left, right, top, or bottom edge of the screen (location randomly selected). The subject fixated the center of the cross and pressed the spacebar when ready. After a random, uniformly distributed delay between 500 and 1500 ms, the cross disappeared and the randomly sampled face image was displayed at the center of the monitor. Note that in an earlier control experiment, we found that the pattern of individual differences in preferred fixation behavior on centrally presented faces were conserved when faces were presented at unpredictable locations (Peterson & Eckstein, 2013). During the delay period the subject was required to maintain fixation at the cross, with a deviation of more than 1.0° resulting in an error message and restarting of the trial. The face image remained visible for 500 ms, during which eye movements were allowed, and was then replaced with a 500-ms high-contrast white noise mask. A response screen then appeared consisting of two columns of five names each (the correct name of the face they had just seen and nine randomly sampled foils of the same gender, positions randomized). The subject used the mouse to click on the name they thought was correct after which the correct answer was highlighted for 500 ms before commencing the next trial (Figure 1).

Analysis

Identification performance was quantified as the proportion of trials with a correct identification (*PC*). Individual's face-fixation behavior was quantified by computing the mean location of the first into-face fixation (i.e., the location at the end of the first into-image saccade as defined above in Methods (Stage I): Eye tracking) across the 160 image presentations. We then defined an individual's relative fixation metric, γ, as the distance of his or her mean fixation upward from the mouth relative to the total distance between the mouth and eyes:

$$\gamma = \frac{y_{\text{fixation}} - y_{\text{mouth}}}{y_{\text{eyes}} - y_{\text{mouth}}} \quad (1)$$

Methods (Stage II: Real world)

Participants

"Looking groups" were defined before the current study based on independent data from 250 subjects who had participated in similar face identification studies at the University of California, Santa Barbara (Or et al., 2015; Peterson & Eckstein, 2012, 2013, 2014). As in the current study, the previous work measured the mean location of subjects' first into-face fixation. Interindividual variation was found to be large and consistent along the vertical dimension, ranging from the eyebrows to the mouth (Gurler et al., 2015; Mehoudar et al., 2014; Or et al., 2015; Peterson & Eckstein, 2013). Using these data, we defined criteria to categorize people into three looking groups: Upper Lookers (ULs) were the 15% of the sample who looked highest up on the face, Lower Lookers (LLs) were the 15% who looked lowest, and Middle Lookers (MLs) were everybody in between. We used these predefined criteria to categorize the original 70 subjects from Stage I of the current study into looking groups based on the average location of their first into-face fixation from Stage I. The current sample yielded 11 ULs (15.7%), 45 MLS (64.3%), and 14 LLs (20.0%). For each looking group, we recalled the 10 subjects with the highest calibration scores, as measured by the EyeLink, to participate in Stage II (10 ULs, 10 MLs, and 10 LLs). As with Stage I, subjects received $20 for participation, provided informed consent, and had normal or corrected-to-normal vision. The study was approved by the MIT Committee on the Use of Humans as Experimental Subjects.

Procedure

Subjects were told only that we were interested in assessing everyday, natural visual experience. Critically, we did not mention any specific interest in faces or people. Subjects were first fitted with the mobile eye tracker glasses and GoPro camera (Figure 2A) before initial calibration, validation, and registration (see below and Figure 2B). The experimenter then accompanied the subject for 8–12 min around the lab and nearby hallways of the Brain and Cognitive Sciences Building and the Stata Center across the street, engaging in conversation aimed toward making them feel comfortable with the apparatus. Subjects were then instructed to walk unaccompanied across campus walkways, courtyards, a long hallway, and a busy city street to a predesignated location (12–15 min). The experimenter met the subjects at the location and accompanied them back to the calibration room (5 min), concluding the study (25–30 min total). Each subject followed a similar path that exposed them to a representative sample of environmental settings (indoor locations like hallways, rooms, corridors, etc., and outdoor locations like streets, yards, etc.) and social contexts (no people, engaged in one-on-one interaction, watching others interact, etc.; Figure 2C). Subjects were all run at a similar time of day to maximize the between-subjects consistency of environmental and social conditions.

Real-world eye tracking: Overview

Measuring and analyzing eye movements in unconstrained real world environments poses multiple challenges. Here, we detail a standardized framework that allows the experimenter to reliably collect and analyze accurate data. The framework focuses on standardized routines that maximize the consistency, precision, and retention of data, while avoiding possible subject- and task-specific biases. It also allows for frequent validation across time, a critical aspect as data from mobile eye trackers can be marred by subject/apparatus motion and changing environmental (e.g., lighting) and eye (e.g., pupil size) states that can dramatically compromise initial calibration. Finally, the framework develops a combination of automatic algorithms and novel crowdsourcing techniques for analysis and interpretation.

Apparatus

Real-world gaze direction was measured at 60 samples per second with a pair of Applied Science Laboratory (ASL) Mobile Eye-XG Eye Tracking Glasses (ASL Eye

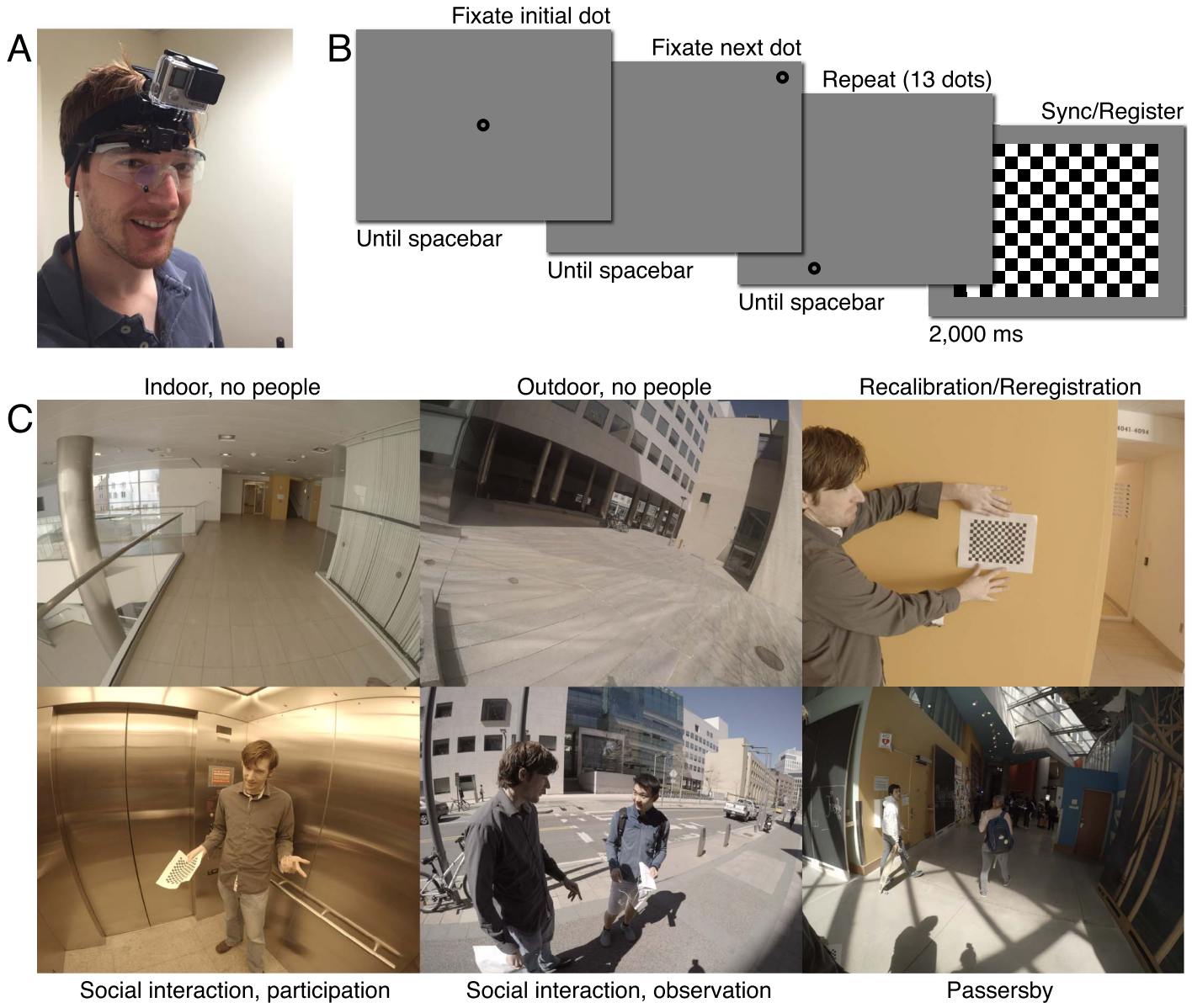

Figure 2. Real-world eye-tracking paradigm (Stage II). (A) Subjects were fitted with a pair of ASL eye tracking glasses. A supplemental GoPro camera enhanced the quality and FOV of the recorded video of the subject's visual environment. (B) Calibration (moving dot) and ASL-to-GoPro video synchronization and registration (checkerboard) were automated and standardized across participants. (C) Each subject walked a similar route through the uncontrolled environments around the MIT campus. Routes and times were chosen to ensure that a variety of locations and social settings were sampled.

