Vision Research 112 (2015) 11–25

Contents lists available at ScienceDirect

Vision Research

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

How does image noise affect actual and predicted human gaze allocation in assessing image quality?

Florian Röhrbein [a], Peter Goddard [b], Michael Schneider [a], Georgina James [b], Kun Guo [b,*]

[a] Institut für Informatik VI, Technische Universität München, Germany
[b] School of Psychology, University of Lincoln, UK

ARTICLE INFO

Article history:
Received 1 October 2014
Received in revised form 27 February 2015
Available online 14 May 2015

Keywords:
Natural scene
Image distortion
Image quality
Gaze behaviour
Visual attention model

ABSTRACT

A central research question in natural vision is how to allocate fixation to extract informative cues for scene perception. With high quality images, psychological and computational studies have made significant progress to understand and predict human gaze allocation in scene exploration. However, it is unclear whether these findings can be generalised to degraded naturalistic visual inputs. In this eye-tracking and computational study, we methodically distorted both man-made and natural scenes with Gaussian low-pass filter, circular averaging filter and Additive Gaussian white noise, and monitored participants' gaze behaviour in assessing perceived image qualities. Compared with original high quality images, distorted images attracted fewer numbers of fixations but longer fixation durations, shorter saccade distance and stronger central fixation bias. This impact of image noise manipulation on gaze distribution was mainly determined by noise intensity rather than noise type, and was more pronounced for natural scenes than for man-made scenes. We furthered compared four high performing visual attention models in predicting human gaze allocation in degraded scenes, and found that model performance lacked human-like sensitivity to noise type and intensity, and was considerably worse than human performance measured as inter-observer variance. Furthermore, the central fixation bias is a major predictor for human gaze allocation, which becomes more prominent with increased noise intensity. Our results indicate a crucial role of external noise intensity in determining scene-viewing gaze behaviour, which should be considered in the development of realistic human-vision-inspired attention models.

1. Introduction

When exploring natural surroundings, we do not direct our attention evenly or randomly to different parts of the scene. Instead, we make a series of saccades to direct a limited number of fixations to local regions that are informative or interesting to us. The preferred regions within a scene are often inspected earlier and attract more fixations (Henderson, 2007). Such gaze allocation provides a real-time behaviour index of on-going perceptual and cognitive processing and is a sensitive index of our attention, motivation, and preference, especially when exploring scenes of high ecological validity (Henderson, 2007; Isaacowitz, 2006). One central research question in this active visual exploration process is to understand how we choose the fixated local regions in the scene.

Many empirical studies have suggested that both bottom-up local saliency computation and top-down cognitive processes are actively involved in determining our fixations in scene exploration. Specifically, the choice of foveated local region is heavily influenced by local low-level image saliency (e.g., local image colour, intensity, contrast, spatial frequency, and structure). We tend to avoid low-contrast and homogeneous 'predictable' regions in natural scenes, and bias our fixation to local features with high-contrast, high spatial frequency, high edge density, and complex local structure (e.g., curved lines, edges and corners, as well as occlusions or isolated spots) (Acik et al., 2009; Krieger et al., 2000; Mannan, Ruddock, & Wooding, 1995, 1996; Parkhurst & Niebur, 2003; Reinagel & Zador, 1999), or to local regions deviated from surrounding image statistics (Einhäuser et al., 2006). On the other hand, top-down factors, such as expectation, memory, semantic and task-related knowledge, could significantly modulate gaze allocation in scene exploration (Guo et al., 2012; Henderson, 2007; Pollux, Hall, & Guo, 2014; Tatler et al., 2011).

These experimental findings complement the development of computational models for predicting where people look in natural

* Corresponding author at: School of Psychology, University of Lincoln, Lincoln LN6 7TS, UK. Fax: +44 1522 886026.
E-mail address: kguo@lincoln.ac.uk (K. Guo).

http://dx.doi.org/10.1016/j.visres.2015.03.029
0042-6989/

vision. Closely resembling our knowledge about neural processing in early visual system, the widely cited bottom-up saliency model (Itti & Koch, 2000) compares local image intensity, colour and orientation through centre-surround filtering at eight spatial scales, combines them into a single salience (conspicuity) map with a winner-take-all network and inhibition-of-return, and then produces a sequence of predicted fixations that scan the scene in order of decreasing salience. To improve its relatively low level of predictive power (e.g., 57–68% correct fixation prediction in some scene free-viewing tasks, Betz et al., 2010), some top-down processing such as scene context (contextual guidance model, Torralba et al., 2006; context-aware saliency, Goferman, Zelnik-Manor, & Tal, 2012), object detection (Judd et al., 2009) and natural statistics (Kanan et al., 2009) are later incorporated into the model. Incorporating these top-down cues does not necessarily sacrifice the computational precision of the original saliency map model, or even alter the basic structure of the approach (Navalpakkam & Itti, 2005). Specifically, combining both bottom-up saliency-driven information and top-down natural scene understanding would greatly improve gaze predictions in a real-word image search task (Kanan et al., 2009). It seems that humans utilise both local image saliency and global scene understanding in guiding eye movements to efficiently sample scene information.

