Individual differences in visual salience vary along semantic dimensions

Benjamin de Haas³,²,¹, Alexios L. Iakovidis⁶, D. Samuel Schwarzkopf⁵,⁶, and Karl R. Gegenfurtner³

¹Department of Psychology, Justus Liebig Universität, 35394 Giessen, Germany; ²Experimental Psychology, University College London, WC1H 0AP London, United Kingdom; and ³School of Optometry & Vision Science, University of Auckland, 1142 Auckland, New Zealand

What determines where we look? Theories of attentional guidance hold that image features and task demands govern fixation behavior, while differences between observers are interpreted as a “noise-ceiling” that strictly limits predictability of fixations. However, recent twin studies suggest a genetic basis of gaze-trace similarity for a given stimulus. This leads to the question of how individuals differ in their gaze behavior and what may explain these differences. Here, we investigated the fixations of >100 human adults freely viewing a large set of complex scenes containing thousands of semantically annotated objects. We found systematic individual differences in fixation frequencies along six semantic stimulus dimensions. These differences were large (>twofold) and highly stable across images and time. Surprisingly, they also held for first fixations directed toward each image, commonly interpreted as “bottom-up” visual salience. Their perceptual relevance was documented by a correlation between individual face salience and face recognition skills. The set of reliable individual salience dimensions and their covariance pattern replicated across samples from three different countries, suggesting they reflect fundamental biological mechanisms of attention. Our findings show stable individual differences in salience along a set of fundamental semantic dimensions and that these differences have meaningful perceptual implications. Visual salience reflects features of the observer as well as the image.

Significance

To see clearly, humans have to constantly move their eyes and bring objects of interest to the center of gaze. Vision scientists have tried to understand this selection process, modelling the “typical observer.” Here, we tested individual differences in the tendency to fixate different types of objects embedded in natural scenes. Fixation tendencies for faces, text, food, touched, and moving objects varied up to two- or threefold. These differences were present from the first eye movement after image onset and highly consistent across images and time. They predicted related perceptual skills and replicated across observer samples. This suggests individual gaze behavior is organized along semantic dimensions of biological significance, shaping the subjective way in which we see the world.

Author contributions: B.d.H. designed research; B.d.H. and A.L.I. performed research; B.d.H. contributed new reagents/analytic tools; B.d.H. analyzed data; D.S.S. provided equipment and advice on study design; K.R.G. provided equipment and advice on study design and analysis; and B.d.H., A.L.I., D.S.S., and K.R.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: The data reported in this paper have been deposited the Open Science Framework at https://osf.io/nv/7t/.

To whom correspondence may be addressed. Email: benjamindehaas@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1820553116/-/DCSupplemental.

Published online May 28, 2019.

www.pnas.org/cgi/doi/10.1073/pnas.1820553116

PNAS  Jun 11, 2019  vol. 116  no. 24  11687–11692
salience (37–39), especially for short saccadic latencies (40, 41). They may therefore reveal individual differences with a deep biological root. We tested the reliability of such differences across random subsets of images and across retests after several weeks. We also tested whether and to which degree individual salience models can improve the prediction of fixation behavior along these dimensions beyond the noise ceiling of generic models. To test the generalizability of salience differences, we replicated their set and covariance pattern across independent samples from three different countries. Finally, we explored whether individual salience differences are related to personality and perception, focusing on the example of face salience and face recognition skills for the latter.

Results

Reliable Salience Differences Along Semantic Dimensions. We tracked the gaze of healthy human adults freely viewing a broad range of images depicting complex everyday scenes (8). A first sample was tested at the University College London, United Kingdom (Lon; n = 51), and a replication sample at the University of Giessen, Germany (Gi_1; n = 51). This replication sample was also invited for a retest after 2 wk (Gi_2; n = 48). Additionally we reanalyzed a public dataset from Singapore [Xu et al. (8); n = 15].