Tracking, Billerica, MA). The ASL tracker uses two cameras to estimate fixation position relative to the central region of the visual world in front of the wearer (Figure 2A). The first camera, termed the scene camera, rests on the top rim of the glasses and records video at 60 frames per second (fps), with a field of view (FOV) spanning 64° horizontally and 48° vertically (640 × 480 pixels). The scene camera was adjusted to align the center of its FOV with that of the subject's. The second camera, termed the eye camera, records an infrared (IR) image of the subject's right eye reflected off a partially IR-reflective coated lens that protrudes from the main lens. This allows the eye camera to detect both the subject's pupil and the corneal reflection of a pattern of three dots produced by an IR emitter (with one dot selected as the primary). The position and orientation of both the eye camera and the IR-reflective lens were adjusted for each subject so that the pupil was centered in the eye camera's FOV and the three IR dots were near the pupil center when the subject looked straight ahead. The eye camera lens was then focused to maximize pupil and IR dot sharpness.

To improve upon the scene camera's FOV, resolution, and image sensor quality (contrast sensitivity,

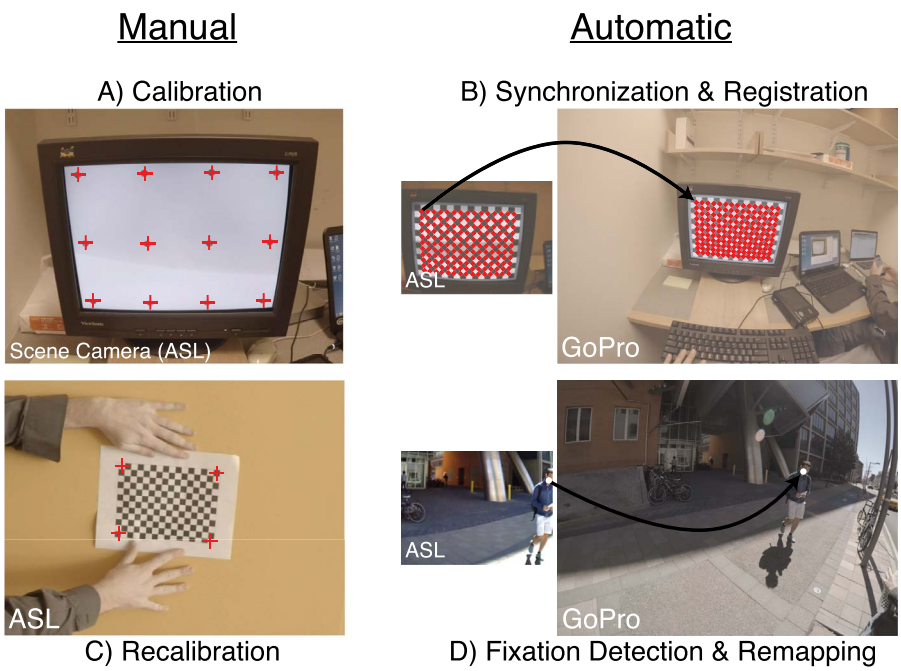

Figure 3. Post processing of eye tracking and video data. (A) A subject-specific function that estimates gaze direction is learned by registering the location of each calibration dot (relative to the ASL scene camera) to the position of the pupil center and corneal reflection (from the eye camera). (B) The vertices of the postcalibration checkerboard pattern are automatically detected in both scene recordings, allowing for automatic synchronization and coordinate registration between videos. (C) Data quality was validated every 3 min by having the subject fixate the corners of a checkerboard pattern. (D) Saccade and fixation events were automatically detected and their spatial coordinates mapped to the high-resolution, wide FOV GoPro video.

temporal properties, etc.), subjects wore a supplementary GoPro Hero4 Black camera (FOV spanning 110° horizontally and 90° vertically; 2704 × 2028 pixels; 30 fps; GoPro, San Mateo, CA). The GoPro was positioned just above the eye tracker glasses and adjusted so that its FOV center aligned with that of the ASL's (Figure 2A). A substantial fisheye distortion was present at the extreme edges of the GoPro FOV. However, the fixations analyzed in the study were mainly restricted to the central region where distortion was minimized.

Calibration

The ASL estimates gaze position by learning the mapping between specific locations in the world (in *x*-*y* coordinates relative to the scene camera) and the displacement vector from the pupil center to the primary IR dot registered by the eye camera. To minimize head movements during calibration, subjects placed their heads on a chin rest located 42 cm from an 18-in. CRT monitor centered in the subject's FOV with a resolution of 1024 × 768 pixels (spanning 50° horizontally and 37.5° vertically). To maximize calibration accuracy and reliability, subjects completed a standardized calibration task written in MATLAB (MathWorks, Natick, MA) and PsychToolbox 3.0.10 (Brainard, 1997). Subjects first fixated on a centrally presented black dot (outer radius 1.0°) with a small gray circular center (inner radius 0.15°). When the subjects were confident they were fixating steadily as close to the dot center as possible, they pressed the spacebar. The dot then relocated randomly to one of 12 positions arranged in a 4 × 3 grid, spaced 14.0° apart horizontally and 15.8° vertically (spanning 42.0° × 31.6°; Figure 2B). The subjects would then fixate the new dot location and again press the spacebar, proceeding through the 13 locations (12 grid plus initial central). After all dots were fixated, an image of the entire array appeared, during which the subjects were instructed to look at the center of each dot, starting from the upper left and moving left to right and row by row for post hoc validation.

This data was used after the testing session for manual calibration using ASL's EyeXG software (ASL Eye Tracking). Independent raters viewed the scene camera video in slow motion (8 fps) with an image of the pupil and displacement vector from the eye camera superimposed. For each calibration dot transition event, the raters waited for the subject's eye to move and stabilize on the new location as ascertained by an abrupt shift in the overlaid pupil/displacement vector. The rater used a mouse to manually select the location of the center of the current calibration dot on the scene camera image (Figure 3A). The ASL EyeXG software

then computed a function that mapped the displacement vectors (eye camera) to the dot locations (scene camera) for the 13 calibration dots for each subject.

Gaze location and fixation event detection

Subjects' gaze location (in x-y coordinates) relative to the scene camera image for each valid frame was estimated by the ASL EyeXG software using the mapping function learned during calibration (Figure 3D). Frames were defined as invalid if the corneal reflection was lost during saccades, blinks, large eccentricity fixations, or extreme external IR illumination and were not included in the analysis. Across all subjects, 67.3% ± 3.4% (mean ± standard error of the mean) of frames were classified as valid, with no significant difference in the percentage of valid frames between looking groups (ULs: 69.3% ± 5.5%, MLs: 67.3% ± 7.3%, LLs: 65.4% ± 5.4%; $p = 0.91$).

Fixations were defined by the automated ASL algorithm as events where six or more consecutive samples (100 ms) were measured within 1° of the sample group centroid. Fixation events were terminated when three consecutive samples measured greater than 1° from the fixation centroid or when pupil data was lost for 12 or more samples (200 ms; Figure 3D). To check the accuracy of this automated algorithm, we reanalyzed the data using two techniques shown to be robust for fixation detection in noisy eye tracking data with significant flicker: (a) a modified automatic fixation detection protocol that incorporates bilateral filtering, and (b) human rater hand-coding (Holmqvist et al., 2011; Wass, Smith, & Johnson, 2012). While both techniques yielded fewer and longer fixations than the ASL algorithm, there was close agreement between all three regarding fixation positions (see Figures 9 and 10 and the Appendix for detailed methods).

