Visual representations supporting category-specific information about visual objects in the brain

Simon Faghel-Soubeyrand (s.faghel-soubeyrand@umontreal.ca)
Department of Psychology, U. de Montréal
Montréal, Québec, Canada

Arjen Alink (a.alink@uke.de)
Center for experimental Medicine,
Institute of Systems Neuroscience, U. of Hamburg

Eva Bamps (e.bamps@bham.ac.uk)
School of Psychology, U. of Birmingham
Birmingham, United Kingdom

Frédéric Gosselin (frederic.gosselin@umontreal.ca)
Department of Psychology, U. de Montréal
Montréal, Québec, Canada

Ian Charest (i.charest@bham.ac.uk)
School of Psychology, U. of Birmingham
Birmingham, United Kingdom

Abstract

Over recent years, multivariate pattern analysis ("decoding") approaches have become increasingly used to investigate "when" and "where" our brains conduct meaningful processes about their visual environments. Studies using time-resolved decoding of M/EEG patterns have described numerous processes such as object/face familiarity and the emergence of basic-to-abstract category information. Surprisingly, no study has, to our knowledge, revealed "what" (i.e. the actual visual information that) our brain uses while these computations are examined by decoding algorithms. Here, we revealed the time course at which our brain extracts realistic category-specific information about visual objects (i.e. emotion-type & gender information from faces) with time-resolved decoding of high-density EEG patterns, as well as carefully controlled tasks and visual stimulation. Then, we derived temporal generalization matrices and showed that category-specific information is 1) first diffused across brain areas (250 to 350 ms) and 2) encoded under a stable neural pattern that suggests evidence accumulation (350 to 650 ms after face onset). Finally, we bridged time-resolved decoding with psychophysics and revealed the specific visual information (spatial frequency, feature position & orientation information) that support these brain computations. Doing so, we uncovered interconnected dynamics between visual features, and the accumulation and diffusion of category-specific information in the brain.

Keywords: vision; categorical representation; EEG; multivariate pattern analysis; psychophysics

Introduction

Recognizing a visual object requires the timely processing of visual information, from fine grained low-level features, to more abstract category-relevant information. In the recent years, a wealth of object recognition neuroimaging results has been obtained using multivariate pattern analyses ("decoding") to characterize "when" and "where" our brains process information about specific stimuli and tasks. To resolve the temporal dynamics of object recognition, studies have applied linear classifiers to time-resolved Magneto/electro-encephalography (M/EEG) activity patterns (see Carlson et al. 2019 for a review). This decoding approach yielded important understandings in objects/face familiarity (Dobs et al. 2019), object memorability (Mohsenzadeh et al. 2019), the emergence of basic-to-abstract category information (Cichy et al. 2014; Contini et al. 2017) to name only but a few studies.

To form a richer understanding of object categorization in the brain at a mechanistic level, such decoding approaches need to be combined with psychophysical procedures that enable revealing the specific content of the information that is available to the brain during a cognitive task. Along this line, a recent study using MEG and psychophysics revealed the processing of task-relevant information in the brain (Zhan et al. 2019), with increasing importance of behaviourally relevant information along the visual ventral stream. These results culminated from over a decade of work using psychophysical techniques similar to reverse correlation (e.g. see Bubbles, Gosselin and Schyns 2001) with behavioral or brain imaging data to decipher the specific visual information (e.g. "detailed" vs. "coarse" spatial frequency information)

underlying object (Caplette et al., 2016), scene (Willenbockel, Gosselin & Võ, submitted) and face (Smith et al. 2008) recognition.

Combining tightly controlled psychophysical paradigms with decoding algorithms would not only provide a way to investigate categorical representations while characterizing the visual features that support information processing in the brain. If used properly, it would also enable neuroscientists to control for low-level processing interpretations behind the disclosed computations. Indeed, it can be a complex endeavour to untangle computations from low-level areas such as V1, which are sensitive to contrast and edge differences between face images from different gender, from decoded whole-brain information. Ultimately, even output from the retina could dissociate between a human face and a giraffe's; the goal of a neuroscientist studying high-level cognition, thus, is to describe how or when *category-specific* (e.g. human vs. animal) information is implemented in the brain.

Here, in an effort to bridge the gap between decoding studies and reverse correlation approaches, we combined diagnostic feature mapping —a carefully controlled psychophysical paradigm that reveals the features supporting recognition (Alink and Charest 2018)— with multivariate pattern analyses applied to concurrently measured EEG data. Similar to classical decoding approaches, we first reveal the time course at which our brain extracts realistic category-specific information about visual objects (i.e. emotion-type & gender information from faces). Then, we go a step further by revealing what visual information (spatial frequency, feature position & orientation information) supports category-specific computations in the brain.

