The Allocation of Visual Attention in Multimedia Search Interfaces

by
Edith Allen Hughes
May 2017

Presented to the
Division of Science, Information Arts, and Technologies
University of Baltimore

In Partial Fulfillment
of the Requirements for the Degree of
Doctor of Science in Information and Interaction Design

Approved by: _____
Kathryn Summers, Dissertation Advisor

_____
Greg Walsh, Committee Member

_____
Deborah Kohl, Committee Member

Abstract

Multimedia analysts are challenged by the massive numbers of unconstrained video clips generated daily. Such clips can include any possible scene and events, and generally have limited quality control. Analysts who must work with such data are overwhelmed by its volume and lack of computational tools to probe it effectively. Even with advances made in machine-learning, humans are still needed to review and assess multimedia clips to determine if the video contains an event of interest or if it provides an actionable insight. This study seeks to provide a better understanding of the influence of the user interface of a multimedia player on the performance of tasks in terms of accuracy and the allocation of visual attention, especially whether the customizations that users make to the interface can enhance their performance of the tasks. The results provide insights that reinforce the human ability – and preference – to discern "gist" from static images as surrogates to viewing video clips, and identifies characteristics of user interface and interaction design of multimedia players to leverage this human ability. Although the findings do not indicate strong correlations between accuracy and false alarm rates with how subjects allocate their visual attention, the qualitative results reinforce that different event types, especially event types that are more difficult to semantically define, may require a greater reliance on the video playback, and therefore an overall different strategy to the triage and video search process. Qualitative analysis reinforces this finding, and provides specific suggestions for interface design and visual search strategies that can strengthen accuracy while maintaining a high rate of throughput.

Acknowledgments

"A picture is worth a thousand words. An interface is worth a thousand pictures." -- *Ben Schneiderman, 2003.*

Several colleagues, for whom I am grateful, helped to make this dissertation possible, either through encouragement to expand the human baseline study methodology to include eye-tracking data collection and analysis and through direct support on the project in the role of test moderator and as data analysis of the human baseline performance study. I am also grateful to my department head at the time who nominated me for the Accelerated Graduate Degree Program, which made it easier for me to both work and attend school simultaneously.

I would also like to acknowledge and thank my dissertation advisor, Kathryn Summers, from whom I learned so much during my doctoral program at the University of Baltimore, for her advice and guidance throughout the program and the dissertation process. I would also like to thank Deborah Kohl and Greg Walsh for serving on my committee and reminding me that the best dissertation is a finished dissertation.

The acknowledgements of support would not be complete without thanking my loving husband Michael and daughter Madeleine. Their love, support, and encouragement sustained me throughout the journey.

The author can be reached via email at eahughes@mitre.org or elahughes@gmail.com.

Table of Contents

List of Tables 4

List of Figures 5

Chapter 1: Introduction 1

Introduction to Multimedia Event Detection 1

Purpose and Results of the Human Performance Baseline Study 2

Purpose of this Dissertation Research 4

Hypotheses Statements 7

Chapter 2: Literature Review 10

Multimedia Information Retrieval and Event Detection 10

Visual Scene Processing 11

Objects and Global Scene Context for Determining Gist 12

Use of Visual Surrogates in Multimedia Event Detection 15

Chapter 3: Methodology 19

Subjects 19

Sample Size and Statistical Power 20

Stimuli Source and Preparation 22

Stimuli Source 22

Stimuli Preparation 22

Apparatus 25

Custom-Developed Windows Explorer/Media Player Application 25

Eye-Tracking Sensor and Recording Software 28
Procedure 28
Test Preparation 29
Test Protocol 29
Multimedia Player Interface Errors 30
Post-Test Survey and Interviews 31
Analysis 31
Data Preparation 31
Analysis of Rates of Accuracy and Visual Attention Allocation 34
Analysis of Visual Attention Between Object Events and NoObject Events 34
Analysis of Visual Attention Between Difficult and LessDifficult Events 35
Chapter 4: Results 37
Influence of Allocation of Visual Attention on Accuracy and Throughput 37
Accurate Target Detections 37
False Alarm Rates 40
Additional Results: Accuracy and Event Type 46
Rates of Throughput and Allocation of Visual Attention 47
Influence of Objects in Semantic Definition of an Event on the Allocation of Visual Attention 53
Influence of Perceived Event Difficulty on the Allocation of Visual Attention 54

Additional Findings: Effect of Size of Thumbnail Area of Interest and Thumbnail Image on Visual Attention Patterns 59
Influence of Size of Thumbnail Viewing Area and Thumbnail Image on Accuracy Rates 64
Influence of Size of Thumbnail Viewing Area and Thumbnail Image on Allocation of Visual Attention 67
Additional Findings: Mean Ratios of Visual Attention and Test Order 77
Additional Findings: Order of Fixation in Areas of Interest 78
Qualitative Results 84
Chapter 5: Conclusions 91
Recommendations for Further Study 93
References 96
Appendix A: Acronyms and Definitions 102
Appendix B: NIST Multimedia Event Detection Event Descriptions 104
Appendix C: Post-Test Questionnaire and Results 113
Appendix D: Post-Test Interview Questions 122

List of Tables

Table 1 Order of Event Tests for Each Subject 20

Table 2 MED'11 Events and Classifiers Used for Ranked Set Filtering..... 23

Table 3 Target event densities in random and rank-ordered test sets. 24

Table 4 Comparison of Accurate Target Detections and Visual Attention Metrics 37

Table 5 Comparison of Accurate Target Detections and Visual Attention Metrics 40

Table 6 Comparison of False Alarm Rates and Visual Attention Metrics 41

Table 7 Correlation of False Alarm Rates to Visual Attention..... 43

Table 8 Correlations of Throughput and Allocation of Visual Attention..... 48

Table 9 Repeated Measures, Within Subjects ANOVA for Visual Attention Measures between Object Events and NoObject Events..... 54

Table 10 Repeated Measures, Within Subjects ANOVA for Visual Attention Measures between Difficult and LessDifficult Events..... 55

Table 11 Correlations of Thumbnail Image Size and Visual Attention..... 74

List of Figures

Figure 1: Custom Windows Explorer/Media Player application interface 26

Figure 2: Sample output data from the custom multimedia player application 27

Figure 3: Overlay of Areas of Interest on Multimedia player Interface 32

Figure 4: Correlation of Accurate Detections to Total Fixation Duration Ratio 38

Figure 5: Correlation of Accurate Detections to Fixation Count..... 38

Figure 6: Correlation of Accurate Detections to Total Visit Duration Ratio..... 39

Figure 7: Correlation of Accurate Detections to Visit Count Ratio..... 39

Figure 8: Correlation of False Alarms Rates to Total Fixation Duration Ratio 41

Figure 9: Correlation of False Alarms Rates to Fixation Count Ratio 42

Figure 10: Correlation of False Alarms Rates to Total Visit Duration Ratio 42

Figure 11: Correlation of False Alarms Rates to Visit Count Ratio 43

Figure 12: Correlation of Mean Rate of False Alarms to Mean Total Fixation Duration Ratio by Subject 44

Figure 13: Correlation of Mean Rate of False Alarms to Mean Fixation Count Ratio 45

Figure 14: Correlation of Mean Rate of False Alarms to Mean Total Visit Duration Ratio 45

Figure 15: Correlation of Mean Rate of False Alarms to Mean Visit Count Ratio 46

Figure 16: Mean Number of Videos per Test by Subject 48

Figure 17: Correlation of Number of Videos Processed to Total Fixation Duration Ratio 49

Figure 18: Correlation of Number of Videos Process to Fixation Count Ratio..... 50

Figure 19: Correlation of Number of Videos Process to Total Visit Duration Ratio 50
Figure 20: Correlation of Number of Videos Processed to Visit Count Ratio 51
Figure 21: Mean Visual Attention Ratio by Number of Videos Processed 52
Figure 22: Correlation of Number of Videos Process to Total Visit Duration Ratio without Subject P05 Outliers 53
Figure 23: Mean ratios of Total Fixation Duration by Subject and Difficulty of Event .. 57
Figure 24: Mean ratios of Fixation Count by Subject and Difficulty of Event 57
Figure 25: Mean ratios of Total Visit Duration by Subject and Event Type Difficulty ... 58
Figure 26: Mean ratios of Visit Count by Subject and Difficulty of Event 58
Figure 27: Default settings of custom video player interface for small thumbnails 59
Figure 28: Default custom video player interface for medium-sized thumbnails 60
Figure 29: Default custom video player interface for large-sized thumbnails..... 60
Figure 30: AOI percentages for the Default Custom Video Player Interface Settings 61
Figure 31: Thumbnail viewing area widened to a width of six small thumbnails 62
Figure 32: Thumbnail AOI is 30.6 percent of total image area 62
Figure 33: Configuration of Thumbnail AOI set to width of one large thumbnail image 63
Figure 34: Configuration of thumbnail viewing area set to width of four medium thumbnail images 64
Figure 35: Correlation of Thumbnail AOI Size to Accurate Detections 65
Figure 36: Correlation of Thumbnail AOI Size to False Alarms 66
Figure 37: Correlation of Thumbnail Image Size to Accurate Detections 66

Figure 38: Correlation of Thumbnail Image Size to False Alarms..... 67

Figure 39: Correlation of Size Thumbnail AOI to Total Visit Duration Ratio..... 68

Figure 40: Correlation of Size of Thumbnail AOI to Percent of Videos Played..... 69

Figure 41: Correlation of Size of Thumbnail AOI to Average Amount of Video Viewed 70

Figure 42: Correlation of Size of Thumbnail AOI to Fixation Count Ratio..... 71

Figure 43: Correlation of Size of Thumbnail AOI to Total Fixation Duration Ratio..... 72

Figure 44: Correlation of Size of Thumbnail AOI to Visit Count Ratio 72

Figure 45: Correlation of the Size of the Thumbnail AOI to the Number of Videos Processed..... 73

Figure 46: Correlation of Thumbnail Image Size to Total Fixation Duration Ratio 75

Figure 47: Correlation of Thumbnail Image Size to Fixation Count Ratio 75

Figure 48: Correlation of Thumbnail Image Size to Total Visit Duration Ratio..... 76

Figure 49: Correlation of Thumbnail Image Size to Visit Count Ratio..... 76

Figure 50: Mean Visual Attention by Test Order 78

Figure 51: Area of Interest for First Fixation by Subject..... 79

Figure 52: Area of Interest for First Fixation by Test Order 80

Figure 53: Area of interest for first fixation by event type 81

Figure 54: Correlation of AOI of First Fixation to Thumbnail Size..... 82

Figure 55: Correlation of AOI of First Fixation to Size of Thumbnail AOI 83

Figure 56: Number of videos played, and amount of each video played, by event type .. 85

Figure 57: Post-test survey scores on "How Difficult to Identify Events" by event type 114

Figure 58: Post-test survey results for "How Successful in Identifying Events" by Event Type 115

Figure 59: Post-test survey scores for "How difficult to match events to description" by Event Type 116

Figure 60: Post-test survey scores for "How difficult to visually search for events" 117

Figure 61: Post-test survey scores for "How difficult was it to stay focused" 118

Figure 62: Post-test survey results for "Rate the amount of mental/perceptual effort" .. 119

Figure 63: Post-test survey scores for "Amount of time pressure felt" 120

Figure 64: Post-test survey score for "Amount of frustration felt"..... 121

Chapter 1: Introduction

Introduction to Multimedia Event Detection

A multimedia analyst's task is to review media in multiple formats from a wide variety of sources to identify trends, patterns, and relationships that provide actionable insights. However, massive numbers of video clips are generated daily on many types of consumer electronics and uploaded to the Internet. In contrast to videos that are produced for broadcast or from planned surveillance, the volume of unconstrained video clips produced by anyone who has a digital camera presents a significant challenge for both manual and automated analysis. Such clips can include any possible scene and events, and generally have limited quality control. Analysts who must work with such data are overwhelmed by the volume and lack of tools to probe it.

However, the ability for machines to identify user-defined (or semantically-defined) events is difficult. Multimedia analysts are interested in a wide and constantly changing range of events, making it difficult to establish special filters that are either pre-defined or based on machine learning algorithms. Instead, the goal is to design automated systems that can accept a human-generated definition of an event to search a multimedia collection. But even with advanced machine learning, humans are still needed to review and assess multimedia clips to determine if the video contains an event of interest or if it provides an actionable insight.

I collected the eye-tracking and post-test interview data for this dissertation research study in conjunction with a human performance baseline study that a research team, of which I was a member, conducted to assess the accuracy and speed of humans in identifying video clips that contain a semantically defined event. The team conducted the baseline study on behalf of a government agency in parallel to the National Institute of Science and Technology (NIST) Text Retrieval Conference Video Retrieval Evaluation (TRECVID) 2011 conference, which is a conference series sponsored by NIST devoted to research in automatic segmentation, indexing, and content-based retrieval of digital video. The purpose of the baseline study was to establish metrics of how accurately and

quickly humans could identify multimedia (i.e., video) clips that contained a specific semantically defined event from among a large set of randomly selected multimedia files. NIST (Over et al., 2011) defines a semantically-defined event as an event that:

- is a complex activity occurring at a specific place and time;
- involves people interacting with other people and/or objects;
- consists of several human actions, processes, and activities that are loosely or tightly organized and that have significant temporal and semantic relationships to the overarching activity;
- is directly observable.

For this dissertation study, I added the collection of eye-tracking data from the subjects as they conducted the multimedia event detection tests so that I could analyze their use of the multimedia player interface – specifically, how they allocated their visual attention between the static thumbnail viewing area and the video playback viewing area – to determine if any patterns emerged that would correlate in the allocation of visual attention in the multimedia player interface to either accuracy or speed of human performance.

Purpose and Results of the Human Performance Baseline Study

The purpose of the baseline study was to collect performance metrics for accuracy in precision (measured inversely in terms of the number of false alarms, or videos selected that do not contain the specified event) and recall (measured inversely in terms of the number of missed detections, or video clips with relevant events that the user did not find), throughput (the total number of videos a user needs to review to identify relevant events), and speed (in terms of time to the completion of reviewing all videos). Researchers will use these baseline metrics as a comparison for automated video retrieval systems under development that will assist with video event detection and analysis, such as the systems that researchers develop and test as part of the TRECVID conference series.

The results of the baseline study for multimedia event detection experiment indicate that, overall, humans are fast and accurate in identifying video clips that contain

target events, especially when the event can be detected based on the visual gist of the thumbnail image. The overall accuracy rate for detecting target events was 0.88, which means that subjects correctly identified target events 88 percent of the time. The overall rate of false alarms was 0.013, indicating that subjects falsely identified target events at a rate of 1.3 percent. This study reinforces that humans can determine visual gist very quickly (Rayner, 2009), and use a variety of attributes, such as objects, people, setting/environment, and action/activity to determine gist (Henderson & Ferreira, 2004; Yang & Marchioni, 2005). When subjects can easily confirm the presence of the target event based on the thumbnail image, or infer its absence based on the lack of the relevant environment and objects in the thumbnail, then their accuracy and throughput remain very high. The baseline study results also show, however, that certain event types and video characteristics can make event detection more difficult and reduce the level of throughput. The following sections summarize those characteristics.

Missed detections. The baseline study results revealed that the three leading causes of missed detections seem to be 1) reliance on a misleading thumbnail image (i.e., does not show a scene or object associated with the event), 2) the defining object or action occurring in only a small portion of the video that the subject does not play, and 3) the ambiguity of the event definition.

False Alarms. The baseline study results also indicated that, although false alarms can be caused by either deciding based on the thumbnail alone or by the ambiguity of the event definition and annotation, the latter reason tends to play a larger role in false alarms, especially where events are defined by actions or activities. Thumbnail images may display objects (e.g., a group of people) or a setting that suggests the event, but when the event is defined by activity or actions (e.g., marching in a parade, dancing in a flash mob) and may or may not have an associated object (i.e., people present for grooming animals; cake, candles, or signs for a birthday party), then subjects need to play more of the video and pay more attention to the total set of contextual clues. The type of events that had higher numbers of false alarms (E006 – Birthday Party, E008

– Flash Mob, E009 – Getting a Vehicle Unstuck, and E012 – Parade) reinforce this finding.

Objects. Events that require the presence of specific objects as part of their definition are easier to detect than events where the objects may help define the event, but are not a required part of the event. For example, the event types of repairing an appliance, getting a vehicle unstuck, or working on a sewing project all require the presence of an object (e.g., an appliance, a vehicle, sewing supplies) as part of the event definition. A birthday party, however, may be identified by a cake with candles, balloons, or party goods that say "Happy Birthday." These objects help to quickly identify the event as a birthday party; however, the lack of these objects does not preclude it from being a birthday party. Other clues, such as audio clues of singing or someone saying "happy birthday," can also define the event.

Event Ambiguity. In some cases, a lack of clarity in the written description or a lack of consistency between the written definition (See Appendix B for the written definitions for each event, as provided by NIST) and the pre-test sample videos caused some confusion among subjects as to whether an observed event qualified as a target event. Subject familiarity with the event also influenced their perceived clarity about the event definition. Some subjects were more familiar with some event types that matched their interests or experiences, such as changing a tire or getting a vehicle unstuck (car aficionados and do-it-yourselfers) or parkour and flash mobs (younger subjects), and experienced less ambiguity with those events than subjects who were not as familiar with the objects or activities.

Purpose of this Dissertation Research

The goal of this dissertation research is to explore whether the user interface has any influence on the human performance of tasks, especially whether the customizations that users make to the interface enhance their performance of the tasks. The baseline study established metrics for the accuracy and throughput of humans in multimedia event detection and identified certain characteristics of the stimuli that may influence their performance. The baseline study, however, did not address questions about how the video

player interface, the subjects' configurations of the thumbnail component of the video player interface, and their use of the two visual components (thumbnails and video playback) for deciding about each video, may have influenced their accuracy and throughput. Specifically, do subjects who have a higher rate of accuracy allocate their visual attention differently in their use of the visual search components of the interface? Understanding how multimedia information retrieval (MIR) system search interface features support or inhibit visual scene processing, especially in the absence of being able to apply text queries to help identify events, will help multimedia interface designers ensure that they provide the flexibility and options that multimedia analysts need to support their work. The findings may also contribute to training and performance improvement guidance to multimedia analysts that will help them refine their search skills and improve their accuracy and throughput by configuring their search interfaces for optimal performance.

The goal of this dissertation research is to isolate the influence of specific components of the user interface used for viewing visuals – whether thumbnails or playing the video clips – to determine how subjects allocate their visual attention to these areas, and how the allocation of visual attention may influence their performance in terms of accuracy and throughput. I used eye-tracking equipment and research methods to obtain data that will identify whether users allocate their visual attention differently based on their multimedia search task, whether certain patterns of visual attention help or hinder either their accuracy or throughput, and under what conditions, such as the characteristics of the type of multimedia event for which they are searching.

Because the results of the baseline study suggested that certain types of events may have influenced accuracy and throughput, I also questioned whether visual scene processing and video processing behaviors of subjects, as indicated by eye movement patterns, vary between different types of semantically defined events. More specifically, do subjects allocate their visual attention differently between the areas of the video search interface when the semantic definition of an Event includes an Object (referred to as Object Events) versus when the semantic definition of an Event does not include an

object (referred to as NoObject Events)? The results of addressing this question may help inform requirements for multimedia search interfaces and recommendations or guidance to media analysts on their approach to using the interface to assess multimedia for target events of interest.

Based on the NIST definitions of the Events (see Appendix A), The Event Types with semantic definitions that include objects include the following Events:

- E007 – Changing a Vehicle Tire
- E009 – Getting a Vehicle Unstuck
- E010 – Grooming an Animal
- E011 – Making a Sandwich
- E014 – Repairing an Appliance
- E015 – Sewing Project

The Events with semantic definitions that do not necessarily include objects include the following Events:

- E006 – Birthday Party
- E008 – Flash Mob
- E012 – Parade
- E013 – Parkour

Another more specific question about differences of behavior within subjects based on differences in Event Types is do subjects allocate their visual attention differently among the areas of interest (AOIs) based on the perceived difficulty of event type? The results of the post-test survey and post-test interviews identified four of the 10 event types as being more difficult to assess than the other six event types. Based on the results of those interviews, I identified the following Events as being the "Difficult" Event types:

- E006 – Birthday Party
- E008 – Flash Mob
- E009 – Getting Vehicle Unstuck
- E012 – Parade

I considered the remaining Events as the "LessDifficult" event types. Comparing eye movements of subjects across and within the Difficult Event types and the LessDifficult event types will determine if any consistent differences in eye movement patterns result from the difficulty of the event type – and perhaps what factors contribute to the difficulty of the event type (i.e., completion of an event to a certain state, such as confirming that a vehicle was unstuck, versus a subtle distinction between a parade and a protest).

Hypotheses Statements

This section presents each of the hypotheses statements that this study intends to address to answer the questions raised by the human performance baseline study.

Influence of Allocation of Visual Attention on Accuracy and Throughput.

To understand how subjects' use of the two visual components (thumbnails and video playback) of the search interface affect their accuracy and throughput, the researcher used eye-tracking data to compare how subjects allocated their visual attention between the two components, and then reviewed for any correlation between measures of accuracy – in terms of correct identification of targets and false alarms – and throughput, measured in terms of number of videos in the set on which the subjects made a decision about the presence of a target.

Null hypothesis. Subjects will not display any difference in visual attention patterns in their use of the search interface that correlate to their level of accuracy. There is no correlation between patterns in the allocation of visual attention to areas within the search interface and the rate of accuracy as measured by Accurate Target Detection or False Alarms.

Hypothesis. Subjects who have high Accurate Target Detection rates or low rates of False Alarms will have different behaviors in how they allocate their visual attention in the search interface.

Null hypothesis. Subjects will not display any difference in visual attention patterns in their use of the search interface that correlate to their level of throughput.

There is no correlation between patterns in the allocation of visual attention to areas within the search interface and the number of videos processed by the subject.

Hypothesis. Subjects who have high throughput, as measured by the number of videos they process, will have different behaviors in how they allocate their visual attention in the search interface.

Influence of Objects in Semantic Definition of an Event on the Allocation of Visual Attention.

To determine if subjects allocate visual attention differently between the thumbnail and video playback components of the interface based on whether the semantic definition includes objects, I conducted a within-subjects, repeated measures analysis of variance to determine if subjects differed in visual attention between Objects Events and NoObject Events.

Null Hypothesis: There is no difference in how subjects allocate their visual attention among search interface AOIs between searching for Events with a semantic definition that includes specific objects (Objects) and searching for Events with a semantic definition that does include specific objects (NoObjects).

Hypothesis: Subjects will allocate their visual attention patterns differently among search interfaces when searching for Events with a semantic definition that includes specific objects (Objects) and searching for Events with a semantic definition that does not include specific objects (NoObjects).

Influence of Perceived Event Difficulty on the Allocation of Visual Attention.

To determine if subjects allocate visual attention differently between the thumbnail and video playback components of when they perceive the event to be more difficult to detect, I conducted a within-subjects, repeated measures analysis of variance to determine if subjects differed in how they allocated their visual attention between Difficult Event Types and LessDifficult Event types.

Null Hypothesis. There is no difference in how subjects allocate their visual attention among the search interface AOIs between searching for Events that they perceive as difficult to identify and Events that they perceive as less difficult to identify.