Synchronization and registration

The ASL EyeXG software outputs an estimated gaze location for each frame in x-y coordinates relative to the ASL native scene camera, but ultimately we wanted to map these fixation coordinates to the higher resolution, larger FOV GoPro video. To do this, we presented a 16×12 checkerboard pattern on the monitor immediately after validation (Figure 3B). After the fact, we implemented an automatic routine in MATLAB that searched for the first frame in the native scene camera video in which a 16×12 checkerboard pattern could be detected. The time in the video was recorded and the coordinates of the checkerboard vertices (192 points) automatically detected (Figure 3B). The same was done with the GoPro video. The video streams were then synchronized by aligning the checkerboard onset times. Then, we computed the projective linear transform matrix, T, that mapped the 192 vertex points from ASL to GoPro coordinates with the minimum mean-square error. The transform matrix was then used to map gaze coordinates for each frame and each fixation event from the ASL video to the GoPro (Figure 3D).

Recalibration and reregistration

To ensure data validity over the course of the study, subjects regularly performed a recalibration and reregistration routine. Every 3 min, the subject was instructed to stop and hold at arm's distance a calibration/registration checkerboard pattern centered at eye level. While keeping the head steady, the subject would fixate, in turn, the extreme upper-left, upper-right, lower-left, and lower-right corners of the checkerboard for two seconds each before resuming their walk (Figure 3C). Similar to the initial calibration, independent raters viewed each recalibration at 8 fps. For each of the four corner fixations, the raters waited until the subject's eye moved and stabilized on the new location indicated by a sudden shift and stabilization of the overlaid pupil/displacement vector. The rater selected the location of the center of the current recalibration target on the scene camera image (Figure 3C), which the ASL EyeXG software used to augment the displacement vector to gaze-location mapping function. Similarly, the 16×12 checkerboard pattern and its corresponding vertices were automatically detected in both videos and any necessary adjustments to the transform matrix were applied.

Analysis: Automatic fixation event filtering

On average, we obtained 24.2 min (87,165 frames) of data per subject (Figure 4A). For this study, we were interested only in the fixation location targeted by saccades. This information is contained completely in the image and gaze position corresponding to the first frame of each detected fixation event. This allowed us to greatly reduce our data set by automatically selecting, for each fixation, a single video frame and eye position for further analysis (average of 3,023 frames/subject; Figure 4B).

Analysis: Crowdsourcing face-fixation events

One of the primary difficulties with studies conducted outside traditional laboratory environments is the decreased ability to control subjects' sensory

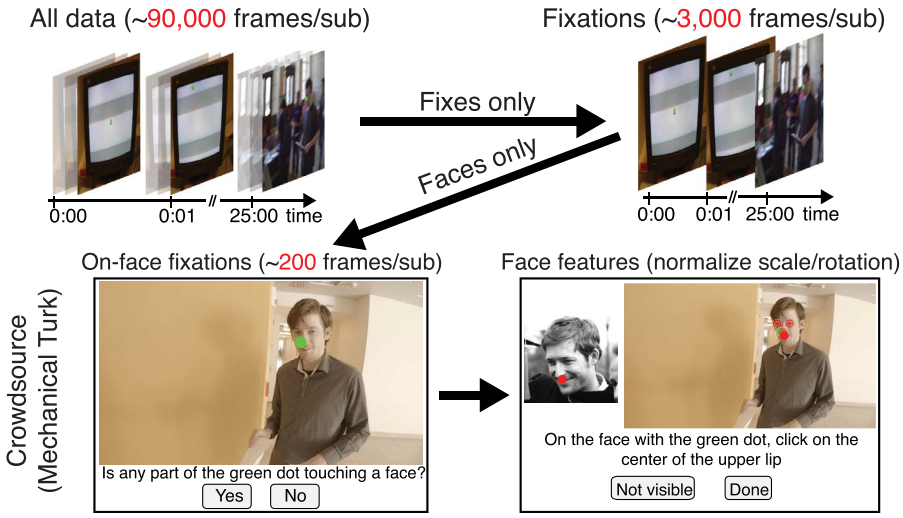

Figure 4. Analysis and interpretation of fixation data. The current study is concerned only with the locations of distinct fixation events, greatly reducing the amount of data to be analyzed (from 60 to around two samples per second). Since only on-face fixations were relevant here, data was further refined with the help of human raters on Mechanical Turk. Finally, human raters were again enlisted to determine the location of the on-face fixations relative to the eyes and mouth.

input. In the lab, the experimenter precisely determines the spatial and temporal characteristics of visual stimulation. Thus, the position (x, y) of gaze at some time (t) unambiguously maps to known stimulus properties. Unconstrained environments do not provide this level of control, as the spatiotemporal properties of the visual stimulus are not known a priori. This situation makes measurements of gaze timing and position necessary but not sufficient for mapping to meaningful stimulus properties. The difficulty of this mapping is determined by the stimulus properties the experimenter is interested in, the quality of the visual recording, and the complexity of the visual environment.

In this study we are interested in how people look at faces. This goal requires the ability to reliably determine whether a fixation is on a face given only the recorded video image and the associated x-y gaze position. While advances in algorithms and computing resources have led to impressive gains in automatic face detection within complex images (Phillips & O'Toole, 2014; Taigman, Yang, Ranzato, & Wolf, 2014), the combination of high-resolution video and unconstrained environmental uncertainty poses a serious challenge to even the most advanced computer face detection systems. In this type of scenario, humans remain the gold standard for face detection accuracy. However, this advantage comes at a cost of processing capacity: An individual can accurately detect faces only up to a certain speed.

To maximize accuracy and throughput, we developed a simple crowdsourcing algorithm using Amazon Mechanical Turk. By drawing on the judgments of many individuals in parallel, crowdsourcing greatly increases the bandwidth of human-based face recognition. Turk raters were shown a series of randomly sampled single video frames corresponding to fixation onsets as described in the previous section. For each image (trial), a bright green dot was overlaid at the measured fixation location, and the rater responded whether any portion of the green dot was touching a face (Figure 4C). To ensure raters were real humans who understood and were actively attending to the task, each image was rated by multiple people. If the first two raters agreed, the response was taken as truth and the image was removed from the rating pool. If the first two raters did not agree, the image was shown to a third tie-breaking rater. Individual raters' performance was monitored by calculating their miss (responding No Face when two separate raters responded Face) and false alarm rates (responding Face when two other raters responded No Face). For online quality assurance, each trial had a one in 30 chance of being a probe. The probe set was a mixture of 80 author-verified images and an expanding set of images that had already been successfully rated by two other raters (who had not themselves been excluded because of low concordance with other raters), with author-verified images more likely to be sampled on earlier trials. If the rater disagreed with the consensus, they would be given a warning message. Raters were allowed two mistakes; a third disqualified them from further participation and all of their rating data was discarded from final analyses. Post hoc manual verification by the authors of a random sample of rated images revealed no false positives or negatives.

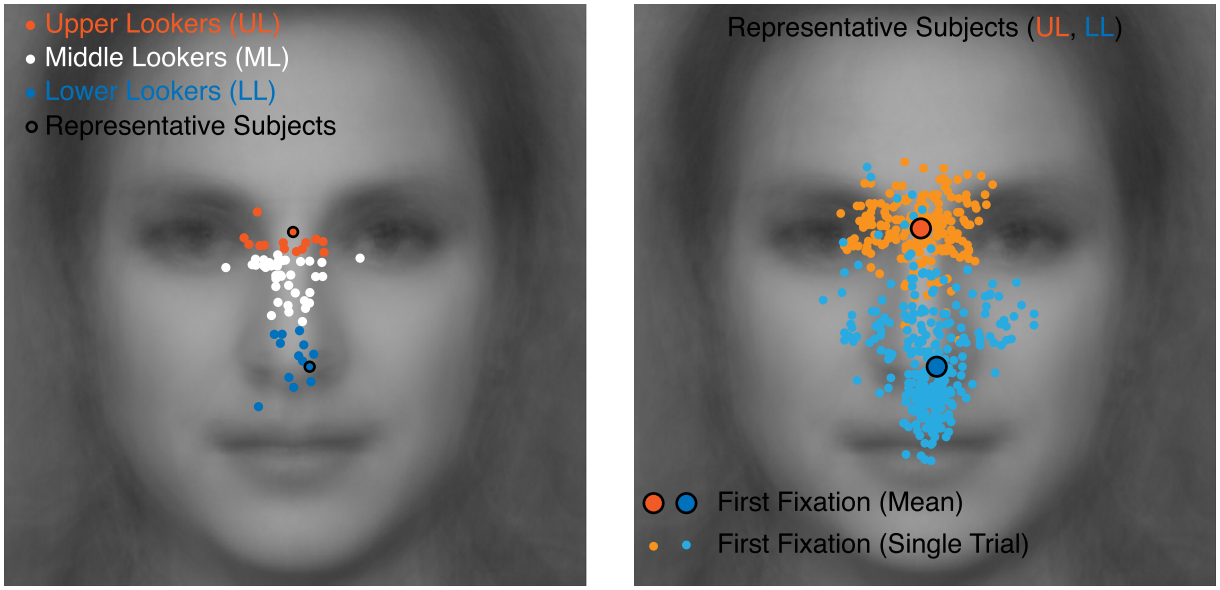

Figure 5. Stage I (in-lab) initial-fixation behavior for face identification. On the left, each dot represents the mean location, across trials, of the initial on-face fixation for one subject. Subjects were categorized as UL (orange), ML (white), or LL (teal) according to predetermined criteria based on previous work. On the right, fixations for each trial (small dots) and the mean across trials (large dots) for one UL (orange) and one LL (teal).