First, we probed the individual tendency to fixate objects with a given semantic attribute, measuring duration-weighted fixations across a free-viewing period of 3 s. We considered a total of 12 semantic properties, which have previously been shown to carry more weight for predicting gaze behavior (on an aggregate group level) than geometric or pixel-level attributes (8). To test the consistency of individual salience differences across independent sets of images, we probed their reliability across 1,000 random (half-) splits of 700 images. Each random split was identical across all observers, and for each split individual differences seen for one-half of the images were correlated with those seen for the other half. This way we tested the consistency of relative differences in fixation behavior across different subsets of images, without confounding them with image content (e.g., the absolute frequency of faces in a given subset of images). We found consistent individual salience differences ($r > 0.6$) for 6 of the 12 semantic attributes: Neutral Faces, Emotional Faces, Text, objects being Touched, objects with a characteristic Taste (i.e., food and beverages), and objects with implied Motion (Fig. 1, gray scatter plots).

Observers showed up to twofold differences in the cumulative fixation time attracted by a given semantic attribute and the median consistency of individual differences across image splits for these six dimensions, ranged from $r = 0.64, P < 0.001$ (Motion) to $r = 0.94, P < 0.001$ (Faces; $P$ values Bonferroni-corrected for 12 consistency correlations) (SI Appendix, Table S1, left hand side).

Previous studies have argued that extended viewing behavior is governed by cognitive factors, while first fixations toward a free-viewed image are governed by “bottom-up” salience (37–39), especially for short saccadic latencies (40, 41). Others have found that perceived meaning (9, 42) and semantic stimulus properties (8, 43) are important predictors of gaze behavior from the first fixation. We found consistent individual differences also in the proportion of first fixations directed toward each attribute. The range of individual differences in the proportion of first fixations directed to each of the six attributes was up to threefold, and thus even larger than that for cumulative fixation time. Importantly, these intersubject differences were consistent for all dimensions found for cumulative fixation time except Motion ($r = 0.34$, not significant), ranging from $r = 0.57, P < 0.001$ (Taste) to $r = 0.88, P < 0.001$ (Faces; $P$ values Bonferroni-corrected for 12 consistency correlations) (green scatter plots in Fig. 1 and SI Appendix, Table S1, right hand side).

![Fig. 1. Consistent individual differences in fixation behavior along six semantic dimensions. For each semantic attribute, the gray scatter plot shows individual proportions of cumulative fixation time for the odd versus even numbered images in the Lon dataset. The green scatter plot shows the corresponding individual proportions of first fixations after image onset. Black inset numbers give the corresponding Pearson correlation coefficient. For each dimension, two example images are given and overlaid with the fixations from one observer strongly attracted by the corresponding attribute (orange frames) and one observer weakly attracted by it (blue frames). The overlays show the first fixation after image onset as a green circle; any subsequent fixations are shown in purple. The two data points corresponding to the example observers are highlighted in the scatter plot, corresponding to the color of the respective image frames. All example stimuli from the COE dataset, published under the Massachusetts Institute of Technology license (8). Black bars were added to render faces unrecognizable for display purposes only (participants saw unmodified stimuli).](www.pnas.org/cgi/doi/10.1073/pnas.1820553116)
These salience differences proved robust for different splits of images (Fig. 2A) and replicated across datasets from three different countries (SI Appendix, Fig. S1 and Table S1). For the confirmatory Gi_1 dataset, we tested the same number of observers as in the original Lon set. A power analysis confirmed that this sample size yields >95% power to detect consistencies with a population effect size of r > 0.5. For cumulative fixation time (gray histograms in Fig. 2A), the six dimensions identified in the Lon sample, closely replicated in the Gi_1, Gi_2 samples, as well as in a reanalysis of the public Xu et al. (8) dataset, with consistency correlations ranging from 0.65 (Motion in the Gi_1 set) to 0.95 [Faces in the Xu et al. (8) dataset] (SI Appendix, Table S1, left column, and Fig. S1). A similar pattern of consistency held for first fixations (green histograms in Fig. 2A), although the consistency correlation for Emotion missed statistical significance in the small Xu et al. dataset (8) (SI Appendix, Table S1, right column, and Fig. S1).

