NeuroImage 81 (2013) 273–282

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

SWIFT: A novel method to track the neural correlates of recognition

Roger Koenig-Robert *, Rufin VanRullen

Centre de Recherche Cerveau et Cognition, Université Paul Sabatier, Université de Toulouse, Toulouse, France
CNRS, CerCo, Toulouse, France

ARTICLE INFO

Article history:
Accepted 28 April 2013
Available online 9 May 2013

Keywords:
Conscious recognition
Object representation
High-level vision
Visual dynamics
Frequency tagging
Consciousness

ABSTRACT

Isolating the neural correlates of object recognition and studying their fine temporal dynamics have been a great challenge in neuroscience. A major obstacle has been the difficulty to dissociate low-level feature extraction from the actual object recognition activity. Here we present a new technique called semantic wavelet-induced frequency-tagging (SWIFT), where cyclic wavelet-scrambling allowed us to isolate neural correlates of object recognition from low-level feature extraction in humans using EEG. We show that SWIFT is insensitive to unrecognized visual objects in natural images, which were presented up to 30 s, but is highly selective to the recognition of the same objects after their identity has been revealed. The enhancement of object representations by top-down attention was particularly strong with SWIFT due to its selectivity for high-level representations. Finally, we determined the temporal dynamics of object representations tracked by SWIFT and found that SWIFT can follow a maximum of between 4 and 7 different object representations per second. This result is consistent with a reduction in temporal capacity processing from low to high-level brain areas.

Introduction

How visual objects are represented as meaningful items in our brains and become part of our conscious experience is one of the most fascinating questions in neuroscience. Current models, largely inspired by invasive studies in monkeys, propose a view of the visual system where object representations emerge progressively from a hierarchical cascade of processing stages (Felleman and Van Essen, 1991; Riesenhuber and Poggio, 1999). Early stages are devoted to extracting simple visual features such as luminance (Amthor et al., 2005), contrast (Sclar et al., 1990), contours (Hubel and Wiesel, 1968) and intersecting lines (Hegdé and Van Essen, 2000). Downstream in the ventral pathway, the integration of these simple features implies that neurons become selective to more and more complex forms, e.g. in area V4 (Gallant et al., 1993). At the highest purely visual area in the ventral stream, the inferotemporal cortex (IT), neurons can be selective to single object categories (Kobatake and Tanaka, 1994; Tanaka, 1996).

While neuronal selectivities in the ventral stream of the monkey visual system are well understood, their associated semantic value is difficult to access. Where and when do meaningful object representations emerge? Non-invasive techniques have been developed to track visual stimulus representations in the human brain – for which perceptual meaning can be more readily assessed. Functional magnetic resonance imaging (fMRI) has played a major role in understanding the human brain areas engaged in object representations. For example, fMRI has revealed that some regions of the temporal lobe are selective to faces in the FFA (Kanwisher et al., 1997), scenes in the PPA (Epstein et al., 1999) or body parts in sub-regions of the LOC (Downing et al., 2001), and there is good evidence that these regions respond more strongly when the corresponding stimuli are consciously perceived by the subjects (Bar et al., 2001; Grill-Spector et al., 2000; Hesselmann and Malach, 2011; Tong et al., 1998). However, the temporal dynamics of object representations on the scale of a few tenths of a second are unattainable to the slower temporal resolution of fMRI. Electroencephalography (EEG) has been extensively used to explore these temporal dynamics in humans. More particularly, steady-state visual evoked potentials (SSVEP) can track the activity elicited by a given visual stimulus in near-real time. This method, also known as frequency tagging, involves the modulation of a stimulus' intensity over time at a fixed temporal frequency *f0*; a neural response is evoked at the same frequency *f0* (and usually its harmonics), thus providing a frequency label (or tag) for the stimulus representation in the brain (Appelbaum and Norcia, 2009; Regan, 1977; Srinivasan et al., 2006). The frequency-tagged response has been found to depend on attention (Ding et al., 2006; Kim et al., 2007; Morgan et al., 1996; Müller et al., 1998) and on the subject's perceptual state (Kaspar et al., 2010; Srinivasan and Petrovic, 2006; Sutoyo and Srinivasan, 2009; Tononi et al., 1998). One limitation of SSVEP is that they normally rely on the modulation of stimulus contrast or luminance; as a result, both semantic object-representations and low-level feature extraction mechanisms are simultaneously tagged

* Corresponding author at: School of Psychology and Psychiatry, Faculty of medicine, Nursing and Health Sciences, Monash University, Clayton, Australia.
E-mail address: rogkoenig@gmail.com (R. Koenig-Robert).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.04.116

at the modulation frequency. As a result, previous studies reported non-consistent effects of object recognition on SSVEP amplitude across tagging frequencies, with recognized images sometimes leading to higher and sometimes to lower SSVEP amplitudes than unrecognized ones (Kaspar et al., 2010). In order to try to disentangle low-level feature extraction processes from semantic object-representations, we developed a novel technique called SWIFT (semantic wavelet-induced frequency tagging) in which we equalized low-level physical attributes (luminance, contrast and spatial frequency spectrum) across all frames of a sequence, while modulating, at a fixed frequency *f0*, the mid- and higher-level image properties carried by the spatial configuration of local contours.