Analysis: Crowdsourcing face-fixation location

To quantitatively compare within face-fixation location between the laboratory and the real world, we need to compute the relative fixation metric, γ (see Equation 1 in the Analysis section of Methods [Stage I]). In the lab, this calculation is simple, as the position of the eyes and mouth are set and known by the experimenter. For the mobile section, we need to estimate these locations on the video frames where faces could be present at any combination of location, pose and size. We again turned to crowdsourcing with a second Mechanical Turk task. Raters were shown random frames that were determined from the first Turk task to have on-face fixations (again signified by a green dot). If the rater determined that the image was originally misclassified as face-present in the first Turk task, a No Face option was available that recycled the image back to the previous Turk task pool. Otherwise, raters were first asked to rotate the image until the face with the dot on it was upright and then clicked on the center of one of the visible eyes and the center of the upper lip (the upper lip was chosen so as to minimize the variability in estimated mouth position due to plastic changes arising from talking, expressions, etc.; Figure 4D). The γ was then computed as before (Equation 1). Each image was scored by two raters. If the raters disagreed by more than 10° of rotation and/or more than 10% of the eye-to-mouth distance, a third rater scored the image and the two most similar ratings were averaged. After the fact, manual verification of a random sample of rated images showed good agreement by the raters and no systematic biases.

Results

In-lab initial face-fixation behavior

Across subjects, the initial into-face saccade landed on average below the eyes (mean ± standard error of the mean: $\gamma = 0.757 \pm 0.025$, $t[69] = 9.86$, $p < 0.001$) and left of the midline ($\chi = 0.041 \pm 0.014$, $t[69] = 3.02$, $p = 0.0035$; Figure 5). Consistent with past literature, individuals varied greatly and consistently in the their preferred face-fixation behavior along the vertical dimension, ranging from the eye brows ($\max[\gamma] = 1.11 \pm 0.061$) to just above the mouth ($\min[\gamma] = 0.17 \pm .065$; Figure 5; Gurler et al., 2015; Mehoudar et al., 2014; Peterson & Eckstein, 2013).

An existing independent sample of face-looking behavior ($n = 275$) was used to predefine criteria to categorize the current subject sample into three groups. ULs fixate higher on the face than 85% of the total previously sampled population ($\gamma_{UL} = 0.93$), LLs fixate lower than 85% ($\gamma_{LL} = 0.55$), with MLs constituting everybody else. Using this criteria, 11 of 70 subjects were categorized as ULs (15.7%), 14 as LLs (20.0%), and 45 as MLs (64.3%; Figure 5). The 10 subjects with the best Stage I EyeLink calibration scores from each group were recalled for the mobile

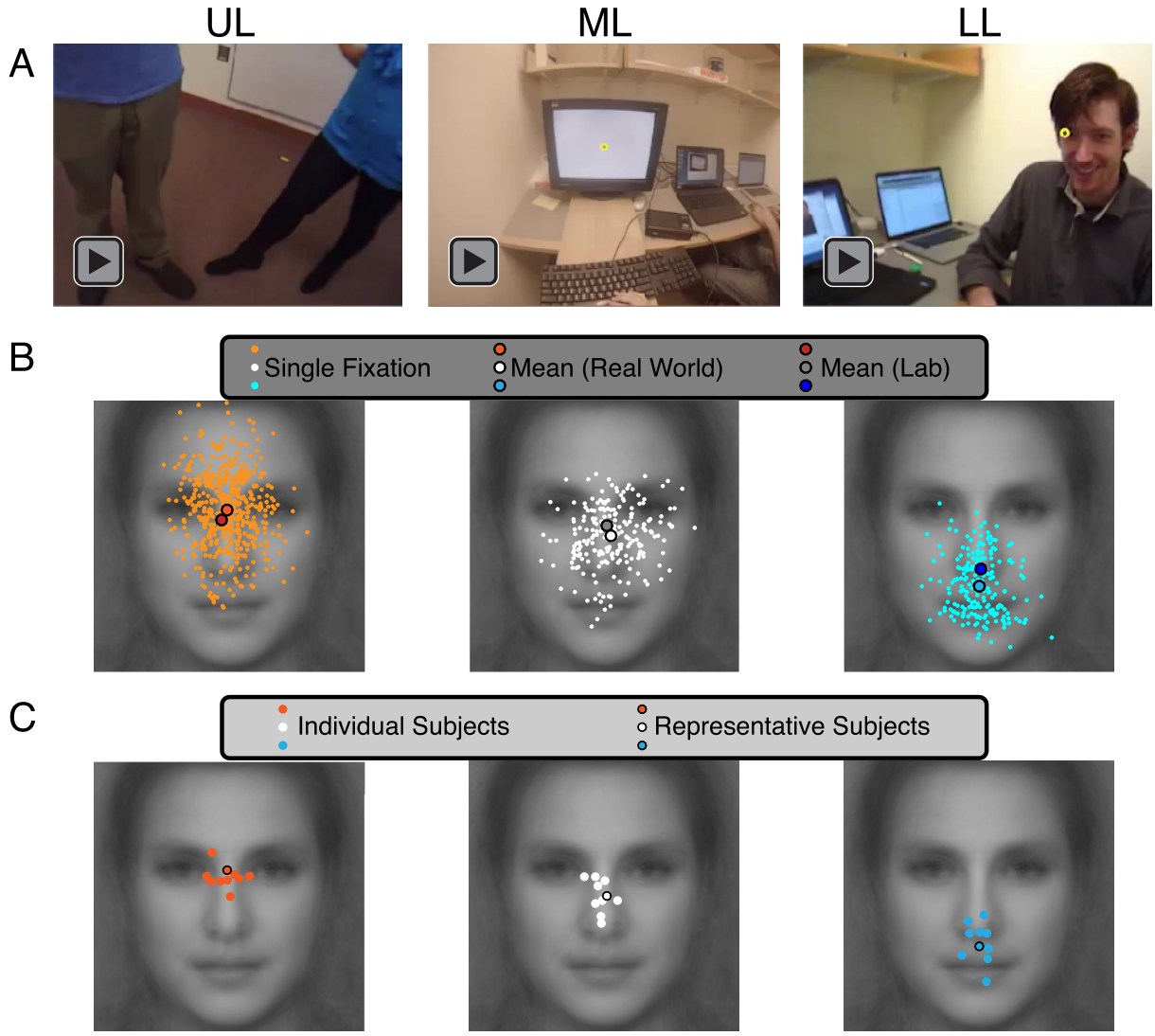

Figure 6. Real-world face-fixation behavior for lab-defined ULs, MLs, and LLs. (A) Each video is from one representative subject from each group, with the white dot denoting gaze position. (B) Dots represent individual on-face fixation events for the same subjects as (A). (C) Each dot represents the mean location across all on-face fixations for a single subject.

condition, resulting in the following γ values (mean $\pm$ standard deviation) for each group: ULs: $\gamma_{UL} = 0.995 \pm 0.098$, MLs: $\gamma_{ML} = 0.815 \pm 0.133$, LLs: $\gamma_{LL} = 0.326 \pm 0.183$ (Figure 7A).

Real-world face-fixation behavior

Subjects' distinctive preferred face-fixation behavior can be appreciated in the example subject videos from each looking group (Figure 6A): Individuals fixated predominantly at their preferred region, with occasional fixations on other face regions quickly followed by a return to the preferred region. Most importantly, individuals' preferred real-world fixation regions were consistent with their laboratory fixations (Figure 6B). Grouping subjects according to their in-lab behavior, the data from real-world viewing showed that ULs ($\gamma_{UL} = 0.921 \pm 0.040$) looked significantly higher than MLs ($\gamma_{ML} = 0.735 \pm 0.056$, $t[18] = 2.91$, $p = 0.005$) who looked significantly higher than LLs ($\gamma_{LL} = 0.267 \pm 0.066$, $t[18] = 5.69$, $p < 0.001$; Figures 6C, 7A).