The individual salience differences we found were consistent across subsets of diverse, complex images. To test whether they reflected stable observer traits, we additionally tested their retest reliability for the full image set across a period of 6–43 d (average 16 d; Gi_1 and Gi_2 datasets). Salience differences along all six semantic dimensions were highly consistent over time (Fig. 2B). This was true for both cumulative fixation time [retest reliabilities ranging from r = 0.68, P < 0.001 (Motion) to r = 0.85, P < 0.001 (Faces)] (gray bars in Fig. 2B and left column of SI Appendix, Table S1) and first fixations [retest reliabilities ranging from r = 0.62, P < 0.001 (Taste) to r = 0.89, P < 0.001 (Text)] (green bars in Fig. 2B and right column of SI Appendix, Table S1).

Additional control analyses confirmed that individual salience differences persisted independent of related visual field biases (SI Appendix, Supplementary Results and Discussion and Fig. S6).

**Individual Differences in Visual Exploration.** Previous studies reported a relationship between trait curiosity and a tendency for visual exploration, as indexed by anticipatory saccades (31) or the dispersion of fixations across scene images (32). The latter was hypothesized to be a “content neutral” measure, independent of the type of salience differences we investigated here. Our data allows us to explicitly test this hypothesis. We ran an additional analysis, testing whether the number of objects fixated is truly independent of which objects an individual fixates preferentially.

First, we tested whether individual differences in visual exploration were reliable. The number of objects fixated significantly varied across observers, with a maximum/minimum ratio of 1.4 [Xu et al. (8)] to 1.9 (Lon) within a sample. Moreover, these individual differences were highly consistent across odd and even images in all four datasets (all r > 0.98, P < 10^{-11}) and showed good test-retest reliability (r = 0.80, P < 10^{-11} between Gi_1 and Gi_2).

Crucially, however, we observed no significant relationship between the individual tendency for visual exploration and the proportion of first fixations landing on any of the six individual salience dimensions we identified (SI Appendix, Table S2, right hand side). For the proportions of cumulative dwell time, there was a moderate negative correlation between visual exploration and the tendency to fixate emotional expressions, which was statistically significant in the three bigger datasets (Lon, Gi_1, and Gi_2; all tests Holm-Bonferroni-corrected for six dimensions of interest) (SI Appendix, Table S2, left hand side). This negative correlation was not a mere artifact of longer dwelling on emotional expressions limiting the time to explore a greater number of objects. It still held when the individual proportion of dwell time on emotional expressions was correlated with the number of objects explored in images not containing emotional expressions (r = -0.52, P < 0.001 for all three datasets).

**Individual Predictions Improve on the Generic Noise Ceiling.** We took a first step toward evaluating how individual fixation predictions may improve on generic, group-based salience models. If individual differences were noise, then the mean of many observers should be the best possible predictor of individual gaze behavior. That is, the theoretical optimum of a generic model is the exact prediction of group fixation behavior for a set of test images, including fixation ratios along the six semantic dimensions identified above. Could individual predictions improve on this generic optimum?

We pooled fixation data across the 117 observers in the Lon, Gi, and Xu et al. (8) samples and randomly split the data into training and test sets of 350 images each (random splitting was repeated 1,000 times, with each set serving as test and training data once, totaling 2,000 folds). For each fold, we further separated a target individual from the remaining group, iterating through all individuals in a leave-one-observer-out fashion. For each fold and target observer, the empirical fixation ratios of the remaining group served as the (theoretical) ideal prediction of a generic salience model for the test images. We compared the prediction error for this ideal generic model to that of an individualized prediction.

The individual model was based on the assumption that fixation deviations from the group generalize from training to test data. It thus adjusted the prediction of the ideal generic model, based on the target individual’s deviation from the group in the training data. Specifically, the target individual’s fixation ratios for the training set were converted into units of SDs from the group mean. These z-scores were then used to predict individual fixation ratios for the test images, based on the mean and SD of the remaining group for the test set. Note that the individual model should perform worse than the ideal generic one if deviations from the group are random (see SI Appendix, Supplementary Methods for details).