Hypothesis. Subjects will allocate their visual attention patterns differently among search interfaces when searching Events that they perceive as difficult to identify in the video and Events that they perceive as less difficult to identify in the video.

I can expect to see correlations between the overall success of the participants in terms of speed and accuracy, and patterns that emerge in the eye-tracking results and their visual attention allocation patterns in search and retrieval. I also expect to see differences between the perceived task complexity as reported by the subjects and their allocation of attention between results of each search test. The difference between the two interface tests may provide insights into what human capabilities and behaviors and what aspects of the interface and interaction influence the degree of success. The eye-tracking data will provide a level of granularity into the human cognitive processes and the influence of user interface features on the multimedia search process that is difficult to attain through traditional usability methods. This level of granularity in understanding the factors that influence visual attention, especially in terms of how they configure and use their multimedia player interfaces will help inform how automated MIR systems can best leverage human-in-the-loop interaction to support fast and accurate multimedia analysis.

Chapter 2: Literature Review

Multimedia Information Retrieval and Event Detection

Multimedia information retrieval is a complex, multimodal process that involves images, audio, and text retrieval and understanding. The technologies that address semantic retrieval of these various modes are not equally robust in their ability to return accurate results automatically. Despite ongoing efforts in video analytics and multimedia event detection challenges, studies still support the view that, when humans are fully engaged, they remain unsurpassed by machines in their ability for most visual search tasks (Eckstein, 2011). Tzoukerman et al. (2012) describe the challenges associated with semantic multimedia extraction, concluding that "the automatic segmentation and annotation of multimedia content is challenging, because most of the information that is of interest is not directly accessible, but has to be inferred using additional contextual information." A recent Video Browser Showdown evaluation found that, for some tasks, humans using a common video player could outperform the expert tools (Schoeffmann, et al., 2014). Several studies have demonstrated that humans can discern the gist, or general semantic category of a scene from the first fixation of the eyes, or in approximately 30 to 50 milliseconds (Henderson & Ferreira, 2004).

Thus, leveraging the human ability to quickly obtain gist from viewing an image or video surrogate is often a critical component of successful multimedia retrieval and is even more important for detecting events that have a semantic definition. Due to the basic functionality of the multimedia player used in the human performance baseline study – that is, no ability to enter any text query, but instead only visual processing of the thumbnails and video clips – the dynamics of visual scene processing applied significantly to this study. The next few sections of this literature review focus on the research of what influences how humans visually process scenes.

Visual Scene Processing

When viewing a natural scene, humans use both bottom-up and top-down approaches to quickly determine the "gist" of a scene. The bottom-up approach involves building patterns from visual signals and the top-down approach uses goal-driven, attentional processes to determine which visual information is relevant (Ware, 2008). During the bottom-up approach, the visual cognitive system translates images from the retina into low-level features, such as color, size, orientation, and motion. From these low-level features, the visual cognitive process then detects patterns that it translates into meaningful visual objects. The visual working memory can hold about three objects at one time (Ware, 2008). The top-down approach to the visual cognitive process refers to how a human directs attention to a visual scene. Humans have a relatively small window for foveal vision, or the clear signals detected from the center of the eye. This window of perception is roughly 2.6 degrees (Nelson & Loftus, 1980) to 4 degrees (Henderson, Williams, Castelhano, & Falk, 2003) from the center of pupil. The attentional aspect of the visual cognitive system determines where a human fixates his eyes to obtain the needed visual information, and thus is goal-driven. Viewers' eyes are drawn to important or informative aspects of a scene, and their goals (or tasks) in viewing a scene will influence what they determine is important to view (Yarbus, 1967; Rayner, 2009).

Humans can obtain the visual "gist" of a scene very quickly – within a single glance or fixation (Rayner, 2009; Castelhano & Henderson, 2007; Nelson & Loftus, 1980). Research studies have shown that humans perceive natural scenes through both bottom-up and top-down approaches. Findings from research conducted by Itti & Koch (2000) and Fine & Minnery (2009) reinforce the theory that saliency (which is typically defined by the low-level features of color, contrast, intensity, brightness, and spatial frequency) is a significant driver of scene perception and can influence spatial working memory. Henderson & Ferreira (2004) summarize the results of several studies to suggest that four cognitive factors influence where viewers look in a scene: 1) *short-term episodic scene knowledge*, or a person using recent scene knowledge to look for a specific object in a given location, 2) *long-term episodic scene knowledge*, such as a

person's familiarity with how his own kitchen is laid out so that he knows where to look for his coffee-maker, 3) *scene schema knowledge*, which refers to general knowledge of a semantically defined scene based on multiple episodes with those types of scenes, such as knowing what a city skyline might look like, and 4) *task knowledge*, which is how to move one's eyes to support a particular task or goal.

Objects and Global Scene Context for Determining Gist

One of the questions that this study aims to address is whether subjects allocate their visual attention differently between static thumbnail images and video playback when searching videos for events that include an object in their semantic definition than when the event does not include an object, or that is defined more by the overall scene than by a specific object. Several research studies have explored the potential influence of objects on visual attention in scene perception versus the overall perceptual characteristics of the scene, such as openness, naturalness, urban versus country, etc. Henderson & Hollingworth (1999) concluded that the first few fixations of a scene are controlled more by the global semantic characteristics of the scene than by the local scene regions, or objects. Oliva and Torrelba (2001) describe a "spatial envelope" model and conclude that semantic categorization of a scene is not dependent upon specific information about object shape or identity but instead can result from a more holistic representation of the scene. Henderson & Ferreira (2004) summarize several studies to conclude that initial fixation placement in a *complex true scene* – that is, a photograph or representation of a real-word scene as opposed to a sketch or other ersatz scene – is influenced by the visual properties of objects, but not by the semantic properties of objects. The results of other studies indicate that humans gain a significant amount of perceptual recognition from both objects within a scene and the overall global context of a scene. Research findings by Oliva et al. (2003) conclude that computational models based on contextual analysis of scene perception resembled human patterns more than saliency models did, concluding that the contextual information of a scene provides an essential shortcut for object detection systems.

Malcolm, Rattinger, & Shomstein (2016) found that the semantic influences of an object can bias visual attention, even if the object is not relevant to the visual search task or even if more predictive influences are present. They recommend that future models of visual attention consider the semantic properties of objects. Castelhano & Witherspoon (2016) found that the search for objects was faster for objects that were in a function-congruent location – that is, a location where a user would expect to find the object in the scene based on how they use the object. Their findings suggest that the knowledge of an object can guide where a user directs visual attention in a scene. In a similar study, Ahmad (2016) confirmed his hypothesis that the associative relevance – that is, the process of associating relevance among entities within a scene and the context of the scene – facilitate the decision-making process by allowing eye fixations to follow the optimal route for human viewing to support cognitive processes.

Belke, Humphreys, Watson, Meyer, & Telling (2008) found that, during a search for a specified object, semantic knowledge about the target is triggered and modulates visual selection. However, this semantic knowledge does not seem to affect the perceptual workload of the working memory, reinforcing the view that humans can discern "gist" rapidly and with little cognitive workload. Van Rullen and Thorpe (2001) identified two distinct processing stages for processing visual information from a scene – the perceptual, task-independent stage and a task-related stage. Their findings indicate that the perception stage occurs very quickly (within 75-80 milliseconds), again reinforcing the notion of the ability to obtain the "gist" of a scene very quickly. Research has identified nine visual gist attributes that users can quickly grasp: 1) object, 2) people, 3) setting/environment, 4) action/activities/events, 5) theme/topic, 6) time/period, 7) geographical location, 8) plot, and 9) visual perception (Yang & Marchionini, 2005). Biederman (1972) found that jumbling a picture affected the ability of viewer to identify a single object, which suggests that the context of the object overall, and not only low-level features, influence perceptual cognition. Neider & Zelinsky (2006) and Torralba, Oliva, Castelhano, & Henderson (2007) found that scene context determines where viewers will direct their attention when conducting a visual search of a scene. Neider &

Zelinsky (2006) determined that viewers can quickly adopt a “look everywhere” strategy and use scene context (although not with strict adherence) to guide their visual search. When targets are in expected areas in a scene, viewers can locate an object as much as 19 percent faster (Neider & Zelinsky, 2006).

Research conducted by Malcolm, Groen, and Baker (2016) demonstrated that different goals use similar properties of a scene, and conversely, that many scene properties can support different goals. Thus, they recommend a more comprehensive framework for scene understanding and conduct experiments that measure the use of multiple scene properties supporting multiple tasks. De Groot, Huettig, & Olivers (2016) suggested that visual and semantic representations and objects occur independently from each other. Specifically, they reported that the substantial bias toward semantically related objects in a visual scene occurs later than other biases toward visually matching objects. They also reported that the semantic bias appears to evolve independently of any visual bias from initial scene recognition. Hwang, Wang, & Pomplun (2011) also argued for the development of a multi-level model of the control of visual attention. They suggest that it is likely that observers start their visual search using the guidance of a generic visual template based on a verbal description of the target (like the written description of the target events in this study). They suggest that this mentally-created template may dominate the search process until the mental template fails to detect the target or until semantic context information becomes cognitively available, or both occur.

These studies suggest that many different factors contribute to cognitive and physical processes of visual scene processing and to the difficulty of establishing models that effectively define the factors that influence the process. The results of these studies suggest it might be difficult to detect whether significant differences occur in how subjects allocate their visual attention when searching for semantically defined events with objects or without objects, or for between targets perceived as difficult to identify and targets perceived as less difficult to identify.

Use of Visual Surrogates in Multimedia Event Detection

When searching for video or multimedia, humans can use their ability to quickly grasp the visual gist of a scene to determine the relevance of a video clip based on a visual surrogate (i.e., a static thumbnail) of the video. The ability of humans to quickly identify a visually relevant image remains far superior to any automated algorithms developed to date.

How a system creates and presents surrogates influences its effectiveness for users. Videos are multimodal in their signals (audio, text, images), therefore video search and retrieval systems should determine relevance based on the different modalities. Also, search is a very task-dependent process (Järvelin & Ingwerson, 2004; Wilson, Kules, Schraefel, & Schneiderman, 2010; Hearst, 2011), thus a system's interface should consider different information needs and use based on the task. Thumbnail images may provide immediate visual feedback on video content, but may not be descriptive of all the contents of a video. Salient stills (composite images that show temporal changes in a shot or a video) or selected key frames may provide more insight into a video's content, but may not be descriptive enough for an open-ended search for a concept such as "funny" or "humor" (Cunningham & Nichols).

Song & Marchioni (2007) found that users who were searching for relevant video clips made more effective use of, and strongly preferred, surrogates that use both audio and video over video-only or audio-only surrogates. They found that spoken descriptions lead to better understanding than visual storyboards alone, but that people like to have visual surrogates to confirm interpretations and to add context. A similar study involving the use of eye-tracking data found that participants looked at and fixated on titles and text descriptions that describe the video clips more often than on the thumbnail images of the clips, but that they used the images to confirm which video clips were relevant to their search query (Hughes, Wilkens, Widemuth, & Marchionini, 2003). If a user selects a relevant image, the system should use the associated text from audio processing, metadata, or surrounding text as additional input to refine the overall search results. (Mei, Yang, Hua, & Li, 2011). The results of these studies on the use of surrogates suggest that

the data may indicate a preference in subjects' visual attention toward the use of thumbnails over the use of video, except perhaps when sound would be a strong factor in event detection (e.g., hearing "happy birthday" being sung at a gathering of people or a type of music that often accompanies parkour).

Designers of MIR systems should leverage the unique human ability to quickly grasp visual gist when designing interfaces and system interaction. Schneiderman's (1996) principle of "overview first, zoom and filter, then details-on-demand" applies to the interface and interaction design of MIR search results. Much research has demonstrated that browsing is a very important task in the review of MIR search results (Amir, Berg, & Permuter, 2005; Cunningham & Nichols, 2008; Hauptmann, Lin, Yan, Yang & Chen, 2006) and the search results display should enable users to obtain a visual overview first, and provide the users with the control to zoom and filter, and then obtain details on demand according their specific information needs and what they determine is relevant. Interactive system performance strongly correlates with a system's ability to allow the user to efficiently survey many candidate video clips or representations to find the relevant ones (Hauptmann, Lin, Yan, Yang & Chen, 2006). Interfaces should include a range of visual surrogates that afford users several cues to build visual gist. Visual surrogates must be easily viewable without consuming the entire screen. Providing an adjustable image size can help users if the controls are intuitive and require low effort (Song & Marchionini, 2007).

> If you have a lot of graphics for users to see at one time, thumbnails should be your choice. They give users an overview of all the visual content you want to present without requiring them to open all of the files. This saves users download time and uses up less Internet bandwidth. (Xu, 2012)

A search results display interface should integrate browsing and playback capabilities, enabling the user to play a portion of the video directly from the surrogate presented in the summary view (Lee & Smeaton, 2002). Because time is an important factor in video content, playback interaction should provide at least a time bar. Other

methods for indicating time include indicating small quantifiable objects in a video abstraction or visualizing a time length as the depth of each surrogate view (Lee & Smeaton, 2002). However, in following Nielsen & Molich's (1990) guideline of a "simple and natural dialogue," using the interface that most users understand (i.e., the time slider), is the most intuitive approach (Lee & Smeaton, 2002).

Playback capabilities should consider the audio track as well. Coutrot, Guyader, Ionescu, & Caplier (2012) found that gaze is impacted by the related soundtrack, even without spatial auditory information. Their results indicate that adding sound to the visual saliency models might improve overall efficiency of video searching. Defining the audio equivalent of a thumbnail or key frame is challenging because humans can only listen to one audio stream at a time. However, speech can be sped up to a factor of 4, while preserving the ability of the user to understand the content. Time-scale modification research improvements have led to the ability to modify long sounds such as vowels or quiet portions of the speech to play back faster (Ponceléon & Slaney, 2011, p. 613). However, because humans are much more efficient at browsing visual content than audio content, applying speech transcription and showing the transcribed text in context with the relevant key frames is sometimes more effective than providing separate audio playback.

With MIR, no single interface meets all user needs. Because different users have different information-seeking needs, which result in the use of different information seeking strategies, designers of user interfaces for video retrieval systems should consider the varying intentions and needs of different users and the characteristics of the repository or scope of content the system is retrieving (Järvelin & Ingwerson, 2004; Wilson, Schraefel, & White, 2009; Hearst, 2011). A recent review of video browsing interfaces and applications summarized that the user interfaces of video browsing applications are very diverse, as are the methods for indexing and analyzing the video content (Schoeffmann, Hopfgartner, Marques, Boeszoermenyi, & Jose, 2010). This review, however, identified several heuristics that apply across multimedia systems:

- Leverage the multimodal aspects of multimedia by enabling users to search using text, images, and audio content.
- Leverage the human ability to browse visual objects and scenes quickly by providing a way to browse retrieved results visually; then support drill-down into a specific video clip as needed.
- Provide user flexibility in the interface to move between the multimodal search and browse capabilities. User interfaces "should provide a way to switch between different features flexibly. … Being able to customize the interface, the provision of options, and ultimately the intelligent provision of the appropriate interfaces to the user performing a particular task is ideal" (Lee & Smeaton, 2002, p.14).
- Provide users with control over the experience through playback controls that allow users to rewind, fast forward, keyframe zoom, and use a timeline to understand the temporal context of a video.
- Enable users to select from explicit relevance feedback options.
- Support multi-level browsing capabilities of videos clips, segments, and individual shots.

Several factors account for the increased complexity of MIR. One key reason is that multimedia is multimodal in terms of the content available as data for indexing and retrieval. Multimedia, which includes video, can include audio, images, embedded text, surrounding text, and metadata associated with an item. Humans process audio, video, and text in different ways using different sensory and cognitive capabilities. Therefore, an effective MIR system needs to provide an interface and interaction design to allow users to leverage each of these modalities per how the human cognitive system best processes the information.

Chapter 3: Methodology

Subjects

For the baseline study, I recruited 20 subjects (11 female) from existing university email distribution lists and colleagues. I interviewed candidate subjects to determine if they met the criteria for the test, which were that subjects must have completed one year of college and must have ongoing experience in viewing multimedia (videos) on the Internet. The ages of subjects ranged from 20 to over 50 years, with 14 of the subjects being older than 25 years. I also screened the subjects based on their amount of time spent viewing videos each week (we sought subjects who view videos via the web at least 1-2 hours per week) and screened for whether they wore glasses (specifically, bifocals) that might interfere with the eye-tracking recordings.

To assign the subjects to the tests, I used a Latin Square design to randomize the assignment of the test sequence to create a fully within-subject experiment using two different data sets: randomly selected and ordered videos, and pre-filtered and rank-ordered videos. Table 1 provides the Event Test order for the subjects whose data I used for this analysis of the allocation of visual attention to video player interface components.

Table 1

Order of Event Tests for Each Subject

Subject	Gender	Age	Event Test Order									
			1	2	3	4	5	6	7	8	9	10
P01	Male	23-	6	7	10	8	14	11	12	15	13	9
P02	Female	23-	7	8	6	14	10	12	13	11	9	15
P03	Male	23-	8	14	7	10	6	13	9	12	15	11
P05	Male	25+	10	6	14	7	8	15	11	9	12	13
P06	Female	25+	9	13	15	12	11	14	8	10	7	6
P07	Female	25+	10	14	6	8	7	15	9	11	13	12
P12	Female	25+	12	13	9	6	15	7	8	11	14	10
P13	Female	23-	13	9	12	15	11	8	14	7	10	6
P17	Female	25+	15	9	11	13	12	10	14	6	8	7
P18	Male	25+	11	15	12	9	13	6	10	7	14	8
P20	Female	23-	13	12	9	11	15	8	7	14	6	10

Red numbers indicate Rank-Ordered Tests; shaded cells indicate no eye-tracking data collected. NOTE: Although the 20 participants of the human baseline test had participants in the age range of 23-25, all participants with eye-tracking data were in the younger than 23 (23-) or older than 25 (25+) range. No participants with eye-tracking data fell in the 23-25 age range.

Sample Size and Statistical Power

Due to schedule and equipment limitations, I could collect eye-tracking data on only 112 of the 200 tests conducted for the baseline study of accuracy and throughput. To ensure that the eye-tracking data could support both within-subject and between-subject analysis, the analysis included eye-tracking data only from participants who had at least seven of the tests recorded using the eye-tracking sensors. Table 1 also shows the participants and tests that were recorded with the eye-tracking sensor. The shaded boxes indicate the tests that did not obtain a successful eye-tracking recording; all other tests had successful eye-tracking recordings for a total of 96 tests. Eleven participants had seven or more tests with successful eye-tracking recordings; eight of the 11 had nine or more successful eye-tracking recordings.

For each of the ANOVA comparisons, 10 subjects had enough tests in both levels of each of the two-factor analyses (e.g., Object Events versus NoObject Events, Difficult Events versus LessDifficult Events) to provide reasonable comparisons of the effect. Gravetter & Wallnau (2014) define the power of a test as "the probability that the test will correctly reject a false null hypothesis" or the "probability that the test will identify a treatment effect if one really exists." This dissertation study has three null hypotheses statements that the results will either reject or fail to reject. The results should address the probability of rejecting a false null hypothesis – that is, to fail to identify a valid difference in the comparison, as recommended by Sauro & Lewis (2012). For the hypotheses, I used a combination of correlations across all the individual tests and, where relevant, repeated measures, within-subject ANOVAs to determine if I can reject the null hypothesis. For both types of analysis, I used a probability value of $p < .05$, two-tailed, as the threshold for indicating a significant effect or correlation. Although this research study does have hypothesis statements, I am not assessing a specific treatment, or comparing against a benchmark set of metrics, but instead I am identifying potential patterns of visual attention behavior that should be confirmed through additional, targeted research. The results of a smaller sample size have relevance for this purpose.

For hypotheses tested with correlations, however, the small sample size is not as significant a concern. The overall purpose of this study is to identify potential patterns of eye-movement behaviors when searching multimedia to detect semantically defined events. As such, the study is formative in nature, and not as dependent upon statistical certainty to indicate areas for further investigation or consideration. The correlations I used to assess hypotheses statements and to identify other relevant results compare all tests with valid eye-tracking data to each other in terms of accuracy, false alarms, and the metrics of the allocation of visual attention, so the sample size for these correlations is in the range of 96 and above. Because the purpose of a correlation is to identify whether a pattern exists, but does not prove causation, patterns identified in this study through strong correlations will need further study with larger sample sizes and control groups to

determine or validate causation. I provide more specific recommendations for further study to address the shortcomings of the small sample size in the Conclusions section.

Stimuli Source and Preparation

Stimuli Source

The baseline study used the NIST MED'11TEST video data, which is a collection of approximately 1,200 hours of Internet video consisting of 31,817 individual clips of publicly available, user-generated content posted to various Internet video hosting sites, as the stimuli. All the NIST video clips were in the MPEG-4 format. The Linguistic Data Consortium collected and annotated the clips to identify which events appear in the clips, and which clips have "near-miss" events (defined as content closely related to the event, but lacking critical evidence for a human to declare the event occurred), and which clips have "related" events. The baseline study also used the NIST MED'11 Training and Testing Event Kits, which include written definitions and descriptions of the event types and example video clips.

Stimuli Preparation

The baseline study assembled two types of tests: random-ordered and rank-ordered. The government agency researchers who requested the baseline study wanted to compare human performance on a completely random set of videos – that is, a set of videos with no previous machine-applied filtering or ranking – and a set of videos that a multimedia analysis and retrieval tool had filtered based on text descriptions applied to the video annotations. Multimedia analysts may have to identify semantically defined events under either of those two circumstances, so having each participant test on a set of random ordered video clips and rank-ordered clips would establish the comparison.

Each test set of both types had 1,000 video clips that included both target event clips and background clips selected from the larger NIST MED '11 Evaluation corpus of 31,817 annotated clips. In each of the 10 randomly ordered test sets, the baseline study specified the numbers of target and background clips based on the percentage of target

and background clips in the larger set, to preserve the original densities. The baseline study selected target clips and background clips using uniformly distributed random numbers to index the larger set. The baseline study randomly sorted the set of 1,000 clips sorted to create thirty-five permutations of random orderings so that each participant would have a different permutation.

To create the rank-order (or machine-filtered) tests, the baseline study applied combinations of generic classifiers to the full set of 31,817 annotated clips from the NIST MED corpus. The baseline study used the IBM Multimedia Analysis and Retrieval System (IMARS) to pre-filter and rank order the video clips for the test sets with a combination of classifiers that best described each of the specific MED ’11 evaluation events. Table 2 shows the mapping of the classifiers to the MED’11 evaluation events.

Table 2

MED’11 Events and Classifiers Used for Ranked Set Filtering

Event	Event Description	IMARS Classifiers for Ranked Sets
E006	Birthday party	Party U Group of People
E007	Changing a vehicle tire	Vehicles U Outdoors
E008	Flash mob gathering	Outdoors U Group of People U Urban Scene
E009	Getting a vehicle unstuck	Vehicles U Outdoors
E010	Grooming an animal	(Pet U Individual) appended with Indoors U Outdoors)
E011	Making a sandwich	Indoors U Outdoors
E012	Parade	Group of People
E013	Parkour	Urban Scene U Outdoors
E014	Repairing an appliance	Indoors U Individual

E015	Working on a sewing project	Indoors U Individual

(U: union)

The baseline study saved the top 1,000 video clips from the resultant ranked list as the test set. Although none of the combinations of classifiers is descriptive of its target event, for all but one (E010) of the target events the classifiers produced sets of clips that included instances of the target event. In these cases, the numbers of targets in the ranked sets exceeded the numbers of targets in the corresponding random sets, although target clips were not consistently ranked at the tops of their respective test sets. Table 3 provides the number of target events contained in each test set, random and ranked.

