Vision Research 119 (2016) 42–49

Contents lists available at ScienceDirect

Vision Research

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

A proto-object based saliency model in three-dimensional space

Brian Hu [a,b,*], Ralinkae Kane-Jackson [a,1], Ernst Niebur [a,c]

[a] Zanvyl Krieger Mind/Brain Institute, Johns Hopkins University, Baltimore, MD 21218, United States
[b] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, United States
[c] Department of Neuroscience, Johns Hopkins University, Baltimore, MD 21218, United States

ARTICLE INFO

Article history:
Received 24 September 2015
Received in revised form 16 December 2015
Accepted 20 December 2015
Available online 19 January 2016

Keywords:
Visual attention
Saliency map
Depth saliency
Proto-object
Gestalt

ABSTRACT

Most models of visual saliency operate on two-dimensional images, using elementary image features such as intensity, color, or orientation. The human visual system, however, needs to function in complex three-dimensional environments, where depth information is often available and may be used to guide the bottom-up attentional selection process. In this report we extend a model of proto-object based saliency to include depth information and evaluate its performance on three separate three-dimensional eye tracking datasets. Our results show that the additional depth information provides a small, but statistically significant, improvement in the model's ability to predict perceptual saliency (eye fixations) in natural scenes. The computational mechanisms of our model have direct neural correlates, and our results provide further evidence that proto-objects help to establish perceptual organization of the scene.

1. Introduction

The brain receives large amounts of visual information that it must make sense of in real-time. Processing the entire visual field with the same level of detail present at the fovea would be an exceedingly complex and costly task requiring much greater computational resources than are available (Tsotsos, 1990). As a result, primates select only the most relevant information and discard the rest, a process known as selective attention. Visual attention is controlled by both bottom-up and top-down mechanisms, which interact to influence the organism's behavior (Yarbus, 1967). Bottom-up attention is involuntary and signal-driven, largely due to the fact that some stimuli are more conspicuous and able to stand out from their surroundings. Top-down attention is task-dependent, and can take into account semantic information such as the familiarity or interestingness of an object, which biases the organism's attention based on its internal state or goals.

Many models of visual attention are constructed with a bottom-up architecture and rely on local contrast in low-level features such as intensity, color, orientation, or motion. Biologically-plausible center-surround differences across different feature channels of an input image can be used to compute a "saliency map" whose maxima indicate where selective attention is deployed (Koch & Ullman, 1985; Niebur & Koch, 1996; Itti, Koch, & Niebur, 1998). However, there is both psychophysical (Einhäuser, Spain, & Perona, 2008) and neurophysiological (Zhou, Friedman, & von der Heydt, 2000; Qiu, Sugihara, & von der Heydt, 2007) evidence that attention relies not only on these simple image features, but also on the perceptual organization of the visual scene into tentative objects, or proto-objects (Rensink, 2000). A biologically-inspired model of proto-object based saliency has been proposed to take into account these recent findings (Craft, Schütze, Niebur, & von der Heydt, 2007; Mihalas, Dong, von der Heydt, & Niebur, 2011; Russell, Mihalas, von der Heydt, Niebur, & Etienne-Cummings, 2014). The model includes border ownership selective cells (referred to as border ownership cells in the following) and grouping cells, which interact to achieve figure-ground segmentation of the image into proto-objects (figures) and the background (ground). Border ownership cells have been found in primate visual cortex, with the majority of neurons in area V2 having this property. These cells signal in their neural activity the one-sided assignment of an object border to the region perceived as figure (Zhou et al., 2000). Border ownership cells are also modulated by attentional influences (Qiu et al., 2007). Grouping cells integrate global context information about proto-objects in the scene according to Gestalt principles such as closure, continuity, convexity, *etc.* Importantly, grouping cells act at intermediate stages of vision and do not require higher-level information about object identity, semantic knowledge, *etc.* They send feedback to border ownership cells *via* fast white matter projections, which bias the activity of border ownership cells to reflect the correct

* Corresponding author at: Zanvyl Krieger Mind/Brain Institute, Johns Hopkins University, 3400 N. Charles Street, 338 Krieger Hall, Baltimore, MD 21218-2685, United States.

E-mail addresses: bhu6@jhmi.edu (B. Hu), rkanejac@terpmail.umd.edu (R. Kane-Jackson), niebur@jhu.edu (E. Niebur).

[1] Present address: Department of Electrical Engineering, University of Maryland, College Park, MD 20742, United States.

http://dx.doi.org/10.1016/j.visres.2015.12.004
0042-6989/

figure-ground segmentation of proto-objects. In this framework, visual saliency is a function of grouping cell activity, which represents the size and location of proto-objects within the image.

Border ownership cells have been shown to respond to figure edges defined by a variety of image features, *e.g.* luminance edges, color edges, *etc.* When no monocular edge information is present (*i.e.* when the figures are defined by random dot stereograms using only binocular disparity), border ownership selectivity is also imparted by stereoscopic edges (Qiu & von der Heydt, 2005). Critically, their response to these different figural cues is typically the same in the two-dimensional (2D) and three-dimensional (3D) cases – the preferred side-of-figure of border ownership cells is consistent for all cues that define the figure. The activity of border ownership cells thus provides an interpretation of the visual scene in terms of depth-ordered surfaces that correspond to objects in 3D space.