Averaged across folds and cumulated across dimensions, the individual model reduced the prediction error for cumulative dwell time ratios for 89% of observers (t_{116} = 11.39, P < 0.001) and for first fixation ratios for 77% of observers (t_{116} = 8.32, P < 0.001).
Across the group, this corresponded to a reduction of the mean cumulative prediction error from 10.09% (± 0.42% SEM) to 5.33% (± 0.10% SEM) for cumulative dwell time ratios and from 14.31% (± 0.56%, SEM) to 9.61% (± 0.10% SEM) for first fixation ratios. Individual predictions explained 74% of the error variance of ideal generic predictions for cumulative dwell time ratios and 58% of this error variance “beyond the noise ceiling” for first fixation ratios (again, averaged across folds and cumulated across dimensions) (see Fig. 3 and SI Appendix, Fig. S2 for individual dimensions).

Covariance Structure of Individual Differences in Semantic Salience.

Having established reliable individual differences in fixation behavior along semantic dimensions, we further explored the space of these differences by quantifying the covariance between them. For this analysis we collapsed neutral and emotional faces into a single Faces label, because they are semantically related and corresponding differences were strongly correlated with each other (r = 0.74, P < 0.001; r = 0.81, P < 0.001 for cumulative fixation times and first fixations, respectively). Note that we decided to keep these two dimensions separated for the analyses above because the residuals of fixation times for emotional faces still varied consistently when controlling for neutral faces (r = 0.73, P < 0.001), indicating an independent component (however, the same was not true for first fixations, r = 0.24, not significant).

The resulting five dimensions showed a pattern of pairwise correlations that allowed the identification of two clusters (Fig. 4B). This was illustrated by the projection of the pairwise (dis)similarities onto a 2D space, using metric dimensional scaling (SI Appendix, Fig. S3). Faces and Motion were positively correlated with each other, but negatively with the remaining three attributes: Text, Touched, and Taste. Interestingly, Faces, the most prominent dimension of individual fixation behavior, was strongly anticorrelated with Text and Touched, the second and third most prominent dimensions [Text: r = −0.62, P < 0.001 and r = −0.47, P < 0.001 for cumulative fixation times and first fixations, respectively (Fig. 4A, Upper); Touched: r = −0.58, P < 0.001 and r = −0.80, P < 0.001 (Fig. 4A, Lower)]. These findings closely replicated across all four datasets (SI Appendix, Fig. S3). Pairwise correlations between (z-converted) correlation matrices from different samples ranged from 0.68 to 0.95 for cumulative fixation times and from 0.91 to 0.98 for first fixations.

Perceptual Correlates of Salience Differences. If salience differences are indeed deeply rooted in the visual cortices of our observers, then this might have an effect on their perception of the world. We aimed to test this hypothesis by focusing on the most prominent dimension of salience differences: Faces, as indexed by the individual proportion of first fixations landing on faces (which is thought to be an indicator of bottom-up salience). Forty-six observers from the Gi sample took the Cambridge Face Memory Test (CFMT) and we tested the correlation between individual face salience and face recognition skills. CFMT scores and the individual proportion of first fixations landing on faces correlated with r = 0.41, P < 0.005 (SI Appendix, Fig. S4, Right). Interestingly, this correlation did not hold for the individual proportion of total cumulative fixation time landing on faces, which likely represents more voluntary differences in viewing behavior (r = 0.21, not significant) (SI Appendix, Fig. S4, Left).

Additionally, we explored potential relationships with personality variables, but found no significant correlations between gaze behavior and standard questionnaire measures (SI Appendix, Supplementary Results and Discussion and Fig. S5).