Table 3

Target event densities in random and rank-ordered test sets.

	E006	E007	E008	E009	E010	E011	E012	E013	E014	E015
Random	4	3	4	3	2	4	4	3	2	2
Ranked	12	4	29	8	0	9	28	10	3	3

The combinations of available classifiers were far too generic to yield a relevant ranking, so they primarily served to filter the larger set into a subset with a higher probability of containing target events. For example, for “Flash Mob Gathering” and “Parade” are both events that must involve a Group of People and applying that classifier helped to increase the target density.

Apparatus

Custom-Developed Windows Explorer/Media Player Application

To streamline the user tasks of reviewing videos and deciding whether a video clip contained the target Event, the baseline study used a custom application to support the video review, playback, and decision-making. This application combines the capabilities of Windows Explorer to provide thumbnail images and Windows Media Player for playing/skimming the videos. The application generated the thumbnail images using a "best frame" approach, which is to select the first non-black, non-solid color frame that occurs in the video clip.

In addition, the interface includes six buttons for indicating the subject's decision with a level of confidence associated with it. Once a subject decides, the application removes a video from the list of thumbnails and videos available for review. The application also aggregates structured data from each subject's session. Figure 1 shows a screen capture of the user interface of the custom Windows Explorer/Media Player application.

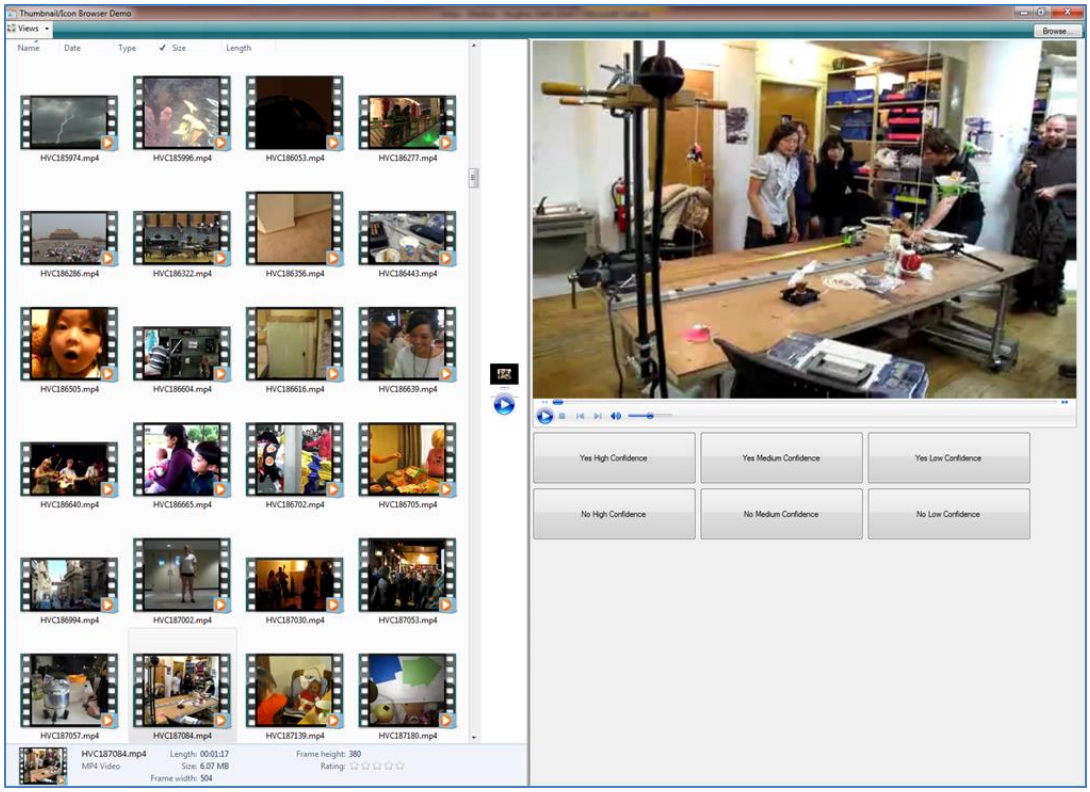

Figure 1: Custom Windows Explorer/Media Player application interface

The application also records and aggregates structured data from each subject's session into a comma-separated values (.csv) file. Figure 2 shows an example of the data that the application generated.

P05	P05_E006	20111117	123916		
P5_0000	FALSE	3	TRUE	0:50	00:04.9
P5_0001	FALSE	3	TRUE	4:39	00:03.9
P5_0004	FALSE	2	TRUE	0:15	00:04.4
P5_0005	FALSE	3	TRUE	0:57	00:03.6
P5_0006	FALSE	3	TRUE	1:47	00:03.1
P5_0002	FALSE	3	TRUE	1:28	00:03.5
P5_0003	FALSE	3	TRUE	8:35	00:04.3
P5_0011	FALSE	3	TRUE	0:58	00:03.7
P5_0012	FALSE	3	TRUE	6:00	00:02.2
P5_0013	FALSE	3	TRUE	1:47	00:07.7
P5_0014	FALSE	3	TRUE	0:40	00:03.7
P5_0017	FALSE	3	TRUE	2:24	00:08.2
P5_0020	FALSE	3	TRUE	0:36	00:02.4
P5_0021	FALSE	3	TRUE	3:01	00:02.1

Figure 2: Sample output data from the custom multimedia player application

The data recorded from each individual event test includes the following metrics:

- The top row shows, from left to right, the following information:
    - Participant number
    - Test number
    - Date
    - Start time
- The data columns for each decision include the following data
    - Video File Name – the name of the video file, customized to the specific subject test
    - Determination – user contributed data as to whether the subject believes the video clip contains the target event type (TRUE/FALSE)

- Confidence rating – user-contributed data as to how confident the subject is in his or her determination.
  - 1 = Low confidence
  - 2 = Medium confidence
  - 3 = High confidence
- Video Played – whether the subject played the video before making the determination (TRUE/FALSE)
- Length of video – the total length of the video file (Minutes/Seconds)
- Amount of video played – how long did the subject play the video (Minutes/Seconds/hundredths of seconds)

The baseline study used a trial version of the IBM Multimedia Analysis and Retrieval System (IMARS)[1] to create the rank-ordered data sets. The trial version provided 25 general classifier terms for indexing videos.

Eye-Tracking Sensor and Recording Software

To record the eye movements of participants during the baseline study, I used two Tobii X120 eye-tracking sensors and Tobii Studio Software version 2.3.0.0. The eye-tracking data that I collected during the baseline study is the source of the visual attention data that forms the basis of this dissertation research.

Procedure

Subjects conducted a total of 10 tests; five tests using the random data sets and five tests using the rank-ordered sets. The baseline study divided the 10 event types into two groups, Group A (E006-E010) and B (E011-E015). One-half of the subjects searched for events in Group A using the random data sets and Group B using the rank-ordered data sets. The other half of the subjects searched for events in Group A using the rank-ordered data sets and events in Group B using the random data sets. The baseline study randomized the order of the tests within each type of data set (random and rank-ordered)

[1] IBM developerWorks: IBM Multimedia Analysis and Retrieval System Communities, https://www.ibm.com/developerworks/mydeveloperworks/groups/service/html/communityview?communityUuid=7dc62548-8bc8-42c4-b2e9-150dde7c649a.

using a Latin Square formula. All subjects completed the random data set tests first and then conducted the rank-ordered tests. This approach provided a within-subject analysis of the data to compare the two types of video sets.

For each baseline study test, subjects had a test moderator who introduced the subject to the test event type, provided the subject with the sample videos for the event type, calibrated the subject on the eye-tracking sensor (if available for the test), timed the test to ensure that the subject did not exceed the maximum allowed time of two hours and to note the time of completion, and conducted the post-test interview. Test moderators also informed the subjects when they started the series of rank-ordered tests. The experiments were set up as double-blind: neither the subjects nor test moderators had access to or knowledge of the ground truth details, so they did not know how many targets and near-misses were in each of the data sets.

Test Preparation

Prior to the start of each test, the test moderator reviewed the specific test protocol with each subject, based on the specific event of the test and whether the videos were in random order or in rank order. Test moderators had two different scripts to use in preparing participants for a test; one script for random-ordered videos and one for rank-ordered videos. Subjects could read the NIST definition and description of the event before viewing the sample videos. Subjects also had ten minutes to review a set of videos that contained examples of the event. Subjects were also allowed to adjust the multimedia player interface to change the size of the thumbnail; subjects could choose between small, medium, and large thumbnail images. Subjects were also able to adjust the width of the thumbnail portion of the window. Subjects were not able to adjust the size of the video display, the controls, and the voting buttons; their size remained constant in the interface.

Test Protocol

Subjects conducted 10 tests, each test focusing on one event type. For each test, a subject had two hours to review a data set of 1,000 videos to identify the positives of one Event Type (e.g., Birthday Party). Subjects used the custom application to view the video thumbnail images and to select video clips to play/skim/review in a media player that

provided a slide bar and fast-forward/rewind capabilities. Subjects selected a video for review by clicking on the thumbnail image, which then displayed the static image in a larger window of the media player. The subject could decide whether the video contained the target event based on the thumbnail or static image in the video playback area alone, or after reviewing some, or all, of the video in the media player. Subjects were asked to decide as to whether each video clip contains the target event type for the session by clicking one of six buttons. The buttons included a confidence level associated with the Yes/No determination. Specifically, the buttons available were:

- Yes – High Confidence
- Yes – Medium Confidence
- Yes – Low Confidence
- No – High Confidence
- No – Medium Confidence
- No – Low Confidence

When the subject clicked the button to record the determination, the thumbnail of the video clip was moved into a "Viewed Videos" folder to indicate that the subject had already decided whether the video contained the target and would not need to review it again.

Multimedia Player Interface Errors

Occasionally users encountered a problem with the customized multimedia player interface that resulted from the video buffer not properly clearing. This problem would cause the player to either freeze during video playback or display a solid green line on a black background instead of playing the video. When this error occurred, subjects would notify the test moderator to stop the recording and timer, restart the media player, and resume the test. Fortunately, the multimedia player would save the users work up to that point. That is, the media player would keep track of which videos the user had already decided on and would resume the test with only the remaining videos that the user had not yet indicated whether they contained the target event. Although this occurrence would create multiple eye-tracking recordings for some tests, I was able edit the

recordings to remove any eye movements that occurred once the media player froze. I then aggregated the eye-tracking data across the recordings for the test, after editing.

Post-Test Survey and Interviews

After each test, subjects completed an online post-test survey to report their perception of task complexity and task load. I developed a modified version of the NASA Task Load Index to capture subjects' perceptions of the demands of the task, and their frustration, effort, and performance on the task. The NASA Task Load Index, often abbreviated as NASA-TLX, is a multi-dimensional scale created by NASA to obtain work-load estimates from operators while they are performing a task, or immediately after they completed a task (Hart, 2006). The NASA-TLX assesses task load on six dimensions: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration (Hart & Staveland, 1988). Appendix B contains the full set of questions that I developed for the post-test survey. The test moderators conducted a brief post-test interview to obtain the subjects' feedback on unique strategies or approaches they used for the specific event type and to clarify their responses to the Post-Test survey. Appendix B also contains the full set of post-test interview questions. This data indicates whether subjects found some event types to be easier or harder to search for and evaluate than others, and to learn whether (and how) the event type influenced their approach to the task.

Analysis

Data Preparation

Eye-Tracking Data. To extract the relevant eye-tracking data from the Tobii Studio recordings for analysis, I edited each recording using the following steps:

1. ***Scene Creation.*** For each eye-tracking recording, I created a scene to represent the complete video review process that the subject conducted during the test. The starting point for each scene was when the subject finished adjusting the multimedia player application interface. The ending point for each scene was one of three points: 1) when the subject finished reviewing all the videos, 2) when the subject reached the two-hour time limit for each test, or 3) when the subject

encountered an error with the multimedia player interface and needed to have the software restarted. In cases where subjects encountered an error with the multimedia player application, the test may have multiple eye-tracking recordings that I edited individually to extract the data, but then aggregated the data from the recordings into one set of numbers to represent the test for analysis.

2. ***Designate Areas of Interest (AOI).*** To assess the subject's eye movements within the multimedia player interface, I used the Tobii Studio Visualizations capabilities to create four Areas of Interest (AOIs) for each recording: 1) Thumbnail AOI, 2) Video Player AOI, 3) Video Controls AOI, and 4) Voting Buttons AOI. Figure 3 shows an example of how the four areas of interest overlay the multimedia player interface. As depicted in Figure 3, these AOIs did not overlap. The Thumbnail AOI was the only AOI that varied in size, because the Thumbnail viewing frame was the only portion of the interface that subjects could customize.

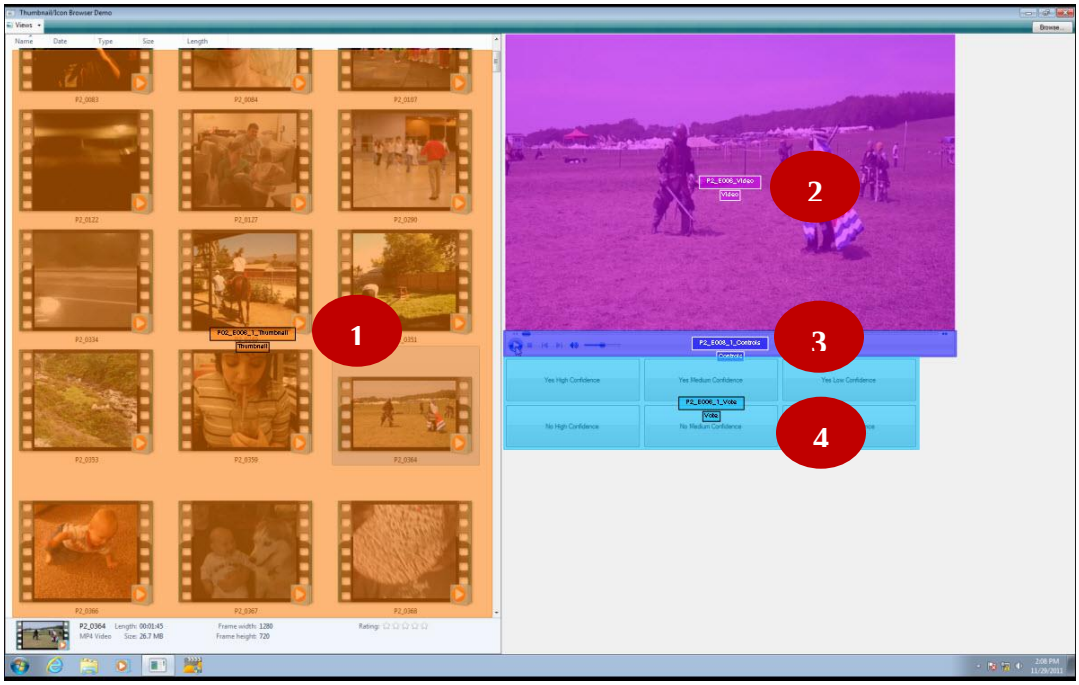

Figure 3: Overlay of Areas of Interest on Multimedia player Interface

3. ***Establish and Assign AOI Groups.*** Each AOI created in each test recording had its own unique name. To analyze AOIs across participants and events, I established a group for each of the four AOIs in the multimedia player interface, and assigned the individual AOIs to the appropriate group.
4. ***Export raw data.*** After I established the AOIs in a recording, I exported the raw data for each recording, to support any further detailed analysis of fixations and their sequence in relation to events, as needed to identify or substantiate eye movement patterns.
5. ***Generate AOI Statistics.*** I used the Tobii Studio Statistics capability to extract the AOI statistical data. I used the capability to generate and extract the following metrics for each AOI in each recording (Tobii Technology, 2010):
    - **Total Fixation Duration:** This metric measures the sum of the duration for all fixations within an AOI, thus the N value used to calculate the descriptive statistic is based on the number of recordings.
    - **Fixation Count:** This metric measures the number of times the participant fixates on an AOI or an AOI group.
    - **Total Visit Duration:** This metric measures the duration of all the visits within an AOI or AOI group. In this case the N value used to calculate descriptive statistics is based on the number of recordings. A visit is defined as the interval of time between the first fixation on the AOI and the next fixation outside the AOI. This measurement accounts for total saccade (i.e., eye movements) time in addition to fixation time.
    - **Visit Count:** This metric measures the number of visits within an AOI or AOI group. Each individual visit is defined as the interval of time between the first fixation on the AOI and the next fixation outside of the AOI.

Use of Ratios for Visual Attention Metrics. Because the duration of a test varied by subject, due to the search strategy that each subject developed as he or she progressed through the series of tests and whether the subject was motivated more by speed than by accuracy, I used the ratio of thumbnail AOI to video AOI to measure each of the four

visual allocation metrics of Total Fixation Duration, Fixation Count, Total Visit Duration, and Visit Count. Although I mapped four Areas of Interest for each of the recordings – Thumbnail AOI, Video AOI Video Controls AOI, and Voting Buttons AOI – the visual attention to the Video Controls AOI and the Vote Buttons AOI do not directly support the visual search process. The ratios of Thumbnail AOI use to Video AOI use provides a measure of how much time a subject spends in each AOI used for visual search relative to their use of the other AOI. The larger the ratio, the more visual attention allocated to the Thumbnail AOI (the numerator) relative to the visual attention allocated to the Video AOI (the denominator).

User Preferences Data. In addition to the eye-tracking data, I also reviewed each test recording to note the customizations, if any, that a user made to the thumbnail size and thumbnail display area of the multimedia player interface. I also noted any search behaviors related to the "triage" of the thumbnail images, such as setting some video clips aside to view at the end of the test, or patterns in playing video segments. I added this data to the subject data for each test so that I could identify correlations between user search interface preferences and behaviors with eye movements, accuracy, and throughput.

Analysis of Rates of Accuracy and Visual Attention Allocation

I used two methods to analyze the data to determine if subjects display any correlation between their levels of accuracy and their allocation of visual attention within the search interface. One method was to compare the accuracy rates with each of the four different visual attention metrics (Total Fixation Duration, Fixation Count, Total Visit Duration, and Visit Count) for each of the 100 individual tests to identify any correlations. The second approach was to compare the subjects' mean rate of accuracy across all tests with the mean ratio for each of the four visual attention metrics across all tests.

Analysis of Visual Attention Between Object Events and NoObject Events

I used a Repeated Measures Within-Subjects Analysis of Variance to determine if subjects allocate their visual attention differently when searching for Events with a

semantic definition that includes an Object (Object Events) from when they are searching for Events with a semantic definition that do not necessarily include an Object (NoObject Events). The Event Types with semantic definitions that include objects include the following Events:

- E007 – Changing a Vehicle Tire
- E009 – Getting a Vehicle Unstuck
- E010 – Grooming an Animal
- E011 – Making a Sandwich
- E014 – Repairing an Appliance
- E015 – Sewing Project

The Events with semantic definitions that do not necessarily include objects include the following Events:

- E006 – Birthday Party
- E008 – Flash Mob
- E012 – Parade
- E013 – Parkour

To prepare the data for the ANOVA, I took the mean of the ratios for each of the visual attention measures – Total Fixation Duration, Fixation Count, Total Visit Duration, and Visit Count – for the Object Events and for the NoObject Events. The ANOVA then compared those two means within subjects to determine if any significant differences exist.

Analysis of Visual Attention Between Difficult and LessDifficult Events

I also used a Repeated Measures Within-Subjects Analysis of Variance to determine if subjects allocate their visual attention differently when searching for Events that they perceived to be Difficult from when they are searching for Events that they perceive to be LessDifficult. I based whether a test was deemed Difficult or LessDifficult on the aggregated results of how subjects rated the difficulty of the test during the post-test survey and reaffirmed in the post-test interviews. The Event Types that subjects perceived to be Difficult to detect included:

- E006 – Birthday Party
- E008 – Flash Mob
- E009 – Getting a Vehicle Unstuck
- E012 – Parade

By default, that placed the other six Event Type into the LessDifficult set, which included:

- E007 – Changing a Vehicle Tire
- E010 – Grooming an Animal
- E011 – Making a Sandwich
- E013 – Parkour
- E014 – Repairing an Appliance
- E015 – Sewing Project

Chapter 4: Results

Influence of Allocation of Visual Attention on Accuracy and Throughput

Accurate Target Detections

The results of the comparison of the rates of accuracy and the allocation of visual attention across the set of the individual tests (that is, not grouped by subject or event type) did not identify any significant correlations between rates of accuracy and allocation of visual attention. Table 4 shows the Pearson correlation score (*r*), the significance value of the correlation (*p*), and the correlation direction and significance summary for the comparison of Accurate Target Detections and Visual Attention Metrics.

Table 4

Comparison of Accurate Target Detections and Visual Attention Metrics

Visual Attention Metric	$r(99) =$	$p =$	Correlation
Total Fixation Duration ratio	.06	.54	Positive, Not significant
Fixation Count ratio	.08	.43	Positive, Not significant
Total Visit Duration ratio	.06	.54	Positive, Not significant
Visit Count ratio	.07	.49	Positive, Not significant

Figures 4 through 7 show the scatterplot diagrams for each of the above correlations.