These experimental findings and computational models of visual attention in scene perception are derived mainly from studies using high-quality images in laboratory settings. Real-world scene perception, however, often involves selecting, extracting and processing diagnostic information from a noisy environment (e.g., due to bad weather condition). Typically, the images and videos we view daily are subject to a variety of distortions during acquisition, compression, storage, transmission and reproduction, any of which will degrade visual quality. It is proposed that most distortion processes would disturb natural image statistics (Sheikh, Bovik, & de Veciana, 2005) and may attract attention away from local regions that are salient in undistorted images. Furthermore, our perceptual processing strategy tends to change with the level of external noise, independent of the observer's internal noise (Allard & Cavanagh, 2012).

Considering that our visual system has evolved and/or learned over time to process visual signals embedded in natural distortions, it is reasonable to assume that we should have developed a near-optimal processing strategy for visual signals corrupted by these distortions. So far only a handful of psychophysical and computational studies have attempted to investigate our perceptual sensitivity to image blur (e.g., Watson & Ahumada, 2011) and image resolution (e.g., Castelhano & Henderson, 2008; Torralba, 2009). These studies have shown that we could essentially classify natural scenes or understand scene gist at a very low resolution (up to 16×16 pixels depending on image complexity), suggesting that we might use the same diagnostic visual cues in low- and high-resolution scenes. One recent eye-tracking study further showed that although low-resolution images attracted fewer fixations with shorter saccade length, the location of fixations on low-resolution images tended to be similar to and predictive of fixations on high-resolution images (Judd, Durand, & Torralba, 2011). On the other hand, some studies have observed that viewing of noisy images (e.g., applying masking, low- or high-pass spatial frequency filters to different image regions) was associated with shorter saccade amplitudes and longer fixation durations (Loschky & McConkie, 2002; Mannan, Ruddock, & Wooding, 1995; Nuthmann, 2013; Pomplun, Reingold, & Shen, 2001; van Diepen & d'Ydewalle, 2003), indicating human fixation distribution in image viewing may change with image noise.

These findings are potentially very significant to refine models of visual attention in scene perception. However, the generalisation of them is limited by methodological issues such as use of a narrow range of scenes (different categories of natural scenes have different scene statistics which may be subject to different impact by the same distortion type, e.g., the appearance of high spatial frequency stimuli is more affected by blur than low spatial frequency stimuli), and concentration on the manipulation of image parameters (e.g., resolution) rather than perceptually perceived image quality. It is unclear how different types and levels of image distortion would impact on perceived image quality, gaze pattern used to assess image quality, and predictive power of visual attention models. As we always assume that our brain has evolved to efficiently code and transmit information from natural surroundings, to determine what would be an efficient code in natural vision, it is essential to know how variance in image noise would affect scene saliency computation, and cognitive processes involved in sampling and encoding degraded scene information. Such research also meets strong and present interest in computer vision and signal processing to develop human-vision-inspired foveated active artificial vision systems and image/video quality assessment algorithms (e.g., Winkler, 2012) that will benefit numerous applications, such as enhancing the multimedia experience of human consumers and improving the efficiency of surveillance systems.

In this study we combined psychophysical, high-speed eye-tracking and computational approaches to investigate how different image distortions affected our gaze behaviour in assessing the perceived image qualities and the predictive power of computational saliency models. In the eye-tracking experiment, we applied a Gaussian low-pass filter, circular averaging filter and additive Gaussian white noise to systematically distort both man-made and natural landscape scenes, and recorded participants' gaze patterns in evaluating the perceived quality of the distorted images. In the following computational experiment, we applied various state-of-the-art computational models of visual attention, such as Judd model (Judd et al., 2009), Erdem model (Erdem & Erdem, 2013), Graph-based visual saliency model (Harel, Koch, & Perona, 2007) and Adaptive whitening saliency model (Garcia-Diaz, Fdez-Vidal, et al., 2012; Garcia-Diaz, Leborán, et al., 2012), to these natural images of varying distortion, and systematically compared their performance in predicting human gaze allocation in viewing of degraded images.

2. Experiment 1: Eye-tracking study

2.1. Methods

Twenty-four undergraduate students (16 female, 8 male), age ranging from 18 to 25 years old with the mean of 20.67 ± 2.48 (Mean ± SEM), volunteered to participate in this study. All participants had normal or corrected-to-normal visual acuity, and normal colour vision (checked with Ishihara's Tests for Colour Deficiency, 24 Plates Edition). The Ethical Committee in School of Psychology, University of Lincoln approved this study. Written informed consent was obtained from each participant, and all procedures complied with the British Psychological Society Code of Ethics and Conduct and with the World Medical Association Helsinki Declaration as revised in October 2008.

Digitised colour scene images were presented through a ViSaGe graphics system (Cambridge Research Systems, UK) and displayed on a non-interlaced gamma-corrected colour monitor (30 cd/m^2 background luminance, 100 Hz frame rate, Mitsubishi Diamond Pro 2070SB) with the resolution of 1024×768 pixels. At a viewing distance of 57 cm, the monitor subtended a visual angle of $40 \times 30^\circ$.

10 man-made scenes and 10 natural landscape scenes were sampled from the author's collection based on the DynTex database (Péteri, Fazekas, & Huiskes, 2010) (Fig. 1). The original high