Relationship between in-lab and real-world face-fixation behavior

A repeated measures two-way analysis of variance found significant main effects of looking group ($F[2, 27] = 65.45$, $p < .001$) and modality (laboratory vs. real world; $F[2, 27] = 9.62$, $p = 0.004$) on fixation behavior (γ), but not a significant interaction ($F[2, 27] = 0.28$, $p = 0.76$; Figure 7A).

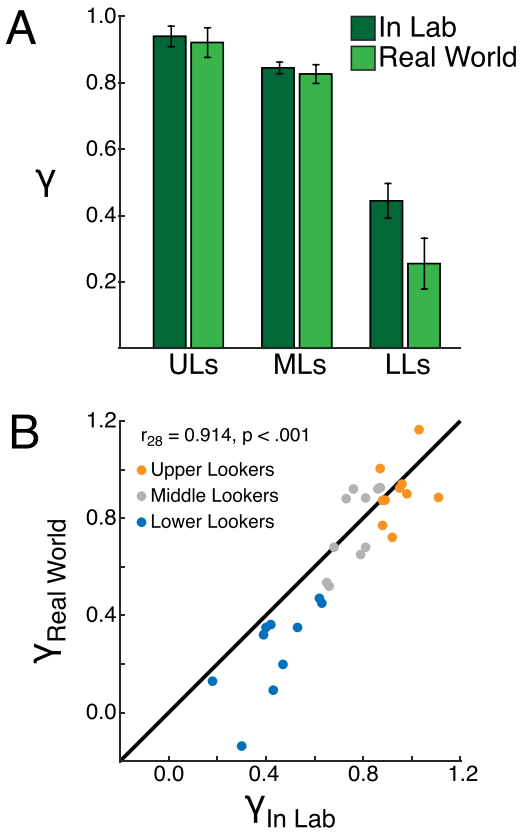

Figure 7. Relationship between real-world and in-lab face-fixation behavior. (A) The lab-measured group differences in the mean location of the initial on-face fixation, from 0 (center of the mouth) to 1 (center of the eyes), are also observed under real-world conditions. (B) The conservation of face-gaze patterns between the lab and the world is consistent at the individual level across the range of observed behavior.

Across the sample, correlational analysis showed that an individual's real-world fixations were strongly predictive of their laboratory behavior ($r[28] = .914$, $p < 0.001$; Figure 7B). This relationship was near ceiling given the reliability of the each modality's measurements. For each of 1,000 bootstrap samples, we randomly split each subject's data in half, computed γ for each half, and calculated the correlation between the two halves. The average split-half reliabilities were $r = 0.996$ and $r = 0.909$ for the in-lab and real-world measurements, respectively, with an average split-half correlation of $r = 0.905$ between them (correlation value lower than for the full data set due to smaller sample sizes).

Discussion

Here we tested whether individual differences in face-looking behavior, observed previously only in restricted lab conditions, generalize to the real world. To answer this question, we measured subjects' fixation positions on faces both under controlled lab conditions and while they walked around the MIT campus. Our main finding is that face-fixation patterns are remarkably similar in the two situations, with an individual's lab fixation behavior strongly predicting their real-world gaze, nearly as well as possible given measurement reliability (Figure 7). These results demonstrate that the prior lab-based finding of individual differences in face-fixation behavior generalizes to real-world vision. They further imply that the superior face recognition performance when an individual fixates his or her preferred location (Peterson & Eckstein, 2013) both reflects, and optimizes, that person's real-world face recognition behavior. Taken together, these results suggest that real-world face recognition entails two qualitatively distinct stages: face detection in the periphery and face recognition at the fovea. Finally, the methods developed here provide a rich dataset of images that humans have actually experienced during real-world viewing, including the viewer's fixation position on each image, opening up important new avenues for investigation of the statistics of the images landing on peoples' retinas during natural behavior (RIS), and the tuning of human behavior and neural representations to those statistics.

Comparing real-world and in-lab eye movements

The work presented here builds on previous studies that have sought to characterize how people move their eyes in naturalistic real-world environments and how these eye movements relate to those observed under controlled laboratory conditions. Most mobile eye-tracking studies have assessed fixation behavior while subjects execute specific tasks, generally within a single location (making tea or sandwiches: Hayhoe, 2000; Hayhoe & Ballard, 2005; Land, Mennie, & Rusted, 1999; driving: Land, 1992; Land & Lee, 1994; visual search: Foulsham, Chapman, Nasiopoulos, & Kingstone, 2014; Mack & Eckstein, 2011; gaze-cueing: Macdonald & Tatler, 2013, 2015; social: Einhäuser et al., 2009; Laidlaw, Foulsham, Kuhn, & Kingstone, 2011; Risko, Laidlaw, Freeth, Foulsham, & Kingstone, 2012). A smaller number of studies have assessed eye movements in unconstrained natural environments and behavior (Cristino & Baddeley, 2009; Foulsham, Walker, & Kingstone, 2011; Hart et al., 2009). In general, these studies have assessed coarse statistical trends across groups of subjects (e.g., tendency to fixate the image center in the lab versus a "world-center," the horizon, outside the lab). The improved reliability of data collection and efficiency of data analysis provided by the techniques developed here allow for a significant

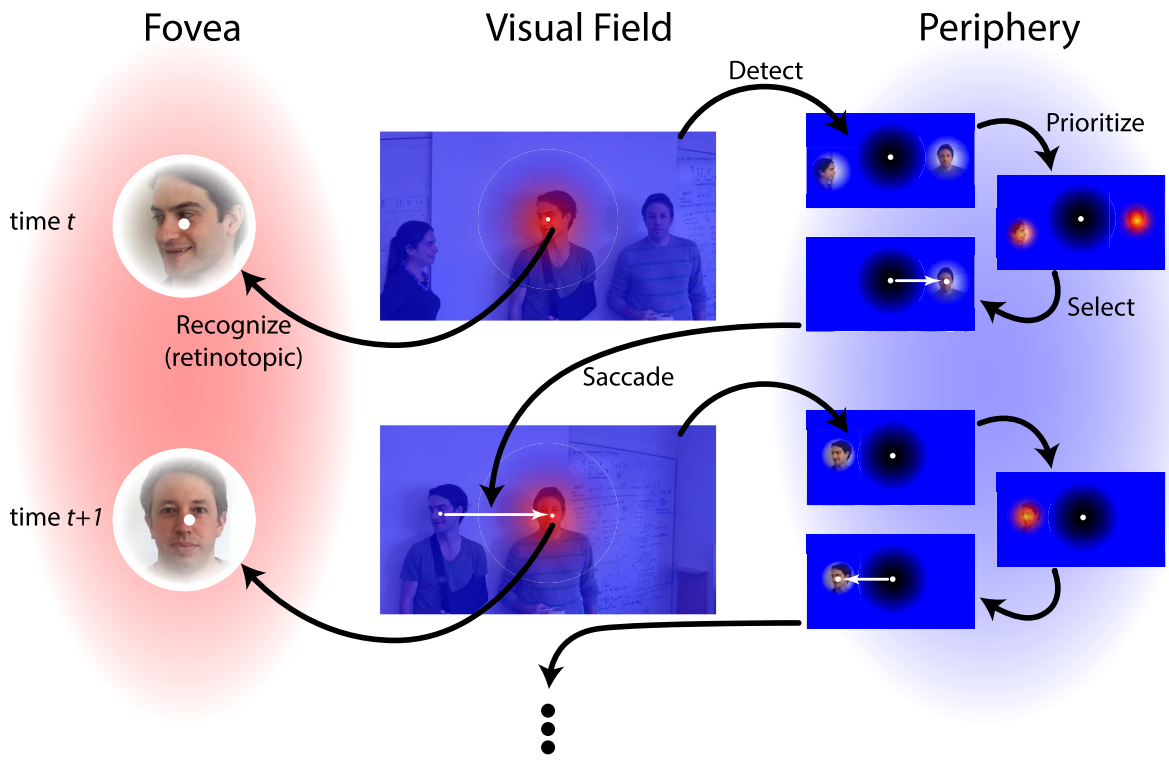

Figure 8. Schematic of a parallel peripheral-detection/foveal-recognition model. At any given time, *t*, the foveal (red) and peripheral (blue) retinal images are determined by the position of the body, head, and eyes. Peripheral mechanisms are tuned to image properties that support face detection. Faces likely to contain important visual information are selected and targeted with eye movements, providing powerful foveal resources for detailed recognition tasks. The eye movement is precise and individual-specific, eliminating image translation variance and possibly matching retinotopic face representations.

expansion of the type and scope of real-world eye tracking studies (Figures 2 through 4).

