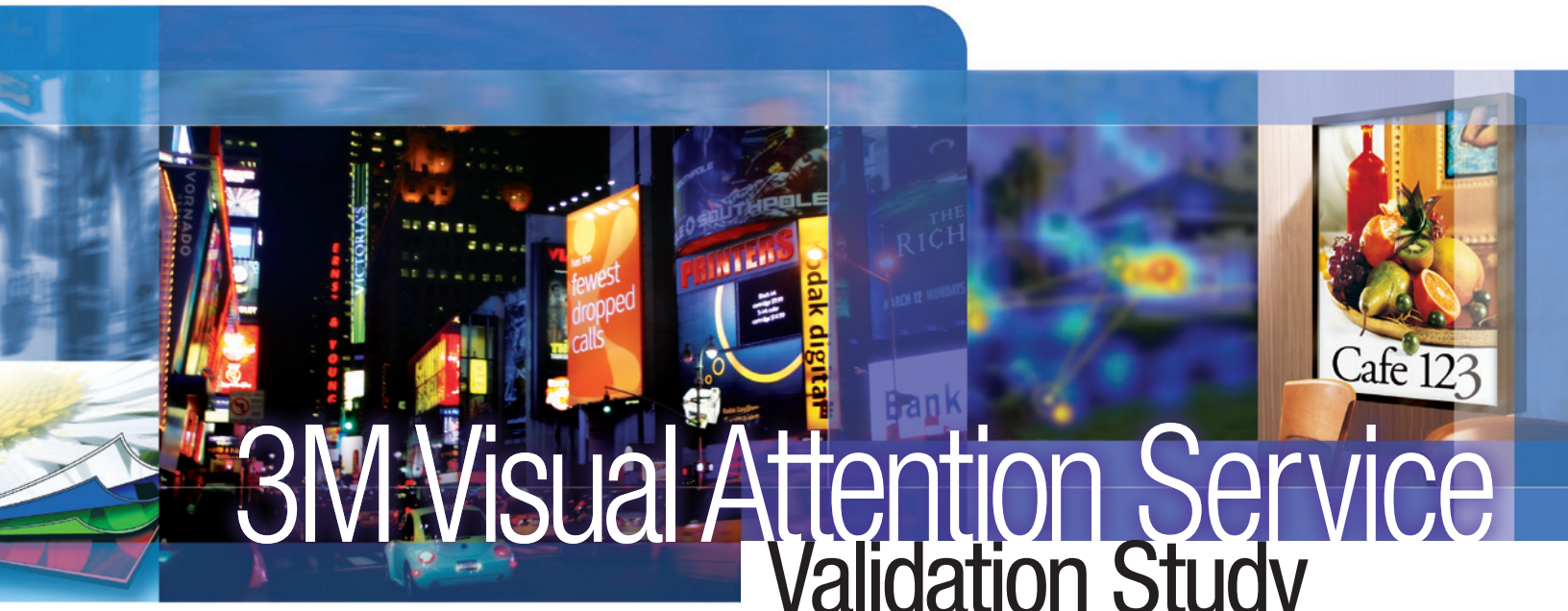


3M Visual Attention Service Validation Study

Abstract

Advances in both Behavioral Vision Science and Computational Vision are the basis for computational models of human visual attention that take as input an image or video and make predictions about where people will fixate in the first 3 to 5 seconds of viewing. Initial fixations play an important role in acquiring an understanding of a scene and/or content and serve as the gateway to further visual analysis. 3M Visual Attention Service (3M VAS) uses a computational model of visual attention to make predictions about where initial fixations will occur in an image. Validation studies are presented in which eye-tracking results are compared to the predictions made by 3M VAS on the same set of images. Using signal detection Response-Operator-Characteristics (ROC) we found that the model was able to predict human fixations at around 85% of the theoretical limit (theoretical limit being eye-tracking). If biases are removed from the eye-tracking data (e.g., the central bias which is a function of collecting data on computer monitors), the model performs at approximately 90% of the theoretical limit.

