Cogn Comput (2011) 3:1–3
DOI 10.1007/s12559-011-9096-1

Saliency, Attention, Active Visual Search, and Picture Scanning

John G. Taylor · Vassilis Cutsuridis

Published online: 24 February 2011

How is a complex visual scene processed? How is the selection of one particular location in a visual scene accomplished? Does it involve bottom-up, sensory-driven cues or top-down world knowledge expectations or both? How the decision is made *when* to terminate a fixation and move the gaze? How the decision is made *where* to direct the gaze in order to take the next sample? The goal of the special issue is to attempt to answer these questions by advancing our understanding of the state-of-the-art on bottom-up and top-down approaches to active visual search and picture scanning. The manner in which attention is involved was considered a highly relevant topic to the special issue.

The special issue starts with a paper by Mital, Smith, Hill, and Henderson who investigated the contribution of low- to mid-level visual features to gaze location during free viewing of a large dataset of videos ranging in content and length. Their results showed that temporal features such as flicker, motion, and their respective contrasts were the most predictive of gaze location. Motion and mid-level visual features may influence gaze allocation in dynamic scenes, but remain unclear whether this influence is involuntary or due to correlations with higher-order factors such as scene semantics.

Fuchs, Ansorge, Redies, and Leder investigated experimentally whether visual salience attracts attention in a bottom-up manner. Their results from participants viewing abstract and depictive paintings as well as photographs in free viewing and target-search experimental setups showed that local feature contrasts in color, luminance, and orientation drive attention in all conditions that a characteristic short-lived temporal profile of the salience-driven effect on fixations exists.

Urban, Follet, Chamaret, Meur, and Baccino investigated the importance of spatial scale to predict visual attention. They proposed a fast attentional model and studied which frequency band predicts the best fixation locations during free viewing task. They found that medium frequencies (0.7–1.3 cycles per degree) globally allowed the best prediction of attention, with variability among categories. Fixation locations were found to be more predictable using medium to high frequencies in man-made street scenes and low to medium frequencies in natural landscape scenes.

Foulsham and Underwood tested the hypothesis of whether visual saliency can be overridden in a search task with saccades being planned on the basis of target features, rather than being captured by saliency. Their results from three experimental conditions seem to indicate that saliency effects on search may have nothing to do with bottom-up saccade guidance.

Humphrey and Underwood explored whether scanpath similarity could be improved by verbalising one's memory of the picture and whether the previously observed domain-specific advantage was still present when no bottom-up information was available. Results showed recognition accuracy to be increased by post-stimulus verbalisation, and experts demonstrated an advantage for stimuli that contained domain-relevant information. Saliency influenced both verbal feedback and eye movements, but was moderated by domain expertise. Scanpaths were more

J. G. Taylor (✉)
Department of Mathematics, King's College,
London WC2R 2LS, UK
e-mail: taymore2002@aol.co.uk

V. Cutsuridis (✉)
Center for Memory and Brain,
Boston University, Boston, MA 02215, USA
e-mail: vcut@bu.edu

similar when pictures were described compared to when imagined, and specialists produced more similar scanpaths when describing domain-specific pictures, compared to control pictures and control participants.

Vig, Dorr, Martinetz, and Barth investigated the predictive nature of eye movements when viewing different types of natural or realistic scene sequences. Cross-correlation analysis of image-based saliency maps with an empirical saliency measure derived from eye movement data revealed the existence of predictive mechanisms responsible for a near-zero average lag between dynamic changes of the environment and the responding eye movements. They also showed that the degree of anticipation is reduced when moving away from natural scenes by introducing camera motion, jump cuts, and film editing.

Yates and Stafford addressed the issue of whether the shape of the saccadic latency–eccentricity function is a by-product of diminished stimulus representation in the periphery. Using a target-elicited saccade experimental paradigm, ten subjects were shown M-scaled and unscaled Gaussian targets over a horizontal range of $\pm 40°$. Saccadic latencies increased at an equal rate for peripheral targets regardless of whether targets were M-scaled or not, suggesting that the changes of latency with eccentricity are not a by-product of resources devoted to stimulus representation, but instead are a functional adaptation, which takes account of the likelihood of saccades of each amplitude in the natural environment.

Yanulevskaya, Geusebroek, Marsman, and Cornelissen attempted to characterize the contrast distribution of natural images with a two-parameter Weibull distribution. Their results show that as few as two values per image region are already enough to achieve a performance comparable with the state-of-the-art in bottom-up saliency prediction.

Jacob and Hochstein presented a mathematical model which provided evidence that the information available to the observer at any given moment when observing a visual scene depends on previous fixations and the eye movement scanpath. Results showed that the amount of available information grows and fluctuates stochastically over time, so that the larger the scene, the less information is available, on average, for each scene part.

Rebhan and Eggert present a system architecture of task-dependent representation of a visual scene with a control mechanism that decides which properties need to be extracted and how the independent processing modules should be combined, based on the knowledge stored in the system's long-term memory. They evaluate their system in a real-world table scene and show that the amount of data processed and stored by the system is considerably lower compared with processing regimes used in other systems.

Gepperth, Rebhan, Hasler, and Fritsch focused on the question of how the autonomous learning of invariant models can be embedded into a performing system and how such models can be used to define object-specific attentional modulation signals. To answer this question, they developed a large-scale hierarchical system for object detection fusing bottom-up (signal-driven) processing results with top-down (model or task-driven) attentional modulation. They then tested the performance of their system on approximately 3,500 annotated video images from more than 1 h of driving.

