Neuroscience and Biobehavioral Reviews 74 (2017) 44–57

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

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

Understanding natural scenes: Contributions of image statistics

Andrea De Cesarei [a,*], Geoffrey R. Loftus [b], Serena Mastria [a], Maurizio Codispoti [a]

[a] Department of Psychology, University of Bologna, Bologna, Italy
[b] Department of Psychology, University of Washington, Seattle, WA, USA

ARTICLE INFO

Article history:
Received 29 September 2016
Received in revised form 5 January 2017
Accepted 9 January 2017
Available online 12 January 2017

Keywords:
Natural scenes
Attention
Learning
Emotion

ABSTRACT

Visual processing of natural scenes is carried out in a hierarchical sequence of stages that involve the analysis of progressively more complex features of the visual input. Recent studies have suggested that the semantic content of natural stimuli (e.g., real world photos) can be categorized based on statistical regularities in their appearance, which can be detected early in the visual processing stream. Here we review the studies which have investigated the role of scene statistics in the perception of natural scenes, focusing on both basic visual processing and specific tasks (visual search, expert categorization, emotional picture viewing). Visual processing seems to be adapted to visual regularities in the visual input, such as the amplitude-frequency relationship. Moreover, scene statistics can aid performance in specific tasks such as distinguishing animals from artifactual scenes, possibly by modulating early visual processing stages.

Contents

The visual system continuously translates retinal inputs into meaningful semantic representations. Everyday experience indicates that understanding the meaning of incoming visual information can usually be done without any overt feeling of effort. A single glance is sufficient for us to recognize a typical scene with cars in a street, surrounded by buildings, as "a town". Furthermore, we can train ourselves to recognize abstract stimuli; for instance, experienced radiologists can efficiently identify many features of medical images that lay people cannot.

In spite of its subjective ease, vision is computationally highly complex. The visual input that hits our eyes is constantly changing in terms of position, shading, and illumination; in addition, the visual input is often incomplete, for instance due to the presence of occluding objects. One of the main problems that the visual system faces is that of converting the ever-changing, incomplete, and often ambiguous retinal information into a stable semantic representation. This is achieved through a combination of a bottom-up hierarchical analysis of the visual input, and top-down feedback based on prior or contextual information.

As a result of bottom-up sensory stimulation and of the task context, visual input is processed and a stable semantic representation is eventually attained. According to several authors, the inherent uncertainty of visual information requires the visual system to take a probabilistic approach which involves multiple stages of processing. In the domain of object recognition, it has been suggested that multiple possible interpretations of the input are initially generated, and subsequent processing guides the choice towards the most likely candidate (Bar, 2004; Sanocki, 1993; Hochstein and Ahissar, 2002). Similarly, in the domain of visual search it has been

* Corresponding author at: Department of Psychology, University of Bologna, Viale Berti Pichat, 5, 40127 Bologna, Italy.
E-mail address: andrea.decesarei@unibo.it (A. De Cesarei).

http://dx.doi.org/10.1016/j.neubiorev.2017.01.012
0149-7634/

suggested that a first, efficient but unselective stage is followed by a later stage of processing, which can operate on more complex visual representation at a cost of more attentional resources (Di Lollo et al., 2001; Hoffman, 1979; Wolfe et al., 1989). In the literature that is reviewed here, it is suggested that the earliest visual processing stage may build on the analysis of the statistical structure of the visual input, and this can prove to be beneficial to the system as it exploits regularities in the appearance of natural scenes.

Although visual processing happens continuously, its products (percepts) are often not accessible to awareness. Rather, they may become accessible if specific conditions occur, e.g. when attention or memory require an overt focus on the perceptual outcome. Consistently, the manipulation of relevance through attention has been often used as a way to investigate visual perception (Mack and Rock, 1998). Several factors affect the relevance of a scene content. Here we will consider three contexts in which some scenes acquire a more relevant status compared to other scenes: visual search, expert categorization, and emotional picture viewing. When we are looking for our keys on a cluttered desktop, we become aware that the keys are there. We take less time to identify a specific dog if we have learned to identify it, as in the case of our own dog. Finally, emotional stimuli such as pictures of erotic couples and mutilated bodies are implicitly more relevant than neutral scenes—and, even in the absence of task demands, responses are elicited which can be interpreted as evidence that the content of a scene was processed.

In the following sections, we provide a conceptual review of some findings in the field of natural scene understanding. After briefly focusing on the hierarchical processing of the visual input which is done in the earliest stages of visual processing, we move to studies which have investigated the relationship between natural scene statistics and visual processing of natural scenes. Specifically, we examine the role of visual statistics in target detection, expert categorization, and viewing of emotional stimuli.

