# A Biologically Inspired Focus of Attention Model

Master's Thesis Proposal

Daniel I. Harris
dih0658@gmail.com
Computer Science Department
Rochester Institute of Technology

September 14, 2007

Committee Members:
Chair: Dr. Roger Gaborski
Reader: Dr. Carl Reynolds
Observer: Thomas Borrelli

**Abstract**

A problem afflicting modern object recognition systems relates to the size of the input in which they recognize objects. Given a real world scene taken from a digital camera, a target object to be recognized may be very small compared to the entire input image and be surrounded by a large amounts of clutter or unnecessary data. Clutter and extra data needs to be removed so that object recognition systems can perform proficiently and efficiently. To remove clutter, a focus of attention model is developed which will assist complex object recognition systems in finding small, interesting, sub-regions of a large input which may need to be processed in further detail.

# 1 Introduction

As computer vision continues to be a highly researched topic in computer science, object recognition has remained one of the most studied aspects of computer vision. Many recognition systems have become extremely proficient at recognizing a varying number of objects that the system has been previously trained to identify. However, one of the common problems of these recognition systems is that the scene which contains a target object to be identified, needs to be relatively simple or uncluttered to reduce error and processing time when performing the recognition task. The inability of object recognition systems to function both proficiently and efficiently on complex and cluttered scenes is a primary concern that must be addressed before such systems can be put to use in a broad range of tasks.

Since one of the goals of computer vision is to build computational models which perform the recognition task as well or better than humans, researchers have began looking at the underlying biology of the human recognition task to solve the cluttered scene problem. A current theory to explain human object recognition describes that the recognition task takes place in two distinct stages: a pre-attentive stage and an attentive stage of recognition [4]. Object recognition models comprise the attentive, or second, stage of recognition in which complex processing is performed for the purpose of identifying or recognizing an unknown objects in a scene. As shown by results from such systems [5], object recognition is usually computationally expensive and does not always perform optimally on large and cluttered scenes. The pre-attentive stage of recognition has been described as a stage in which many simple features are quickly computed and compared in parallel over an entire input image for the purpose determining a much smaller focus of attention that a more computationally expensive second stage of recognition can process.

Psychological research has been the basis of this two stage model of recognition. Anne Treisman, a psychologist from Princeton University, has done extensive research into the field of visual search. Through experimentation, Treisman developed the idea that human attention functions like a "spotlight", illuminating different areas of the visual field for further, more intense, processing. Treisman has also shown that this spotlight of attention does not simply "sweep" across the entire visual field processing everything iteratively, but instead jumps around, focusing on the most conspicuous areas first. In 1980, Treisman published a paper along with Gelade describing a feature integration theory of attention [6]. The premise behind the feature integration theory of attention is that individual features (contrast, color, etc.) are processed separately and then later combined by the brain in order to find the most conspicuous area in regards to the combination of all detected individual features. Treisman's attentional spotlight would then jump between the most conspicuous areas for further processing. Another aspect which Treisman has researched relates to perceptual grouping during the pre-attentive stages of recognition. Treisman, as well as others, have found that features are grouped in the pre-attentive stage of recognition and effect focus of attention during a search task [7]. Perceptual

grouping has been found to follow many Gestalt grouping principles, specifically similarity (color, orientation, etc.) and proximity. It is believed that attention during visual search shifts between objects in similar perceptual groups before shifting to a conspicuous object in a different perceptual group.

As described earlier, many models already exist which are responsible for the second stage of the recognition task. To complement these systems, a model describing the pre-attentive stage of recognition is needed which simulates both Treisman's feature integration theory as well as the attentional spotlight and shifting mechanism. Such a system could then be used to determine an area of an input image which requires further processing by a more complex recognition system.

## 2 Previous Work

The task of creating a computer model simulating Treisman's feature integration theory of attention was described by Koch and Ullman in their paper "Shifts in selective visual attention: towards the underlying neural circuitry" [4]. The model that Koch and Ullman describe to find the conspicuous (salient) regions of an image uses a series of filters to process a color input image into a set of individual feature maps. Each feature map is created by processing the input image with a single feature filter such as a contrast or orientation filter. Koch and Ullman also describe how these feature maps should then be combined into a single "saliency map" which describes how conspicuous different regions of the input image are. Finally, Koch and Ullman's research proposes a winner-take-all competitive network to determine which area of the image is most conspicuous. From the initial description proposed by Koch and Ullman a number of researchers have implemented similar saliency models. Itti and Koch have created biologically plausible models by using biologically based Difference of Gaussian and Gabor filters to extract features such as contrast and orientation [3]. Red/Green and Blue/Yellow color differences were also used to encode color saliency since research has suggested that the human brain processes color in a similar fashion. Itti and Koch have also been responsible for researching different ways, some biologically plausible, of combining the many different feature maps into a final saliency map [2].