In a separate line of work, it has been shown that surface representations play a key role in intermediate-level vision, and that visual attention can be deployed at the level of perceptual surfaces (He & Nakayama, 1992; He & Nakayama, 1995, for a model of attention to surfaces see Hu, von der Heydt, & Niebur, 2015). Despite these experimental observations, current models of border ownership do not explicitly use depth information and do not address how traditional 2D Gestalt cues interact with depth cues during the figure-ground segmentation process. An exception is a study by Mishra, Shrivastava, and Aloimonos (2012) who used computer vision methods to compute border ownership from low-level depth information and then performed object segmentation in natural images.

Even though in recent years stereoscopic 3D content has become increasingly prevalent, *e.g.* in the viewing of entertainment programs in cinemas and homes, little is known about how visual attention is deployed within 3D environments. It is thus important to understand how humans allocate their attention when viewing natural images and videos in 3D (Le Callet & Niebur, 2013). Binocular disparity cues, which can be used to generate strong depth percepts, have been shown to alter different aspects of eye movements when participants viewed 3D images (Jansen, Onat, & König, 2009) and videos (Huynh-Thu & Schiatti, 2011). Only recently have 3D eye tracking datasets been made available which can be used to compare human eye movements with predictions of attentional models. The availability of these datasets and the recent explosion in new 3D content makes it possible to design computational models of 3D saliency and evaluate their performance objectively.

The goals of our research are (1) to extend a proto-object based saliency model (Russell et al., 2014) to include depth information, and (2) to evaluate its performance in perceptual saliency prediction. We show that combining 2D Gestalt cues with depth cues improves the performance of our model on three different 3D eye tracking datasets. In the model, depth information along with other 2D features biases grouping cell activity, which then interacts with border ownership cells to represent proto-objects, the tentative objects within the scene. These proto-objects are a first step in figure-ground segmentation of the image, and also give an indication of the salient points within the image. We evaluate the proto-object saliency maps produced by our model against ground truth data in the form of human eye fixations using a battery of different metrics.

2. Related work

2.1. Models of 3D visual attention

Compared to the number of models that have been proposed for 2D visual saliency, relatively few attempts have been made to study how visual attention is deployed within 3D environments. Existing models of 3D visual attention often compute a 2D saliency map which is then combined with the depth information to produce a new saliency map. These models fall into three categories (Wang, DaSilva, LeCallet, & Ricordel, 2013) based on how the depth information is used: stereovision models, depth-weighting models, and depth-saliency models. For a comprehensive review of 3D visual attention models, see Wang et al. (2013), Ma and Hang (2015).

While the depth-weighting and depth-saliency models assume that a depth map has been computed, without specifying how, stereovision models explicitly implement the computation of depth information from the left and right views of the scene, thus replicating the human visual system's stereoscopic perception. An example of this is a study by Bruce and Tsotsos (2005), which extended a 2D selective tuning model of attention to also incorporate binocular information. However, no quantitative assessment of this model was performed.

Depth-weighting models use a base 2D saliency model (computed using one of the existing methods) and then multiplicatively weight the resulting saliency map with the depth information. Regions that are closer to the observer obtain higher weights, corresponding to greater combined saliency. In a model developed by Lang et al. (2012), novel depth priors are learned from a training portion of the data, and these are then combined with the output of a 2D saliency model either using pixel-wise addition or multiplication. With these depth priors, the authors find an increase of performance by 6–7% on their dataset compared to the base 2D model without depth information.

Depth-saliency models come in two flavors. In one, both a depth saliency map, obtained from depth alone, and a more traditional saliency map, obtained from 2D information alone, are computed. The two maps are then linearly combined to generate the final saliency map. Wang et al. (2013) determine depth saliency in a separate experiment involving synthetic stereoscopic stimuli, which allows them to reduce the influence of monocular depth cues, as well as control for the depth of objects and the background. With their experimental results, they propose a probabilistic model of depth saliency, where the probability of a point being fixated in 3D space is related to the magnitude of center-surround differences in depth contrast. Linearly combining these two saliency maps in a 1:1 ratio (50% weight each for 2D features and depth information) results in better performance on their dataset. In the second type of depth-saliency models, depth information is treated as an additional feature channel, on the same footing as intensity, color, orientation, *etc.* The final saliency map is then a function of depth as well as of these other features (Ouerhani & Hügli, 2000; Jost, Ouerhani, von Wartburg, Müri, & Hügli, 2004; Hügli, Jost, & Ouerhani, 2005).

Our approach falls in the latter class of depth-saliency models, where all image features, including depth, interact through linear combination resulting in the final saliency map. Our model is completely integrated – depth information is treated as another cue which interacts with 2D Gestalt cues to influence figure-ground assignment of proto-objects within the scene. This agrees with anatomical and neurophysiological data that show that disparity selective cells, which are important for encoding stereoscopic depth information, are found in the same early cortical areas as neurons representing other features used in typical saliency models, like color and orientation (Hubel & Wiesel, 1962; Poggio, Gonzalez, & Krause, 1988). Different from previous models (Ouerhani & Hügli, 2000; Jost et al., 2004; Hügli et al., 2005), our model is not only based on basic image features (like color, intensity, *etc.*) but it includes elements of perceptual organization, in particular proto-objects. The model is an extension of a previously described 2D model (Russell et al., 2014) and is constructed by including depth information as an additional feature. All features are used to determine proto-object based saliency.