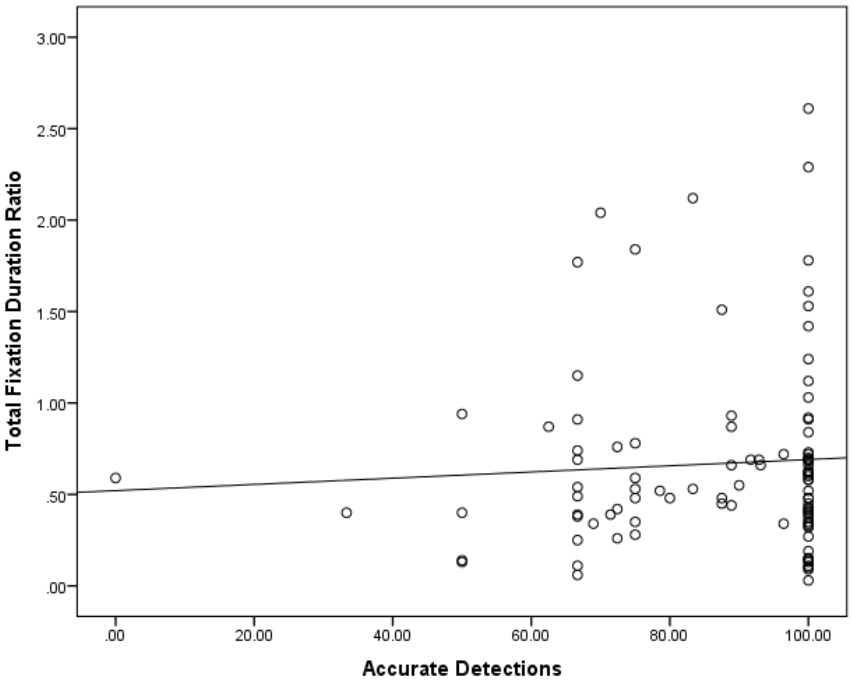

Figure 4: Correlation of Accurate Detections to Total Fixation Duration Ratio

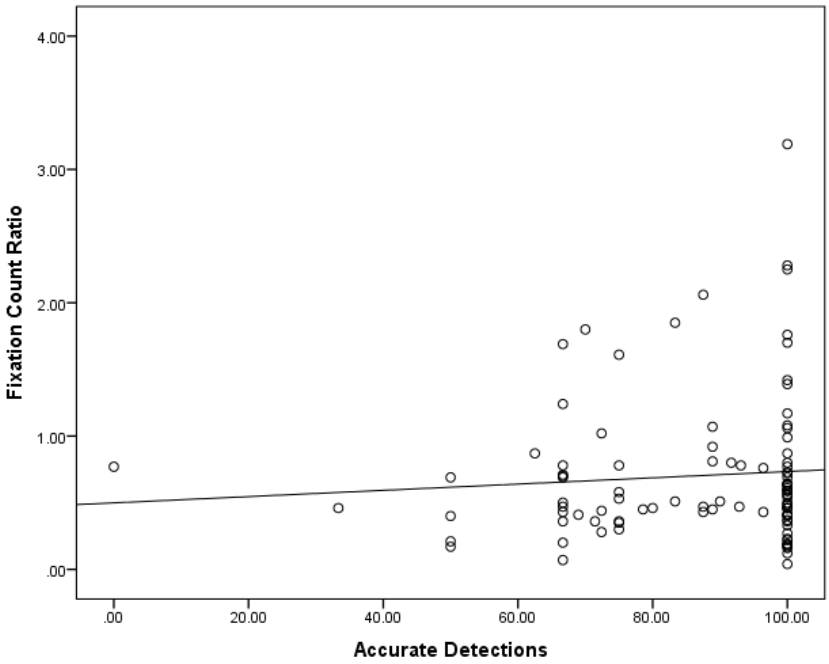

Figure 5: Correlation of Accurate Detections to Fixation Count

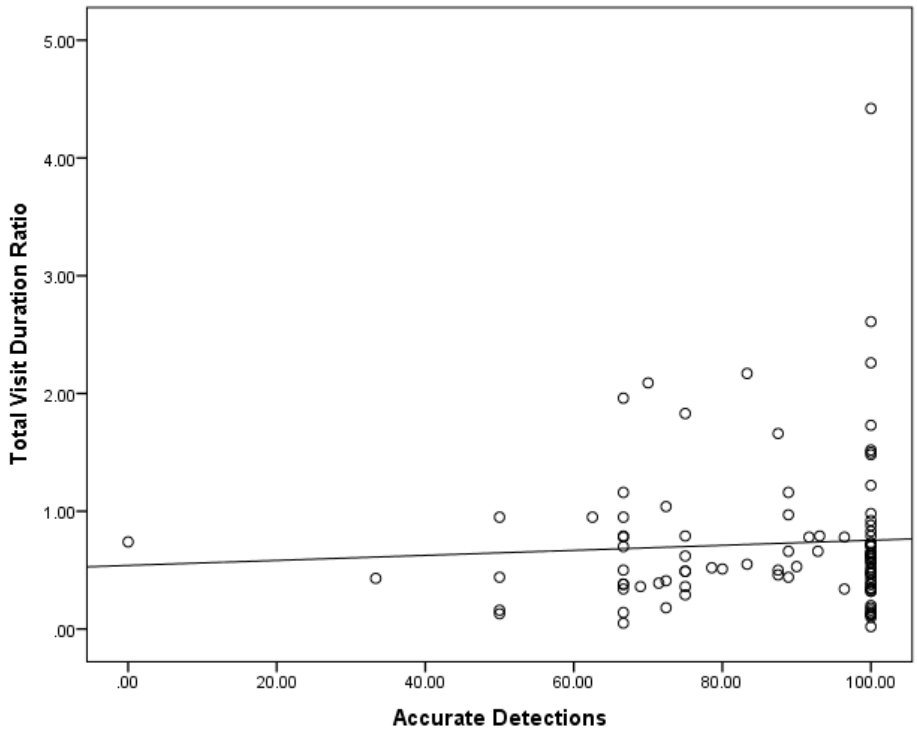

Figure 6: Correlation of Accurate Detections to Total Visit Duration Ratio

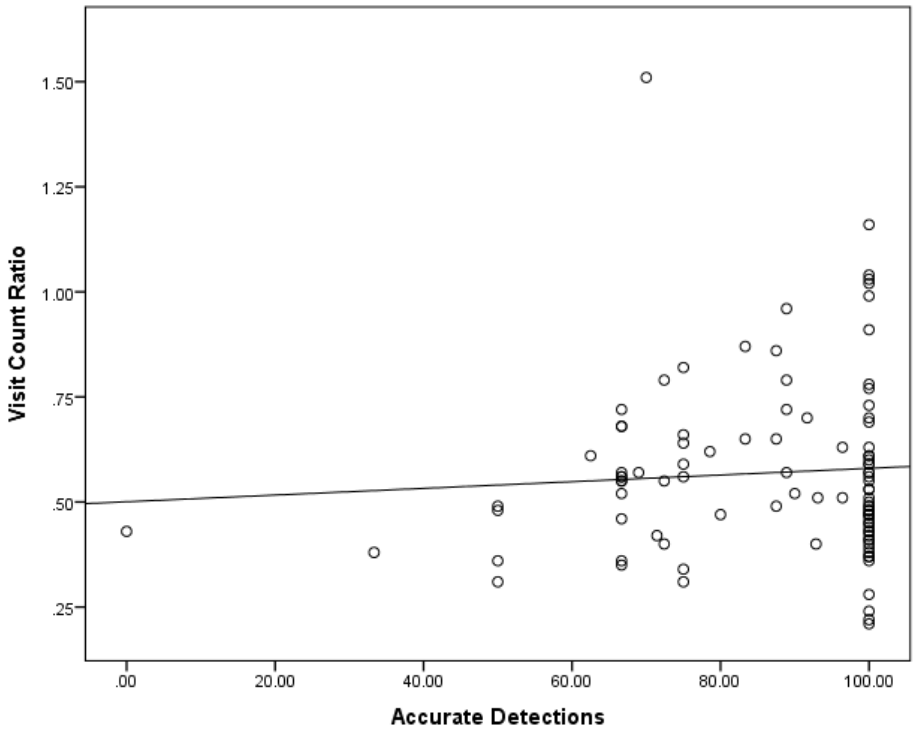

Figure 7: Correlation of Accurate Detections to Visit Count Ratio

The results of the second type of comparison, which was to compare the subjects' mean rate of accuracy across all tests with the mean ratio for each of the four visual attention metrics across all tests, also indicate that there are no significant correlations between eye movements as measured by Total Fixation Duration, Fixation Count, Total Visit Duration, and Total Visit Count within each area of interest (AOI) and the accuracy measures of Accurate Target Detections (Accuracy). Table 5 shows the Pearson correlation score (*r*), the significance value of the correlation (*p*), and the correlation direction and significance summary for the comparison of the subjects' mean Accurate Target Detections and Visual Attention Metrics for each of the 10 event tests.

Table 5

Comparison of Accurate Target Detections and Visual Attention Metrics

Visual Attention Metric	$r(9) =$	$p =$	Correlation
Total Fixation Duration ratio	-.12	.75	Negative, Not significant
Fixation Count ratio	-.08	.83	Negative, Not significant
Total Visit Duration ratio	-.18	.65	Negative, Not significant
Visit Count ratio	-.34	.34	Negative, Not significant

False Alarm Rates

The results of the comparison of the false alarm rates and the allocation of visual attention across the set of individual tests also did not reveal any significant correlations. Table 6 shows the Pearson correlation score (*r*), the significance value of the correlation (*p*), and the correlation direction and significance summary for the comparison of False Alarm Rates and Visual Attention Metrics.

Table 6

Comparison of False Alarm Rates and Visual Attention Metrics

Visual Attention Metric	$r(99) =$	$p =$	Correlation
Total Fixation Duration ratio	-.14	.17	Negative, Not significant
Fixation Count ratio	-.15	.14	Negative, Not significant
Total Visit Duration ratio	-.12	.54	Negative, Not significant
Visit Count ratio	.03	.79	Positive, Not significant

Figures 8 through 11 show the scatterplot diagrams of the above correlations.

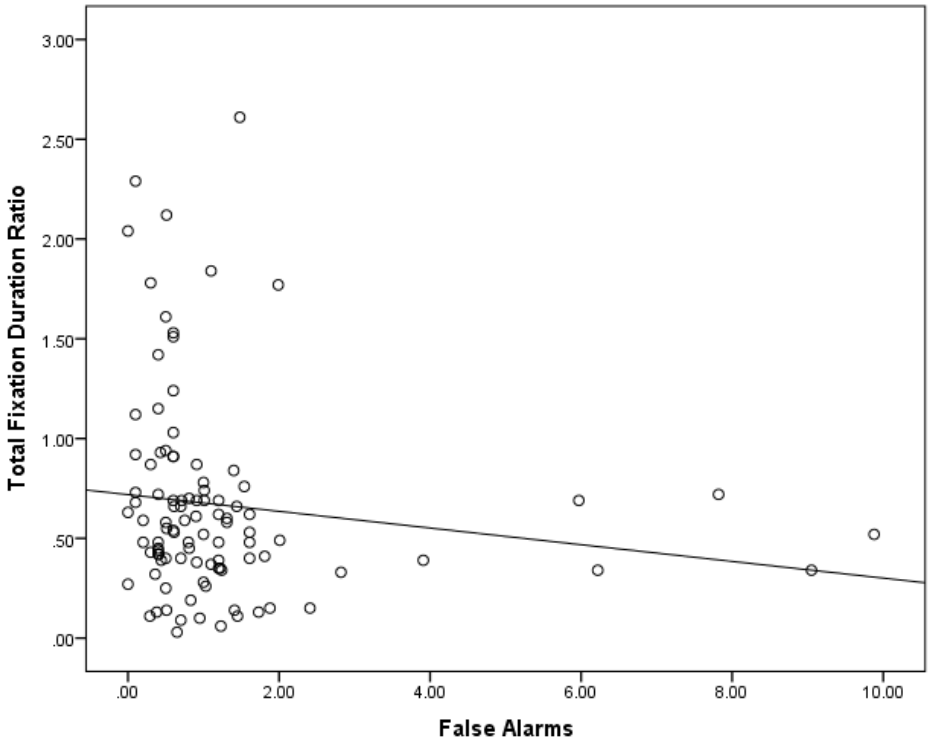

Figure 8: Correlation of False Alarms Rates to Total Fixation Duration Ratio

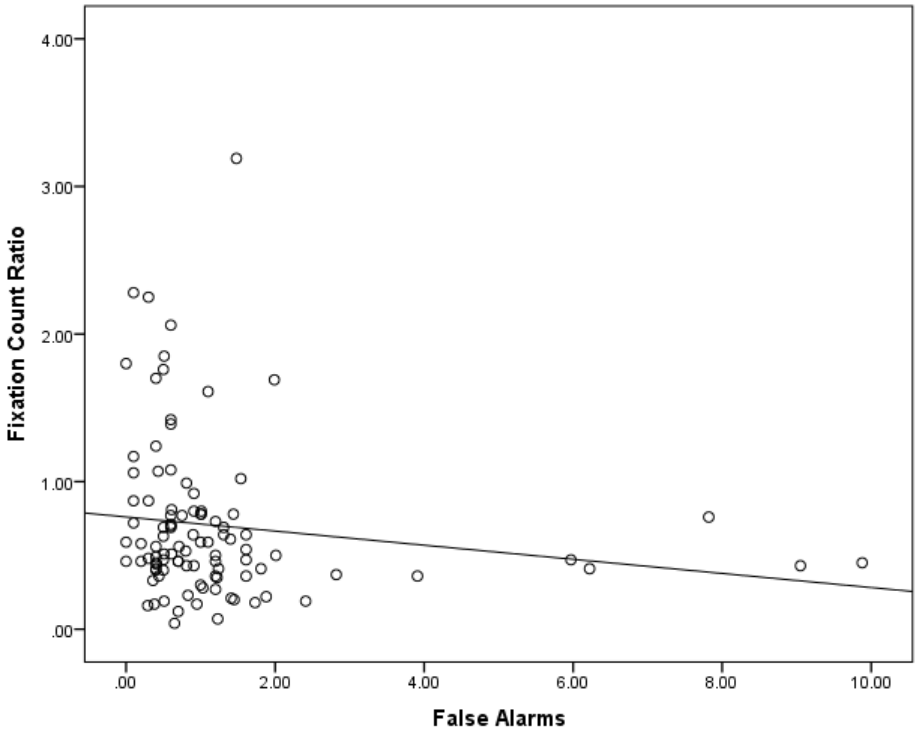

Figure 9: Correlation of False Alarms Rates to Fixation Count Ratio

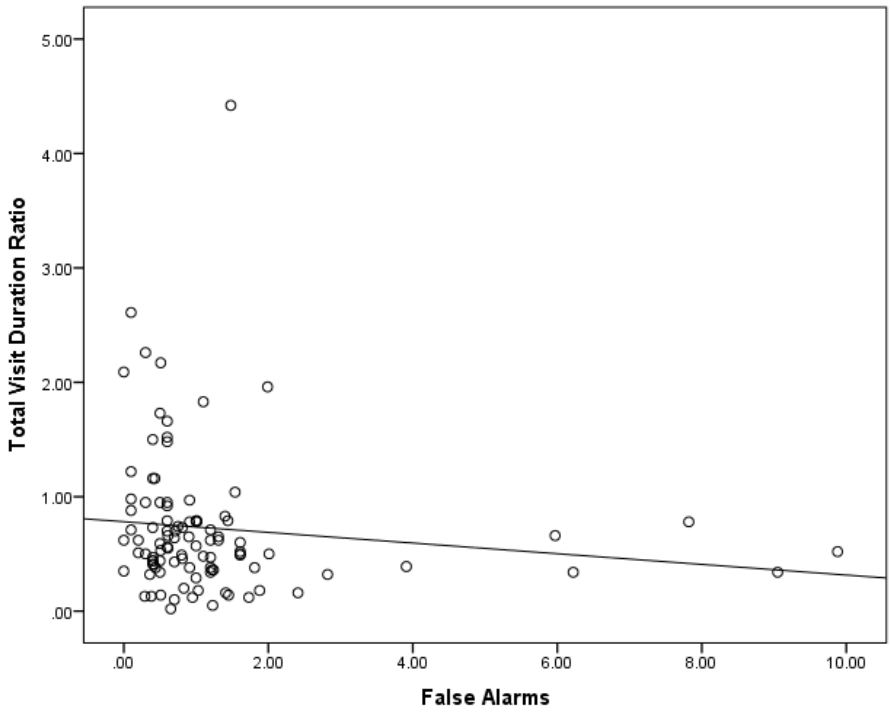

Figure 10: Correlation of False Alarms Rates to Total Visit Duration Ratio

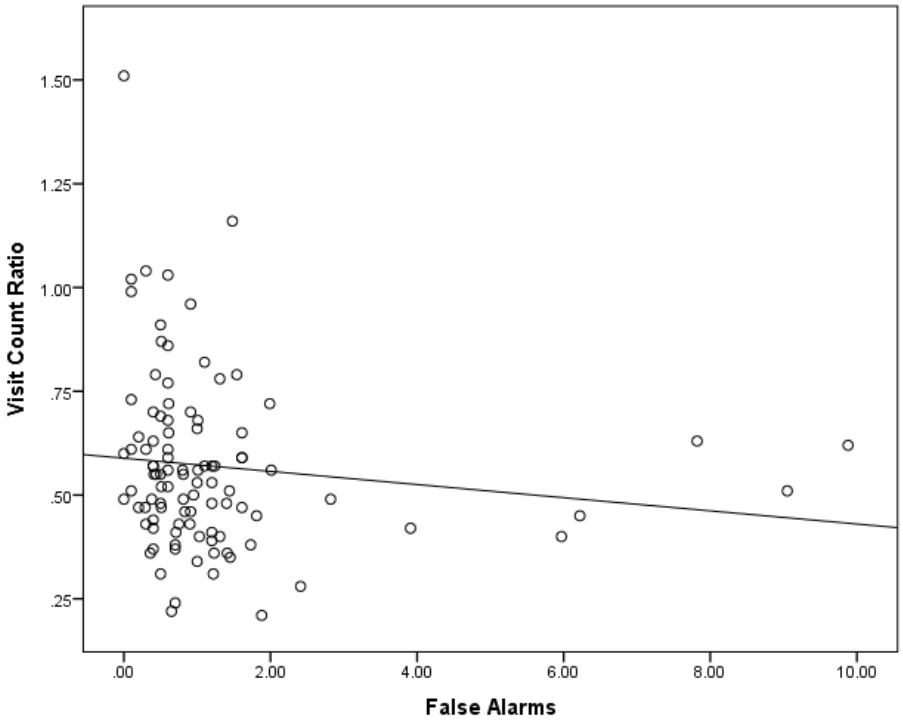

Figure 11: Correlation of False Alarms Rates to Visit Count Ratio

Table 7 shows the correlations of the subjects' mean rates of False Alarms across all tests and the mean ratio of each visual attention metric across all tests. These results also indicate no strong correlation exists between the rate of False Alarms and the allocation of visual attention.

Table 7

Correlation of False Alarm Rates to Visual Attention

Visual Attention Metric	$r(9) =$	$p =$	Correlation
Total Fixation Duration ratio	.34	.34	Positive, Not significant
Fixation Count ratio	.37	.30	Positive, Not significant
Total Visit Duration ratio	.33	.35	Positive, Not significant
Visit Count ratio	.18	.61	Positive, Not significant

Figures 12 through 15 show the scatterplot diagrams of these correlations.

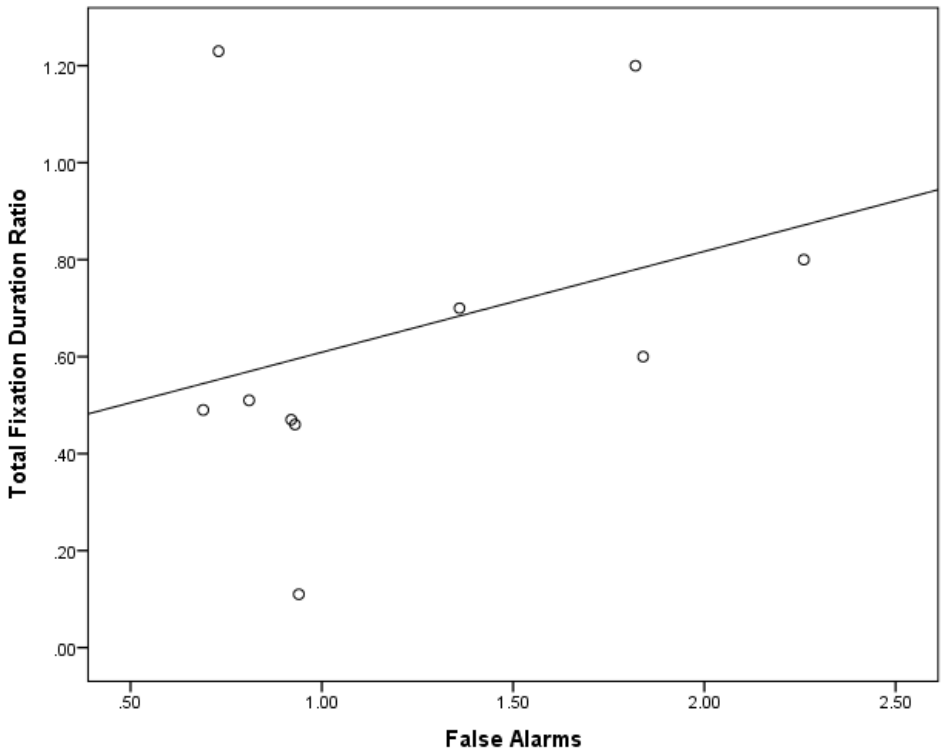

Figure 12: Correlation of Mean Rate of False Alarms to Mean Total Fixation Duration Ratio by Subject

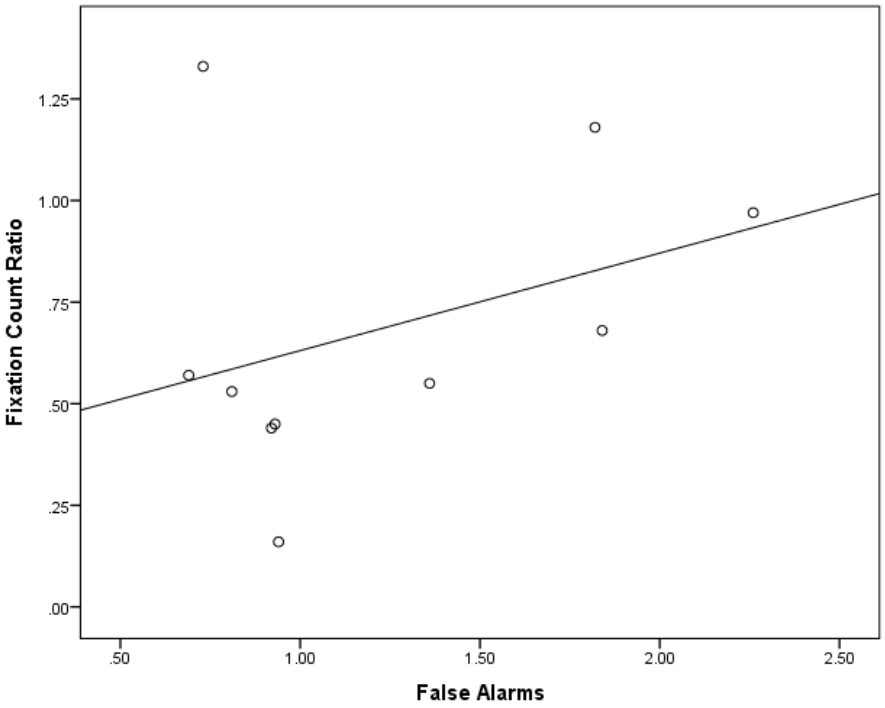

Figure 13: Correlation of Mean Rate of False Alarms to Mean Fixation Count Ratio

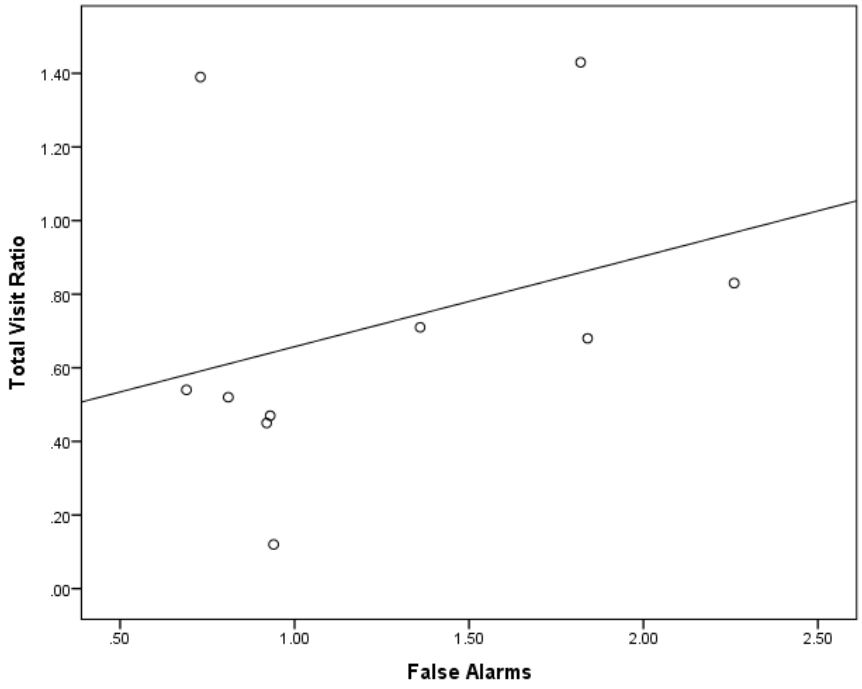

Figure 14: Correlation of Mean Rate of False Alarms to Mean Total Visit Duration Ratio