Method

Experimental procedure & stimuli

Participants (N=5) were asked to categorize the emotion (fear vs. joy) or gender (male vs. female) of faces which physical content were partially revealed on every trial (2560 trials per task, per participant). More specifically, we presented 16 expressive male/female faces that varied in feature position content (e.g. more left-eye information present, no mouth information present), spatial-frequency content (coarseness of the presented information) and orientation content (e.g. more horizontal content) in a pseudo-random fashion on every trial (Figure 1; see Alink & Charest, 2018). This ensured that computations from low-level brain areas sensitive to these information variations (e.g. striate cortex), could not help our decoding algorithms, and, conversely, that higher-level information would be distilled-out. Both time-resolved brain information about emotion-type and gender were obtained from brains receiving this pseudo-random low-level visual stimulation, and the original 16 face images used for this stimulation were identical between tasks. Thus, only task instructions ("discriminate emotion", "discriminate gender") differed.

A typical trial consisted of the following events: a fixation dot (with a jittered duration centered around .5s), and a randomly sampled face (1s) followed by the participant's response. Face images were presented in a random order across trials. The two task-conditions were completed on two separate EEG recordings and lasted ~3 hours each, including EEG headset installation.

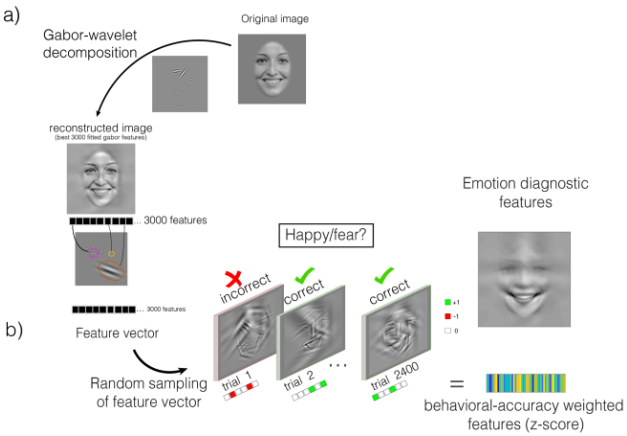

Figure 1. Stimuli creation and experimental paradigm. (a) A face image is fitted with Gabor patches of different orientations (10 to 180 deg), spatial-frequencies (from 2 to 76 cpi), [x,y] image coordinates and sizes. Only the 3k Gabors with the best fits (i.e. the features) are chosen to reconstruct an efficient version of the face image. (b) On a given trial, the feature vector is randomly sampled to create a stimulus.

EEG recording & time resolved decoding

Electroencephalographic (EEG) data were recorded with a BioSemi 128 electrodes headset at 1024 hz. EEG traces were down sampled to 256 Hz and band-passed filtered (.01-30hz). For both task-conditions, the raw EEG patterns were fed to a linear-discriminant classifier which task was to classify the category of the presented stimulus (i.e. either gender in the face gender discrimination task, emotion-type in the emotion face discrimination task). We trained and tested (5-fold cross-validation, 5 repetitions) a different model with EEG patterns from every time point from -100 ms to 1000 ms in 4 ms steps, relative to face-onset. To assess significance, we trained and tested a classifier with identical parameters to categorize shuffled labels, repeated this step 10000 times, and compared our observed classification accuracy to this null-distribution in a time-resolved manner.

These decoding results are presented in **Figure 2a**. Specifically, we display the time course of this category-specific information for emotion-type (while participants completed the face-emotion discrimination task). It shows that emotion-specific information extracted by the brain from faces is present relatively late, starting from 272 ms ($p<.01$

points are underlined in gray; p<.001 in black), and culminates in a large temporal plateau around 550 ms. Furthermore, we observed a second informative window around 900ms after face onset. Traditional univariate event-related potentials (electrodes P6 & P7 average) are overlaid on top of the classification time-course. A brief comparison suggests that the decoded information —which discarded most of low-level information processes due to our controlled visual stimulation— emerges right after face (N170) and attention-related processes (P200).

