2001

Using human observers' eye movements in automatic image classifiers

Timothy Grabowski

Follow this and additional works at: http://scholarworks.rit.edu/theses

Recommended Citation


This Thesis is brought to you for free and open access by the Thesis/Dissertation Collections at RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact ritscholarworks@rit.edu.
Using Human Observers' Eye Movements in Automatic Image Classifiers

Timothy Grabowski

Resume

Center for Imaging Science

Rochester Institute of Technology

May 2001

In association with: Alejandro Jaimes, Jeff Pelz, Jason Babcock, and Shih-Fu Chang (with acknowledgment to Diane Kucharczyk and Amy Silver)

Table of Contents
Using Human Observers' Eye Movements in Automatic Image Classifiers

Timothy Grabowski

Table of Contents

Abstract

Copyright

Acknowledgements

THESIS

Introduction

Background

Experiment

Results

Conclusion

References

Additional Topics of Interest

Resume

Industrial Associates PowerPoint Slide Presentation
Using Human Observers' Eye Movements in Automatic Image Classifiers

Timothy Grabowski

This research thesis explored the way in which people look at images in different semantic categories (e.g., handshake versus not-handshake), and directly related those results to computational approaches for automatic image classification. Although many eye tracking experiments have been performed, to our knowledge, this was the first study that specifically compares eye movements across categories, and that links category-specific eye tracking results to automatic image classification techniques. The hypothesis was that the eye movements of human observers for images in different semantic classes differ (e.g., handshakes, others), and that this information can be effectively used in automatic techniques. We presented eye-tracking experiments that showed the variations in eye movements (i.e., fixations and saccades) across different individuals for images in different categories. Then, we examined how empirical data obtained from eye tracking studies like this one, for a specific class; can be integrated with computational frameworks.

The experiment consisted of 5 sets of images of 50 images each. 10 subjects were tracked while viewing these 250 images and the results were analyzed to ascertain whether or not there is a correlation between different subjects’ eye movements and similar images. 6 male and 4 female subjects were tested, all of whom had normal or corrected to normal vision.

I will discuss the key ideas of the experiment: eye-tracking, vision terminology and experimental set-up. I will discuss the results of the experiment in a general sense. Alexandro Jaimes is responsible for the computational development of the software; I provided him with the raw data and performed some rudimentary analysis

Table of Contents
This work is copyrighted and may not be reproduced in whole or part without permission of the Center for Imaging Science at the Rochester Institute of Technology.

This report is accepted in partial fulfillment of the requirements of the course SIMG-503 Senior Research.
Title: Using Human Observers' Eye Movements in Automatic Image Classifiers
Author: Timothy Grabowski
Project Advisor: Jeff Pelz
SIMG 503 Instructor: Anthony Vodacek

Table of Contents
Using Human Observers' Eye Movements in Automatic Image Classifiers

Timothy Grabowski

I would like to thank my advisor Jeff Pelz, Alejandro Jaimes of Columbia University, the staff in the Vision and Psychophysics lab, and all of my friends who volunteered an hour to be subjects. Thank you!

Table of Contents
Using Human Observers' Eye Movements in Automatic Image Classifiers

Timothy Grabowski

Introduction

This research thesis explored the way in which people look at images in different semantic categories (e.g., handshake versus not-handshake), and directly related those results to computational approaches for automatic image classification. Although many eye tracking experiments have been performed, to our knowledge, this was the first study that specifically compares eye movements across categories, and that links category-specific eye tracking results to automatic image classification techniques. The hypothesis was that the eye movements of human observers for images in different semantic classes differ (e.g., handshakes, others), and that this information can be effectively used in automatic techniques. We presented eye-tracking experiments that showed the variations in eye movements (i.e., fixations and saccades) across different individuals for images in different categories. Then, we examined how empirical data obtained from eye tracking studies like this one, for a specific class; can be integrated with computational frameworks.

The experiment consisted of 5 sets of images of 50 images each. 10 subjects were tracked while viewing these 250 images and the results were analyzed to ascertain whether or not there is a correlation between different subjects’ eye movements and similar images. 6 male and 4 female subjects were tested, all of whom had normal or corrected to normal vision.