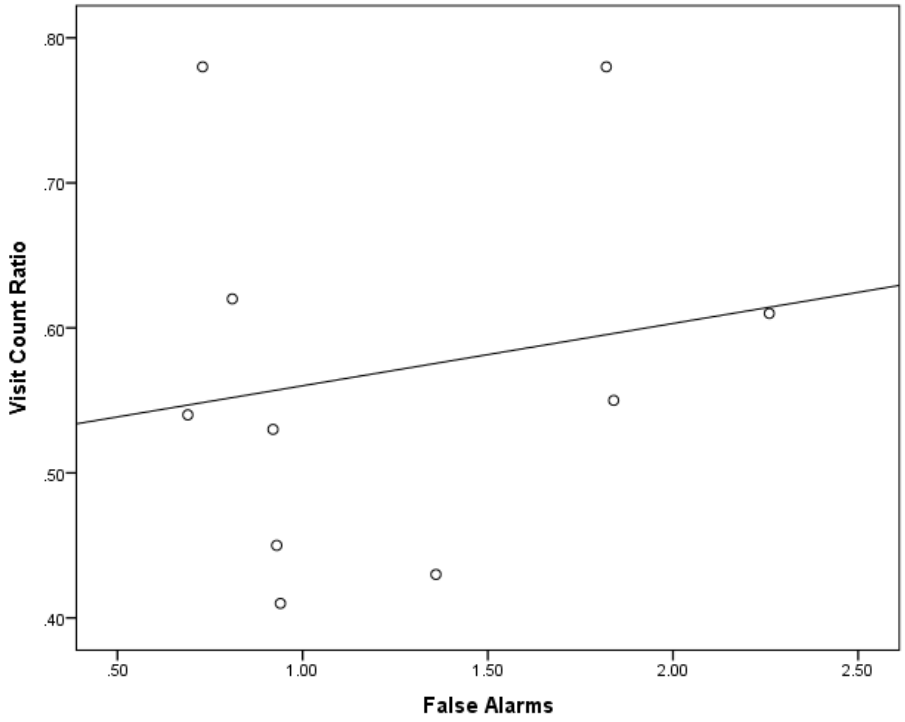

Figure 15: Correlation of Mean Rate of False Alarms to Mean Visit Count Ratio

The lack of any significant correlation between rates of accuracy and visual attention measures, both when assessing correlations across all individual tests and assessing correlations based on the subjects' mean scores for accuracy and visual attention, indicates that I can fail to reject the null hypothesis and conclude that there is no correlation, or strong relationship, between how subjects allocate their visual attention in the search interface and their rates of accuracy in identifying target videos with semantically defined events.

Additional Results: Accuracy and Event Type

The Repeated Measures, Within-Subject ANOVA suggests a moderately significant difference in Accurate Target Detections between NoObject Events and Object Events. NoObject Events have a lower mean of percentage of Accurate Target Detections accuracy (M=82.00, SD = 10) than Object Events (89.89, SD = 4.84), $F(1, 9) = 5.22$, $p = .048$, partial $\eta^2 = .37$.

For the False Alarms measure of accuracy, however, the Within-Subject ANOVA indicates a significant difference in the average rate of False Alarms between NoObject Events and Object Events. Subjects' mean NoObject scores were significantly higher (M=2.06, SD = 1.31) than Object Event false alarm rates (M = .68, SD = .20), $F(1,9)$ = 11.94, p = .007, $\eta^2 = .57$. These results indicate that when a semantic event includes an object as part of its definition, humans find it easier to identify targets than when the definition does not have an object.

Rates of Throughput and Allocation of Visual Attention

The results of the initial analysis of the human baseline study indicated that most subjects refined their search behaviors to improve their throughput of the videos as they progressed through their sequence of tests. That is, even if they did not process (i.e., vote to indicate whether a video contained the target event) all the videos during their first test, they adjusted their search behavior to process most, if not all, of the videos in each test. As Figure 16 indicates, only two subjects for whom I have eye-tracking data have a mean number of videos processed that was below 900 videos per test.

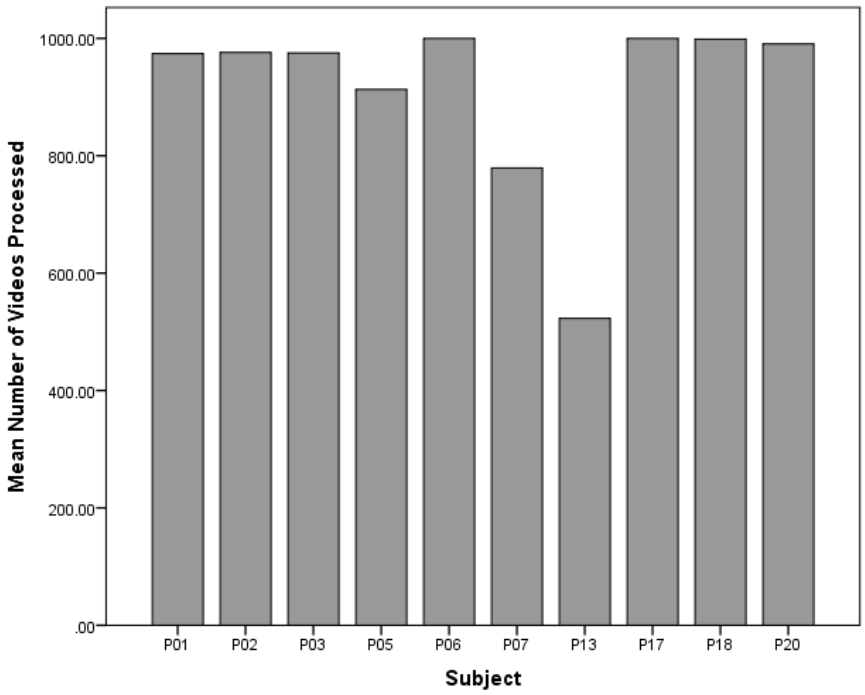

Figure 16: Mean Number of Videos per Test by Subject

Three of the four correlations related to throughput indicate a significant positive correlation between throughput and the ratios of the four Visual Attention measures. Table 7 shows the Pearson correlation score (*r*), the significance value of the correlation (*p*), and the correlation direction and significance summary for the comparison of the throughput, measured as the Number of Videos Processed, and Visual Attention Metrics.

Table 8

Correlations of Throughput and Allocation of Visual Attention

Visual Attention Metric	*r*(99) =	*p* =	Correlation
Total Fixation Duration ratio	.26	.01	Positive, Significant
Fixation Count ratio	.21	.03	Positive, Significant
Total Visit Duration ratio	.16	.12	Positive, Not significant
Visit Count ratio	.30	.00	Positive, Significant

The only correlation that is not significant is the correlation of Total Visit Duration ratio to the Number of Videos Processed, $r(99) = .16$, $p = .12$. Figures 17 through 20 show the scatterplot diagrams for these correlations.

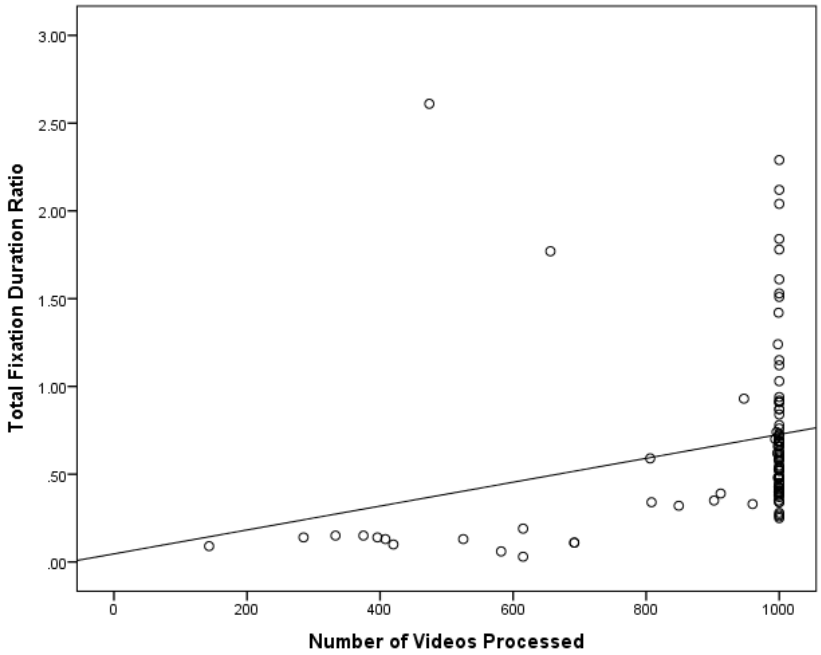

Figure 17: Correlation of Number of Videos Processed to Total Fixation Duration Ratio

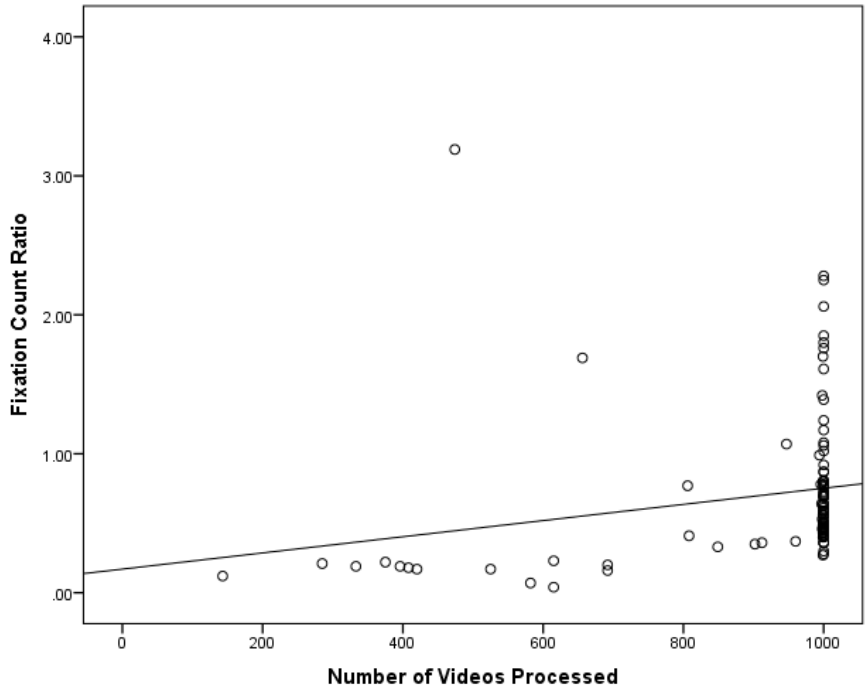

Figure 18: Correlation of Number of Videos Process to Fixation Count Ratio

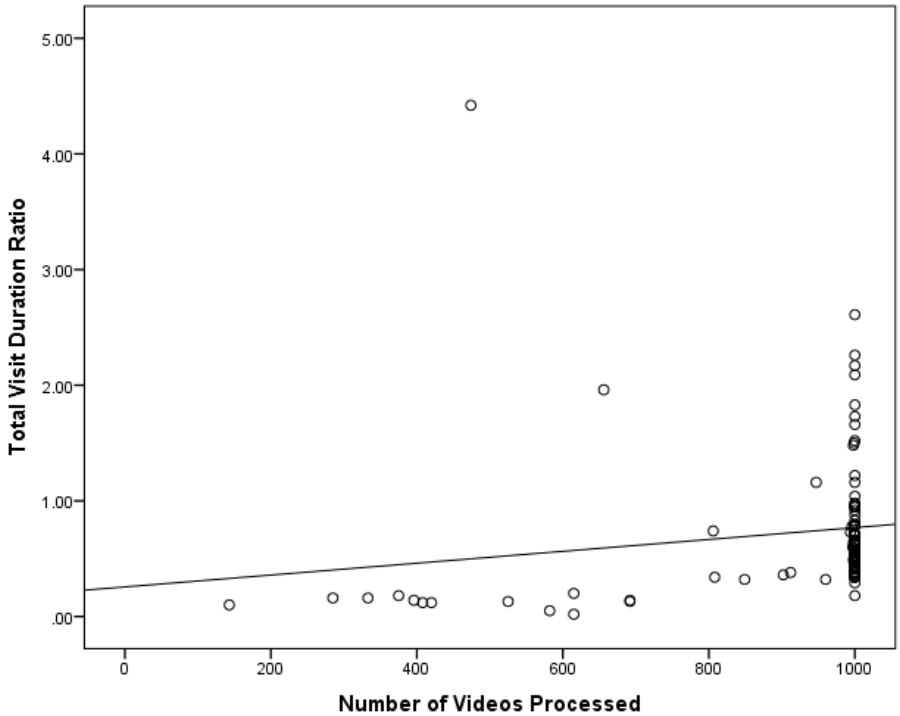

Figure 19: Correlation of Number of Videos Process to Total Visit Duration Ratio

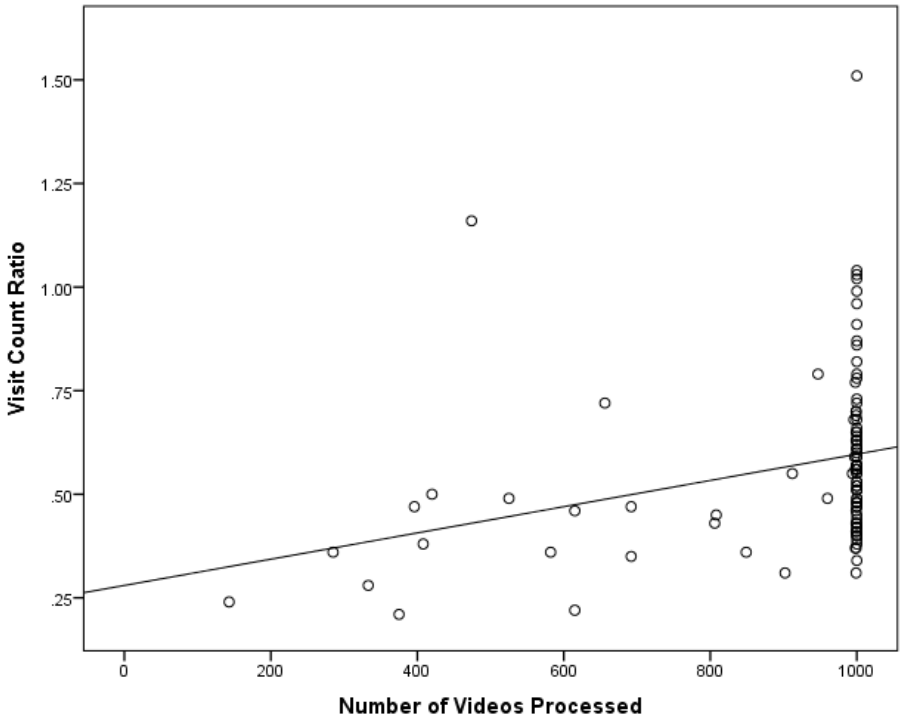

Figure 20: Correlation of Number of Videos Processed to Visit Count Ratio

Of interest, however, is Figure 21, which shows the mean ratio for each of the Visual Attention metrics for the full range of values for Number of Videos Processed during a test. This graph validates the correlation as it reflects a trend in that, as the Number of Videos Processed increases, the ratio increases, indicating more visual attention allocated to the Thumbnail AOI as compared to the Video AOI. I would expect to see this result based on research cited in the literature review on the preference to use static images to search for videos than to play through the videos.

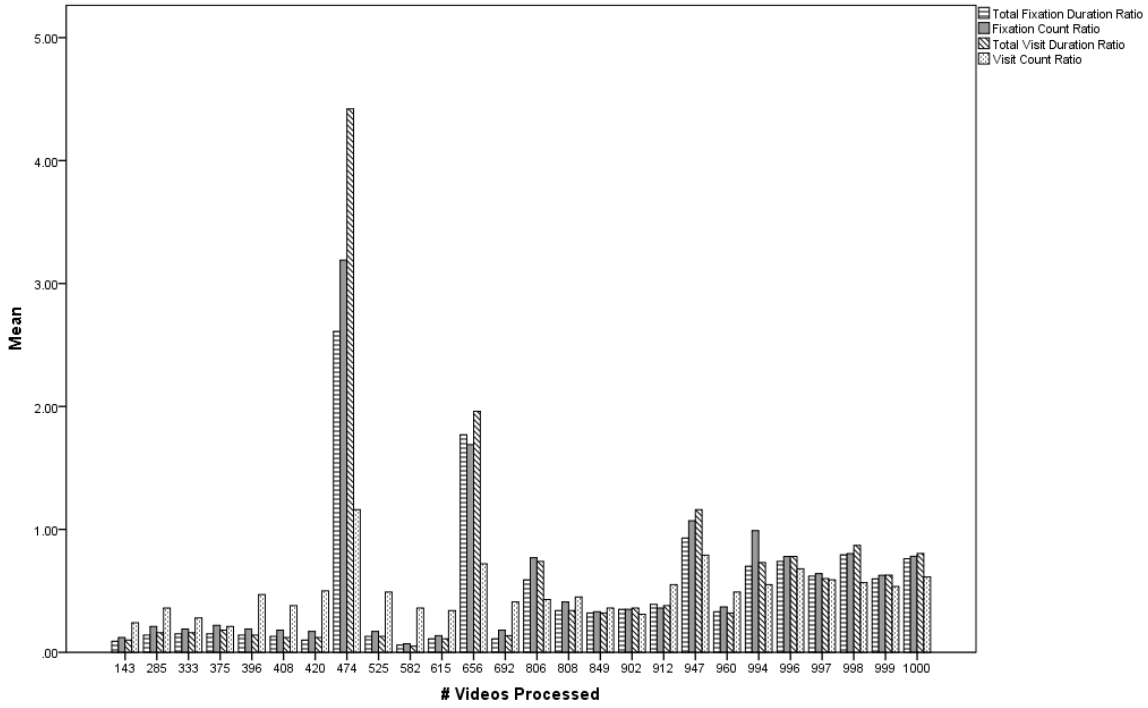

Figure 21: Mean Visual Attention Ratio by Number of Videos Processed

The chart, however, does have two obvious outliers at 474 Videos Processed per test and at 656 Videos Processed per test. These two instances are both Subject P05, and represent his first and second tests in the series of tests. Unlike most of the other subjects who allocated less time to the Thumbnail AOI as compared to the Video AOI in their early tests, and then gradually changed their allocation as the tests progressed, Subject P05 seemed to allocate more time to the Thumbnail AOI when he processed fewer videos. The Total Visit Duration ratio of Subject P05 is exceptionally high compared to the other ratios on both of those tests, which is an outlier from the rest of the data, which indicates that he spent a higher overall total of time in the Thumbnail AOI in early tests, as compared to the Total Visit Duration ratios across the remaining tests. When the correlations are calculated without the two outlier tests of Subject P05, the positive correlation of the Total Visit Duration ratio to the Number of Videos Processed becomes significant, $r(97) = .43$, $p = .00$. Figure 22 shows this scatterplot diagram of this revised, significant positive correlation.

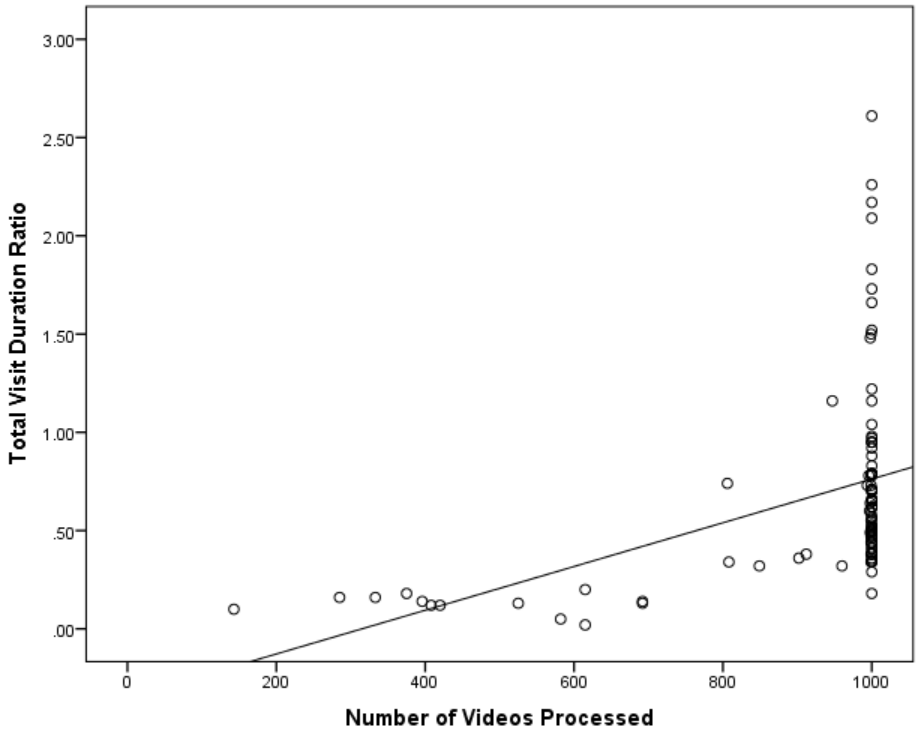

Figure 22: Correlation of Number of Videos Process to Total Visit Duration Ratio without Subject P05 Outliers

Influence of Objects in Semantic Definition of an Event on the Allocation of Visual Attention

The results of the Repeated Measures, Within-Subjects Analysis of Variance indicate that there was not a significant difference between the mean ratios of any of the visual attention measures for Object Events and NoObject Events. Table 8 shows the results of the ANOVA for each of the four visual attention metrics, including the mean and standard deviation for the NoObject Events and Object Events, the *F* ratio (*F*), which measures the mean differences between treatments, the significance of the treatment (*p*), and the partial eta squared (η^2), which indicates the effect size.

Table 9

Repeated Measures, Within Subjects ANOVA for Visual Attention Measures between Object Events and NoObject Events

Visual Attention Metric	NoObject Events	Object Events	$F(1,9)$ =	p =	partial η^2 =
Total Fixation Duration ratio	M = .61, SD = .37	M = .69, SD = .35	1.64	.23	.15
Fixation Count ratio	M = .62, SD = .33	M = .73, SD = .40	2.92	.12	.25
Total Visit Duration ratio	M = .65, SD = .40	M = .75, SD = .44	3.04	.12	.25
Visit Count ratio	M = .55, SD = .14	M = .59, SD = .13	5.20	.05	.37

These results suggest that I can fail to reject the null hypothesis because there is no difference in how subjects allocate their visual attention between Object Events and NoObject Events.

