Cognition 123 (2012) 33–49

Contents lists available at SciVerse ScienceDirect

Cognition

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

Mental visualization of objects from cross-sectional images

Bing Wu [a,*], Roberta L. Klatzky [b], George D. Stetten [c,d]

[a] Cognitive Science and Engineering Program, Arizona State University, Mesa, AZ 85212, USA
[b] Department of Psychology and Human–Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, USA
[c] Robotics Institute, Carnegie Mellon University, Pittsburgh, USA
[d] Department of Biomedical Engineering, University of Pittsburgh, Pittsburgh, USA

ARTICLE INFO

Article history:
Received 18 May 2011
Revised 16 November 2011
Accepted 7 December 2011
Available online 2 January 2012

Keywords:
Visualization
Integration
Spatiotemporal
Anorthoscopic
Cross-section

ABSTRACT

We extended the classic anorthoscopic viewing procedure to test a model of visualization of 3D structures from 2D cross-sections. Four experiments were conducted to examine key processes described in the model, localizing cross-sections within a common frame of reference and spatiotemporal integration of cross sections into a hierarchical object representation. Participants used a hand-held device to reveal a hidden object as a sequence of cross-sectional images. The process of localization was manipulated by contrasting two displays, *in situ* vs. *ex situ*, which differed in whether cross sections were presented at their source locations or displaced to a remote screen. The process of integration was manipulated by varying the structural complexity of target objects and their components. Experiments 1 and 2 demonstrated visualization of 2D and 3D line-segment objects and verified predictions about display and complexity effects. In Experiments 3 and 4, the visualized forms were familiar letters and numbers. Errors and orientation effects showed that displacing cross-sectional images to a remote display (*ex situ* viewing) impeded the ability to determine spatial relationships among pattern components, a failure of integration at the object level.

1. Introduction

While visual object recognition may seem effortless, it is actually a highly constructive process, involving a stream of processing that begins with the retinal image, advances through a sequence of computational stages, and culminates in a match to representations in memory (Biederman, 1987; Marr, 1982; Tarr & Bülthoff, 1998). The necessity for construction is particularly apparent in the situation where information constituting an object is progressively unveiled from spatially distributed exposures. For example, viewing large objects may require integration across saccades, each with its own retinal projection (Irwin, 1991; Melcher & Morrone, 2003; Rayner, 1978), or people may rotate an object to view it from multiple perspectives (Harman, Humphrey, & Goodale, 1999). Under such conditions the ability to integrate information over time and space is critical to the formation of object representations. Here we study spatiotemporal integration in a particular form: Observers use a sequence of 2D cross sections, taken from a virtual object in 3D space, to obtain a representation of its 3D structure. We wish not only to demonstrate the capability for 3D construction from cross-sections, but also to further understand the underlying processes that make it possible.

1.1. Construction of object representations: a spatiotemporal process

Despite substantial progress in understanding visual object recognition, the underlying processing still remains a matter of debate. The various theories that have been formulated can be broadly categorized into two approaches,

* Corresponding author. Address: Cognitive Science and Engineering Program, Arizona State University, 7271 E. Sonoran Arroyo Mall, Mesa, AZ 85212, USA.

E-mail address: Bing.Wu@asu.edu (B. Wu).

0010-0277/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2011.12.004

view-based vs. structure-based, according to their views on mental object representation (for review, see Riesenhuber & Poggio, 2000; Tarr & Vuong, 2002). While the two approaches are quite different, and in many aspects mutually contradictory, they share a fundamental idea: The recognition of an object relies on object representations derived from retinal stimulation. In the context of structure-based approaches, for example, Marr's (1982) computational framework describes how object representation progresses from the retinal image to a ''primal sketch'' and finally to an object-centered 3D model, whereas Biederman's Recognition-By-Components theory (1987) proposes an intermediate volumetric representation called a ''geon.'' In contrast, view-based approaches advocate that retinal images lead to a collection of stored 2D views (Bülthoff, Edelman, & Tarr, 1995; Lawson, Humphreys, & Watson, 1994; Tarr & Bülthoff, 1998), which are compared to incoming targets subject to transformations such as interpolation and mental rotation (Lawson, 1999; Willems & Wagemans, 2001).