I will discuss the key ideas of the experiment: eye-tracking, vision terminology and experimental set-up. I will discuss the results of the experiment in a general sense. Alexandro Jaimes is responsible for the computational development of the software; I provided him with the raw data and performed some rudimentary analysis.

Background

Several studies have examined eye movements of individuals as they perform natural tasks (e.g., Collewijn, et al. 1992, Eppelboim et al, 1997, Pelz et al, 2000) or observe pictures (e.g., photographs in Gould 1976, art in Nodine, et al., 1993 and Solso 1994). The results from similar studies have been used by researchers in various fields in the development of theories of visual perception and recognition (e.g., Land & Furneaux, 1997, Yarbus, 1967). Some of these theories state that humans move their eyes over the most informative parts of an image (Yarbus 1967). The eye movements (i.e., fixations and saccades) are strongly influenced by the visual content of the image (Gaarder, 1975), and by the task being performed by the observer (e.g., classify the image; search for an object; Yarbus 1967).

Various computational approaches that perform automatic object recognition have drawn on theories of the functionality of the human visual system. In order to limit the amount of information to be processed, for example, some techniques detect Regions of Interest (ROIs). Regions that may be relevant to the problem at hand (e.g., face recognition) are first selected. The areas of the image, and the order of processing depend on that initial selection. Such processing has been compared to the processing humans perform: eye movements are performed to select interesting parts of an image (see Privitera & Stark 1998). In spite of the similarities between human processing and automatic techniques, most computational approaches are based on general theories that do not directly link the specific problem at hand with the information obtained from experiments involving human subjects. For example, when we observe the details of a face we move our eyes in a specific path and fixate on areas that we deem important (e.g., eyes, nose, mouth). The areas that we observe, and the order in which we make those observations depend highly on the content of the image (e.g., a face vs. a
landscape), and the task (e.g., recognize the face; determine if there is a face in the image). It is likely, however, that there may be patterns in the way different individuals look at images in the same category. Nonetheless, information on how humans perform these specific tasks is seldom included in computational approaches (e.g., regions of interest in a landscape scene may be selected with "grass" filter).

In this paper, we link an experimental study on how humans observe different classes of photographs (e.g., images in which two people are shaking hands), with a computational approach to detect objects/scenes in the same class of photographs.

**Experiment**

An ASL 504 Remote Eye-tracker made by Applied Science Laboratories is the primary piece of equipment for this experiment. Its purpose is to record the movement of the subject’s eye by using infrared light reflected from the cornea and retina.

![](image)

**ASL 504 Remote Eye-tracker**

The Remote Eye-tracker was placed below a 27" CRT. The primary benefit of this tracking device is to allow the participants to view the images freely, without any headgear to limit natural behavior. To minimize body and head movements (since the focus of the experiments was to determine the areas of the image the participants observed), the participants were asked to sit approximately 36 inches from the computer monitor, and follow the instructions relevant to the experiment. In addition to this equipment two laptop computers were employed as well. One laptop was used to store the image set in PowerPoint format. The other laptop was used for the eye-tracker program. The software was run from the DOS prompt and communicated with the microprocessor.

The final procedure of the experiment differs slightly from the original trials performed previously in order to obtain test data. The procedure used for the trials is as follows:

**Pilot Experiment**

In the main experiment we presented naive subjects with a set of 200 photographs obtained from various on-line news services and image databases containing 40 handshakes and 160 non-handshake images. Participants were given the following instructions: "we will show you a set of images. After you have viewed all of them, we will ask you some questions about the set of images; we won't ask about specific images, or ask you the names of any people in the images." The goal was to ensure that the subjects paid attention, without over-analyzing the photographs. Each image was viewed for four seconds. The fixation density plots for images in the same category (handshake images) were similar across different individuals and different photographs. In particular, the handshake class was compared against the non-handshake classes. The spatial distributions of the fixations for images in the handshake class were similar across the participants in the experiment, and they were different from the plots for the non-handshake images. Note that the handshake images included a variety of compositions, including 'portrait' and 'landscape' formats.
The finalized procedure for testing this hypothesis remains much the same as the pilot experiment. The image set was increased to 250 images and there were five distinct classes or categories of images. The categories were “handshake”, “center”, “landscape”, “crowd”, and “miscellaneous”. The handshake category had strictly two people shaking hands; there were no other people in the image, whether in the background or foreground. The center class consisted of images with a “single”, “main” object in roughly the middle of the image and nothing of importance around it (for example, an airplane flying through the sky). The crowd class consisted of an image without a “single”, “main” subject (for example, a large crowd of demonstrators). The landscape class was simply an image of natural landscapes with no real focus point or main object. The miscellaneous class consisted of various images of people, places and/or objects.