Influence of Perceived Event Difficulty on the Allocation of Visual Attention

Table 9 presents the results of the one-way, Within-Subjects Analysis of Variance comparing Difficult Events to Less Difficult Events of the ANOVA for all four visual attention metrics. The results indicate no significant difference in both the Total Fixation Duration ratio and the Total Visit Duration ratio. The results, however, do indicate a moderately significant difference in the mean ratios of Fixation Count in the Thumbnail AOI to Fixation Count in the Video AOI for Less Difficult Events (M = .79, SD = .50) than for Difficult Events (M = .61, SD = .32), $F(1, 9) = 5.36$, $p < .05$, partial $\eta^2 = .37$. These results indicate that the frequency of fixation – that is, of focused visual attention – within Thumbnail AOI is somewhat higher compared to the frequency of fixation for the Video AOI when the subject is looking for a specific object.

Table 10

Repeated Measures, Within Subjects ANOVA for Visual Attention Measures between Difficult and LessDifficult Events

Visual Attention Metric	Difficult Events	Less Difficult Events	$F(1,9)$ =	p =	partial η^2 =
Total Fixation Duration ratio	M = .58, SD = .29	M = .71, SD = .40	3.88	.08	.30
Fixation Count ratio	M = .79, SD = .50	M = .61, SD = .32	5.36	<.05	.37
Total Visit Duration ratio	M = .59, SD = .09	M = .75, SD = .13	3.99	.08	.31
Visit Count ratio	M = .60, SD = .16	M = .53, SD = .12	5.85	.04	.39

These results suggest that, although subjects did not vary greatly in the overall amount of time they focused (Total Fixation Duration) or in the length of time spent in each AOI (Total Visit Duration) when searching for target Events with Objects and for target Events without Objects, subjects may have increased the frequency of their focused visual attention (Fixation Count) in the Thumbnail AOI for Less Difficult Events as compared to when they were searching for Difficult Events, suggesting that they may consider the thumbnail images more useful for an initial assessment to identify videos with targets of Less Difficult Events than to identify videos with targets of Difficult Events. Another interpretation of the results is that Difficult Events require more visual attention in the video area of the search interface to identify the video that contains a target event. That is, subjects spend more time playing the videos – and therefore fixating on – the video during Difficult Events, which is what I would expect to see in their behavior.

The ANOVA results also indicated a moderately significant difference between the mean ratios for Visit Count of Difficult Event Types and the mean ratios for Less Difficult Event Types. These results suggest that subjects may allocate their visits to both AOIs more equally – or move between the Thumbnail AOI and the Video AOI more frequently – with Difficult Events than with Less Difficult Events. The results do indicate some patterns of difference among subjects, which Figures 23 through 26, and the related discussion, will address.

Figures 23 through 26 show the mean ratios of the four different visual attention metrics by Subject and Difficulty of Event. The ratios for each visual attention metric are the ratio of visual attention allocated to the Thumbnail AOI compared to the visual attention in the Video AOI. The higher the value of the ratio, the greater the allocation of visual attention to the Thumbnail AOI as compared to the Video AOI, because the Thumbnail AOI is the numerator. These charts show that, even considering the outliers for Subject P05, most of the subjects show a slightly higher mean ratio for the LessDifficult Events than for the Difficult Events. The higher ratios indicate that they allocated a higher percentage of visual attention to the Thumbnail AOI than to the Video AOI on LessDifficult Events than they did for Difficult Events. Subject P13 had significantly lower mean ratios for the Total Fixation Duration, Fixation Count, and Total Visit Duration metrics than the other subjects, suggesting that she used the Video AOI significantly more than the other subjects. I could expect to see this significant difference with P013, given that she had a considerably lower level of throughput (i.e., number of videos processed per test) than the other subjects. Subject P13's throughput ranged from 285 to 692 videos per test. Subject P13 also had a significantly higher number of videos played per test compared to the other subjects. Thus, I expect to see her with visual attention allocation ratios that indicate a much higher percentage of time allocated to the Video AOI than to the Thumbnail AOI as compared to other subjects. Subject P17, however, played significantly fewer videos as compared to other subjects, as indicated by her consistently higher ratios across all four visual allocation metrics.

These results may reinforce the findings reported in other research studies that indicate observers tend to favor surrogates (i.e., thumbnails) of video clips over playing the video, at least for the initial assessment of a video clip. The preference of surrogates would correspond to the subjects being motivated by speed, or throughput, more than by accuracy, because they could assess their individual performance on throughput whereas they had no feedback about their accuracy.

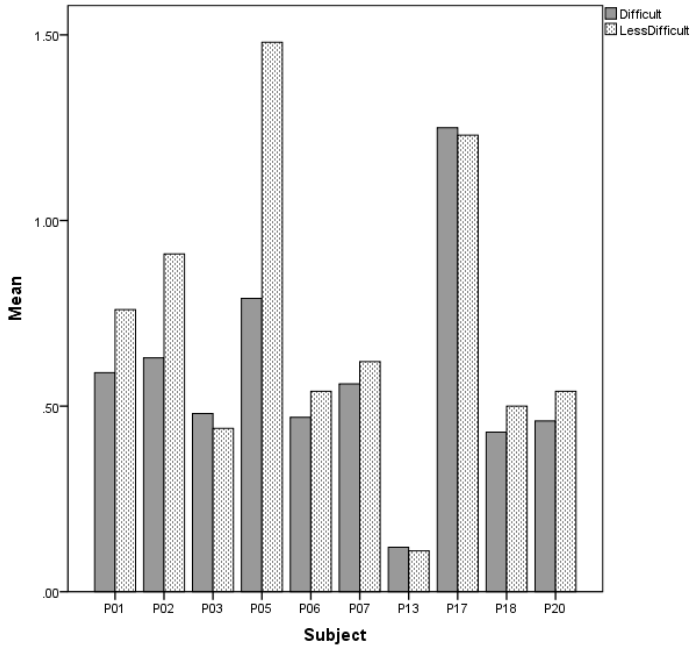

Figure 23: Mean ratios of Total Fixation Duration by Subject and Difficulty of Event

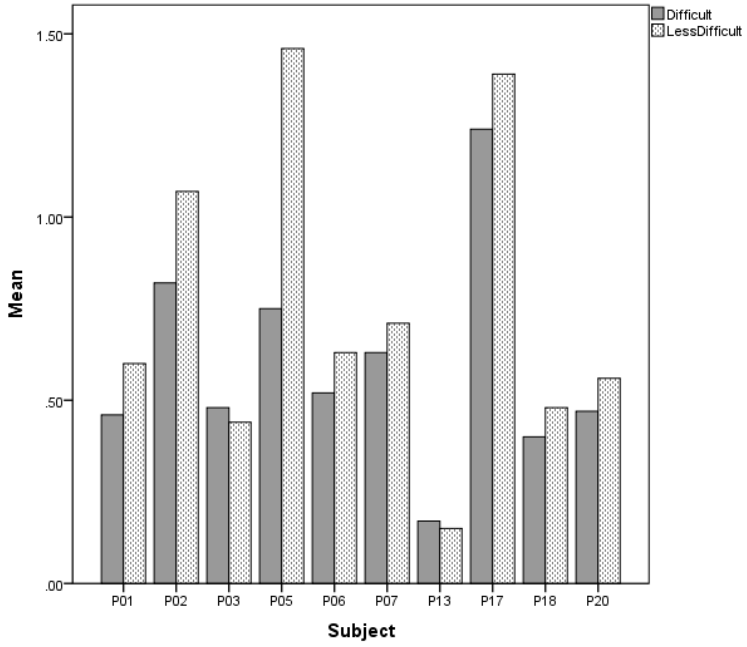

Figure 24: Mean ratios of Fixation Count by Subject and Difficulty of Event

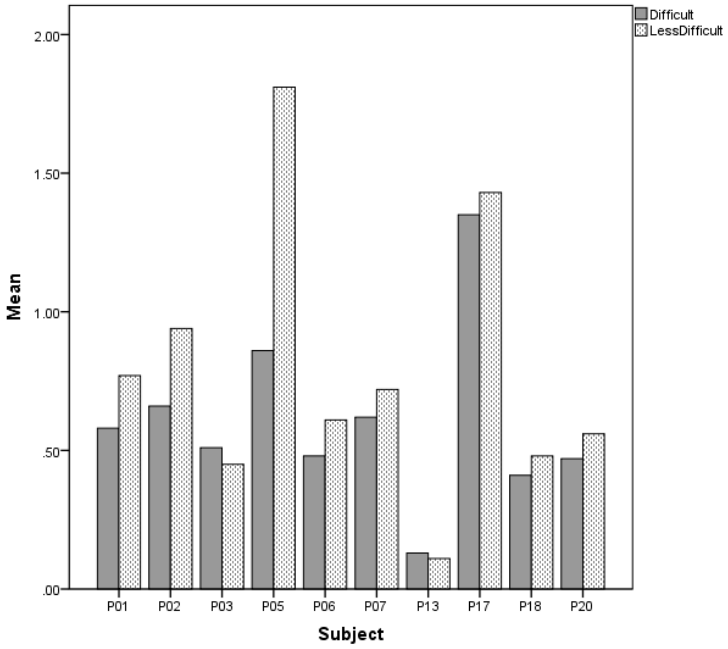

Figure 25: Mean ratios of Total Visit Duration by Subject and Event Type Difficulty

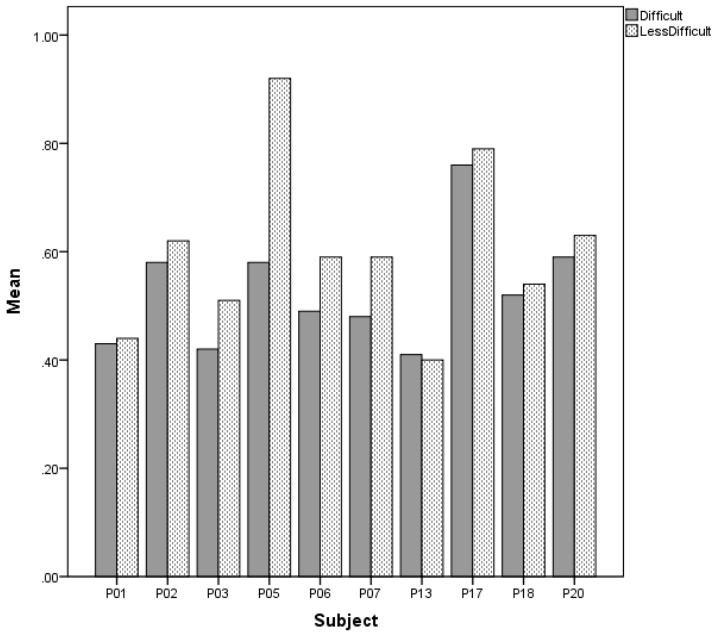

Figure 26: Mean ratios of Visit Count by Subject and Difficulty of Event

Additional Findings: Effect of Size of Thumbnail Area of Interest and Thumbnail Image on Visual Attention Patterns

Before they began the testing process, the test moderator informed subjects that they could modify the search interface to change both the size of the thumbnail image and the width of the thumbnail viewing area. Subjects would make these changes at the start of the test, before they started to view and decide on the video clips. Subjects had three options for the size of the thumbnail image: Small, Medium, and Large. Figures 27 through 29 show the relative size of Small, Medium, and Large thumbnail images in the custom video player. These images display the default settings of the custom video player interface for each of the three options for thumbnail image size.

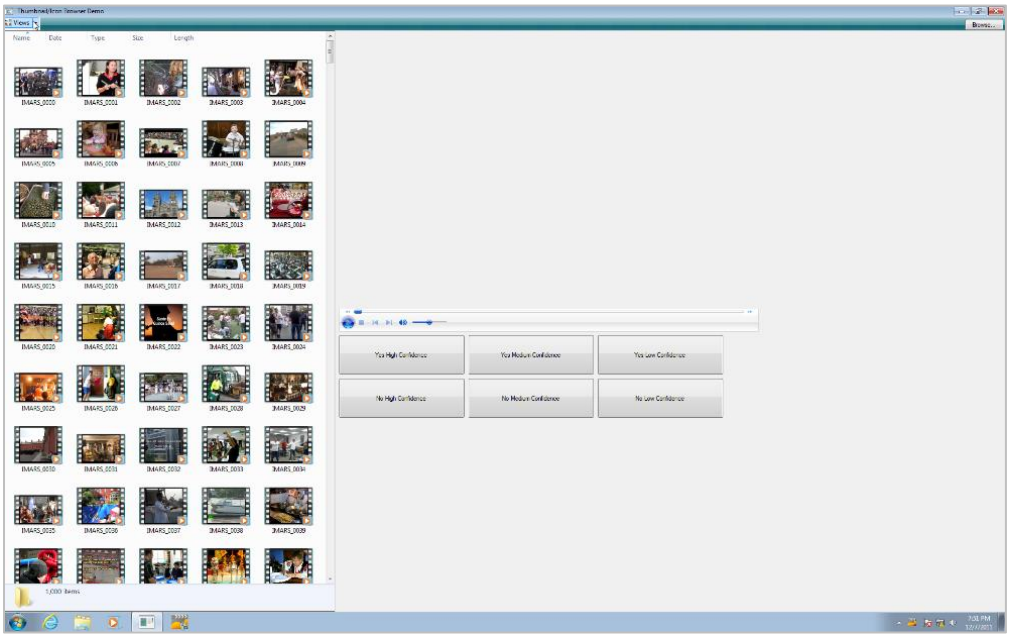

Figure 27: Default settings of custom video player interface for small thumbnails

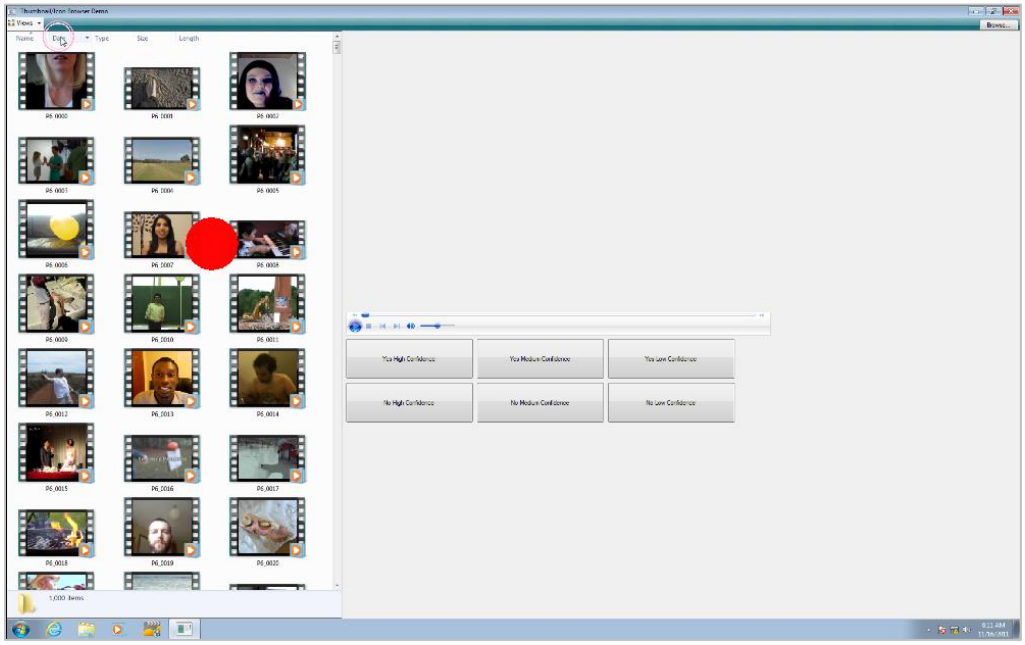

Figure 28: Default custom video player interface for medium-sized thumbnails

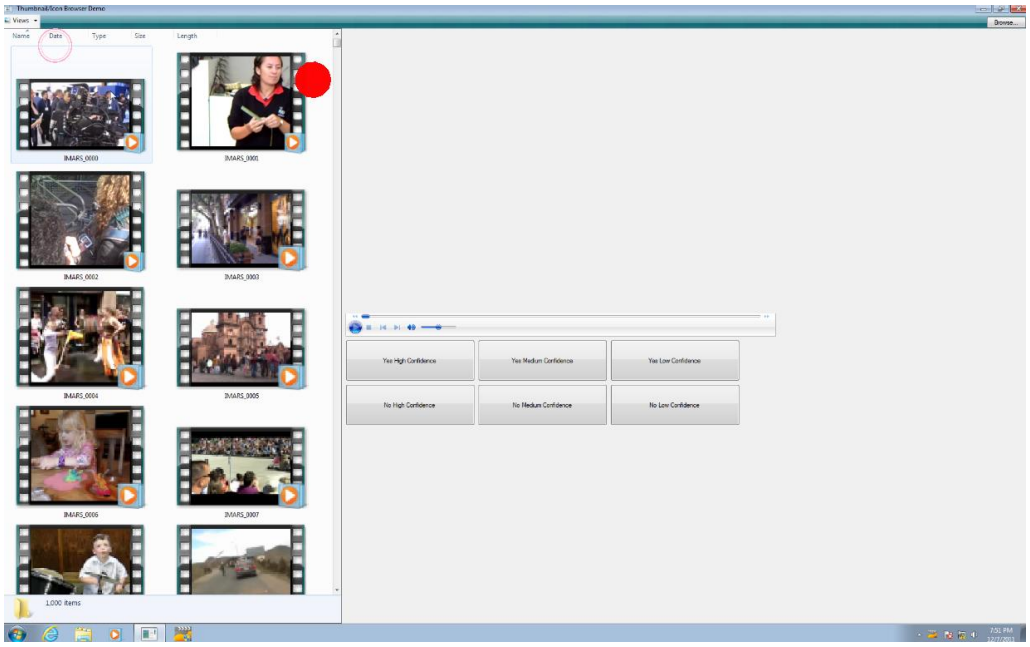

Figure 29: Default custom video player interface for large-sized thumbnails

Subjects could also change the width of the thumbnail viewing area by sliding the bar that divides the thumbnail area from the other three areas (Video, Video Controls, and

Voting Buttons). Because this width was a continuum instead of fixed values, I used the value for the size of Thumbnail AOI as a percentage of the total image area (i.e., the custom video player interface viewing area), as calculated by the Tobii Studio software. When an analyst assigns AOIs within an image, Tobii calculates the percentage of an AOI as a total percentage of the overall image area, which was the custom video player interface. Figure 30 shows the Tobii Studio view that displays the percentage of each AOI for the default settings of the custom video player. The default percentage of the Thumbnail AOI is 27.5 percent of the total image. Only the Thumbnail AOI changed as a percentage of the total image size. The other three AOIs (Video, Video Controls, and Voting Buttons) remained constant because subjects could not change their relative size.

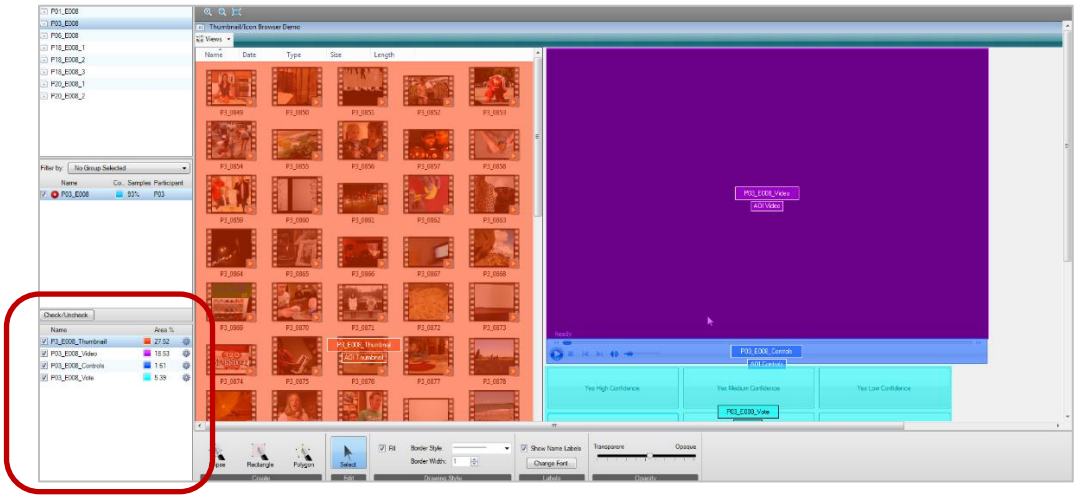

Figure 30: AOI percentages for the Default Custom Video Player Interface Settings

Subjects, however, changed the two variables in different configurations. For example, some subjects extended the width of the Thumbnail area window, but kept the thumbnail size small, so that six or seven thumbnail images were in a single row. Figure 31 shows the custom video player interface in this configuration. Figure 32 shows that the Thumbnail AOI in this configuration is 30.6 percent of the total image area.

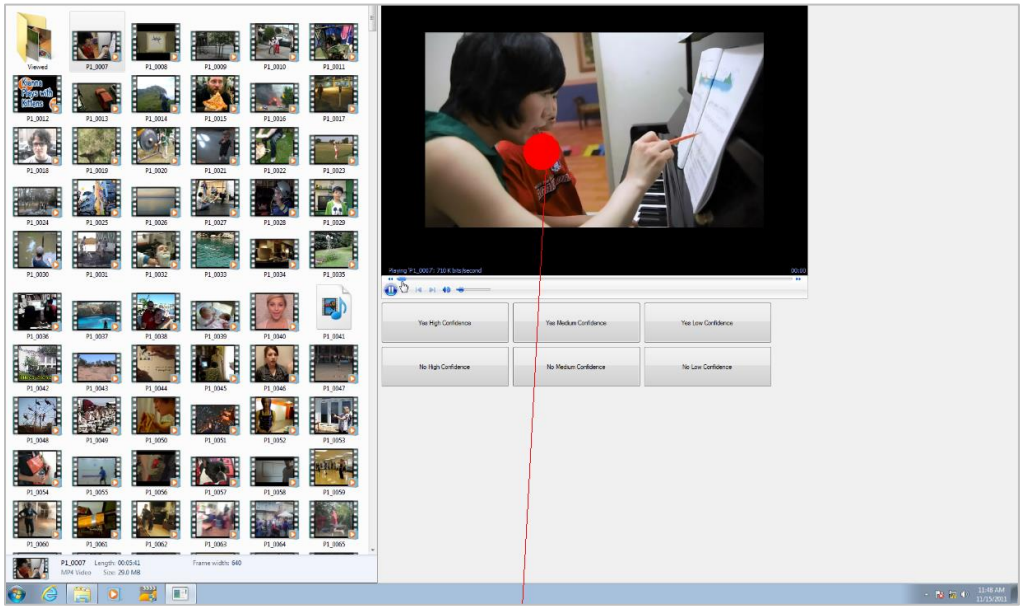

Figure 31: Thumbnail viewing area widened to a width of six small thumbnails

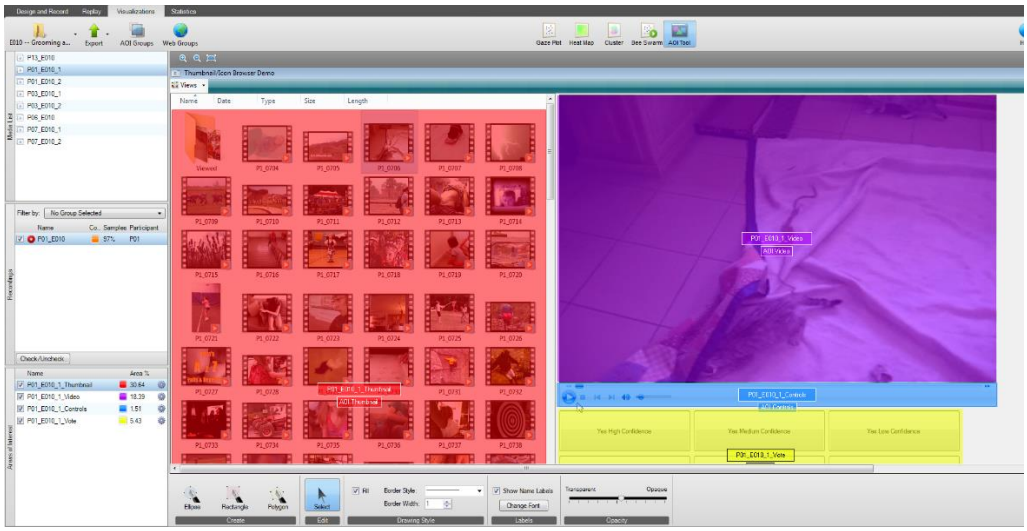

Figure 32: Thumbnail AOI is 30.6 percent of total image area

Other subjects made the thumbnail size large, but then narrowed the thumbnail AOI down so that only one large image appeared in each row. Figure 33 shows the custom video player interface set to this configuration. For this configuration, the size of the Thumbnail AOI is only 12 percent of the total image size.