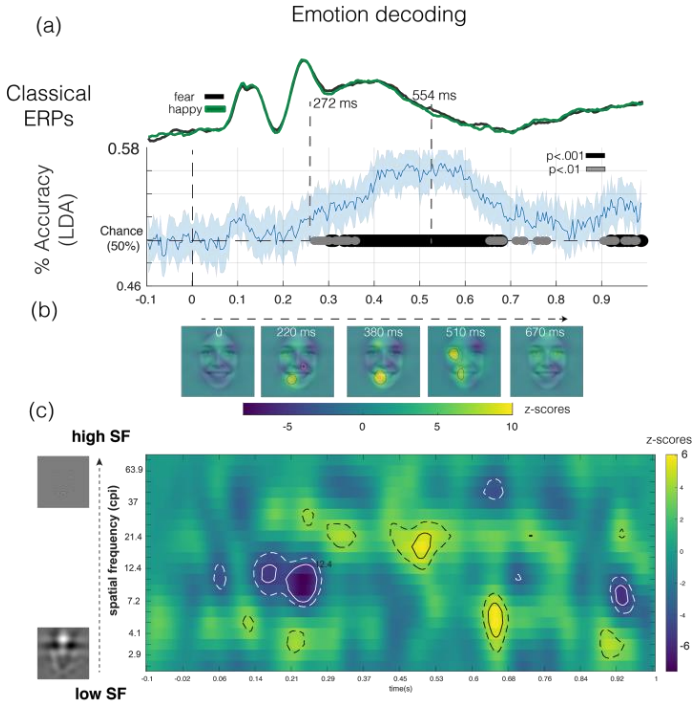

Figure 2. (a) face category-specific information extracted by the brain is unraveled in time using a linear-classifier. **(b, c)** The specific (feature position, spatial-frequency) visual representations supporting the decoded information is obtained by correlating the classifier's confidence-values with the presented features in a time resolved manner.

Temporal generalization

We then asked whether the temporal unfolding of this category-specific information emerged from a stable or varying neural code, i.e. if the EEG pattern dissociating two classes generalized at different points in time (King and Dehaene 2014). We computed a temporal generalization matrix where our lda model was trained on EEG patterns from a specific time *i* and tested on EEG patterns from a different time, *j*, within the -100 to 1000 ms time window. Accurate decoding at a specific coordinate in this matrix indicates that the model was able to generalize it's training at time *i* to time *j*, and therefore that both time points shared a similar (stable) EEG pattern.

The resulting matrix, shown in **figure 3**, indicates that our time-resolved decoding results emerged from at least two distinct brain processes. First, emotion category-specific information travels across brain areas from ~270 to 350 ms after face onset. This is then followed by a stable neural pattern that suggests evidence accumulation from ~350 to 650 ms. The following diagonal pattern suggests that emotion-category specific is finally diffused across brain areas, presumably before the perceptual decision.

Visual features behind category-specific information

But what, exactly, is contained in this category-specific information? To reveal the specific visual features underlying time resolved decoded information in the brain, we first extracted the trial-by-trial distance to the linear discriminant classifier decision criterion *c*, for every time point, participant and tasks. This produced a vector of 2560 distance-values (henceforth referred as *d-vals*) per participant, task and time point. These d-vals could be negative (class 1, e.g. happy category) or positive (class 2, e.g. fear category). To ensure that a positive correlation between visual features and d-vals relate to accurate model classification, we sign-flipped (*-1) the d-vals for trials where class 1 stimuli were presented, such that any positive valued d-val reflects correct identification by the model, and any negative valued d-val reflects misclassification by the model. D-vals distributions were then z-transformed across the 2560 trials for every time point, participant, and task, and smoothed in the time domain with a Gaussian kernel of 4 ms of standard-deviation. Next, we computed the relative presence of the visual features across trials, for either spatial frequency content (from 2.4 cpi to 76.7 cpi in 20 steps), orientation content (10 to 180 degree in 10 degree steps) or feature [x,y] position (3000 image coordinates). As a final final step, we weighted this standardized [FeaturePresence x Trials] matrix with the standardized model confidence-values at every time point, and smoothed this matrix with a 2D Gaussian of 2 standard-deviation. Again, permutation testing was used to assess significance.

This procedure revealed how brain representation of emotion categorical information is supported by spatial [x,y] feature position representations (**Figure 2b**). **Figure 2c** shows how this time-resolved information is supported by spatial-frequency (SF, coarse or fine details) content through time. Positive (yellow-green) values indicates a positive correlation between (accurate) model classification confidence and feature information presence, while a negative (dark-blue) values indicates low-confidence (uncertainty) of the model while the specific visual content was presented. In other words, high positive values indicatethat a specific visual feature helped/supported the brain's representation of category-specific information from a face.