The debate over the representations used in object processing should not obscure an important source of information, namely, the spatiotemporal relationships across patterns of visual stimulation as an object is perceived. In the real world, motion of a viewer or an object is arguably more common than static, passive viewing. As a result, the visual system has access to a sequence of images that exhibit temporal and spatial correlations, which it takes advantage of for object perception and recognition. Previous research has shown that such spatiotemporal information can aid in interpreting biological motion (Johansson, 1973), categorizing familiar objects (Lawson et al., 1994), identifying faces (Wallis & Bülthoff, 2001), and recognizing novel objects (Stone, 1998, 1999). For example, participants in Stone's studies (1998, 1999) saw videos of amoeba-like objects that rotated in one direction in the learning phase but in the reversed direction in the subsequent recognition phase. The rotation-reversal produced a significant reduction in recognition performance, suggesting that spatiotemporal information had been incorporated in object representations. Naming familiar objects is also found to be easier when participants viewed structured sequences compared to random sequences of object views (for review see Lawson, 1999).

Although such studies clearly implicate a role for spatiotemporal information in object recognition, few studies have addressed how object representations can be constructed through spatiotemporal integration, particularly in the domain of 3D structures. A difficulty in constructing tasks to study 3D integration per se is how to preclude alternative strategies. For example, the observed effects of structured sequences in recognizing familiar objects might be related to the learning of association between image features and pre-existing object representations, rather than a spatiotemporal process of building an object representation. In order to avoid such confounds, the present research asked people to visualize objects on the basis of their cross-sectional images. In geometry, a cross section is a 2D sample taken from a 3D object by slicing the object with a plane. Cross-sections are thus ideal for studying spatiotemporal integration in a pure form: They have the important properties that they reveal only a small portion of the 3D object, exclude volume information and, often, preclude perceptual exposure of critical features that might, by themselves, support responses. Moreover, being defined in object-centered space, the internal structures shown in cross-sectional images are perspective independent. The motivation for the current research is not only to demonstrate the capability for 3D construction from cross-sections and its role in object recognition, but also to further understand the underlying processes that make it possible. This work can be considered as a 3D analog of *anorthoscopic* perception, which we next briefly review.

1.2. Anorthoscopic perception through aperture viewing

A classic technique used to study the integration of spatiotemporally distributed visual exposures is anorthoscopic viewing (Zöllner, 1862), in which a large figure is exposed by passing a small aperture over it, or the figure passes behind a stationary aperture. Over many variations in this paradigm, there is broad agreement that such piecemeal exposures can lead to the perception of the whole figure that transcends the limited aperture available at any one point in time, albeit subject to distortion or noise (Fendrich, Rieger, & Heinze, 2005; Hochberg, 1968; Kosslyn, 1980; Palmer, Kellman, & Shipley, 2006; Parks, 1965; Rock, 1981). Three-dimensional objects that translate or rotate behind an aperture (Day, 1989; Fujita, 1990) can also be recognized anorthoscopically.

Although the mechanisms underlying anorthoscopic perception are still not fully understood, it is commonly accepted that the necessary processes include segmentation of portions of the underlying figure as garnered through occluding aperture, storage of piecemeal information acquired over time, localizing the pieces within a common spatial framework, and finally, assembling the pieces into an integrated form. Of particular interest in this study are the last two processes, namely, localizing the piecemeal views and integrating across views according to their spatial relationships. Together these processes constitute a form of *visualization* (McGee, 1979).

Two general hypotheses have been suggested to explain how 2D anorthoscopic perception might be accomplished. The ''retinal painting'' hypothesis, first formulated by Helmholtz (1867), asserts that when an extended scene is viewed through a moving aperture, a representation is constructed by projecting the successive views onto adjacent retinal loci, thereby resulting in an integrated percept. Contravening this hypothesis, however, there is evidence that an anorthoscopic percept can occur when the possibility of retinal painting has been eliminated (Fendrich et al., 2005). Moreover, research has shown that even early integration processes, operating at the level of saccades, make use of a spatiotopic frame of reference (Melcher & Morrone, 2003).

An alternative, the post-retinal storage hypothesis, suggests that the information available through the aperture is stored in a working memory and then combined into a whole figure (Girgus, Gellman, & Hochberg, 1980; Hochberg, 1968; Parks, 1965; Rock, 1981). This theory requires a mechanism for localizing each visible aperture,