Current Biology 22, 191–196, February 7, 2012 ©2012 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2011.11.061

Report

Measuring Internal Representations from Behavioral and Brain Data

Marie L. Smith,[1,*] Frédéric Gosselin,[2] and Philippe G. Schyns[3]
[1]Department of Psychological Sciences, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK
[2]Département de Psychologie, Université de Montréal, CP 6128, Succursale Centre-Ville, Montréal, QC H3C 3J7, Canada
[3]Centre for Cognitive Neuroimaging, Institute of Neuroscience and Psychology, University of Glasgow, 58 Hillhead Street, Glasgow G12 8QB, UK

Summary

The study of internal knowledge representations is a cornerstone of the research agenda in the interdisciplinary study of cognition. An influential proposal assumes that the brain uses its internal knowledge of the external world to constrain, in a top-down manner, high-dimensional sensory data into a lower-dimensional representation that enables perceptual decisions and other higher-level cognitive functions [1–9]. This proposal relies on a precise formulation of the observer-specific internal knowledge (i.e., the internal representations, or models) that guides reduction of the high-dimensional retinal input onto a low-dimensional code. Here, we directly revealed the content of subjective internal representations by instructing five observers to detect a face in the presence of only white noise, to force a pure top-down, knowledge-based task. We used reverse correlation methods to visualize each observer's internal representation that supports detection of an illusory face. Using reverse correlation again, this time applied to observers' electroencephalogram activity, we established where and when in the brain specific internal knowledge conceptually interprets the input white noise as a face. We show that internal representations can be reconstructed experimentally from behavioral and brain data, and that their content drives neural activity first over frontal and then over occipitotemporal cortex.

Results and Discussion

We presented observers with white-noise stimuli and instructed them that on half of the trials a face would be hidden in the noise, when in fact no face was ever presented. To resolve the task, observers must use their own internal knowledge of a face and match it with the incoming white-noise stimulus. A small correlation between the noise and internal knowledge enables the observer to "detect" a face, a process we previously termed "superstitious perception" [10]. Via reverse correlation [11] from behavioral responses, we established the subjective internal representation that each observer used in the face detection task—i.e., we established what information is critical.

This leaves unaddressed the crucial issues of where and when the brain deploys this critical information to conceptually interpret the input white noise. To this aim, on each trial we concurrently recorded the observer's brain activity via electroencephalography (EEG). Using only unstructured noise stimuli, and providing no a priori target face, we can isolate where (in terms of EEG electrodes) and when the observer's brain uses its own internal face information to interpret the high-dimensional input noise as a face.

Behavior: From Face Detection Behavior to an Internal Face Template

Despite the presence of only noise, observers detected a face on 44% (σ = 5.9%) of trials in 1050 ms, with no difference in reaction times between face detection decisions [mean difference: 34.8 ms, σ = 40 ms, t(4) = 2.1, not significant]. Upon completion of the experiment we debriefed observers, and all expressed shock that no face was ever presented.

If an observer's decisions occurred randomly, through stochastic variations in ongoing neural processes (e.g., [12]) or pure guesswork, systematic structures should not emerge from the reverse correlation analysis as white noise averages to gray. However, if perceptual decisions require a systematic matching of internal face knowledge with minimal bottom-up input from the noise, reverse correlation should depict this internal information. The resulting classification images (Figure 1B) did indeed reveal facial structures including the "eyes" for all observers but S5; a nose, mouth, and chin outline for observers S1, S2, and S3; and a hairstyle for S3. Thresholding [13] emphasized the key internally represented features. Note that there is no statistical method ideally suited to perform this thresholding, because all methods assume relatively focal signals resulting in a reduced sensitivity for more distributed features, e.g., face contour and hairstyle. We therefore opted to retain the full classification images for all subsequent analyses.

Internal knowledge representations should demonstrate individual subjectivity. To measure this, we generated a reference template for each observer from the first half of the experimental trials. We then computed the "faceness" of subsequent trials by correlating each noise stimulus with the reference template. We found a significant linear relation between the faceness ratings and the probability of a face detection response (Figure 1C, solid lines; $r^2 > 0.94$, df = 9, $p < 0.0001$ for observers S1, S2, S4, and S5; $r^2 = 0.78$, $p = 0.008$ for observer S3). Critically, this relationship did not hold when using the reference template of another observer to rank the noise stimuli by faceness ($r^2 < 0.66$, not significant for all at $p < 0.05$, Bonferroni corrected)—except for observers S1 and S2 ($r^2 > 0.94$, $p < 0.0001$), where there is a clear overlap in the behavioral information templates.