In order to validate the sensitivity of our technique to high-level visual representations, we reasoned that SWIFT should satisfy 3 criteria that we tested in separate experiments. First, activity elicited by explicitly recognized objects should be clearly differentiated from activity elicited by non-recognized objects: indeed, we found that SWIFT is insensitive to unrecognized objects presented up to 30 s, but is highly selective to the recognition of the same objects once their identity has been explicitly revealed. Second, as a consequence of the top-down transmission of attention signals (Lauritzen et al., 2009; Saalmann et al., 2007), attentional modulation intensity should be greater for high-level visual representations than for lower ones; indeed, we demonstrated that SWIFT responses are strongly modulated by top-down attention – considerably more so than classic SSVEP signals. Third, as a result of a reduction in temporal processing capacity from early visual cortex to higher areas (Gauthier et al., 2012; Holcombe, 2009; McKeeff et al., 2007), high-level representations should be limited in their temporal sensitivity: indeed, we found that SWIFT responses reached a limit between 4 and 7 items per second.

Material and methods

SWIFT sequences creation

SWIFT sequences were created by cyclic wavelet scrambling in the wavelets 3D space. We chose wavelet image decomposition rather than other types of image transformation (such as the Fourier transform) because wavelet functions (contrary to Fourier functions) are localized in space: this allowed us to scramble contours while conserving local low-level attributes. The first step was to apply a wavelet transform based on the discrete Meyer (dmey) wavelet and 6 decomposition levels, using the Wavelet toolbox under Matlab (MathWorks); in other words, the image was converted to a multi-scale pyramid of spatially organized maps. At each location and scale, the local contour is represented by a 3D vector $\vec{v1}$, with the 3 dimensions representing the strengths of horizontal, vertical and diagonal orientations. The vector length $\left|\vec{v1}\right|$ is a measure of local contour energy. In a second step, for each location and scale, two random vectors ($\vec{v2}$ and $\vec{v3}$) were defined that shared the length of the original vector ($\left|\vec{v1}\right| = \left|\vec{v2}\right| = \left|\vec{v3}\right|$), thus conserving local energy. By definition, the 3 vectors describe a unique circular path over an isoenergetic sphere where all surface points share the same energy (i.e., the same Euclidian distance from the origin) but represent differently oriented versions of the local image contour. The cyclic wavelet-scrambling was then performed by rotating each original vector (representing the actual image contour), along the circular path defined above. Some wavelet elements (defined by a specific spatial location and decomposition scale) underwent this rotation once per cycle (i.e., at the fundamental frequency *f0*) while others rotated multiple (integer) times per cycle (i.e., at harmonic frequencies of *f0*, from the 2nd up to the 5th harmonic). The introduction of harmonics was crucial to spread the temporal luminance modulation over a broader frequency band, avoiding low-level evoked activity at the tagging frequency *f0*. The 5 harmonic frequencies were distributed equally and randomly among all the wavelet elements. Finally, the inverse wavelet transform was used to obtain the image sequences in the pixel domain. By construction, the original unscrambled image appeared once in each cycle, with a number of intervening wavelet-scrambled frames that depended on the monitor refresh rate and the tagging frequency *f0*. For each original image, several distinct wavelet-scrambling cycles were computed (5 cycles in experiment 1, 2 cycles in experiment 2 and 4 cycles in experiment 3), with different randomly chosen values for the wavelet-scrambling trajectories and the harmonic rotation frequency at each wavelet element. These different cycles were presented in random alternation during the experimental sequences. Two final normalization steps were necessary in order to ensure that the temporal luminance modulation for every pixel was constant (i.e. without any peaks at the individual harmonics frequencies) within the range of harmonic modulation frequencies, and also to ensure the conservation of the mean luminance across frames. First, we calculated the Fourier transform across frames for every pixel and normalized their luminance modulation spectra. Second, mean frame luminance was equalized over time. (NB: A Matlab script following this procedure to create a wavelet-scrambling sequence based on any given original image is available as Supplementary Material).

Subjects

All subjects gave informed consent to take part in these studies that were approved by the local ethics committee. A total of 49 observers (26 women, aged 22 to 53) participated in the 3 experiments (19 in Experiment 1, 8 in Experiment 2 and 24 in Experiment 3).

Stimuli and procedure

For all 3 experiments, subjects were placed at 57 cm of a CRT screen with a refresh rate of 170 Hz in a dark room.

In Experiment 1, 100 SWIFT sequences containing either grayscale natural images (bodies with faces 29%, bodies with no visible faces 16%, animals 21% and manmade objects 14%, downloaded from the Internet) or low-level matched textures synthesized using the texture synthesis algorithm developed by Portilla and Simoncelli (2000) were shown. The number of images in each category was chosen in order to maximize the number of non-canonical images and thus promote the occurrence of 'unrecognized' images. The image contours were modulated cyclically over time at $f0 = 1.4953$ Hz. The experiment was divided in 4 blocks of 25 trials each. Each trial lasted 42 s (30 s of naïve period + 2 s of steady image presentation + 10 s of cognizant period). Sequences ($10.5° \times 10.5°$ visual angle) were presented at the center of the screen over a gray background. Subjects were asked to keep their fixation over a red cross at the center of the display during the trial. They gave their responses (presence of a non-abstract item) at any time during the first naïve period by pressing the left arrow of the computer keyboard for the low confidence threshold (key 1: "I perceive an object-like item, but I am not sure of which object it is") and the right arrow for the high confidence threshold (key 2: "I see an object and I have identified it confidently"). Trials were classified as 'quickly recognized' when a natural image was presented and the subject recognized an object with high confidence within the first 10 s of the naïve period. Trials were classified as 'tardily recognized' when a natural image was presented but the subject did not recognize an object during the 30 s of the naïve period. Trials were classified as 'no-object' when abstract textures were presented. Two of 19 subjects were not considered in the analysis because they had less than 7 tardily recognized trials. For the 17 remaining subjects, the mean number of quickly recognized trials was 22.2, tardily recognized was 22.4 and 20 no-object trials were presented systematically (the remaining trials, corresponding to incomplete or erroneous recognition, were not included in the analysis). Response time for key 2 ("I see an object and I have identified it confidently") in fast recognized