British Machine Vision Association, Applied Vision Association, and Society for Pattern Recognition
Image Features & Statistics Symposium 2005

The role of natural image statistics in visual memory and recognition

Ian van der Linde[†], Umesh Rajashekar[‡], Lawrence K. Cormack[*], Alan C. Bovik[‡]
[†] Department of Computing, APU, Essex CM1 1JJ (UK) {i.v.d.linde@apu.ac.uk}
[‡] Department of Electrical and Computer Engineering, University of Texas at Austin, Austin TX 78712 (USA)
[*] Department of Psychology, University of Texas at Austin, Austin TX 78712 (USA)

Abstract

Recent years have seen a resurgent interest in eye movements during natural scene viewing. Aspects of eye movements that are driven by low-level image properties are of particular interest due to their applicability to biologically motivated artificial vision and surveillance systems. In this paper, we report an experiment in which we recorded observers' eye movements while they viewed calibrated greyscale images of natural scenes. Immediately after viewing each image, observers were shown a test patch and asked to indicate if they believed it was part of the image they had just seen. The test patch was either randomly selected from a different image from the same database or, unbeknownst to the observer, selected from one of the fixations on the image just viewed. We find that several low-level image properties differed significantly relative to the observers' ability to successfully designate test patches. The goal of the paper was to measure the image properties that facilitate visual memory in a non-cognitive natural setting to test the theory that VSTM and VLTM are implemented iconically rather than as encoded sensory information. Results indicate that the VSTM/VLTM model we tested may need to be re-evaluated since memory for the non-cognitive patches used was shown to survive eye movements.

1. Introduction

Spatial contrast sensitivity varies considerably across the human field of view, being high only at the fovea and dropping rapidly as the visual periphery is approached. This variation in sensitivity occurs as a result of the changing density, type, and interconnection of retinal cells, and necessitates ballistic eye movements to bring areas of interest into incidence with the fovea sequentially as visual attention is shifted. Saccades are made approximately 3 times every second to re-orient the high-acuity fovea at the central retina onto regions of interest in the field of fixation. During the fixations between saccades, eye orientation is relatively static for a period of 200-300ms, and detailed information from the scene may be garnered. Saccades do not rely on continuous visual feedback, but are pre-programmed and engaged during the final 150-175ms of the previous fixation using some selection criteria to determine the most salient target [1]. Internal muscular feedback conveying the eye-globe orientation is the trigger to halt the oculomotor system when the saccade has oriented the visual axis at/near the desired target, with saccades rarely being pre-empted in mid-flight with little or no new visual information being acquired during the saccade, a phenomenon known saccadic suppression.

By analysing image characteristics at fixation coordinates, researchers have postulated some locally conspicuous features that attracted the observer to initiate a saccade and demonstrated that fixation patches bear significantly different properties when compared with patches selected at random from the same image [2]. A matrix of scalar values representing the conspicuity of each image location, commonly referred to as a saliency map, may be constructed. This may be used to predict where an observer will most likely place fixations on a given input image. This has applications for machine vision, auto-foveated video compression, and to improve our ecological understanding of the human visual system.

Four forms of visual memory are generally considered to exist to support the human visual system (HVS): visible persistence, informational persistence, visual short-term memory (VSTM), and visual long-term memory (VLTM). Of these, only VSTM and VLTM are considered to survive eye-movements, and. are believed to operate by abstract object representation rather than as encoded sensory information [8]. In alternative theories, no visual memory for natural scenes is believed to exist, with the scene itself acting as an external memory.

Irrespective of the memory model employed, researchers have shown that information at the point of fixation is preferentially retained in memory compared with image regions that have only been sensed non-foveally [3], despite our ability to attend to one area of the visual field while fixating at another. Existing work has measured the ability of observers to memorise objects in complex scenes, or to memorise symbols in a controlled synthetic environment. This paper extends this work to investigate visual memory performance in natural scenes containing few semantically interesting features. It is theorised that the HVS is specifically optimised for viewing natural images rather than any synthetically generated stimuli or man-made environment. The set of natural images is a tiny subset of all possible images exhibiting some interesting properties, such as their scale invariance, preponderance of horizontal and vertical edges, and smooth harmonics. Natural images are preferable for the study of pre-attentive and low-level/bottom-up visual phenomena, having been the stimuli driving the evolution of the HVS. This study combined eye tracking with a simple visual memory/recognition task.