**Image Classes**
Results

After analyzing the data from all ten subjects, over 400,000 points of data were produced. Since the eye-tracker program captures data every 16 milliseconds, each subject had roughly 36,000 fixation points. The following was found:

1) Stron image dependence:
   When one subject viewed different images, both of similar and dissimilar categories, their scan path and fixation patterns varied. No subject looked at two images the same way.

2) Strong similarity in fixation patterns for some images: When one image was viewed by each subject, the fixation patterns and scan paths were similar. Although the fixation points were not exactly the same, similar Regions of Interest were clearly defined. This supports the hypothesis that people look at an image in much the same way.

3) Wide variation in fixation patterns for some images: When one image was viewed by each subject, the fixation patterns and scan paths were dissimilar. Just as some images had similar scan paths across the group of subjects, other images had no discernible scan path pattern between subjects.

As one can see from these observations, the experiment was effective in proving that people do look at images in much the same way.

However, there isn’t a strong enough correlation at this time to make the assumption that all people look at one class of images in the same way.

Strong Image Dependence (One subject, four images)
Strong Similarity in Fixation Pattern (One image, four subjects)

Original Image

Subject A

Subject B

Subject C

Subject D
Single Image (Handshake), Four Subjects

Original Image
Here is an example of the handshake class.

This is the class that we were hoping to see the most correlation in. Obviously, the regions of interest are primarily the two faces and secondly, the handshake itself. As one can see, each scan path appears that the subject indeed looked at the faces, and even the handshake. However, once these scan paths are overlaid on top of each other, as seen below, the pattern is lost. There still is high density of fixation points around the faces, but as you can see, the extraneous data must be eliminated before any correlation can be seen.

As seen in the above plot, when all the data points are combined onto one plot, there is a great clutter of data points. The following plots are of all fifty images in each class for all ten subjects (roughly 20,000 data points).
Handshake

Centered Object

From these plots, we can see a correlation for each class, the handshake has the two fixation clouds relative for each face, and the centered object has its fixation cloud near the center of the image. This was as to be expected, since each image was carefully selected to display the characteristics of its own class.
Landscape

Miscellaneous

From these plots, no discernable pattern can be seen. This is part in fact due to the nature of the images selected to fill the landscape and miscellaneous classes. Many images in these classes had objects in the center of the image or scattered in random areas of the images.

Conclusion

At this point in time, the results are very ambiguous. There are scan path patterns in some categories (handshake, center), and no scan path patterns in others (miscellaneous, crowd, landscape). It is very difficult to analyze this much data. Especially since it is only ten subjects. A larger subject group would be more helpful, but only at the cost of exponentially increasing the number of data points that would have to be dealt with.

The link to automatic techniques is not easy, the data provided makes it easier, but by no means does it allow us to develop the code necessary for the software program. This difficulty arises from the test subjects. The individuals that performed this experiment; are just that, individuals. Each person has his or her own unique upbringing, schooling, and personal aptitudes. This is good from an experimentation standpoint, but allows us to see just how differently a subject views an image.

The subjects in this experiment were all college students. This was done because of expediency and availability. Perhaps different social classes would view these images in different ways. Artists would almost definitely look at an image differently than mathematicians. This experiment was a good first attempt; though possibly in the future, different social "castes" might be tested to see if there is a different correlation of scan paths.

However, it would not be beneficial to run hundreds of subjects through this experiment, even running an extra forty subjects of a broader social spectrum, would increase the number of data points to over 2 million. Another possibility would be to divide the image into sections, essentially quantizing the image into regions where fixation points near each other could be grouped into basically one fixation point.

While we have proven that people do view images in similar ways, there is not strong enough empirical data to make the broad jump to a computational framework. I believe that this jump is not too far in the future, as continual progress is constantly being made in both the computer science and psychophysical fields. The applications of this project have definite merit; so further experimentation can be almost assured.
Table of Contents
Using Human Observers' Eye Movements in Automatic Image Classifiers

Timothy Grabowski

References