Peripheral detection and foveal recognition as distinct stages of face perception

The evidence presented here suggests that real-world face recognition entails a systematic sequence of processing steps in which detection operates in the periphery in parallel with recognition at the fovea (Figure 8). According to this hypothesis, the detection mechanism continuously monitors for the presence of faces in the visual periphery (Step 1: Detect). Relevant features of peripheral faces that can be computed with adequate precision (e.g., location, size, pose, motion) are then combined to form a retinotopic face priority map, which is integrated with other social and nonsocial priority calculations to form a general attention-guiding priority map (Step 2: Prioritize; Bisley & Goldberg, 2010; Fecteau & Munoz, 2006; Itti, Koch, & Niebur, 1998; Koehler, Guo, Zhang, & Eckstein, 2014). Next, the highest priority location is selected for subsequent fixation (Step 3: Select). When a face is selected for the next fixation, the eye movement system exploits the stereotyped T-shaped configuration of facial features to precisely target saccades to the individual's specific preferred face-fixation position (Step 4: Saccade). This brings the face image to a reliable position on the fovea, where it is processed by specialized recognition mechanisms, shown previously to be highly retinotopically specific (Step 5: Recognize; Peterson & Eckstein, 2012). According to this model, face detection and face recognition are fundamentally different processes, with detection occurring for faces in the periphery at a wide range of eccentricities and positions, and recognition proceeding at the fovea on faces that are usually centered at a single stereotyped retinal location. Note that Steps 1–4 (detection, prioritization, selection, and saccadic targeting of peripheral faces) likely proceed in parallel with Step 5 (recognition of the currently foveated face).

The model of face perception just sketched can be tested using the methods developed in the current study. In particular, we can use our growing database of natural images our observers experienced (including their fixation position on those images) to ask: (a) Where do faces land on the retina in real-world viewing? (b) What are the features of peripherally viewed faces that guide selection for saccadic targeting?

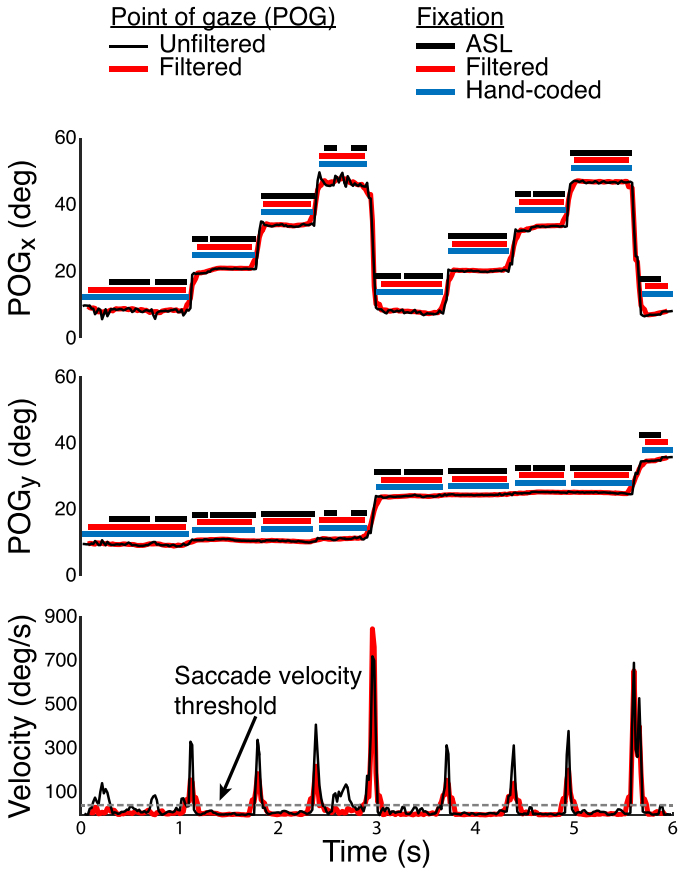

Figure 9. Example of fixation validation for a 6-s segment from one subject's data. Thin black lines in the top and middle plots are the raw *x* and *y* gaze coordinates, respectively, with thin red lines denoting the gaze position after bilateral filtering and interpolation. Above the traces, bars represent the times of individual fixation events detected by the ASL algorithm (black), by the new filtering procedure (red), and by hand (blue). The bottom plot shows instantaneous velocity for the raw (black) and filtered (red) data, with the dotted gray line denoting the threshold used for saccade detection.

(c) Is human size invariance for face recognition tuned to the statistics of retinal face sizes that occur during natural viewing? The general hypothesis that we can now test in detail is that the face detection and face recognition systems are each specifically tuned for task-specific statistics of experienced natural images.

Retinal image statistics

More broadly, this work makes possible a richer and more ecologically valid dataset with which to test the core ideas of Natural Systems Analysis (Geisler, 2008)—that the computations employed by the visual system are the product of evolutionary optimization for the sensory evidence (i.e., images) and tasks critical for survival. A deep understanding of these systems requires knowledge of the properties of the visual environment in which they operate (i.e., natural image statistics; Botvinick, Weinstein, Solway, & Barto, 2015; Olshausen & Field, 1996; Simoncelli & Olshausen, 2001; Torralba & Oliva, 2003). While the study of natural image statistics has provided crucial insights into the computations carried out by the visual system, the degree to which these images faithfully represent real-world visual experience is unclear. Prior studies have typically analyzed sets of narrow FOV static photographs that have not been selected to reflect everyday visual experience. Critically, these images do not have fixation data, a critical missing element given the radically lower visual acuity in the visual periphery. The framework presented here simultaneously collects high resolution, wide FOV video of the visual environment and corresponding eye movements, allowing us to directly measure the retinotopic images people experience in everyday life, which we term RIS. This new database should be applicable to myriad problems of vision beyond face perception.

Real-world face fixations in impaired populations

Finally, the methods developed here enable us to rigorously measure real-world gaze behavior in populations that may have deficits in face recognition. Fixation behavior may be a prime determinant of successful face recognition, yet how those with possible recognition deficits look at faces in the real world is largely unknown.

For example, a deficit in the recognition of faces is frequently reported in Autism Spectrum Disorder (ASD). While the findings in the literature are conflicting, most evidence suggests that face recognition impairments in people with ASD are greater under natural viewing conditions (e.g., static vs. dynamic, computer images vs. real faces; Jemel, Mottron, & Dawson, 2006; Weigelt, Koldewyn, & Kanwisher, 2012). The literature is also conflicting on the question of whether people with ASD differ from typically developing (TD) subjects in the way they look at faces, but avoidance of faces in general and eyes in particular apparently becomes more pronounced with increasing naturalism (Gharib, Adolphs, & Shimojo, 2014; Klin, Jones, Schultz, Volkmar, & Cohen, 2002; Speer, Cook, McMahon, & Clark, 2007). Most importantly, few studies have measured gaze behavior on faces in natural viewing in people with ASD (Magrelli et al., 2013; Vabalas & Freeth, 2015), and none have done so on a large scale during normal behavior in unconstrained environments. Overall, the evidence suggests that any differences in face perception between people

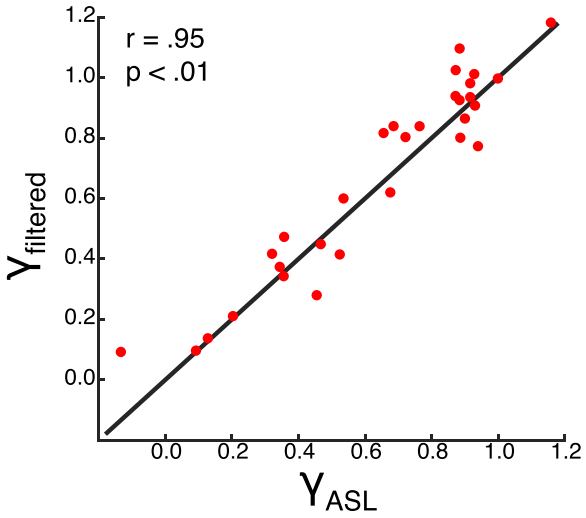

Figure 10. Correspondence between each subject's mean face-fixation location (γ) according to fixations detected by the ASL algorithm (*x*-axis, γ_{ASL}) and the new filtering procedure (*y*-axis, $\gamma_{filtered}$).

with ASD and TDs should be greatest under these conditions, which we can test in the future using the methods developed here.