1. Properties of hierarchical visual processing

At its most basic level, visual processing can be broken into a sequence of hierarchically organized stages which analyze progressively more complex aspects of the visual input (Goldstone and Hendrickson, 2010; Riesenhuber and Poggio, 1999; Serre, 2016; Wallis and Rolls, 1997). At the first stages of cortical visual processing, hierarchical processing is reflected in the organization of simple and complex cells in the visual cortex (Hubel and Wiesel, 1959; reviewed in Hubel and Wiesel, 1998); more specifically, these types of cells differ in their sensitivity to changes in stimulus size, orientation, and position, and are thought to represent hierarchically distinct levels of visual analysis. In subsequent stages, the responses of simple and complex cells are further combined and more complex visual representations can be achieved. Later in the visual processing stream, progressively more complex levels of processing are carried out, until, eventually, object-level identification is reflected in the activation of specific cortical areas, such as the lateral occipital complex (DiCarlo et al., 2012; Grill-Spector et al., 2001; Ungerleider and Bell, 2011;). While a full review of hierarchical visual processing is outside the scope of this review, two important properties will be discussed, namely the attainment of perceptual invariances, and the specialization of the visual system towards specific visual features or classes of stimuli.

Invariant object representations are a fundamental achievement in the visual processing hierarchy. Invariant object representations consist in the mapping of several sensory stimuli to a same perceptual representation. Objects may vary markedly in their appearance (e.g., due to changes in viewpoint or illumination), and may produce radically different retinal and cortical patterns of activation. Mapping these patterns of activation into the same perceptual representation allows observers to perceive an object as one that remains the same over time, despite changes in motion and illumination. Visual invariances, for instance in terms of position- and scale-tolerant representations, are first observed in the inferotemporal cortex (Cauchoix and Crouzet, 2012; Miyashita, 1993 ; Nishijo et al., 1993), a region of the primate brain which is functionally homologous to the human lateral occipital complex (Grill-Spector et al., 2001).

The available evidence concerning the functioning of the visual system indicates that a progressively more complex sequence of stages analyzes the visual input.[1] However, it has been debated whether the same processing sequence is executed for both simple artificial stimuli (e.g., gratings and digits) and more complex ones (e.g., natural scenes). Some studies have challenged this possibility, suggesting that natural stimuli have an advantage compared to artificial ones, in that they require fewer attentional resources for categorization (Li et al., 2002; VanRullen et al., 2005). For example, in one study (Li et al., 2002), participants were asked to report on the stimulus category (simple characters or natural scenes) which were presented outside the fovea, while performing a foveal attentional task. It was observed that natural scenes were accurately categorized even when attentional resources were exhausted by the competing foveal task, while the categorization of simple characters in the same conditions was at chance level (Li et al., 2002). These findings are at odds with a feature integration model, which presumes that all objects which fall in the focus of attention must be analyzed as a conjunction of simpler features, and require attentional resources to be integrated (Treisman and Gelade, 1980).

One possible explanation for the advantage of natural scenes over simple stimuli is that, because the visual system has evolved in the real world, natural scenes may be more engaging stimuli; that is, they may require less directed attention than simpler stimuli such as bars, spots, or sinusoidal gratings (Braun, 2003; Felsen and Dan, 2005; Hasson et al., 2010). This interpretation assumes that object features (such as the presence of an eye) could be efficiently detected in a scene because they are meaningful features in the real world. At a more general level however, this possibility links to the debate about which stimulus features may be associated with efficient processing, and can guide visual attention (Wolfe and Horowitz, 2004). It has been suggested that both object features and statistical regularities in the appearance of a scene guide visual attention during scene viewing; however, when the available evidence for guidance of attention by semantic category was reviewed, object features and statistical regularities were regarded as “probably not guiding attributes” because of the confounding between perceptual and semantic features (Wolfe and Horowitz,

[1] A debated issue in the study of the organization of the visual cortex is whether some areas are specifically devoted to the processing of some stimuli, such as faces. Domain-specific positions posit that specialized areas exist, which are more or exclusively involved in the processing of one stimulus class. One prominent example is a region of the fusiform cortex, which has been named the fusiform face area (FFA), as it responds more strongly to faces than to objects (Kanwisher et al., 1997). Similar observations have suggested that another fusiform region is most responsive to bodies (fusiform body area, FBA; Peelen and Downing, 2005). On the other hand, domain-general accounts of visual processing assume that similar processes analyze all visual information, and little or no specificity for classes of stimulus, such as faces, exist. In this view, content-specific effects may arise because of expertise, familiarity, or because of other factors that make one category or feature more relevant than others (Gauthier et al., 2000; Gauthier et al., 1999). While these views have been subject to close scrutiny and contrast in the past 20 years, it has recently been suggested that focusing on the question of whether an area is or is not specifically dedicated to the processing of a category may be misleading, and an conclusive answer cannot be provided, due to differences in category structure, perceptual features, and data analysis procedures (Gauthier and Tarr, 2016). Rather, these authors suggest a feature-based approach, raising the questions of which features are diagnostic for specific tasks, which cortical mechanisms analyze these features, and what the level of abstraction (invariance) is of each cortical processing stage involved.