This result emphasizes the subjective nature of the internal face representations, their consistency across experimental sessions, and their power to predict perceptual detection responses from input noise. However, there remains an infinitesimally small probability that white noise could depict a reasonably well-structured face on a particular trial. This information could be sufficient to bias subsequent detection responses in a bottom-up manner. To control, we extracted

*Correspondence: marie.smith@bbk.ac.uk

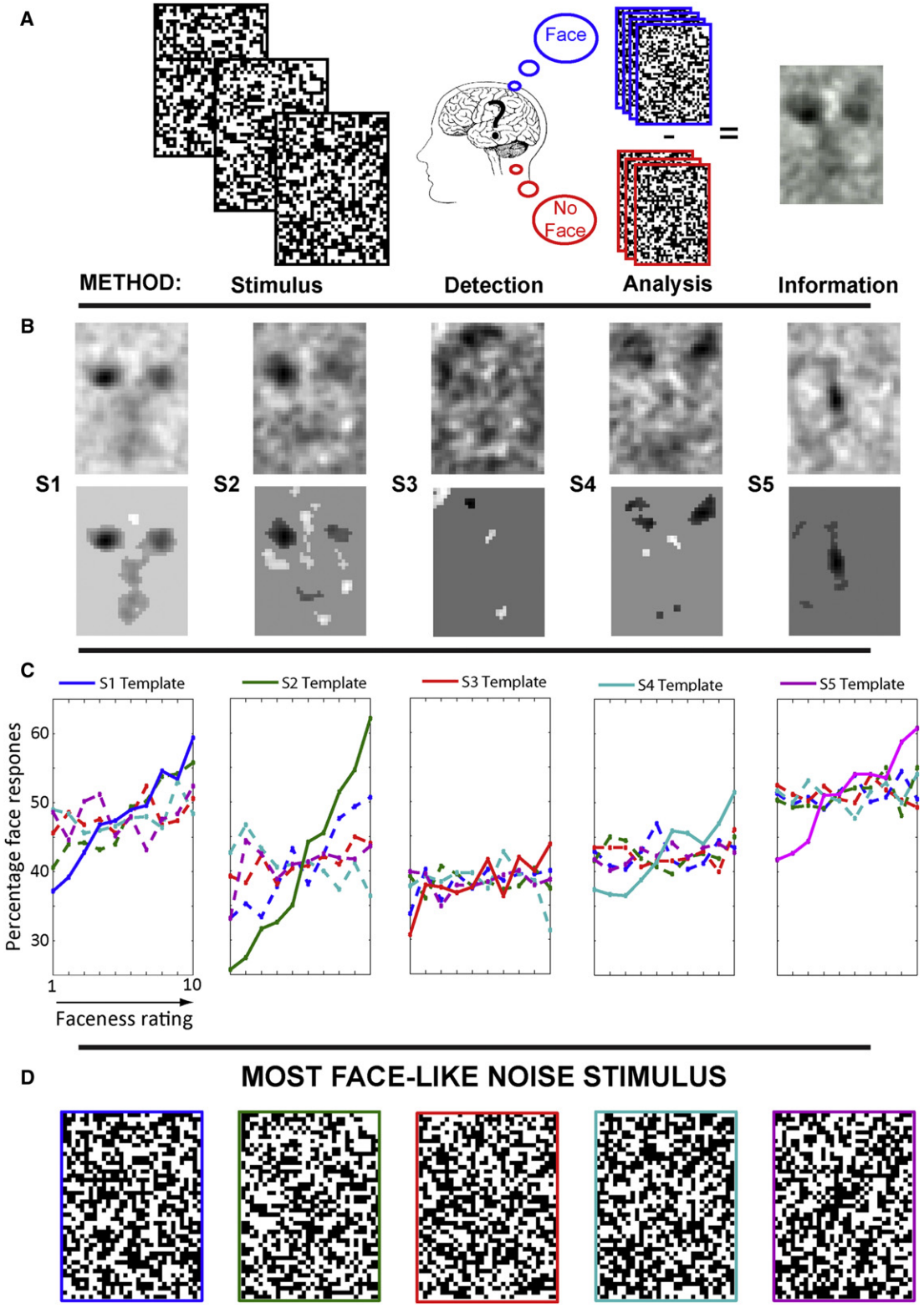

Figure 1. Behavior: From Face Detection Behavior to an Internal Face Template

(A) Illustration of the experimental design and behavioral reverse correlation analysis (illustrated with average of all observers).