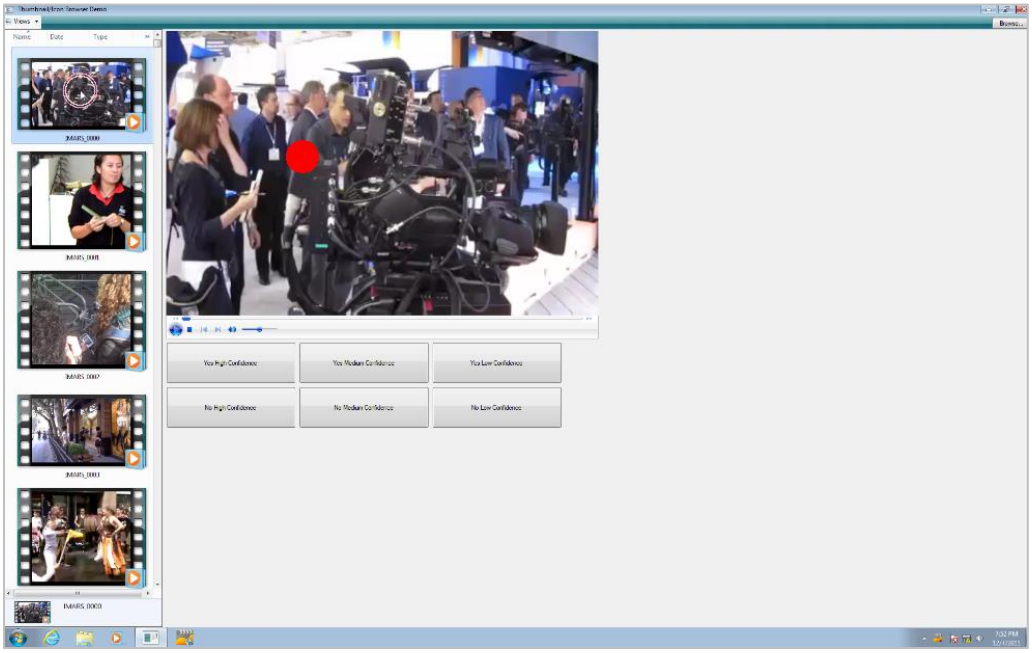

Figure 33: Configuration of Thumbnail AOI set to width of one large thumbnail image

Some subjects selected the medium size for the thumbnail image, and widened the thumbnail viewing area to the width of four medium thumbnail images. Figure 34 shows this configuration. For this configuration, the Thumbnail AOI was 34.7 percent of the total image area. A few subjects would set the thumbnail image size to medium or large, but not necessarily adjust the width of the thumbnail viewing area. Their interfaces would then be the default interfaces as previously presented in Figures 10, 11, and 12.

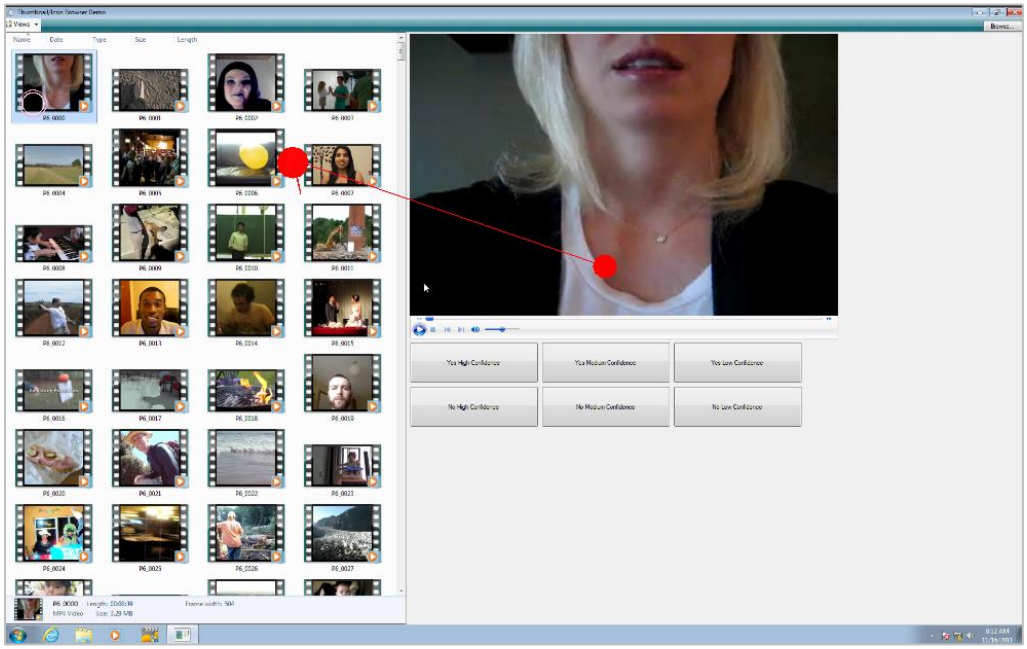

Figure 34: Configuration of thumbnail viewing area set to width of four medium thumbnail images

To understand the influence of the thumbnail viewing area size and the size of the thumbnail images on the visual attention patterns and search success, I performed correlation analysis on the two variables, independent from each other, to determine if any significant correlations exist between the size and rates of accuracy and ratios of visual attention allocation.

Influence of Size of Thumbnail Viewing Area and Thumbnail Image on Accuracy Rates

The results of the analysis of the Thumbnail Viewing Area size and the Thumbnail Image size do not identify any significant, direct correlations between either variable and rates of accuracy. The positive correlation between the Thumbnail AOI size and rate of Accurate Target Detections was not significant, $r(100) = .071$, $p = .48$. The positive correlation between the Thumbnail AOI size and the rate of False Alarms was not significant, $r(100) = .15$, $p = .14$. The positive correlation between Thumbnail Image size and Accurate Target Detections was not significant, $r(101) = .13$, $p = .19$. The positive correlation between Thumbnail Image Size and False Alarms was not significant,

$r(101) = .12$, $p = .23$. Figures 35 through 38 show the scatterplot diagrams of these four correlations.

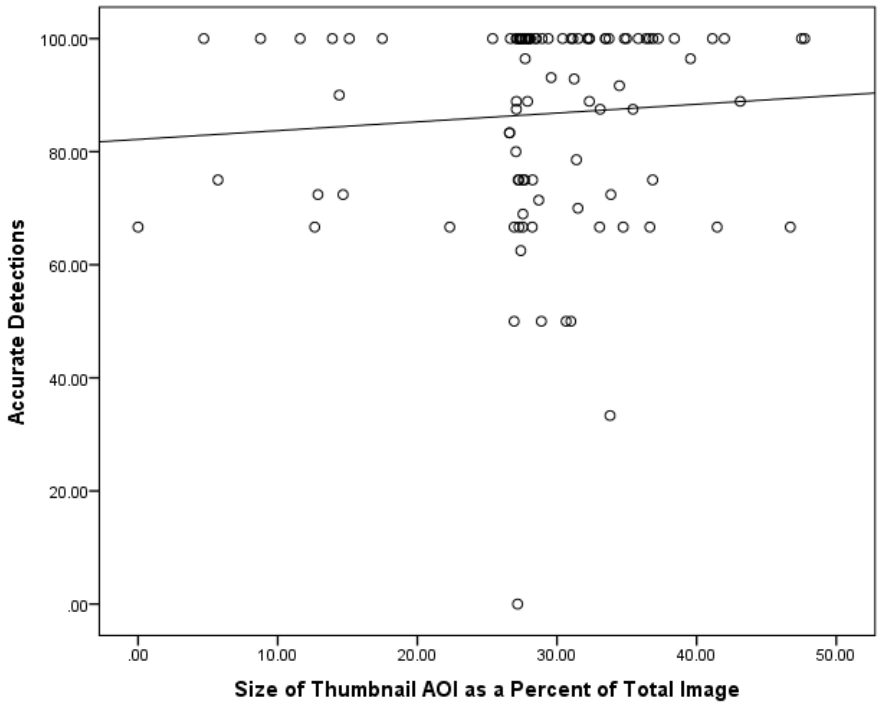

Figure 35: Correlation of Thumbnail AOI Size to Accurate Detections

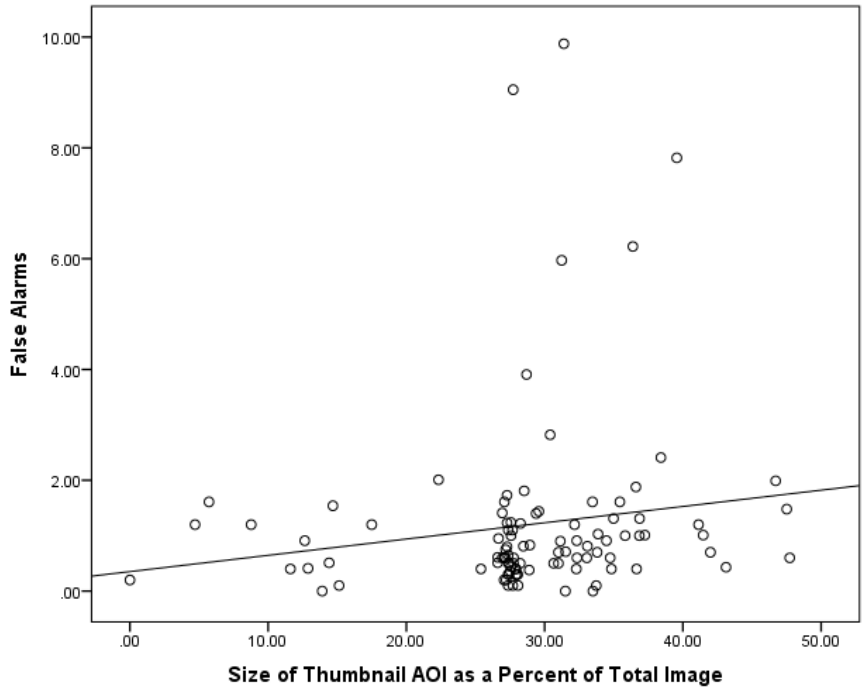

Figure 36: Correlation of Thumbnail AOI Size to False Alarms

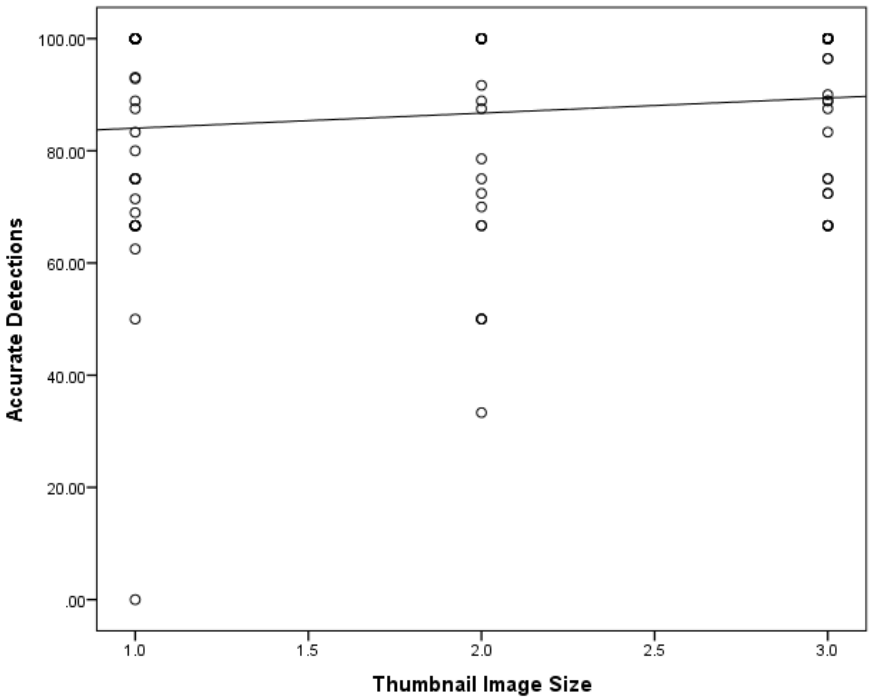

Figure 37: Correlation of Thumbnail Image Size to Accurate Detections

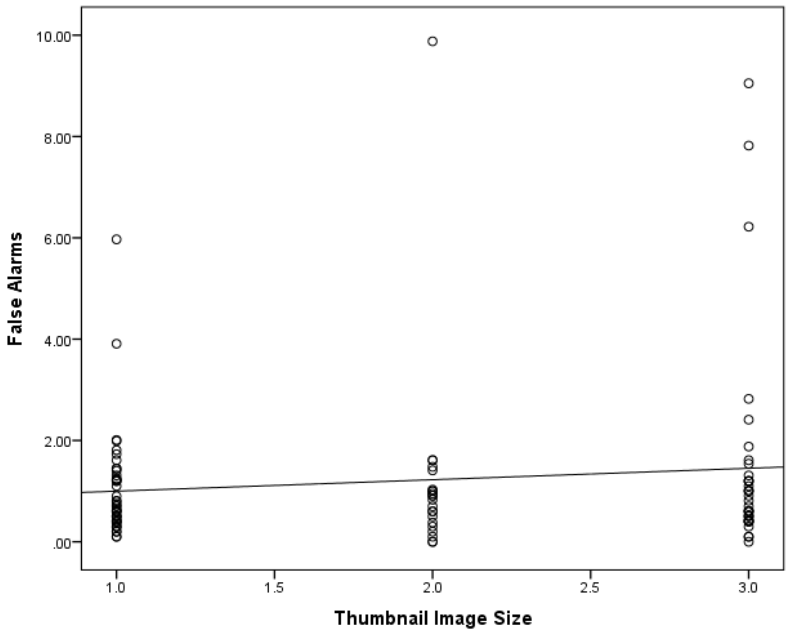

Figure 38: Correlation of Thumbnail Image Size to False Alarms

Influence of Size of Thumbnail Viewing Area and Thumbnail Image on Allocation of Visual Attention

Because subjects could modify both the size of the Thumbnail AOI and the size of the thumbnail images, I cannot readily determine if having a higher number of medium or large thumbnails available for viewing at the same time improves accuracy or throughput for subjects. The following analysis looks at each variable in isolation without considering the other variable. That is, the analysis first looks at the potential influence on the size of the Thumbnail AOI, regardless of the size of the thumbnail image, to see if any patterns of behavior are evident. Then the analysis looks at whether patterns of visual attention emerge with respect to the size of the thumbnail image, regardless of the size of the Thumbnail AOI. The following results provide some insights as to how these variations affected their visual attention.

The results of the analysis of the size of the Thumbnail AOI indicate a significant positive correlation between Thumbnail AOI size (as a percent of total image size, or percent of the total video interface) and the Total Visit Duration ratio, $r(98) = .22$, $p = .03$. This correlation suggests that the larger the Thumbnail AOI (as a percentage of the

total interface area) is during a test, the more time the subject spent viewing the Thumbnail AOI. Figure 39 shows the scatterplot diagram of this correlation.

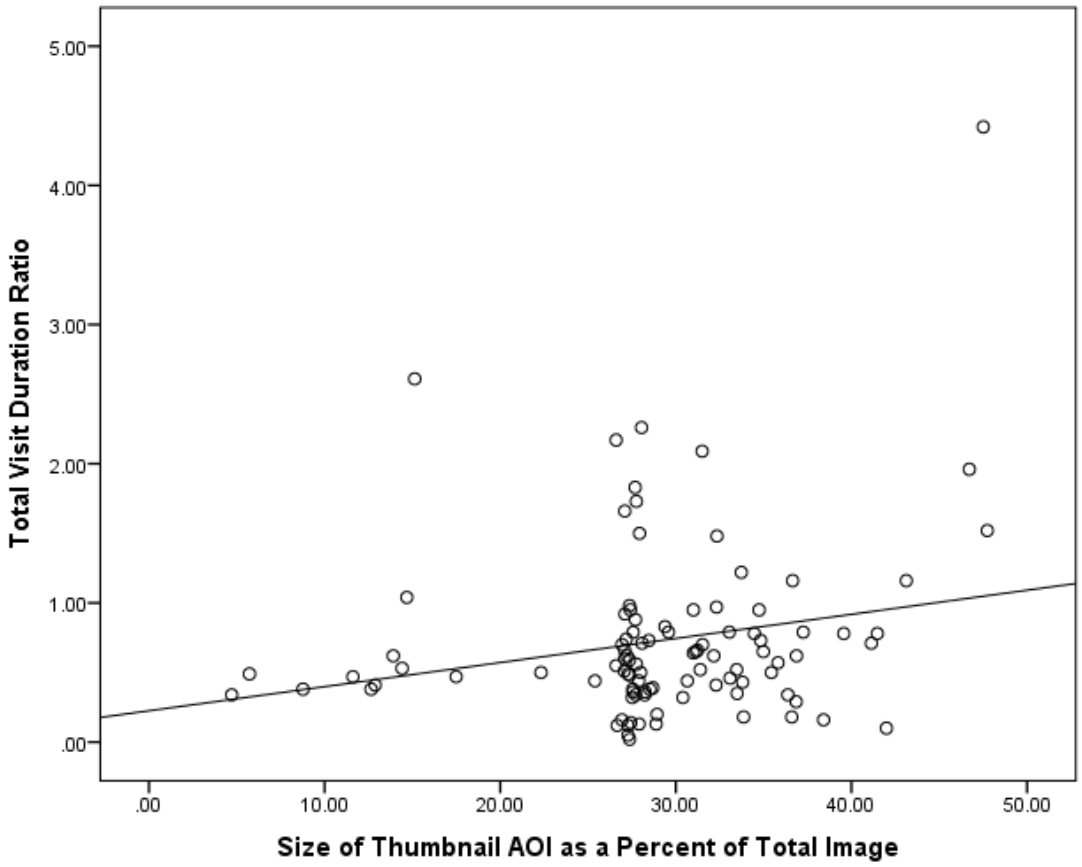

Figure 39: Correlation of Size Thumbnail AOI to Total Visit Duration Ratio

We might expect that if subjects made the Thumbnail AOI larger, it would indicate that they might rely on using the thumbnail images as surrogates to viewing the video, especially if they are motivated by speed more than accuracy.

The significant correlation between Thumbnail AOI size (as a percent of total interface space) and the Percent of Videos played during a test, however, suggest an opposite behavior. That is, the positive correlation of Thumbnail AOI size to the number of videos played, as a percentage of the total number of videos, indicates that the larger the size of the Thumbnail AOI, the more videos the subject played, $r(8) = .33$, $p = .00$ (two-tailed). Figure 40 shows the scatterplot of the correlation of the Thumbnail AOI

Size to the Percent of Videos Played. I would expect that, if subjects were relying on the thumbnails as surrogates to playing the video, they would not need to play as many videos before deciding, but instead might play fewer videos. These results suggest that the larger Thumbnail AOI may not result in an increasing use of Thumbnails as surrogates.

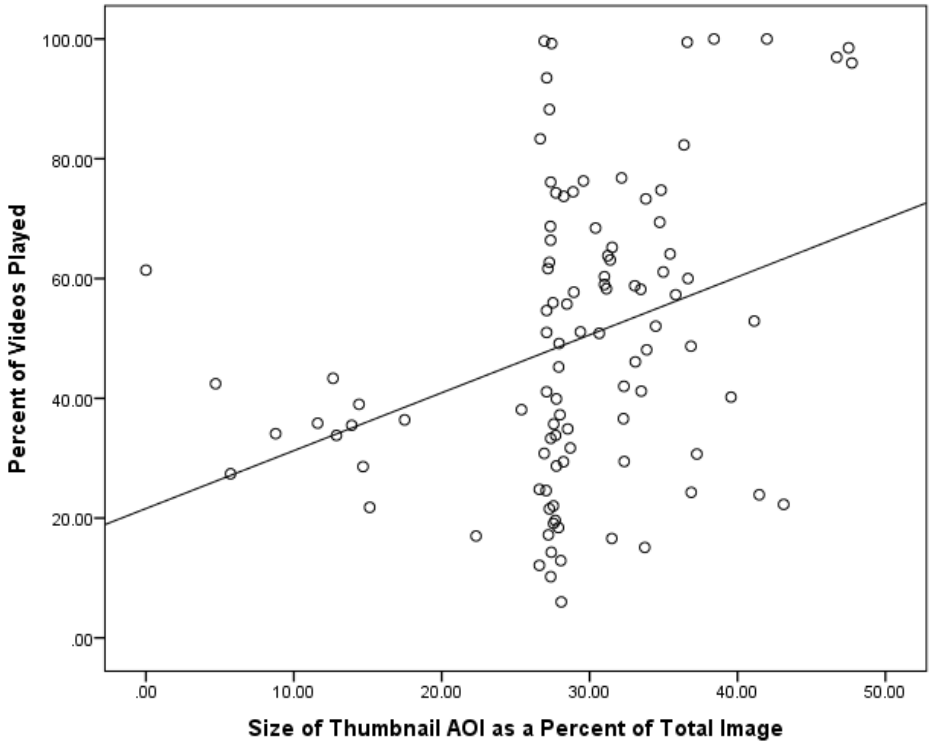

Figure 40: Correlation of Size of Thumbnail AOI to Percent of Videos Played

The analysis of subject behavior in viewing videos is not complete, however, without looking at the correlation of the amount of each video played, as a percentage of the total length of the video, to the size of the Thumbnail AOI. The positive correlation of Thumbnail AOI Size to the Percentage of Video Viewed for each played video is not significant, $r(101) = .12$, $p = .22$. Figure 41 shows the scatterplot diagram of this correlation. These results suggest that, although subjects play a higher number of the videos with a larger Thumbnail AOI, the amount of each video that they played did not have a corresponding increase with the size of the Thumbnail AOI. The overall results

suggest that, for the most part, the average amount of video viewed was consistently around the 10 percent amount, regardless of the size of the Thumbnail AOI.