Many saliency models simply describe still-saliency, or saliency for a single image or instance in time. However, research into visual attention suggests that motion is another of the basic features that can be used to determine saliency. Young and Lesperance have created a series of biologically plausible filters to detect motion by adding the concept of time to a simple Gabor filter [8]. The underlying idea behind their research is to combine the derivatives of three Gaussian functions together which then creates a three-dimensional filter similar to a Gabor filter but with an extra dimension which accounts for shifts in time. By changing the number of derivatives taken and the parameters of each individual Gaussian function the filter can be changed to account for motion in a specific direction, or speed. The "Video Exploitation and Novelty

Understanding in Scenes" (VENUS) system developed at the Rochester Institute of Technology uses a saliency model which incorporates the research of Koch and Ullman along with the research conducted by Young and Lesperance to create a saliency model which utilizes biologically inspired motion filters to create motion feature maps for use in the final saliency map creation. The saliency model created for the VENUS system is used to extract regions for the novelty detection module which detects and catalogs novel events in video [1].

# 3 Goals of the Research

The primary goal of this research is to replicate the VENUS saliency model and enhance it further by developing a focus of attention shifting mechanism. This new model be able to process a video and locate the most salient regions in the input but also determine the current focus of attention at an instant in time as well as the shifts in attention as the video progresses. The information gained from this model will allow an object recognition model to process the more interesting regions of an input first, avoiding the clutter problem and increasing overall processing speed.

In order to accomplish this goal, a biologically plausible still saliency model using Difference of Gaussian and Gabor filters will be created. In addition to the basic still saliency model, Young and Lesperance's spatio-temporal motion filters will be implemented and used to create a motion saliency model which accounts for directional motion. These models will then be combined to create a motion-sensitive saliency model as used in the VENUS system. By utilizing one of the feature map combination strategies proposed by Itti and Koch [2] to avoid feature wash out, the resulting saliency maps produced by model will be more accurate than those produced by the VENUS system.

Once the VENUS saliency model has been duplicated and enhanced, an attention shifting model will be added as a post-processing step to the video processing system. The attention shifting model will account for what the model has previously determined interested and inhibit such regions so that the focus of attention will shift to a new, highly salient, region over a short period of time. In addition to inhibiting regions which have already been focused on, the focus of attention model will also assist in determining the location to focus attention on. Instead of just focusing on the next highest region of salience, the focus of attention model will implement proximity and feature similarity preferences to make the shifts in attention conform to similar perceptual groups as shown in the research conducted by Treisman.

# 4 Software Requirements

The model being implemented for this research will be created in the MATLAB computing environment created by MathWorks. MATLAB was chosen as the primary computing environment for this research due to the number of toolboxes

available to assist with basic computer vision and video processing functions. Any computer system capable of running the MATLAB computing environment will be able to use the focus of attention model developed for this research. It should be noted, however, that the MATLAB environment running on a UNIX platform is not capable of reading compressed video files. In order to process a video file on a UNIX computer system (or variants), uncompressed video is the only type of video that can properly be decoded.

Video creation software is also required to fully utilize the model being developed. Video input files must to be created as input for the model. The output from the model is a series of numbered images (one image per frame of the input video), which can then be compressed into an output video using a software package such as VirtualDub (www.virtualdub.org).

# 5 Deliverables

The deliverables for this research will include:

- Source Code
  - The still saliency model
  - The motion saliency model
  - Focus of attention model
  - Video stream process model (incorporates the above 3 components)
- Documentation
  - Research proposal
  - Final report
  - Presentation slides
  - Usage instructions
- Data
  - Input data (both still images and videos) for final model
  - Output data described in final report
  - Intermediate input and results from the in-progress model and individual model components

# 6 Milestones

- September
  - Thesis proposal approved
  - Still saliency model complete

- October
  - Motion saliency model complete
  - Integration of still and motion saliency models into video stream processing model
- November
  - Basic focus of attention shifting added into video stream processing model
  - Feature similarity shifting added to focus of attention model
- December
  - Testing and recording of sample video data for use with an existing object recognition system
  - Final report complete
- January
  - Defend final report

# References


[1] Gaborski, R. "VENUS: A System for Novelty Detection in Video Streams with Learning" Department of Computer Science, Rochester Institute of Technology (2004)

[2] Itti, L. and Koch, C. "Feature combination strategies for saliency-based visual attention systems" Journal of Electronic Imaging 10: 161-169 (2001)

[3] Itti, L. and Koch, C. "Target detection using saliency-based attention" Computation and Neural Systems Program, California Institute of Technology (1999)

[4] Koch, C. and Ullman, S. "Shifts in selective visual attention: towards the underlying neural circuitry" Human Neurobiology 4: 219-227 (1985)

[5] Mutch, J. and Lowe, D. "Multiclass Object Recognition with Spare, Localized Features" Department of Computer Science, University of British Columbia (2006)

[6] Treisman, A. Gelade, G. "A feature-integration theory of attention" Cognitive Phychology 12: 97-136 (1980)

[7] Treisman, A. "Perceptual Grouping and Attention in Visual Search for Features and For Objects" Journal of Experimental Psychology: Human Perception and Performance Vol. 8 No.2 194-214 (1982)

[8] Young, R.A. and Lesperance, R.M. "The Gaussian Derivative model for spatial-temporal vision: I. Cortical Model" Spatial Vision, Vol. 14 No. 3,4 261-319 (2001)