(B) Behavioral classification images for each of the observers with (bottom row) and without (top row) threshold.

(C) Results of the split-half analysis indicating the relationship between the "faceness" of the noise stimuli (i.e., the correlation between each unique random noise stimulus and the behavioral template of the observer) and the proportion of "face present" responses for each observer's decisions, using their own and each others' behavioral templates to compute the faceness rankings.

(D) The noise stimulus with the highest correlation to the internal face template for each observer.

for each observer the noise stimulus that had the highest correlation with the observer's classification image (maximum r = 0.1). Visual inspection of these maximally correlated stimuli (Figure 1D) revealed no obvious face structure, emphasizing the top-down nature of the face detection task.

Brain: From the Internal Face Template to Its Processing in the Brain

We used the EEG brain measurements to identify where and when the brain uses its internal knowledge (what information) to conceptually interpret the high-dimensional input noise. On average, observers detected a face on 4,134 artifact-free trials ("face present") versus 5,204 no-detection trials ("face absent"). On "face present" trials (compared with "face absent" trials), within the first 500 ms of processing we found significantly increased neuronal responses ($p < 0.005$) over both frontal and lateral occipitotemporal cortex (Figure 2A). But there were no significant differences in onset or duration across frontal and occipitotemporal regions [$t(4) < 0.8$, not significant; see top panel of Figure S1 available online]. Initially this enhanced response on "face present" trials had a topographic distribution of increased negativity over lateral occipitotemporal cortex, similar to a noise-delayed N170 [14], accompanied by an increased positivity over frontal electrodes. Later "face present" trials had increased activation over centroparietal regions with a topography similar to the category-selective P300 component, but with a varying time course across observers (450 to 650 ms).

These event-related potential results are typical of face detection decisions. However, they cannot inform as to where and when the brain uses internal knowledge for detection. To address this, we explored the EEG response of each observer on "face present" trials. We sought to measure systematic associations between single-trial EEG amplitudes and face information in the noise stimuli (see "EEG Analyses" in Experimental Procedures). We found that each observer processed information from their face template consistently over frontal and lateralized occipitotemporal regions in the 200–500 ms following stimulus onset (see bottom panel of Figure S1). There was a direct association between increasing faceness content of the stimuli and enhanced positivity in the single-trial EEG amplitudes over frontal sensors—i.e., the more face-like noise stimuli drove larger neural responses (Figure 2B, $p < 0.01$ highlighted on curves)—and a significant association between increased negative responses over occipitotemporal sensors and the faceness of the noise. Across all five observers, a consistent pattern emerged, pre-400 ms, with significant association peaks over frontal regions preceding the peaks over occipitotemporal regions as predicted [mean difference = 30 ms, $t(4) = 2.3$, $p < 0.05$, one-tailed t test]. Post-400 ms, and with a more variable timing, increased positivity over centroparietal sensors was significantly associated with the faceness of the noise stimuli (see bottom panel of Figure S1). The most face-like stimuli produced the largest P300-like responses, as expected [15, 16].

Finally, we provide direct evidence that the brain uses its internal face knowledge to interpret the noise stimuli by computing EEG classification images (see "EEG Classification Images" in Experimental Procedures). Figure 2B reveals face-like structure in the EEG classification images of each observer. Crucially, the EEG classification images correlate significantly more with the behavioral template of the observer considered than with the behavioral template of the other observers [same observer versus mean of other observers: $t(4) = 5.38$, $p < 0.006$; same observer versus maximum of other observers: $t(4) = 3.74$, $p = 0.02$; Figure 2C].

