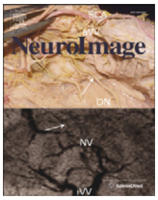
NeuroImage 67 (2013) 213–226

Contents lists available at SciVerse ScienceDirect

NeuroImage

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



Sensory processing during viewing of cinematographic material: Computational modeling and functional neuroimaging

Cecile Bordier *, Francesco Puja, Emiliano Macaluso

Neuroimaging Laboratory, IRCCS, Santa Lucia Foundation, Via Ardeatina 306, Rome 00179, Italy

ARTICLE INFO

Article history:
Accepted 23 November 2012
Available online 29 November 2012

Keywords:
Data-driven
Saliency
Multi-sensory
Cinematographic material
Biologically-inspired vision and audition

ABSTRACT

The investigation of brain activity using naturalistic, ecologically-valid stimuli is becoming an important challenge for neuroscience research. Several approaches have been proposed, primarily relying on data-driven methods (e.g. independent component analysis, ICA). However, data-driven methods often require some post-hoc interpretation of the imaging results to draw inferences about the underlying sensory, motor or cognitive functions. Here, we propose using a biologically-plausible computational model to extract (multi-)sensory stimulus statistics that can be used for standard hypothesis-driven analyses (general linear model, GLM). We ran two separate fMRI experiments, which both involved subjects watching an episode of a TV-series. In Exp 1, we manipulated the presentation by switching on-and-off color, motion and/or sound at variable intervals, whereas in Exp 2, the video was played in the original version, with all the consequent continuous changes of the different sensory features intact. Both for vision and audition, we extracted stimulus statistics corresponding to spatial and temporal discontinuities of low-level features, as well as a combined measure related to the overall stimulus saliency. Results showed that activity in occipital visual cortex and the superior temporal auditory cortex co-varied with changes of low-level features. Visual saliency was found to further boost activity in extra-striate visual cortex plus posterior parietal cortex, while auditory saliency was found to enhance activity in the superior temporal cortex. Data-driven ICA analyses of the same datasets also identified "sensory" networks comprising visual and auditory areas, but without providing specific information about the possible underlying processes, e.g., these processes could relate to modality, stimulus features and/or saliency. We conclude that the combination of computational modeling and GLM enables the tracking of the impact of bottom-up signals on brain activity during viewing of complex and dynamic multisensory stimuli, beyond the capability of purely data-driven approaches.

Introduction

Functional imaging has been used extensively to non-invasively map sensory, motor and cognitive functions in humans. Nonetheless, so far the vast majority of studies have employed simple and repeated stimuli that are in striking contrast with the unrepeated, complex and dynamic signals that the brain has to process in everyday life. Moreover, the neuronal responses in conventional laboratory conditions, i.e. using artificial stimuli, are weaker than those associated with naturalistic stimuli (Mechler et al., 1998; Yao et al., 2007). Thus, recently, a growing interest has risen around the use of more ecologically-valid stimuli during fMRI (e.g. Hasson et al., 2010).

A central issue with naturalistic approaches is that, unlike standard paradigms, there is no straightforward correspondence between the stimuli presented to the subject and any specific sensory, motor or cognitive function. This makes it difficult to use hypothesis-based analysis methods that involve fitting BOLD data with predictors representing specific experimental conditions or processes (general linear model; see Friston et al., 2003). Indeed, many previous studies using complex and dynamic stimuli (e.g. cinematographic material) have resorted to data-driven approaches that do not require any such "a priori" coding.

One of these approaches relies on multivariate analysis based on independent component analysis (ICA) (Bartels and Zeki, 2005; Calhoun et al., 2001a, 2001b; McKeown et al., 1998). ICA performs a blind separation of independent sources from the complex mixture of signal and noise resulting from many different sources. This method does not require any "a priori" specification of the possible causes of the responses (i.e. predictors in a standard GLM analysis) and even no specification of the shape of the hemodynamic response function (HRF). Instead, the method is based on the intrinsic structure of the data. ICA aims to extract a number of unknown sources of signal that are mutually and statistically independent in space or time. Friston (1998) showed the relevance of this idea to biological time-series.

Inter-subject correlation (ISC) analyses is another data-driven approach that has been recently introduced to investigate brain activity associated with the processing of complex stimuli (Hasson et al., 2004; Sui et al., 2012). ISC analyses are based on the idea that presenting the same complex and dynamic sensory input to different subjects will generate

---

* Corresponding author. Fax: +39 06 5150 1213.
E-mail address: c.bordier@hsantalucia.it (C. Bordier).

1053-8119/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2012.11.031

the same pattern of BOLD activity in the brain areas processing the stimuli. Therefore, these areas can be identified by testing for correlated BOLD time-courses between subjects ("synchronization"). This approach has now been employed with a variety of stimuli and tasks (e.g., palm trees task, Seghier and Price, 2009; face processing, Lessa et al., 2011; story comprehension, Lerner et al., 2011; movie watching, Kauppi et al., 2010; representation of action-schemas, Hanson et al., 2009).

However, these data-driven methods have the intrinsic limitation that they require some post-hoc interpretation regarding the processes that generate ICA components (e.g. motion in MT, Bartels and Zeki, 2005) or patterns of inter-subject synchronization (e.g. faces in FFA, Hasson et al., 2004). Here, we propose using an alternative approach where (multi-) sensory stimulus statistics are first extracted via a biologically-plausible computational model and are then used for hypothesis-driven analyses (see also Nardo et al., 2011; Bartels et al., 2008).