2. Method

100 static diurnal images were manually selected from a calibrated greyscale natural image database [4], finding and omitting images containing man-made structures and instinctively attractive features such as animals, faces, and other high-level semantic content. Examples are shown in Fig. 1. Typical images contained natural habitats of trees, grasses, and water. Image linearity was maintained, but brightness

increased such that the brightest point in the image corresponded to the brightest output level of the monitor. The 100 images satisfying the selection criteria were verified by a second reviewer prior to use to ensure the complete absence of cognitive/man-made features, with replacements for rejected images obtained from the same database where necessary.

An SRI Generation V Dual Purkinje eye tracker was used to gather subject fixation coordinates, set to a sampling rate of 200Hz. Bite-bars to keep the maxilla stationary were created for each human subject using a standard dental compound adhered to a mandible sized aluminium frame. Two forehead rests were employed to provide tactile feedback to discourage fore-aft head movements. Image patches harvested for subsequent use in the visual memory task were set to 64×64 pixels, slightly over 1º. This size of the patch corresponds to the size of the high-acuity cone-dominated foveola at the centre of the human fovea.

Fig. 1 Sample Images with Overlaid Fixations

The main experiment consisting of 100 trials per subject was preceded by a dry-run session of 10 trials in order that the subject became familiar with the handheld control box and comfortable in the experimental environment prior to data collection. Stimuli were presented using software written in Matlab 6 using the Psychophysics Toolbox extensions [5][6]. Images were shown in a fixed order for all subjects. Following the presentation of the full-screen stimulus image for 5s, the task required the subject to view a small (64×64, foveola-sized) image patch and respond via the control box to indicate whether they believed the patch was present in the image they had just viewed. To remove the possibility of luminance matching, the patches were brightness jittered by a random amount up or down. The subjects were made aware of this during the dry run session. Subjects were advised that they should free-view the image to take in as much as possible within the 5s display time. To remove the possibility of probability matching, the patch had a 50% probability of having come from the image just viewed, and a 50% probability of coming from another image from the same database.

3. Discussion

On average, image patches corresponding to hit cases showed higher entropy, mean luminance, MSE, SD, number of edges, and unique colours than miss cases. These measures correspond to information content, and theories of eye movements conjecture that regions exhibiting higher information are likely to attract fixations to elucidate. Since these image dimensions may have contributed to regional saliency, it may be postulated that more salient regions are easier to recognise and more memorable, with the caveat to this being that higher contrast (measured both by RMS and Michelson's technique) which is generally considered to contribute to making image regions highly salient (attracting eye movements), did not appear to increase the memorability of the patch. For the eight image statistics applied to the patches we found that those that were recognised scored more highly in 6 of the 8 image statistics than unrecognised patches, with the exception of the two contrast metrics. Details of these results are to be published in a forthcoming paper [7]. We propose that a memorability map of a complex natural scene may be constructed to represent the low-level memorability of local regions in a similar fashion to the familiar saliency map, which records bottom-up fixation attractors.

References

[1] Privitera, C. M. & Stark, L. W. (2000) Algorithms for Defining Visual Regions-of-Interest: Comparisons with Eye Fixations, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 9. 970-982.
[2] Rajashekar, U., Cormack, L.K., Bovik, A.C. (2002), Point of gaze analysis reveals visual search strategies, Human Vision and Electronic Imaging IX, Proc. of SPIE, Vol. 5292, San Jose, CA, January 18-22, 2004
[3] Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. Journal of Experimental Psychology: Human Perception and Performance, 28, 113-136.
[4] van Hateren, J.H. & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. Proc. Royal Society of London B 265:359-366.
[5] Brainard, D.H. (1997). The Psychophysics Toolbox, Spatial Vision 10:433-436.
[6] Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies, Spatial Vision 10:437-442.
[7] van der Linde, I. Rajashaker, U. Cormack, L.K., Bovik, A.C. (2005) A study of human recognition rates for foveola-sized image patches selected from initial and final fixations on calibrated natural images, SPIE Human Vision & Electronic Imaging X.
[8] Hollingworth, A. (2004) Constructing Visual Representations of Natural Scenes: The Roles o f Short- and Long-term Visual Memory, Journal of Experimental Psychology: Human Perception and Performance. Vol. 30 No. 3, 519-537.