Performance Evaluation

Introduction

There has been a great deal of research in Vision Science and Computer Vision to develop theories of visual attention that are converted into mathematical equations and algorithms that make explicit predictions about where people will initially look when viewing complex scenes such as shopping malls, streets, magazine pages, web pages, advertising content, etc.

3M Visual Attention Service is a software tool that takes as input an image or video and as output makes predictions about where people will initially look within that scene. The software is based on 30 years of research in academic institutions and recent research at 3M to better understand what people will initially notice. The goal of this research is to develop a deep understanding of how visual attention is initially allocated in a complex scene—that is where will people look—and to translate that knowledge to a software tool that predicts where these initial fixations will occur.

But how does the visual system make decisions on where to look? The human visual system takes in a large amount of visual information (that is, the entire visual field) at any given time. However, the human visual system expends the bulk of its resources processing only a small fraction of this information—typically where the viewer is fixated. Although the majority of the visual field is not being attended to, it serves as an important part of the visual experience. The human visual system uses the information in the periphery to monitor regions that might be of interest to the viewer—the regions that attract visual attention. If the early perceptual properties (color, motion, contrast etc.) are engaging, the human visual system will move its fixation to that location to gather more visual information. Thus, when one initially enters a room, or initially turns a corner, or looks at a magazine page, the human visual system surveys the scene to become aware of what might be important within the scene. Based on prior research, during this initial "surveillance" period our visual system is attracted to low-level perceptual features such as color, luminance, edges, motion and other early visual processing elements. This initial surveying typically occurs in the first 3-5 seconds. During this survey period the visual system analyzes the visual representation projected on the retina and based on that information the eyes are drawn to particular regions of the scene based on the visual features (e.g., color, contrast, text faces, etc.). What people do after this initial 3-5 seconds will depend on Top-Down processing. These secondary fixations that are primarily driven by Top-Down processing will use the information gathered during these initial fixations to decide

Figure 1

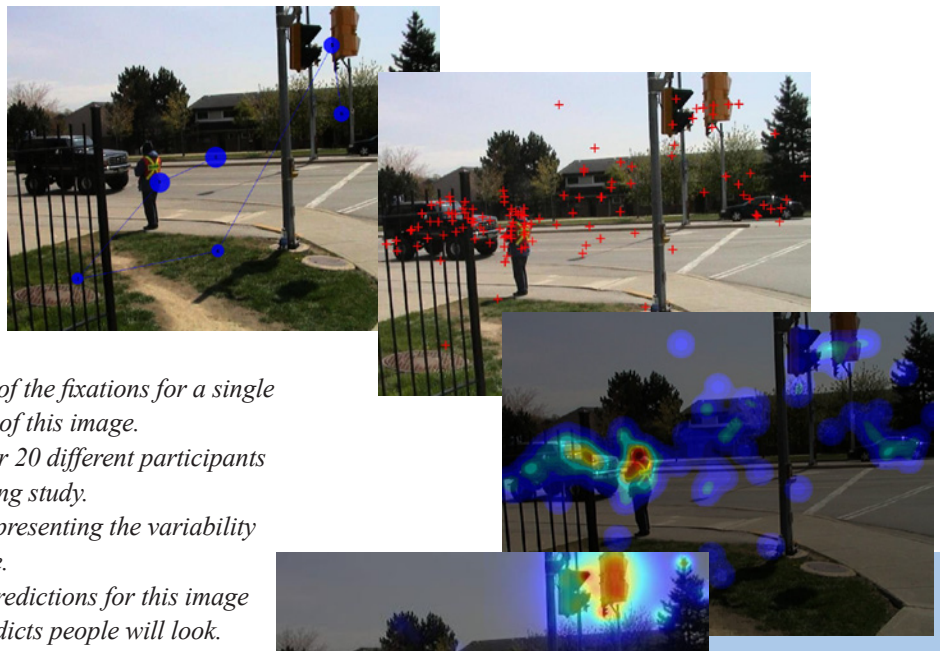


Figure 1. Upper-Left: *An illustration of the fixations for a single subject during a 3 second presentation of this image.*
Left-Center: *The combined fixations for 20 different participants who viewed this image in an eye-tracking study.*
Right-Center: *A Fixation Heat Map representing the variability of the participant's looked at this image.*
Lower Right: *The model's Heat Map predictions for this image showing the regions that the model predicts people will look.*

where to look next based upon the individual's task (e.g., "I need to find the exit" versus "I need to find an elevator") and the relevancy of the items evaluated during these initial fixations (e.g., "I noticed an advertisement for a restaurant on the wall and I need to find someplace to eat tonight with my friends — gather more visual information on that digital sign.").