One of the most successful biologically-plausible computational model of sensory bottom–up processing consists in the computation of "saliency maps" from complex, naturalistic images (Itti and Koch, 2001; Koch and Ullman, 1985). Inspired from the organization of the visual system, saliency maps are based on the extraction of local discontinuities in intensity, color, orientation, motion and flicker (i.e. feature-specific maps). Saliency maps are then computed as a combination of these feature maps. Saliency maps are thought to well-characterize the spatial distribution of bottom–up signals and have been found to predict sequences of fixations during viewing of naturalistic pictures (Parkhurst et al., 2002) and video-clips (Itti, 2005).

While used primarily in eye-movement studies, saliency has been also considered in electrophysiological studies in monkeys (Gottlieb, 2007; Gottlieb et al., 1998; Thompson et al., 2005) and fMRI in humans (Bogler et al., 2011; Nardo et al., 2011). These studies have indicated that visual saliency is represented in visual areas (Li, 2002; VanRullen, 2003), and can influence activity in higher-order parietal/frontal areas as a function of specific attentional operations (i.e. efficacy of the bottom–up signals for spatial orienting, Nardo et al., 2011; winner-take-all mechanism of attentional selection, Bogler et al., 2011). However, studies employing saliency models took into account only the final "saliency map", with little consideration of the possible contribution of the feature-specific maps (Parkhurst et al., 2002; Itti, 2005; Bartels et al., 2008; but see Itti et al., 1998; Liu et al., 2011). Here we used concurrently *both* saliency and feature maps to investigate the impact of bottom–up sensory signals on brain activity during viewing of complex stimuli.

More recently, the use of computational models of saliency has been extended to the analysis of complex signals in modalities other than vision. Using an approach analogous to the original model for vision (Itti et al., 1998; Koch and Ullman, 1985), Kayser and colleagues proposed a method to compute the saliency of complex, naturalistic auditory stimuli (Kayser et al., 2005; see also Altmann et al., 2008; Kalinli and Narayanan, 2007). Again, saliency maps are constructed combining feature maps that, for audition, extract discontinuities in intensity, frequency and time. Here, for the first time, we propose using auditory saliency – as well as auditory features – to investigate brain activity recorded while volunteers were presented with complex auditory stimuli.

Accordingly, the goal of this study was to use computationally-derived indexes of visual and auditory bottom–up signals (i.e. saliency and features) to assess brain activity during the presentation of complex audio-visual stimuli (cinematographic material); and to evaluate this with respect to a conventional block/event condition-based analysis (cf. Exp 1) and a fully data-driven approach (ICA).

Methods

Participants

This study included two experiments. Eight healthy volunteers took part in the first experiment (Exp1; 2 males, range: 21–24, mean age: 23) and seven different volunteers in the second experiment (Exp2a and Exp2b; 1 male, range: 21–23, mean age: 22). All volunteers were Italian speaking, had normal or corrected-to-normal vision, and did not report any neurological impairment. After having received instructions, all participants gave their written consent. The study was approved by the independent Ethics Committee of the Santa Lucia Foundation (Scientific Institute for Research Hospitalization and Health Care).

Stimuli

In Experiment 1, subjects were asked to view half of an episode of the TV-series "24" (21 min 40 s). For this experiment, we manipulated the original video by switching on-and-off color, motion and/or sound at variable intervals. This provided us with a dataset containing a known correspondence between the stimuli and sound/feature-related sensory processes, which enabled us to perform conventional "condition-based" analyses (see below). The three sensory streams (sound: on/off; motion: motion/static; color: color/black-white) were switched on-and-off in blocks with durations ranging between 8.3 and 43.7 s. The time course of the on-off sequences was independent in the three streams, thus including all possible combinations of the three sound/features input. All subjects were presented with the same version of the modified video (i.e. same sequences of on/off-sets).

In Experiment 2, subjects were presented with the original version of the video, without any manipulation of sound, motion or color. Thus, unlike Exp1, now there was a continuous and unknown change of the different sensory input over time. The experiment was divided into two scanning sessions that consisted of the presentation of the first half (Exp2a: 21 min 40 s) and the second half (Exp2b: 20 min 07 s) of the same "24" episode used in Exp1. None of the volunteers of Exp2 had taken part in Exp1, so they watched the TV episode for the first time. The fMRI data analyses were carried out separately for Exp2a and Exp2b, with the aim of assessing the reproducibility of the results.

For both experiments, visual stimuli were back-projected on a semi-opaque screen at the back of the magnet. Participants viewed the screen through a mirror located above their eyes. The size of the visual display was approximately ($24° \times 16°$ visual angle); the sound was presented via MRI-compatible headphones.

Parameterization of visual/auditory features and saliency

Visual saliency model

The visual saliency model allows identification of salient locations within a given image (Itti et al., 1998; Koch and Ullman, 1985). Our current implementation was modified from the software available at: http://www.saliencytoolbox.net. This considers static visual features only (intensity, color and orientation), while here we added motion and flicker contrasts for the analysis of our dynamic visual stimuli (see Itti and Pighin, 2003). The step-by-step detailed description of the procedures and the specific parameters used to compute the visual saliency maps are reported in the Supplementary materials (Table S1).

Modeling consists of decomposing the input image/s into a set of distinct "channels" using linear filters tuned to specific stimulus dimensions. This decomposition is performed at several levels, extracting contrasts at different spatial scales (Gaussian pyramids; Greenspan et al., 1994). The scales were created using pyramids with 9 levels. Together with intensity, color and orientation, here flicker was computed from the absolute difference between the intensity of the current frame and that of the previous frame. Motion was computed from spatially-shifted differences between the intensity pyramids from the current and the previous frame (Itti and Pighin, 2003; Reichardt, 1987). Center-surround interactions are then implemented as differences between fine and coarse scales of the pyramids, here using a set of 6 cross-scale subtractions for each channel of each feature (Itti and Pighin, 2003; Walther