Internal Representations Can Be Accessed Experimentally

Here, we used white-noise stimuli and instructed observers to detect a face that was never actually presented. With reverse correlation methods applied to detection behavior, we revealed the internal subjective information that guided each observer's perceptual decisions. With reverse correlation applied to the EEG signal on "face present" trials, we found where and when each observer's brain deploys internal knowledge to interpret the input noise and revealed the content of this internal, observer-specific knowledge over frontal and occipitotemporal regions.

Internal Representations Drive Neuronal Activation First over Frontal and Then over Occipitotemporal Brain Regions

In a subjective task (i.e., with no "correct" or "incorrect" response), we found robust EEG differences between detection and no-detection trials over frontal and occipitotemporal regions, from 200 ms following stimulus onset. Further analyses revealed the predicted dynamics of frontal before occipitotemporal cortex EEG covarations with the observer-specific face-like nature of the noise [17]. Although frontal cortex has previously been proposed to generate the information driving internal predictions [18, 19], here we revealed the content of this information—i.e., the internal knowledge discussed in predictive coding models of perception in ambiguity [19, 20]. It is worth noting that the expectation of seeing a face can enhance neuronal responses with a similar timing over occipitotemporal cortex ([18], e.g., the N170 component [21]). Face expectations may therefore contribute to the early enhanced activation on face present (versus absent) trials, where processing is driven by factors external to the face-like information from the stimulus per se.

Implications for the Future

We have addressed a cornerstone of the research agenda in cognition: measuring the processing location, timing, and content of the internal visual knowledge that guides reduction of high-dimensional retinal sampling to a low-dimensional conceptual code (a significant advance from initial studies [10, 22]). These representations are formally restricted to the projection of internal knowledge onto a two-dimensional image, and their low-contrast details may be hindered by the use of white noise that has equal power across all spatial frequency bands. Future research should expand the technique for both dimensions of projection and spatial resolution of the represented information. This could be coupled with testing of different categorization tasks (e.g., with faces: identity, gender, and facial expression; or with objects and scenes: car and city versus Porsche and New York) to examine how different categorical knowledge leads to different information reductions of the visual input [23–25].

Our results provide the first direct evidence that the visual information matching an observer's internal knowledge modulates neural activity in frontal cortex. This implies not only that regions of frontal cortex provide the internal knowledge representation to guide processing in ventral temporal cortex, but also that these regions can respond to expected information in advance of any similar modulation over sensory regions.