One may wonder how gender, age, or experience might affect where people initially look. Behavior research has shown that these human attributes have little effect on where people will initially look (assuming equivalent visual acuity and other visual processing capabilities). However, once the visual system has completed the initial surveillance process (usually 3-5 seconds), these other factors (e.g., top-down influences related to personal interest and task) will play a more significant role in where people will look.

Although where people will look within an image is fairly consistent, we still have to acknowledge that there is some variability between individuals as well as some variability in the consistency between individuals for a particular image. Figure 1 shows an illustration of a single subject's fixation sequence, the combined sequences across individuals, a Fixation Map, and the VAS predictions for this image. Notice that there are some regions where most people look (e.g., on the crossing guard) while there are some regions where only a few fixations occur (e.g., base of the lamp post). The between-observer fixation variability (the variability from one subject to another) poses a challenge to validating a computational model of human visual attention.

Data & Methods

L3M VAS was validated using eye-tracking data. Data used for this validation consisted of four different data sets, two of which were collected by academic research labs (MIT & York University) and the third and fourth sets were collected by 3M. Each data set consisted of eye fixations data collected for relatively short periods of time (3-5 seconds) for a variety of images (indoor scenes, outdoor scenes, people, advertisements, etc.). This provides a measurement as to where people will initially look at an image.

All of the data were collected in a similar way: participants were seated in front of a computer monitor and images were presented one at a time for a given period of time. In between each image there was either a short fixed inter-image pause or a wait period in which the subject pressed a button indicating that they were ready for the next image. Participants were instructed to "freely view "the images during that time (that is, they were not given any specific task to complete). During the free viewing period, eye-tracking equipment measured and recorded where the participants looked at the image. Figure 1 provides an illustration of eye-tracking data for a single subject. The blue circles indicate the location and the size of the circle indicates the dwell time for a particular subject. The red "+" in the upper-left illustration of Figure 1 shows the combined location of all 20 participants who looked at this image in the 3M study.

Validating 3M Data Collection Techniques

One of the goals of the current study is to show that the methods 3M uses to collect data match those of outside research institutions. To do this, we collected eye-tracking data using the images from a previous study conducted at York University in which eye-tracking data from their lab is made publically available (http://www-sop.inria.fr/members/Neil.Bruce/eyetrackingdata.zip). The York University images consisted of 120 images of various indoor and outdoor

scenes. One of the purposes of collecting data on the York University images was to replicate their findings and to validate the methods and procedures used at 3M matched that the processes used by other vision scientists. Later we will compare the results on the York University images collected at 3M to those collected at York University.

Figure 2. *Four sample images from the York University image set.*

In addition to collecting data on the York University images, we also collected data in the same study on a series of marketing materials. The purpose of this study is to evaluate how well 3M VAS predicts the initial fixations for advertising and marketing materials. These images consisted of print ads, packaging, outdoor billboards and shelving planograms and were mixed within the set of York images in a random fashion. Figure 2 shows a sample of four images from this data set.

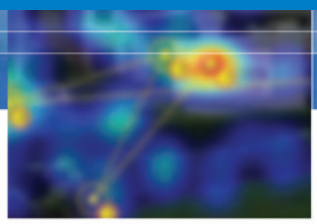
Participants: Three groups of participants were used in the current analysis from three different labs. The York University lab used 20 participants and the MIT study used 15 participants. Eye-tracking data collected at 3M used 20 participants (13 Males and 7 Females) ranging in age from 23 to 60 years old. The participants were 3M employees who were not familiar with the purpose of the study. All of the participants had normal or corrected to normal vision.