Discussion

Individual differences in gaze traces have been documented since the earliest days of eye-tracking (19, 20). However, the nature of these differences was unclear, and therefore traditional salience models have either ignored them or used them as an upper limit for predictability (“noise-ceiling”). Our findings show that what was thought to be noise can actually be explained by a canonical set of semantic salience differences. These salience differences were highly consistent across hundreds of complex scenes, proved reliable in a retest after several weeks, and persisted independently of correlated visual field biases. This shows that visual salience is not just a factor of the image; individual salience differences are a stable trait of the observer, not only the set of these differences, but also their covariance structure replicated across independent samples from three different countries. This may partly be driven by environmental and image statistics (for example, faces are more likely to move than food). But it may also point to a neurobiological basis of these differences. This possibility is underscored by earlier studies showing that the visual salience of social stimuli is reduced in individuals with autism spectrum disorder (33, 44, 45). Most importantly, recent twin studies in infants and children show that individual differences in gaze traces are heritable (33, 34). The gaze trace dissimilarities investigated in these twin studies might be a manifestation of
the salience differences we found here, which would imply a strong genetic component for individual salience differences. Individual differences in gaze behavior have recently gained attention in fields ranging from computer science to behavior genetics (28, 32–35). Previous findings converged to show such differences are systematic, but provided no clear picture of their nature. Our results show that individual salience varies along a set of semantic dimensions, which are among the best predictors of gaze behavior (8). Nevertheless, we cannot exclude the possibility of further dimensions of individual salience. For example, for some of the dimensions for which we found little or unreliable individual differences (watchable, touch, operable, gazed, sound, smell), such differences may have been harder to detect because they carry less salience overall (8). Lower overall numbers of fixations come with a higher risk of granularity problems. However, given that our dataset contains an average of over 5,000 fixations per observer for a wide range of images, it seems unlikely we missed any individual salience dimension of broad importance due to this problem. Future studies may probe the (unlabeled and potentially abstract) features of convolutional neural networks that carry weight for individual gaze predictions and may inform the search for further individual salience dimensions (17, 28).

Recent findings in macaque suggest that fixation tendencies toward faces and hands are linked to the development and processing of corresponding domain-specific patches in the temporal cortex (46, 47). It is worth noting that most of the reliable dimensions of individual salience differences we found correspond to domain-specific patches of the ventral path [as is true for Faces (48–50), Text (51, 52), Motion (53), Touched (46, 54, 55), and maybe Taste (56)]. This opens the exciting possibility that these differences may be linked to neural tuning in the ventral stream.

Our findings raise important questions about the individual nature of visual perception. Two observers presented with the same image can end up with a different perception (5, 6) and interpretation (57) of this image when executing systematically different eye movements. Vision scientists may be chasing a phantom when “averaging out” individual differences to study the “typical observer” (58–60), and vice versa perception may be crucial to understanding individual differences in cognitive abilities (61, 62), personality (63, 64), social behavior (33, 44), clinical traits (65–67), and development (45).

We only took a first step toward investigating potential observer characteristics predicting individual salience here. Individual face salience was moderately correlated with face recognition skills. Interestingly, this was only true when considering the proportion of first fixations attracted by faces. Immediate saccades toward faces can have very short latencies and be under limited voluntary control (68, 69), likely reflecting bottom-up processing. This raises questions about the ontological interplay between face salience and recognition. Small initial differences may grow through mutual reinforcement of face fixations and superior perceptual processing, which would match the explanation of face processing difficulties in autism given by learning style theories (70).

We also investigated potential correlations with major personality dimensions, but found no evidence of such a relationship. Individual salience dimensions also appeared largely independent of the general tendency for visual exploration. However, one exception was the negative correlation between visual salience and attention in fields ranging from computer science to behavioral genetics (28, 32–35). Previous findings converged to show such differences are systematic, but provided no clear picture of their nature. Our results show that individual salience varies along a set of semantic dimensions, which are among the best predictors of gaze behavior (8). Nevertheless, we cannot exclude the possibility of further dimensions of individual salience. For example, for some of the dimensions for which we found little or unreliable individual differences (watchable, touch, operable, gazed, sound, smell), such differences may have been harder to detect because they carry less salience overall (8). Lower overall numbers of fixations come with a higher risk of granularity problems. However, given that our dataset contains an average of over 5,000 fixations per observer for a wide range of images, it seems unlikely we missed any individual salience dimension of broad importance due to this problem. Future studies may probe the (unlabeled and potentially abstract) features of convolutional neural networks that carry weight for individual gaze predictions and may inform the search for further individual salience dimensions (17, 28).