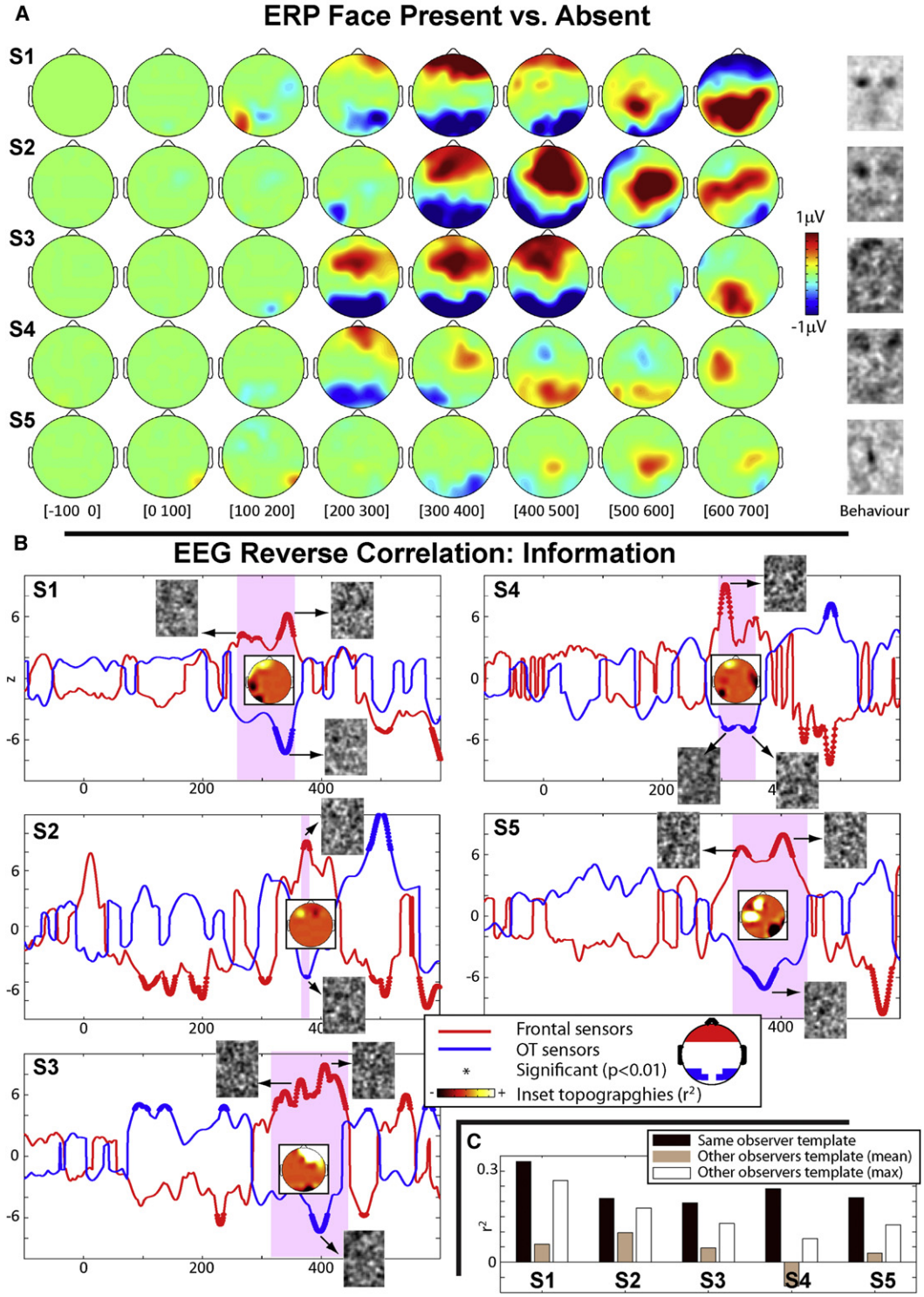

Figure 2. Brain: From the Internal Face Template to Its Processing in the Brain

(A) Significant single-subject differences ($p < 0.005$) in the evoked activity on trials classified as "face present" versus "face absent" (see top panel of Figure S1 for the time course of this activation over frontal and occipitotemporal regions).

(B) Time course of associations (Z-scored, with respect to prestimulus baseline) of single-trial EEG amplitude modulations with face-like information in the noise stimuli on trials classified as "face present" over frontal (red) and occipitotemporal (blue) regions. Significant associations ($p < 0.01$) are highlighted on the curves. Inset topographies depict the topographic distribution of the significant associations in the critical 300–400 ms time interval (see bottom panel of Figure S1 for full set of topographic maps). Inset images depict EEG classification images illustrating the specific visual information driving modulations in neuronal activity.

(C) Correlation of each observer's EEG classification images with their own (same) and the other observers' behavioral information templates (average and maximum single correlation).

Experimental Procedures

Observers, Stimuli, and Task

Stimuli comprised randomly generated black-and-white noise fields with the aspect ratio of a face (Figure 1A; [10]). Each noise field contained a random combination of 32 × 43 black and white pixels expanded up by a factor of 6. Five naive observers (four females, one male; mean age 23.8 years) each saw a different set of 10,500 random noise images. No reference or target face was ever presented. We instructed observers that faces would be present on half of the trials, filling the image space and facing toward them. Observers were instructed to indicate whether they perceived a face or not. We emphasized that the task would be very challenging. Observers gave informed written consent according to the regulations of the Faculty for Information and Mathematical Sciences Ethics Committee at the University of Glasgow.

EEG Recording

We recorded scalp electrical activity (EEG) with sintered Ag/AgCl electrodes mounted in a 62-electrode cap at scalp positions including the standard 10–20 system positions along with intermediate positions and an additional row of low occipital electrodes. Linked mastoids served as initial common reference, and the AFz electrode as ground. Vertical and horizontal electro-oculogram (EOG) was measured by additional electrodes placed around the eyes. Electrode impedance was kept below 10 kΩ throughout. Analysis epochs were generated offline starting 200 ms prior to stimulus onset and continuing for 1,000 ms. Trials containing EEG and EOG artifacts and eye movements were removed using standard artifact detection software. Artifact-free trials were low-pass filtered at 30 Hz, rereferenced to average reference (excluding the EOG channels), and baseline corrected using the mean amplitude in the 200 ms leading up to stimulus presentation.