Upper-Theoretical Performance Limit

In order to fully evaluate the predictive power of a Visual Attention model one needs to account for the natural variation in eye-tracking data—the between-subject variability. Fundamentally, the upper-theoretical boundary for predicting eye-fixations is the ability of one visual system (one person or group of people) to predict the fixations of a second visual system (or group of people). A model of human visual attention cannot outperform an actual visual system and this provides us with an upper-theoretical bound of performance. To measure the upper-theoretical performance boundary we used a split-data design technique. Specifically, we used the fixation data from one half of the subjects to predict the second half of the subjects.

To generate the predictions we used the fixation locations from one half of the subjects and convolved a Gaussian kernel at each location where there was a fixation. The Gaussian kernel was approximately 1-degree of visual angle, which corresponded roughly with the measurement error in eye-tracking equipment. The Upper-Right image in Figure 1 illustrates the output generated by this convolution. We then used the generated representation from this first group of participants to predict the fixations for the second group of participants.

Figure 3. *An illustration of the 3M images used in the validation study.*



Response-Operator-Characteristic

LTo evaluate the predictive performance we calculated ROC values using a split-data method in which we calculated how well one-half of the subjects fixation data predicted the second half of subjects (randomly selected) for each image and then compared the predictive power of the 3M VAS predictions to that of using human data. To do this we generated two predictive maps for each image. One predictive map was generated by taking one half of the subjects and generating a heat map representation from the subjects' fixation locations. The second heat map was generated by using 3M VAS to analyze each image.

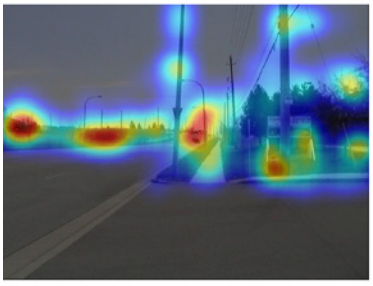
For both sets of predictions (human and 3M VAS) for each pixel there is a continuous value associated with the strength of the prediction for that pixel. To generate an ROC value for each image we varied the threshold of how liberal of a prediction would be considered. When the threshold is high, the model makes very few predictions as to where attention will be allocated, has only a few hits and has very few false alarms (see lower left corner of ROC curve in Figure 4). However, as the threshold decreases, the area in which the model is predicting becomes larger and the model correctly predicts more fixations. However, this increased region also increases the number of false alarms. To evaluate the performance of information available in the Heat Map we used the ROC calculation described above.

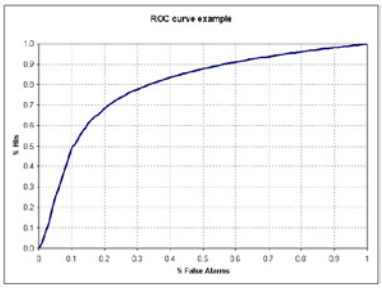
ROC takes into account multiple threshold levels and measures the number of Hits (correct predictions) and False Alarms (incorrect predictions; see Figure 4 for an illustration) for each threshold level. After generating the Hits and False Alarm rates for multiple thresholds, a single metric is generated by calculating the area under the curve. Thus, if the model perfectly predicts the data, the prediction will have a ROC value of 1.0. Figure 5 (next page) shows the distribution of ROC values for the different images for the Human-To-Human comparisons for the York University, MIT and 3M advertising data. The York University data predicted itself with an average ROC value of 0.819. The 3M data predicted the York data with an average ROC value of 0.812. This insignificant difference indicates that the methods and procedures used at 3M closely match those used at York University. The MIT data predicted itself with an average ROC value of 0.89 and the 3M advertising data predicted itself with an average ROC value of 0.93. These values provide us with the upper-theoretical performance boundary by which we will compare 3M VAS performance.

