Perceptive Machines: From Selective Attention and Recognition to Visual Cognition

Deepak Khosla and David J. Huber

HRL Laboratories, LLC
3011 Malibu Canyon Road, Malibu, CA-90265, USA
DKhosla@hrl.com, DJHuber@hrl.com

Abstract

While the task of sensing and perceiving the visual environment as we go about our daily lives is trivial for most humans, attempts to emulate the principles underlying human vision in machine vision systems have only been marginally successful. Attention, mediated by eye movements, acts as the critical gateway to visual cognition by searching for areas with relevant information and selecting the stimuli that will be processed. These stimuli are subsequently recognized as parts of visual world and stitched in spatial and temporal representations eventually leading us to produce an adaptive, composite impression of surroundings in near real-time. In this talk, we will outline recent efforts from our group to translate the knowledge of human vision system into computational models towards realizing intelligent seeing machines. We will describe specific computational models for visual attention, object recognition and spatio-temporal recognition, and present some preliminary results using these models on public domain computer vision datasets.

Keywords: Vision, attention, saliency, object recognition, object segmentation, bio-inspired, perception, intelligent machines, neuro-inspired, spatio-temporal recognition, multi-sensory perception, active control.

Introduction

Humans can analyze a scene very quickly and easily notice objects, even those that have never been seen before. We use our built-in attention and object segmentation systems without a second thought. Likewise, recognizing objects and learning new objects is relatively easy for us. Most current vision models have focused on the passive human visual ability to detect and recognize objects where the camera is supposed to take in the whole scene, attempting to make sense of all that it sees. However, we live a world which is more dynamic and interactive than an art gallery. We do not interact visually with the world by passively absorbing the information like a fixed camera, but rather by actively interacting with the environment and culling information selectively. Computationally, however, paying attention to a scene, extracting regions and objects of interest, and recognizing them has proven to be a great challenge. Because human performance far exceeds that of the best machine vision systems to date, building an artificial system inspired by the principles underlying human vision has been an attractive idea since the field of computer vision was conceived. However, most of the bio-inspired systems only incorporate one aspect of vision, have not been robustly tested on real-world image datasets, and/or are not suited for real-time applications.

Object-Based Attention

We will start by describing a method for object-based saliency that computes attention for a natural scene, attends to the salient objects in the scene, and segments these objects from the background. A vision system must be able to efficiently determine which locations in a scene draw the most attention and then segment the attended object so that it can be identified or interpreted. The segmentation algorithm must be able to cope with scenes containing color gradients caused by different object textures and lighting conditions so that the integrity of an object boundary is maintained. Furthermore, the segmentation boundary must be close enough to the object to minimize background noise that might lead to the misclassification of the object. It should also be able to adjust between scales, to extract large objects or textures and then "zoom in" to examine the constituent parts of the object. The attention algorithm described here addresses each of these problems. The algorithm can work in an unguided bottom-up mode or a biased top-down mode in which the system searches for a specific object. The segmentation step employs a unique region growing algorithm that ignores gradients in color and intensity caused by variations in object texture and lighting. We demonstrate the utility of our system using images from the COIL-100 and CSCLAB data sets, as well as aerial and satellite imagery.

Copyright © 2008, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Object Recognition

We will then describe a neuro-inspired vision system that can (1) learn representations of objects that are invariant to scale, position and orientation; and (2) recognize and locate these objects in static and video imagery. The system uses modularized neuro-inspired algorithms/ techniques that can be applied towards finding salient objects in a scene, recognizing those objects, and prompting the user for additional information to facilitate online learning. The neuro-inspired algorithms are based on models of human visual attention, search, recognition and learning. Most machine vision systems rely on techniques that mandate offline learning and cannot learn new classes without retraining on all prior knowledge. This is computationally inefficient and prohibits interactive learning. The proposed system can be trained in an online manner, meaning that new classes and instances of classes can be learned by the system without retraining on all prior knowledge. The implementation is highly modular, and the modules can be used either as a complete system or independently. The underlying technologies were carefully researched in order to ensure they were robust, fast, and could be integrated into an online system. We evaluated our system's capabilities on the Caltech 101 and COIL 100 datasets, which are commonly used in machine vision, as well as on simulated scenes. Preliminary results are quite promising in that our system is able to process these datasets with good accuracy and low computational times.

Spatio-Temporal Recognition

We will describe a bio-inspired system for spatio-temporal recognition in static and video imagery. An attention and object recognition system is used to determine the location and identities of objects in a scene at a particular time. A working memory model is continuously updated with object location and label data; it uses this data to construct and maintain ordered spatio-temporal sequences. A bio-inspired classifier ascribes event labels to these sequences. A separate network, executed in parallel with the former, classifies any significant spatial relations it observes among the recognized objects. None of the algorithms in our system require offline training, and our system can acquire its entire knowledge base via interactive online learning if desired. The system attempts to classify all events that it observes. If it is not confident in its classification, then it will query the user for the event's correct label and immediately acquire the new knowledge. Unknown spatial relationships can also be interactively learned. Our event recognition system is robust to variations in an object's motion profile. We evaluated the performance of our system on real world video footage of vehicles and pedestrians in a busy street; our system is able to recognize the events in this footage involving vehicles and pedestrians. This system was also used to learn the spatial relationships that compose a particular scene category in static imagery.

References

Khosla, D, Moore, C., and Huber, D. (2007). Bio-inspired visual attention and object recognition. *Proceedings of SPIE,* 6560: 656003.

Khosla, D., Moore, C., and Chelian, S. (2007). A Bio-inspired system for spatio-temporal recognition in static and video imagery. *Proceedings of SPIE,* 6560: 656002.