Behavioral Reverse Correlation Analysis

For each observer, we summed together all of the "face present" noise fields and subtracted from that the sum of all "face absent" noise fields. The resulting classification image represents the information underlying the observer's face detection decisions. To increase the signal-to-noise ratio, we smoothed the classification images with a 2D Gaussian kernel (SD = 1 pixel) and applied the cluster test [13] to isolate individual features.

EEG Analyses

Event-Related Potentials

For each observer, we performed a full analysis of all 58 electrodes over 1,025 time points and used a robust bootstrap methodology [14, 26] to compare the pattern of neuronal activation on trials classed as containing a face versus not. We directly tested H_1 by sampling individual trials (with replacement, on average n = 4,134) independently from the original distributions of the two perceptual response conditions (face versus no face), computing the new mean for each condition, and then storing the differences of the means. Each sample comprised the entire electrode-by-time-point matrix, as these values are not independent. We repeated this process 500 times to generate a distribution of bootstrapped estimates of the mean difference and established a 99.5% confidence interval. Mean differences for which the confidence interval did not include zero were considered significant [14, 26]. The time course of face versus no-face percept differences was estimated as the maximal significant difference over frontal (classed as electrodes FP1, FP2, FPz, AF3, AF4, AF7, AF8, F1, F2, F3, F4, F7, F8, and Fz) and occipitotemporal (electrodes P3, P4, P5, P6, P7, P8, PO3, PO4, PO5, PO6, PO7, and PO8) regions at each time point.

Single-Trial Reverse Correlation

For each observer, we ranked each experimental trial according to its "faceness" by correlating each noise stimulus with the observer's behavioral information template. To establish the electrodes and time intervals where any minimal face-like information in the noise was directly associated with modulations of EEG amplitude, we regressed the single-trial faceness rankings with the corresponding EEG voltages on face percept trials. A random permutation bootstrap [26] established a 99% significance criterion ($p < 0.01$) for the mapping between EEG amplitude modulations and the faceness of the noise stimuli. The time course of this mapping was estimated as the maximal association (Z-scored, with respect to the prestimulus baseline) over frontal (electrodes FP1, FP2, FPz, AF3, AF4, AF7, AF8, F1, F2, F3, F4, F7, F8, and Fz) and occipitotemporal (electrodes P3, P4, P5, P6, P7, P8, PO3, PO4, PO5, PO6, PO7, and PO8) regions at each time point, with significant associations highlighted on the curves.

EEG Classification Images

We generated an EEG classification image for a particular electrode and time point by summing together all noise stimuli associated with high EEG amplitudes and subtracting from that the sum of all noise stimuli associated with low EEG amplitudes, after weighting each image by the voltage that it had elicited on the chosen electrode at the specified time point. Single-trial voltages for each electrode and time point were first Z-scored across trials. EEG classification images were computed for each electrode and time point found to be significantly associated with the processing of face-like information, and the individual classification images were averaged to generate three (two for observer S2) average EEG classification images over frontal and occipitotemporal regions. These average EEG classification images were smoothed as per the behavioral classification images. The averaged frontal and occipitotemporal EEG classification images for each observer were correlated with the behavioral information template of the same observer and the other observers, and the resulting correlations were averaged across regions to generate a measure for each observer of the correlation of their EEG classification images with their own and the other observers' behavioral information templates. Both the maximum single correlation of an observer's EEG templates with any other observer's behavioral template and the average of all correlations of the observer's EEG templates with the other observers' behavioral templates were computed.

Supplemental Information

Supplemental Information includes one figure and can be found with this article online at doi:10.1016/j.cub.2011.11.061.

Acknowledgments

We thank Vaia Lestou for assistance with data collection. This research was supported by Economic and Social Research Council grant R000237901 awarded to P.G.S.