3M VAS Prediction Efficiency

The Human-to-Human analysis provides a valuable metric for evaluating the efficiency of 3M VAS. Because the Human-to-Human analysis provides a way to measure the *Theoretical Limit* of performance — the very best predictive

Figure 4 3M Subjects ROC





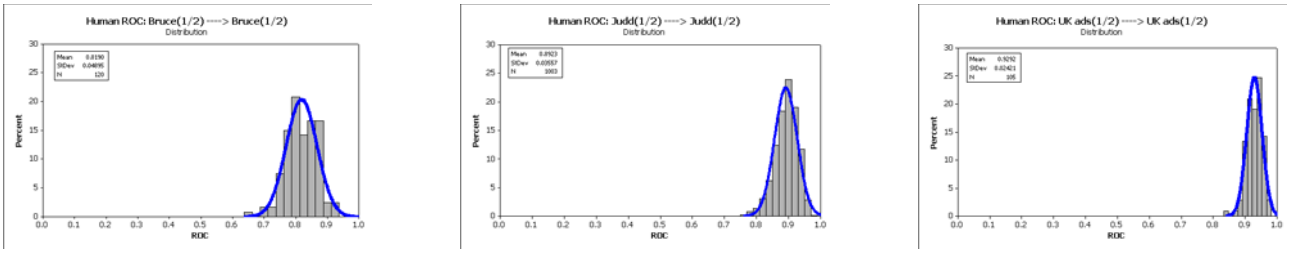


Figure 5. *The distribution of ROC values for the York University Images using the split-data design technique, specifically, providing the upper theoretical performance limit for eye-tracking data. Upper-left is the distribution of values for the York University Data set (Mean ROC=0.819), the Upper-Right shows the distribution for the MIT data set (Mean ROC=0.89) and the Lower plot shows the distribution for the 3M advertising data set (Mean ROC=0.93).*

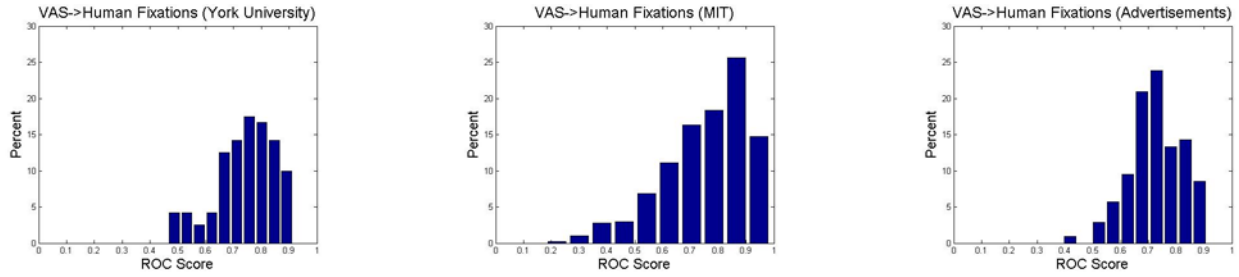
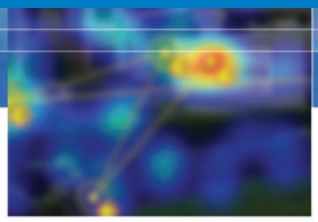


Figure 6. *An illustration of 3M VAS performance on predicting eye-movements for the York University images (Mean ROC=0.74), the MIT images (Mean ROC=0.76) and the 3M advertising images (Mean ROC=0.73).*

performance one can expect if one were to run an eye-tracking study—we can use this performance as our baseline by which to compare the predictions made by 3M VAS to provide a metric of performance relative to human eye-tracking performance. To do this we used the output of 3M VAS on the images in these various studies to predict the eye-fixations using the same ROC data analysis technique described above. The distribution of ROC values are shown in Figure 6. The ROC values for the 3M VAS predictions are 0.74 for the York University images, 0.76 for the MIT images and 0.73 for the 3M advertising images.

Fixation Biases in Eye-Tracking

Within the eye-tracking community it is well known that when collecting eye-tracking data on computer displays there are particular fixation biases that are not determined by the content that is on the computer screen. One well known bias is known as the Center Bias or the Center Fixation Bias. The cause of this bias is due to multiple factors. One known factor is that fixating in the center of the screen provides the most information about the image (given the resolution fall-off in peripheral vision). However, the center bias is due to the fact that people are looking at images on a computer screen. 3M VAS does not predict how people will look at these images on a computer screen but instead how people will look at these images in the real world . Using a technique described in Zhang et al. (2008) we evaluated both the Human-to-Human performance and the 3M VAS performance removing the center bias and any other fixation biases that might have occurred in the data collection.