Another disorder that may be informed by tests of real-world gaze behavior is developmental prosopagnosia (DP), a lifelong deficit in face recognition in the absence of known neurological damage (Behrmann & Avidan, 2005; Duchaine & Nakayama, 2006; Zhang, Liu, & Xu, 2015). The few studies that have examined face-looking behavior in people with DP have incorporated small sample sizes (often a single patient) and lab viewing conditions (Barton, Radcliffe, Cherkasova, & Edelman, 2007; Bate, Haslam, Tree, & Hodgson, 2008; Pizzamiglio et al., 2015; Schmalzl, Palermo, Green, Brunsdon, & Coltheart, 2008; Schwarzer et al., 2006). A natural hypothesis is that some or all of the deficits in face recognition in people with DP result from suboptimal and/or inconsistent looking behavior on faces, which could disrupt the normal development of face representations and/or the ability to enact eye movement strategies that reliably constrain retinotopic input.

Finally, it is of great interest to understand how, why, and when individuals acquire their distinct face gaze behavior. One possibility is that retinotopic tuning of face representations is present at birth, with location tuning varying across the population. This account holds that individuals learn fixation strategies that are optimized for their specific tuning. A second, more likely possibility is that face representations are not strongly tuned to position at birth. Rather, individuals vary, for whatever reasons, in where they look on faces. This early retinotopic visual experience might then guide the learning and development of the basic structure of face representations. This situation could create a positive feedback scenario, such that the performance advantage for fixating a specific region provides an incentive to maintain this looking behavior. On this hypothesis, any early disruption of face-looking behavior could lock in a self-reinforcing cycle of suboptimal face representations and suboptimal face-looking behavior, providing a possible account of developmental prosopagnosia and/or face deficits in people with ASD. This hypothesis could also account for the lifelong face perception deficits in individuals treated early in life for bilateral or left (but not right) lateralized congenital cataracts that deprive face-selective regions in the right hemisphere of patterned visual input for a brief period after birth (Le Grand, Mondloch, Maurer, & Brent, 2001, 2003). A final possibility is that although face representations are retinotopically specific, the general ability to encode new faces is not itself tuned to an individual's particular fixation preference. Rather, consistently fixating the same position causes most face memories to be encoded relative to the individual's specific preferred gaze location. According to this hypothesis, the stability of an individual's specific face-fixation behavior optimizes recognition by matching the retinotopic position of the current face to the retinotopic positions of previously encoded faces. This matching hypothesis predicts that individuals should identify new faces best when they are trained and tested at the same fixation position. Critically, there should be no correlation between individual differences in preferred fixation position and the fixation position during learning that leads to maximum recognition performance during test.

Conclusion

In sum, we found that individual differences in face-fixation behavior reported previously in the lab generalize to real-world viewing. These findings suggest a distinction between two components of face perception: detection of faces in the periphery and recognition of faces in the fovea. These findings also suggest possible causes of lifelong deficits in face perception in people with DP, ASD, and congenital cataracts. Finally, the methods developed here make possible the large-scale collection natural images as seen by humans, including the critical information of fixation position on each image, a dataset that may open up important new constraints on natural systems analysis (Geisler, 2008).

Keywords: mobile eye tracking, eye movements, face recognition, natural systems, retinal image statistics

Acknowledgments

We would like to thank Jason Fischer for his helpful comments on this manuscript. This work was supported by the Center for Brains, Minds, and Machines (CBMM), funded by National Science Foundation Science and Technology Centers award and Computing and Communication Foundations – 1231216. The authors have no financial or proprietary interests.

Commercial relationships: None.
Corresponding author: Matthew F. Peterson.
Email: mfpeters@mit.edu.
Address: Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA.

References

Barton, J. J. S., Radcliffe, N., Cherkasova, M. V., & Edelman, J. A. (2007). Scan patterns during the processing of facial identity in prosopagnosia. *Experimental Brain Research*, *181*(2), 199–211, doi:10.1007/s00221-007-0923-2.

Bate, S., Haslam, C., Tree, J. J., & Hodgson, T. L. (2008). Evidence of an eye movement-based memory effect in congenital prosopagnosia. *Cortex*, *44*(7), 806–819, doi:10.1016/j.cortex.2007.02.004.

Behrmann, M., & Avidan, G. (2005). Congenital prosopagnosia: Face-blind from birth. *Trends in Cognitive Sciences*, *9*(4), 180–187, doi:10.1016/j.tics.2005.02.011.

Bisley, J. W., & Goldberg, M. E. (2010). Attention, intention, and priority in the parietal lobe. *Annual Review of Neuroscience*, *33*, 1–21, doi:10.1146/annurev-neuro-060909-152823.

Botvinick, M., Weinstein, A., Solway, A., & Barto, A. (2015). Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current Opinion in Behavioral Sciences*, *5*, 71–77, doi:10.1016/j.cobeha.2015.08.009.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Cristino, F., & Baddeley, R. (2009). The nature of the visual representations involved in eye movements when walking down the street. *Visual Cognition*, *17*(6/7), 880–903, doi:10.1080/13506280902834696.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341, doi:10.1016/j.tics.2007.06.010.

Duchaine, B. C., & Nakayama, K. (2006). Developmental prosopagnosia: A window to content-specific face processing. *Current Opinion in Neurobiology*, *16*(2), 166–173, doi:10.1016/j.conb.2006.03.003.

Durand, F., & Dorsey, J. (2002). Fast bilateral filtering for the display of high-dynamic-range images. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 257–266). New York: Association for Computing Machinery. doi:10.1145/566570.566574.

Einhäuser, W., Schumann, F., Vockeroth, J., Bartl, K., Cerf, M., Harel, J., & König, P. (2009). Distinct roles for eye and head movements in selecting salient image parts during natural exploration. *Annals of the New York Academy of Sciences*, *1164*(1), 188–193, doi:10.1111/j.1749-6632.2008.03714.x.

Fecteau, J. H., & Munoz, D. P. (2006). Salience, relevance, and firing: A priority map for target selection. *Trends in Cognitive Sciences*, *10*(8), 382–390, doi.org/10.1016/j.tics.2006.06.011.

Foulsham, T., Chapman, C., Nasiopoulos, E., & Kingstone, A. (2014). Top-down and bottom-up aspects of active search in a real-world environment. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *68*(1), 8–19, doi:10.1037/cep0000004.

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, *51*(17), 1920–1931, doi:10.1016/j.visres.2011.07.002.

Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, *110*(2), 160–170, doi:10.1016/j.cognition.2008.11.010.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, *59*(1), 167–192, doi:10.1146/annurev.psych.58.110405.085632.

Gharib, A., Adolphs, R., & Shimojo, S. (2014). "Don't look": Faces with eyes open influence visual behavior in neurotypicals but not in individuals with high-functioning autism. *Journal of Vision*, *14*(10), 681, doi:10.1167/14.10.681. [Abstract]

Gurler, D., Doyle, N., Walker, E., Magnotti, J., & Beauchamp, M. (2015). A link between individual differences in multisensory speech perception and eye movements. *Attention, Perception, & Psychophysics*, *77*(4), 1333–1341, doi:10.3758/s13414-014-0821-1.

't Hart, B. M., J. Vockeroth, F. Schumann, K. Bartl, E. Schneider, P. König, & Einhäuser, W. (2009).

Gaze allocation in natural stimuli: Comparing free exploration to head-fixed viewing conditions. *Visual Cognition*, *17*(6–7), 1132–1158, doi:10.1080/13506280902812304.

Hayhoe, M. (2000). Vision using routines: A functional account of vision. *Visual Cognition*, *7*(1–3), 43–64, doi:10.1080/135062800394676.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194, doi:10.1016/j.tics.2005.02.009.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). Eye tracking: A comprehensive guide to methods and measures. Oxford, UK: Oxford University Press.

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

Jemel, B., Mottron, L., & Dawson, M. (2006). Impaired face processing in autism: Fact or artifact? *Journal of Autism and Developmental Disorders*, *36*(1), 91–106, doi:10.1007/s10803-005-0050-5.

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry*, *59*(9), 809, doi:10.1001/archpsyc.59.9.809.

Koehler, K., Guo, F., Zhang, S., & Eckstein, M. P. (2014). What do saliency models predict? *Journal of Vision*, *14*(3):14, 1–27, doi:10.1167/14.3.14. [PubMed] [Article]

Laidlaw, K. E. W., Foulsham, T., Kuhn, G., & Kingstone, A. (2011). Potential social interactions are important to social attention. *Proceedings of the National Academy of Sciences*, *108*(14), 5548–5553, doi:10.1073/pnas.1017022108.