Received: September 6, 2011
Revised: October 28, 2011
Accepted: November 28, 2011
Published online: January 19, 2012

References

1. Schyns, P.G., Gosselin, F., and Smith, M.L. (2009). Information processing algorithms in the brain. Trends Cogn. Sci. (Regul. Ed.) *13*, 20–26.
2. Schyns, P.G., Goldstone, R.L., and Thibaut, J.P. (1998). The development of features in object concepts. Behav. Brain Sci. *21*, 1–17, discussion 17–54.
3. Ullman, S., Vidal-Naquet, M., and Sali, E. (2002). Visual features of intermediate complexity and their use in classification. Nat. Neurosci. *5*, 682–687.
4. DiCarlo, J.J., and Cox, D.D. (2007). Untangling invariant object recognition. Trends Cogn. Sci. (Regul. Ed.) *11*, 333–341.
5. Bruner, J.S., and Postman, L. (1949). On the perception of incongruity; a paradigm. J. Pers. *18*, 206–223.
6. Harnad, S. (1987). Categorical Perception: The Groundwork of Cognition (Cambridge: Cambridge University Press).
7. Kersten, D., and Yuille, A. (2003). Bayesian models of object perception. Curr. Opin. Neurobiol. *13*, 150–158.
8. Bar, M. (2004). Visual objects in context. Nat. Rev. Neurosci. *5*, 617–629.
9. Friston, K.J., and Kiebel, S.J. (2009). Predictive coding under the free-energy principle. Philos. Trans. R. Soc. Lond. B Biol. Sci. *364*, 1211–1221.
10. Gosselin, F., and Schyns, P.G. (2003). Superstitious perceptions reveal properties of internal representations. Psychol. Sci. *14*, 505–509.
11. Ahumada, A.J., and Lovell, J. (1971). Stimulus features in signal detection. J. Acoust. Soc. Am. *49*, 1751–1756.
12. Wild, H.A., and Busey, T.A. (2004). Seeing faces in the noise: stochastic activity in perceptual regions of the brain may influence the perception of ambiguous stimuli. Psychon. Bull. Rev. *11*, 475–481.
13. Chauvin, A., Worsley, K.J., Schyns, P.G., Arguin, M., and Gosselin, F. (2005). Accurate statistical tests for smooth classification images. J. Vis. *5*, 659–667.

14. Rousselet, G.A., Pernet, C.R., Bennett, P.J., and Sekuler, A.B. (2008). Parametric study of EEG sensitivity to phase noise during face processing. BMC Neurosci. *9*, 98.
15. Smith, M.L., Gosselin, F., and Schyns, P.G. (2004). Receptive fields for flexible face categorizations. Psychol. Sci. *15*, 753–761.
16. Smith, M.L., Gosselin, F., and Schyns, P.G. (2006). Perceptual moments of conscious visual experience inferred from oscillatory brain activity. Proc. Natl. Acad. Sci. USA *103*, 5626–5631.
17. Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J., Schmid, A.M., Dale, A.M., Hämäläinen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., and Halgren, E. (2006). Top-down facilitation of visual recognition. Proc. Natl. Acad. Sci. USA *103*, 449–454.
18. Righart, R., Andersson, F., Schwartz, S., Mayer, E., and Vuilleumier, P. (2010). Top-down activation of fusiform cortex without seeing faces in prosopagnosia. Cereb. Cortex *20*, 1878–1890.
19. Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., and Hirsch, J. (2006). Predictive codes for forthcoming perception in the frontal cortex. Science *314*, 1311–1314.
20. Zhang, H., Liu, J., Huber, D.E., Rieth, C.A., Tian, J., and Lee, K. (2008). Detecting faces in pure noise images: a functional MRI study on top-down perception. Neuroreport *19*, 229–233.
21. Bentin, S., Allison, T., Puce, A., Perez, E., and McCarthy, G. (1996). Electrophysiological studies of face perception in humans. J. Cogn. Neurosci. *8*, 551–565.
22. Hansen, B.C., Thompson, B., Hess, R.F., and Ellemberg, D. (2010). Extracting the internal representation of faces from human brain activity: an analogue to reverse correlation. Neuroimage *51*, 373–390.
23. Schyns, P.G. (1998). Diagnostic recognition: task constraints, object information, and their interactions. Cognition *67*, 147–179.
24. Gosselin, F., and Schyns, P.G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. Vision Res. *41*, 2261–2271.
25. Schyns, P.G., Bonnar, L., and Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. Psychol. Sci. *13*, 402–409.
26. Wilcox, R.R. (2005). Introduction to Robust Estimation and Hypothesis Testing (New York: Academic Press).