Rapantzikos, Avrithis, and Kollias propose a Gestalt law-inspired method to compute visual saliency from video sequences, where the visual input is represented by a volume in space–time and decomposed into a set of feature volumes in multiple resolutions. Feature competition is used to produce a saliency distribution of the input implemented by constrained minimization. The authors also provided comparisons of their method against already established models on public datasets as well as data from three experiments of their own.

Heinke and Backhaus applied the Selective Attention for Identification model (SAIM) to visual search (VS) experiments involving simple lines and letters. Simulation results for asymmetric and symmetric searches and for the influence of background line orientations suggested that a large part of the empirical evidence can be explained by competitive processes in the brain, which are modulated by the similarity between the target and the distractor as well as by the featural properties of distractors. They concluded that search asymmetries may be the outcome of interactions between top-down (knowledge about search items) and bottom-up (feature of search items) processing. This interaction in their VS-SAIM is dominated by a novel mechanism, the knowledge-based on-center-off-surround receptive field, where the exact shape of the receptive field is modulated by both, top-down and bottom-up processes.

Dempere-Marco, Hu, and Yang presented a conceptual framework able to acquire expert knowledge from eye-tracking data of skilled individuals solving complex visual tasks. They argued that a feature space is a domain in which each dimension is defined as a mathematical construct, which may correspond to perceptually meaningful visual cues and which can take either numerical or categorical values. Similarities between visual search patterns are essential to characterize the stereotypical visual behavior of observers and thus, expert knowledge. Since knowledge representation is closely related to the feature domain in which the search is analyzed, feature relevance measures become central to knowledge gathering. Finally, they demonstrated a proof of concept of their proposed framework by applying it to expert knowledge gathering in lung radiology.

Kootstra, de Boer, and Schomaker proposed a saliency model that predicts eye fixations on mirror symmetric forms. Experimental results from an eye-tracking paradigm with participants viewing complex photographic images showed that their symmetry model predicted human eye fixations significantly better on a wide variety of images including many that are not selected for their symmetrical content than the Itti and Koch model. Additionally, their results showed that early fixations are on highly symmetrical areas of the images. They concluded that symmetry is a strong predictor of human eye fixations and that it can be used as a predictor of the order of fixation.

Yubing, Cheikh, Guraya, Konik, and Trémeau investigated saliency and attention in video sequences and propose a novel spatiotemporal saliency model devoted for video surveillance applications. The video saliency maps computed from their model were compared with the gaze maps obtained from subjective experiments with SMI eye tracker for surveillance video sequences. The results show strong correlation between the output of the proposed spatiotemporal saliency model and the experimental gaze maps.

de Croon, Postma, and van den Herik proposed a novel gaze control model (ACT-DETECT) that uses the information from local image samples in order to shift its gaze toward object locations. The model was evaluated on a face-detection task using a publicly available image set.

Fix, Rougier, and Alexandre presented a dynamic field theory of visual exploration of a scene where selection of the next locus of attention, memorizing the previously attended locations, anticipating the consequences of eye movements, and integrating bottom-up and top-down information are the necessary mechanisms to perform a active visual search task.

Wasserman, Bolbecker, Li, and Lim-Kessler proposed a new theory of visual attention based on past research in horseshoe crab, where a top-down component involving efferents that go from the brain to the photoreceptors and a bottom-up component involving consequent neuromodulatory alterations of the timing of the afferent photoreceptor potentials evoked by light stimuli interact to shift attention to relevant regions in the visual field. The validity of their theory was subsequently tested by the newly developed multifocal electroretinogram (mfERG) technology, which provides objective measures of the timing of human retinal potentials while obtaining quantitative behavioral measures of both the observer's state of attention and of visual performance. Finally, suggestions were made for the applicability of their theory and the experimental technique in the objective diagnosis and the quantitative evaluation of treatments in a variety of attention disorders.

Fellenz proposed a neurobiological model for object-based visual attention of the ventral visual pathway. The attentional system comprised of the pulvinar nucleus for attentional selection, the superior colliculus (SC) for the selection of a saccade target, the frontal eye field (FEF) as a working memory map for inhibition of return (IOR), and the posterior parietal cortex (PPC) as a spatial map for representing locations in a head-centered coordinate frame. The model was applied to covert attentive picture perception.

Wichert presented a computational model of the mechanism of retrieval of associations from associative memory during visual scene analysis. In the model, the retrieval phase of the associative memory was divided into two stages: the attention stage and the binding stage. In the attention stage, an attention window selected patterns representing objects for further access. In the binding stage, the selected patterns formed an address vector. The behavior of the model was supported by theoretical analysis and empirical experiment.

Finally, Neokleous, Avraamides, Neocleous, and Schizas presented a neurocomputational model of visual selective attention aimed to investigate much debated connections of attention and consciousness.

We hope this special issue will inspire readers to actively research visual salience, attention, active visual search, and picture scanning using both experimental and computational techniques. As it was clearly shown in this special issue, a multidisciplinary approach is needed to decipher how low-level processing of visual features and high-level cognitive processing can interact in gaze control (see also [1–3]). We would like to thank all authors who have provided us with exciting and forefront contributions to this special issue.

References

1. Cutsuridis V. A cognitive model of saliency, overt attention and picture scanning. Cognit Comput. 2009;1:292–9.
2. Findlay JM, Gilchrist I. Active vision: the psychology of looking and seeing. Oxford: Oxford University Press; 2003.
3. Taylor JG, Hartley M, Taylor N, Panchev C, Kasderidis S. A hierarchical attention-based neural network architecture, based on human brain guidance, for perception, conceptualisation, action and reasoning. Image Vis Comput. 2009;27(11):1641–57.