Land, M. F. (1992). Predictable eye-head coordination during driving. *Nature*, *359*(6393), 318–320, doi:10.1038/359318a0.

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature*, *369*(6483), 742–744, doi:10.1038/369742a0.

Land, M. F., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, *28*(11), 1311–1328, doi:10.1068/p2935.

Le Grand, R., Mondloch, C. J., Maurer, D., & Brent, H. P. (2001). Neuroperception: Early visual experience and face processing. *Nature*, *410*(6831), 890–890, doi:10.1038/35073749.

Le Grand, R., Mondloch, C. J., Maurer, D., & Brent, H. P. (2003). Expert face processing requires visual input to the right hemisphere during infancy. *Nature Neuroscience*, *6*(10), 1108–1112, doi:10.1038/nn1121.

Macdonald, R. G., & Tatler, B. W. (2013). Do as eye say: Gaze cueing and language in a real-world social interaction. *Journal of Vision*, *13*(4):6, 1–12, doi:10.1167/13.4.6. [PubMed] [Article]

Macdonald, R. G., & Tatler, B. W. (2015). Referent expressions and gaze: Reference type influences real-world gaze cue utilization. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(2), 565–575, doi:10.1037/xhp0000023.

Mack, S. C., & Eckstein, M. P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of Vision*, *11*(9):9, 1–16, doi:10.1167/11.9.9. [PubMed] [Article]

Magrelli, S., Jermann, P., Noris, B., Ansermet, F., Hentsch, F., Nadel, J., & Billard, A. (2013). Social orienting of children with autism to facial expressions and speech: A study with a wearable eye-tracker in naturalistic settings. *Frontiers in Psychology*, *4*, 840, doi:10.3389/fpsyg.2013.00840.

Mehoudar, E., Arizpe, J., Baker, C. I., & Yovel, G. (2014). Faces in the eye of the beholder: Unique and stable eye scanning patterns of individual observers. *Journal of Vision*, *14*(7):6, 1–11, doi:10.1167/14.7.6. [PubMed] [Article]

Olshausen, B. A., & Field, D. J. (1996). Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, *7*(2), 333–339, doi:10.1088/0954-898X_7_2_014.

Or, C. C.-F., Peterson, M. F., & Eckstein, M. P. (2015). Initial eye movements during face identification are optimal and similar across cultures. *Journal of Vision*, *15*(13):12, 1–25, doi:10.1167/15.13.12. [PubMed] [Article]

Peterson, M. F., & Eckstein, M. P. (2012). Looking just below the eyes is optimal across face recognition tasks. *Proceedings of the National Academy of Sciences*, *109*(48), E3314–E3323, doi:10.1073/pnas.1214269109.

Peterson, M. F., & Eckstein, M. P. (2013). Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. *Psychological Science*, *24*(7), 1216–1225, doi:10.1177/0956797612471684.

Peterson, M. F., & Eckstein, M. P. (2014). Learning optimal eye movements to unusual faces. *Vision*

Research, *99*, 57–68, doi:10.1016/j.visres.2013.11.005.

Phillips, P. J., & O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, *32*(1), 74–85, doi:10.1016/j.imavis.2013.12.002.

Pizzamiglio, M. R., Luca, M. D., Vita, A. D., Palermo, L., Tanzilli, A., Dacquino, C., & Piccardi, L. (2015). Congenital prosopagnosia in a child: Neuropsychological assessment, eye movement recordings and training. *Neuropsychological Rehabilitation*, e-pub ahead of print, doi:10.1080/09602011.2015.1084335.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025, doi:10.1038/14819.

Risko, E. F., Laidlaw, K. E. W., Freeth, M., Foulsham, T., & Kingstone, A. (2012). Social attention with real versus reel stimuli: Toward an empirical approach to concerns about ecological validity. *Frontiers in Human Neuroscience*, *6*, 143, doi:10.3389/fnhum.2012.00143.

Schmalzl, L., Palermo, R., Green, M., Brunsdon, R., & Coltheart, M. (2008). Training of familiar face recognition and visual scan paths for faces in a child with congenital prosopagnosia. *Cognitive Neuropsychology*, *25*(5), 704–729, doi:10.1080/02643290802299350.

Schwarzer, G., Huber, S., Grüter, M., Grüter, T., Groß, C., Hipfel, M., & Kennerknecht, I. (2006). Gaze behaviour in hereditary prosopagnosia. *Psychological Research*, *71*(5), 583–590, doi:10.1007/s00426-006-0068-0.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–426, doi:10.1109/TPAMI.2007.56.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216, doi:10.1146/annurev.neuro.24.1.1193.

Speer, L. L., Cook, A. E., McMahon, W. M., & Clark, E. (2007). Face processing in children with autism: Effects of stimulus contents and type. *Autism*, *11*(3), 265–277, doi:10.1177/1362361307076925.

Stampe, D. M. (1993). Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. *Behavior Research Methods, Instruments, & Computers*, *25*(2), 137–142, doi:10.3758/BF03204486.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1701–1708). Columbus, OH: IEEE.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*(3), 391–412, doi:10.1088/0954-898X_14_3_302.

Vabalas, A., & Freeth, M. (2015). Brief report: Patterns of eye movements in face to face conversation are associated with autistic traits: evidence from a student sample. *Journal of Autism and Developmental Disorders*, *46*(1), 1–10. doi:10.1007/s10803-015-2546-y.

Wass, S. V., Smith, T. J., & Johnson, M. H. (2012). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, *45*(1), 229–250, doi:10.3758/s13428-012-0245-6.

Weigelt, S., Koldewyn, K., & Kanwisher, N. (2012). Face identity recognition in autism spectrum disorders: A review of behavioral studies. *Neuroscience & Biobehavioral Reviews*, *36*(3), 1060–1084, doi:10.1016/j.neubiorev.2011.12.008.

Zhang, J., Liu, J., & Xu, Y. (2015). Neural decoding reveals impaired face configural processing in the right fusiform face area of individuals with developmental prosopagnosia. *Journal of Neuroscience*, *35*(4), 1539–1548, doi:10.1523/JNEUROSCI.2646-14.2015.

Appendix

Fixation results were validated by comparing the position and timing data of the fixations detected by the ASL algorithm (see Methods [Stage II], Gaze location and fixation event detection) to fixations detected using modified versions of two techniques that have been shown to be robust to high levels of position noise and frequent periods of lost or unreliable data (Wass et al., 2012).

First, we reanalyzed all data following a modified version of the fixation-detection algorithm for unreliable eye tracking data described in Wass et al. (2012). The procedure was as follows: (a) Samples labeled as missing data, or with out-of-range coordinates (x more than 32° and/or y more than 24° from the scene camera center), were labeled as invalid; (b) Valid data was smoothed with a bilateral filter to reduce small within-fixation jitter while preserving large saccadic displacements (Durand & Dorsey, 2002; Frank, Vul, & Johnson, 2009; Stampe, 1993); (c) The mean absolute

deviation (MAD) in gaze position was calculated within a six sample (100 ms) sliding window; (d) Windows with a MAD less than 50°/s were classified as potential fixations, with consecutive qualifying windows concatenated into longer potential fixations; (e) Potential fixations separated by fewer than nine consecutive invalid samples (150 ms) were concatenated if they were displaced by less than 1°, with invalid samples assigned the mean position of the preceding potential fixation; and (f) Potential fixations were labeled as valid fixations if they were immediately preceded and followed by a likely saccade event (MAD of three preceding/succeeding samples greater than 100°/s) and displaced from the mean of the preceding and succeeding potential fixations by at least 1°.

Second, the authors hand-coded fixation events for a random sample of data through visual inspection of gaze position versus time plots (Holmqvist et al., 2011; Wass et al., 2012) and fixation-overlaid scene camera video (Figure 6). Fixations were defined as epochs of relatively stable gaze position preceded and succeeded by abrupt shifts in gaze.

The two validation procedures were in good agreement with the automatic ASL algorithm (see Figure 10 for an example of fixations detected by each procedure). Both of the validation procedures detected fewer, and longer, fixations than the ASL, largely due to the merging of shorter fixations interrupted by brief false saccade events into longer fixations (Figure 9). Fixation position did not differ across procedures, and, critically, the positions of on-face fixations were unaffected, with strong across subject correlations between the ASL procedure's γ values and both the filtering procedure ($r = 0.95$, $p < 0.01$; Figure 10) and hand-coding ($r = 0.89$, $p < 0.01$).