Modified ROC

The modified ROC (mROC) calculates performance in much the same way as the standard ROC. As a reminder, the standard ROC calculates the number of hits as a function of the number of false alarms (incorrect predictions) by varying the levels of threshold (threshold independent analysis). The right side illustration in Figure 4 illustrates a sample curve. The mROC, by contrast, evaluates how well the predictions generated for a particular image (e.g., Image-1) predict the fixations for that image against the fixations for all of the other images. As with the standard ROC approach, the mROC calculates the percentage of hits for the target image (i.e., correctly predicting the fixations for the target image for a particular threshold) against the percentage of hits for the fixations for all other images in the study.

It is well known that when collecting data on a computer there is a bias for fixating in the center of the screen (Tseng et al. 2009). There are a number of reasons for this center bias, which were discussed and evaluated by Tseng et al. (2009), that we will not go into in this manuscript. Removing this center bias from the analysis is important since the performance of the model will be reduced because the model predicts what people will see in a real environment (free viewing with no frame), not how perform in an eye-tracking study on a computer. It should be noted that one way to improve the model's performance is to put an explicit center bias within the model's predictions. However, this improves the model's performance for predicting fixations on a computer monitor but not necessarily in the real world.

One way to think about the mROC calculation is that the mROC evaluates how well the model is able to predict the unique fixations for a target image relative to all other images. More accurately, the more unique the correctly predicted fixation positions for a particular image, the more weight that is given to the score. Therefore, if there are numerous fixations in the center of the computer screen and the model correctly predicts these fixations, these scores will be weighted less than a prediction made to fixations where, overall, there were very few fixations across all of the images.

Results

When taking out the bias, the upper-theoretical performance measure (Human-to-Human) for the York University data is 0.71 and the 3M VAS performance is 0.66 with a Prediction Efficiency value of 93%. The MIT data had an upper-theoretical limit of 0.70 and the 3M VAS performance was 0.66 with a prediction efficiency of 94%. The 3M advertising data had an upper-theoretical limit of 0.76 and the 3M VAS performance was 0.65 with a prediction efficiency of 85%.

Summary & Conclusions

We presented a validation study of 3M VAS by comparing the saliency predictions made by 3M VAS to eye tracking on three different data sets. Two of the data sets were collected by outside academic institutions (York University and MIT) and one was collected at 3M. A variety of images were used, which included indoor scenes, outdoor scenes, and advertising content.

[1] One may consider web pages as content that is naturally viewed on a computer screen. Interestingly enough web pages also have biases that go beyond a center bias.

Performance was evaluated using Signal Detection ROC metric that takes into account how well the model correctly predicts a fixation (hit) along with incorrect predictions (false alarms). To evaluate how well the model is able to account for the intrinsic variability in the data (inter-subject variability), we compared the model's predictive power to that of actual eye-tracking data to predict eye-tracking data. This split-data approach provided an upper theoretical limit on eye-fixation performance.

To evaluate 3M VAS we compared the model's predictions to that of the upper theoretical limit produced by actual eye-fixations (split-data analysis) to provide a measure of predictive efficiency for 3M VAS. 3M VAS predictive efficiency is 90% for the York University images, 85% for the MIT images and 79% for the 3M advertising images. The predictive efficiency increased dramatically when fixation biases (e.g., center fixation bias when viewing images on a computer screen) were removed from the data set. With the biases removed, 3M VAS predictive efficiency was 93% for the York University images, 94% for the MIT images and 85% for the 3M advertising images.

References

Tseng, P.-H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. Journal of Vision, 9(7):4, 1-16, http://journalofvision.org/9/7/4/, doi:10.1167/9.7.4.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. Journal of Vision, 8(7):32, 1-20, http://journalofvision.org/8/7/32/, doi:10.1167/8.7.32.

Commercial Graphics Division
3M Digital Out of Home
3M Visual Attention Service
3M Center, Building 207
St. Paul, MN, 55144-1000
USA
3M.com/vas