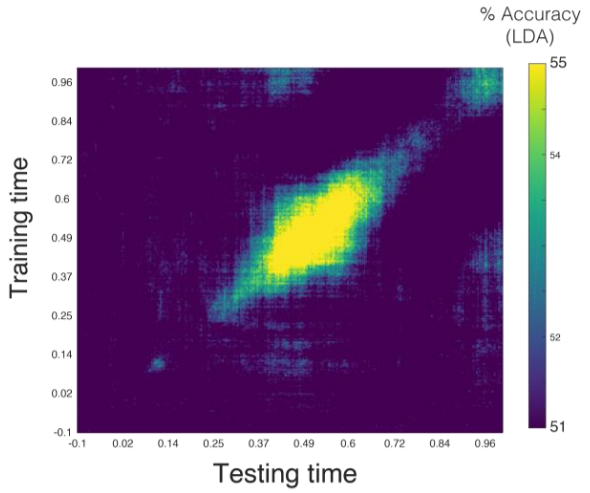

Figure 3. Generalization of EEG patterns through time. Above chance accuracy on the diagonal indicates variable brain representations, while off-diagonal decoding indicates a stable neural encoding of category-specific information.

A number of things can be distilled from the planes shown in figure 2 b and c. First, the peak decoding plateau observed from time-resolved classification (around 550 ms) is helped by the presence of mid-high-frequency, fine detailed visual information (~12-35cpi). Note that this 300-500 ms time window also coincides with the stable accumulation of evidence found in the temporal generalization matrix (figure 3). Second, a coarse-fine-coarse pattern seems to coincide with the diffusion-accumulation-diffusion of information that was shown in figure 3. Third, and finally, feature position maps indicate that high-level information about emotion is supported by clearly defined representation of facial attributes: mouth representation supporting decoding in the 200-400 ms window, followed by a more complete mouth + left-eye representation coinciding with peak decoding accuracy.

Conclusion

Past studies have focused on describing where perceptual and cognitive representations were encoded or when they unfolded in time, but had yet to explicitly describe, and actually *see* the specific visual content that supports such representations. Here, we fill the gap between time-resolved decoding and visual psychophysics and reveal the visual features underlying the decoding of realistic, category-specific information in the brain through time. The visual features we reveal with this method match psychophysical behavioral data (e.g. coarse-to-fine processes, e.g. Caplette et al., 2016; mouth for emotion, Smith et al., 2008). Doing so, we show that seemingly simple linear steps of evidence processing by the visual system in fact engage multiple and interconnected dynamics between visual features, and the accumulation and diffusion of category-specific information in the brain.

Acknowledgments

This work was supported by an European Research Council (ERC) Starting Grant ERC-2017-StG 759432 (to I.C.), and a Natural Sciences and Engineering Research Council of Canada (NSERC) scholarship (to S.F.S).

References

Alink, Arjen, and Ian Charest. n.d. "Individuals with Clinically Relevant Autistic Traits Tend to Have an Eye for Detail." https://doi.org/10.1101/367532.

Caplette, Laurent, Bruno Wicker, and Frédéric Gosselin. 2016. "Atypical Time Course of Object Recognition in Autism Spectrum Disorder." *Scientific Reports*. https://doi.org/10.1038/srep35494.

Cichy, Radoslaw Martin, Dimitrios Pantazis, and Aude Oliva. 2014. "Resolving Human Object Recognition in Space and Time." *Nature Neuroscience* 17 (3): 455–62.

Contini, Erika W., Susan G. Wardle, and Thomas A. Carlson. 2017. "Decoding the Time-Course of Object Recognition in the Human Brain: From Visual Features to Categorical Decisions." *Neuropsychologia* 105 (October): 165–76.

Gosselin, F., and P. G. Schyns. 2001. "Bubbles: A Technique to Reveal the Use of Information in Recognition Tasks." *Vision Research* 41 (17): 2261–71.

Carlson, T. A., Grootswagers, T., & Robinson, A. K. (2019). An introduction to time-resolved decoding analysis for M/EEG. *arXiv preprint arXiv:1905.04820*.

Dobs, Katharina, Leyla Isik, Dimitrios Pantazis, and Nancy Kanwisher. 2019. "How Face Perception Unfolds over Time." *Nature Communications* 10 (1): 1258.

King, J. R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, *18*(4), 203-210.

Mohsenzadeh, Y., Mullin, C., Oliva, A., & Pantazis, D. (2019). The perceptual neural trace of memorable unseen scenes. *Scientific reports*, *9*(1), 6033.

Smith, F. W., Muckli, L., Brennan, D., Pernet, C., Smith, M. L., Belin, P., Gosselin, F., Hadley, D. M., Cavanagh, J. & Schyns, P. G., (2008). Classification images reveal the information sensitivity of brain voxels in fMRI.

Willenbockel, V., Gosselin, F. & Võ (submitted). Spatial frequency tuning for indoor scene categorization

Zhan, Jiayu, Robin A. A. Ince, Nicola van Rijsbergen, and Philippe G. Schyns. 2019. "Dynamic Construction of Reduced Representations in the Brain for Perceptual Decision Behavior." *Current Biology: CB* 29 (2): 319–26.e4.