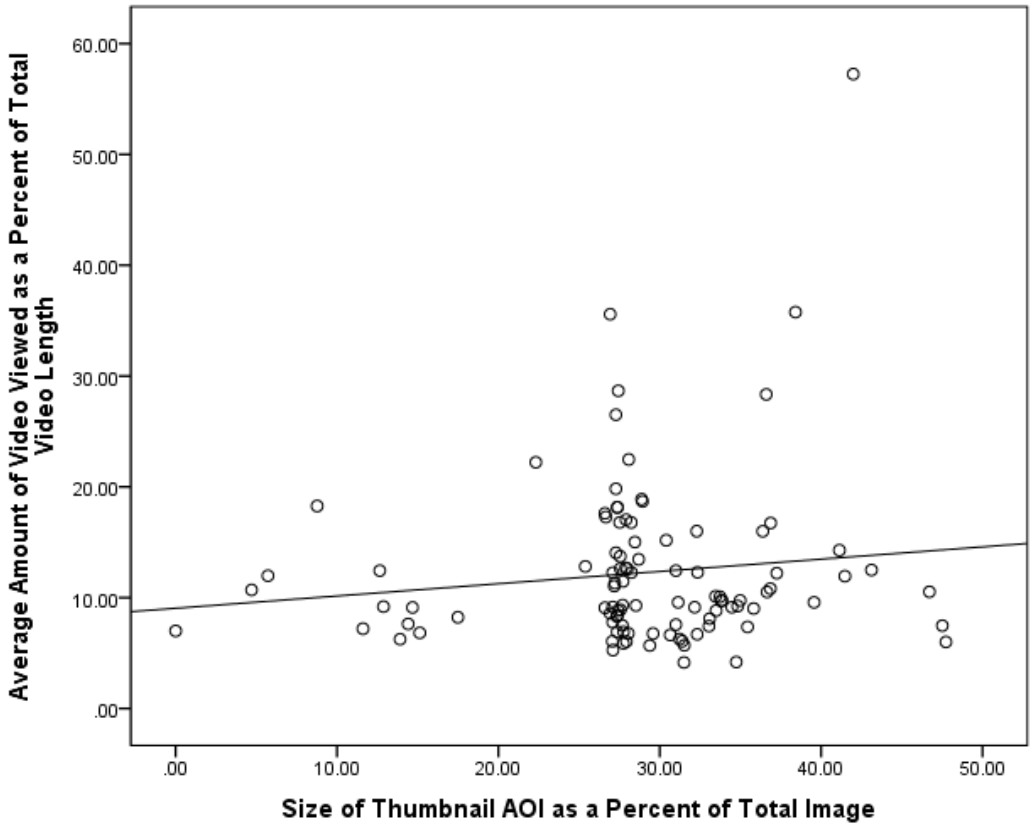

Figure 41: Correlation of Size of Thumbnail AOI to Average Amount of Video Viewed

The correlations of Thumbnail AOI size to the remaining three visual attention metrics also indicate no strong correlations between Thumbnail AOI size and an increase in visual attention – especially focused visual attention, as measured by the Fixation Count and Total Fixation Duration ratios. The Thumbnail AOI size does not have a significant positive correlation to the Fixation Count ratio, $r(98) = .19$, $p = .06$ or to Total Fixation Duration ratio, $r(98) = .18$, $p = .08$, or to the Visit Count ratio, $r(98) = .03$, $p = .79$. If subjects were allocating more visual attention to the Thumbnail AOI when its size is larger, than the correlations would be stronger. That is, as the size of the Thumbnail AOI increases, the ratios for each of the visual attention metrics would also increase in a corresponding manner because the larger the ratio, the more visual attention is allocated

to the Thumbnail AOI. The lack of strong correlations, however, suggest that the visual attention is more evenly allocated between the Thumbnail AOI and the Video AOI, regardless of the size of the Thumbnail AOI. Figures 42, 43, and 44 show the scatterplot diagrams of each of the above three correlations of Thumbnail AOI size to the visual attention metrics.

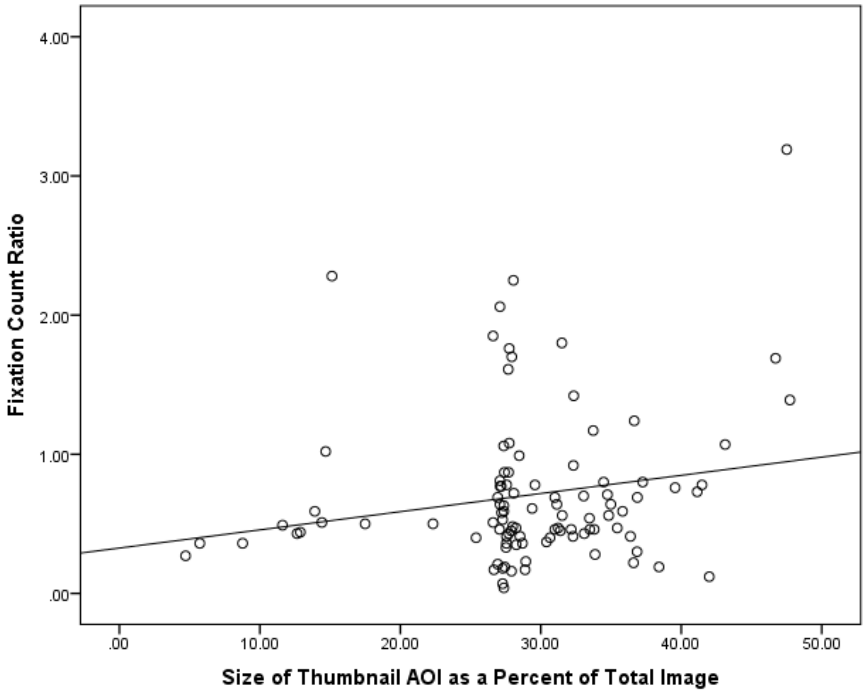

Figure 42: Correlation of Size of Thumbnail AOI to Fixation Count Ratio

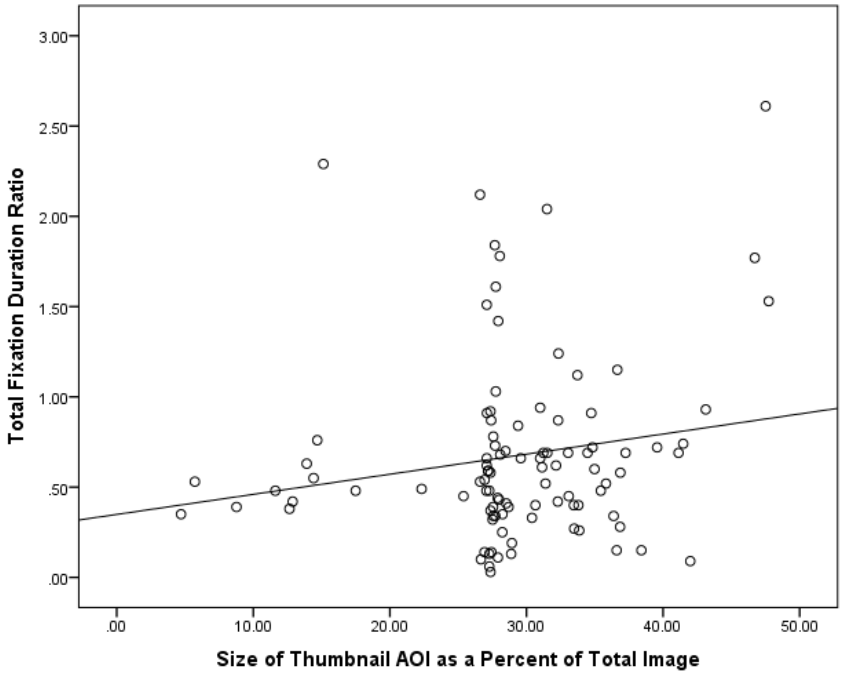

Figure 43: Correlation of Size of Thumbnail AOI to Total Fixation Duration Ratio

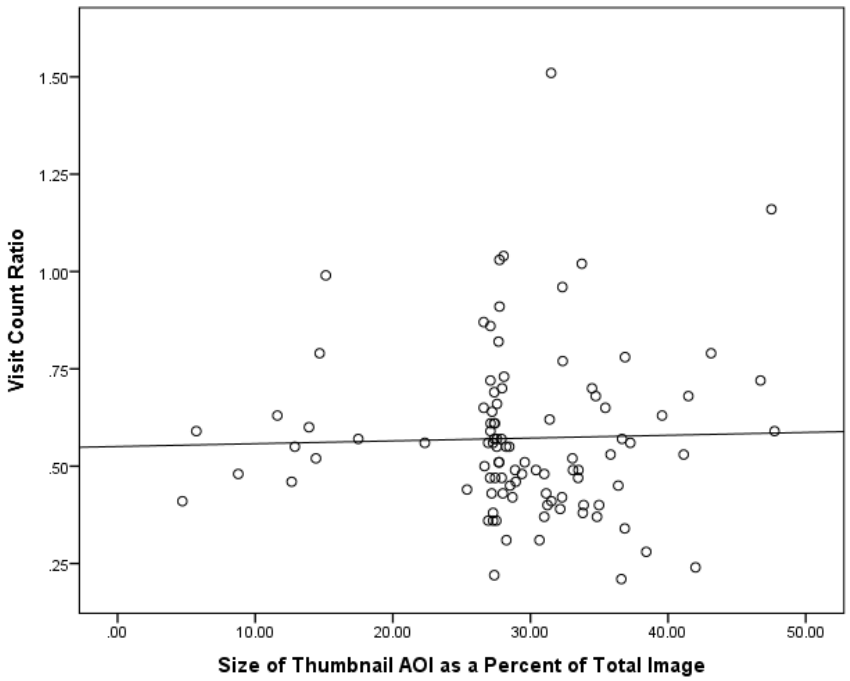

Figure 44: Correlation of Size of Thumbnail AOI to Visit Count Ratio

As mentioned earlier, subjects may rely on thumbnail images as surrogates to reviewing the videos if they are focused on speed. Thus, I would expect to see a strong correlation between Thumbnail AOI size and the number of videos processed if the subject is changing the Thumbnail AOI size to increase his or her speed. The results, however, show a significant negative correlation of Thumbnail AOI Size and Number of Videos Processed, $r(101) = -.21$, $p = .04$, two-tailed. Figure 45 shows the scatterplot diagram of this correlation.

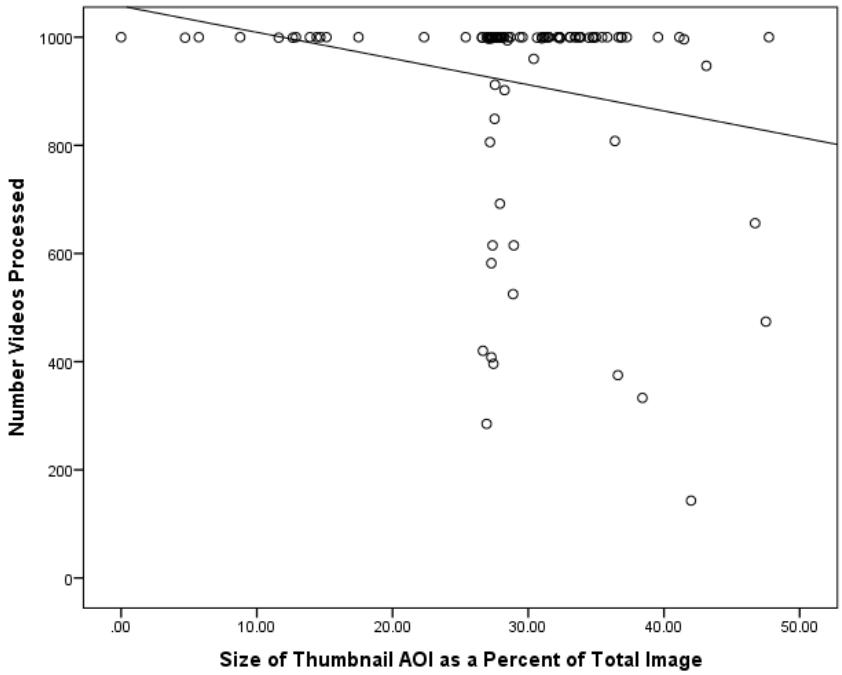

Figure 45: Correlation of the Size of the Thumbnail AOI to the Number of Videos Processed

These results indicate that the size of the Thumbnail AOI does not likely have a strong influence on whether a subject relies on the thumbnail image as a surrogate to playing the video – or at least the Thumbnail AOI size is a not a good indicator or predictor of subjects using thumbnail images as surrogates as a means of increasing their throughput. Although a larger Thumbnail AOI area has a higher Total Visit Duration ratio – that is, subjects spend more cumulative visual attention time in the AOI – subjects

do not allocate more focused visual attention with the Thumbnail AOI, as indicated by the Total Fixation Duration or Fixation Count metrics.

Thumbnail image size (i.e., Small, Medium, or Large), however, does appear to have a strong relationship to visual attention. The results indicate that Thumbnail Image Size (i.e., Small, Medium, Large) has strong positive correlations to all four visual attention ratios. That is, the larger the thumbnail size (i.e., Medium or Large), the larger the visual attention ratio, indicating a higher use of the Thumbnail AOI as a percentage of the Video AOI when the thumbnail size is larger. Table10 shows the correlations between Thumbnail Image Size and each of the four visual attention metrics.

Table 11

Correlations of Thumbnail Image Size and Visual Attention

Visual Attention Metric	$r(99) =$	$p =$	Correlation
Total Fixation Duration ratio	.25	.01	Positive, Significant
Fixation Count ratio	.32	.00	Positive, Significant
Total Visit Duration ratio	.24	.02	Positive, Significant
Visit Count ratio	.29	.00	Positive, Significant

Figures 46 through 49 show the scatterplot diagrams of each of these correlations.

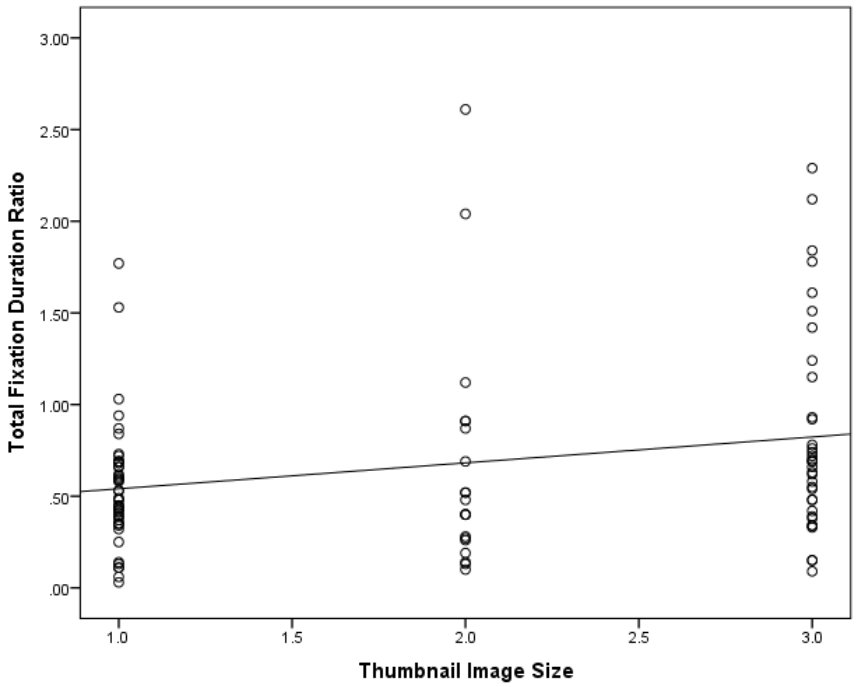

Figure 46: Correlation of Thumbnail Image Size to Total Fixation Duration Ratio

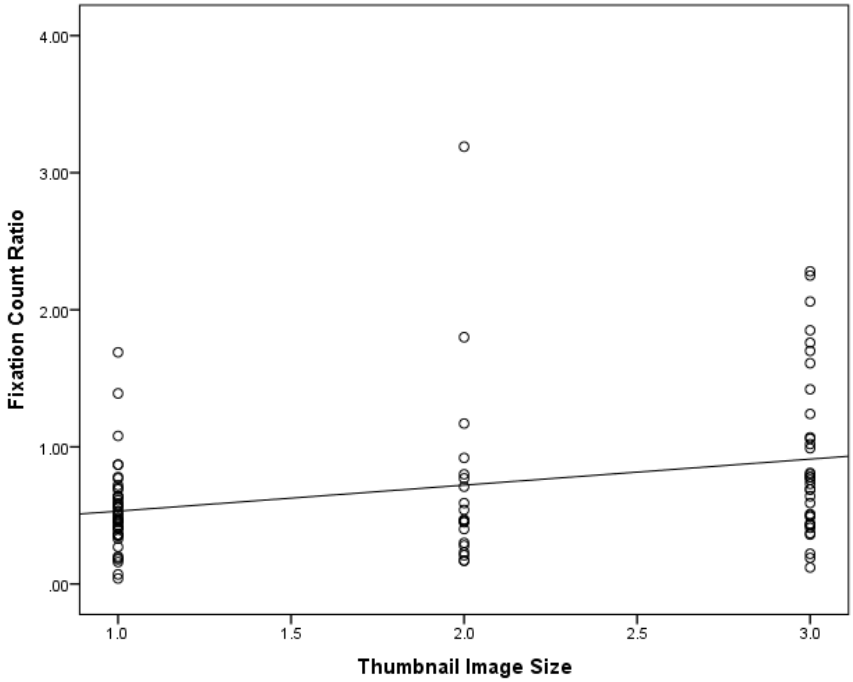

Figure 47: Correlation of Thumbnail Image Size to Fixation Count Ratio

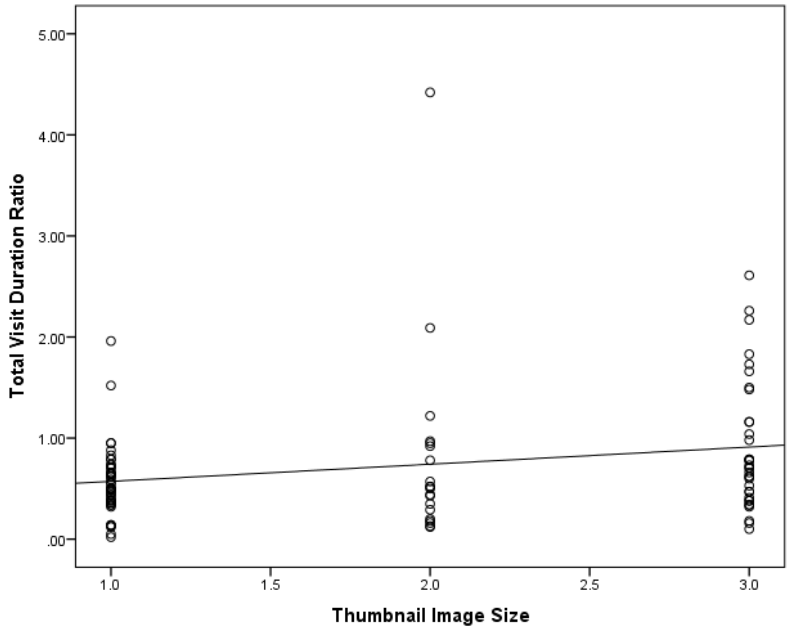

Figure 48: Correlation of Thumbnail Image Size to Total Visit Duration Ratio

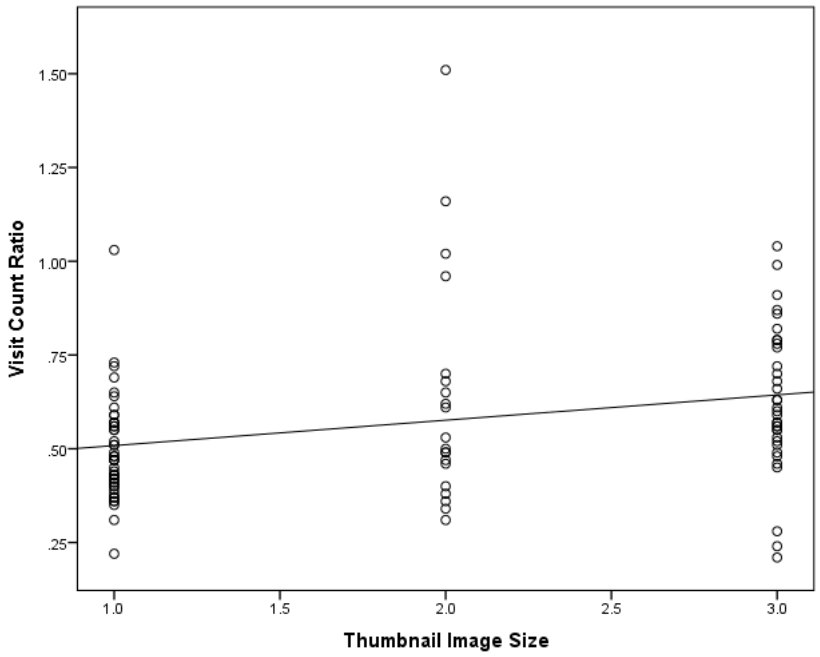

Figure 49: Correlation of Thumbnail Image Size to Visit Count Ratio

This collective set of results on the Thumbnail Area size as a percentage of the total interface and the Thumbnail Image size suggest that the size of the Thumbnail

Image has greater determination over search behaviors and allocation of visual attention than does the size of the Thumbnail AOI. Specifically, the larger the thumbnail, the more likely the subject will allocate focused visual attention to decide whether he needs to view the video to make an accurate determination about the video. Because the percentage of the thumbnail area can be large compared to the total interface area but can still have small thumbnails within the area, the subjects may be working harder to make an initial determination on the video without viewing a portion of the video, or spending more time skimming the thumbnails. Thus, subjects are spending more time during each visit to the Thumbnail Area, but are not having as many occurrences of focused visual attention (i.e., as measured by Fixation Counts or Total Fixation Duration).

Additional Findings: Mean Ratios of Visual Attention and Test Order

Figure 50: Mean Visual Attention by Test Order, reflects a pattern of visual attention as subjects progressed through the sequence of tests that I would expect to see with the omission of the Subject P05 outlier tests. The pattern reflects that, as subjects progressed through the sequence of tests, they gradually increased their ratio of visual attention allocation more toward the use of the Thumbnail AOI, suggesting that they relied less on video to confirm that a target Event was included. The pattern also shows an increase in the ratio of Thumbnail AOI attention to Video AOI attention when the tests shifted from the set of randomly selected videos (Regular) to ranked selected videos (IMARS), which included more target Events. This pattern suggests that, with the Ranked ordering of videos, and with the increased number of targets, subjects relied even less on the Video AOI to decide whether a video contained a target. This finding is what I would expect based on the results of post-test interviews, during which subjects indicated that they realized that, with the rank-ordered sets, the target videos were likely to occur early in the set. Therefore, they could identify the targets more quickly and make determinations faster on the remaining thumbnail images.

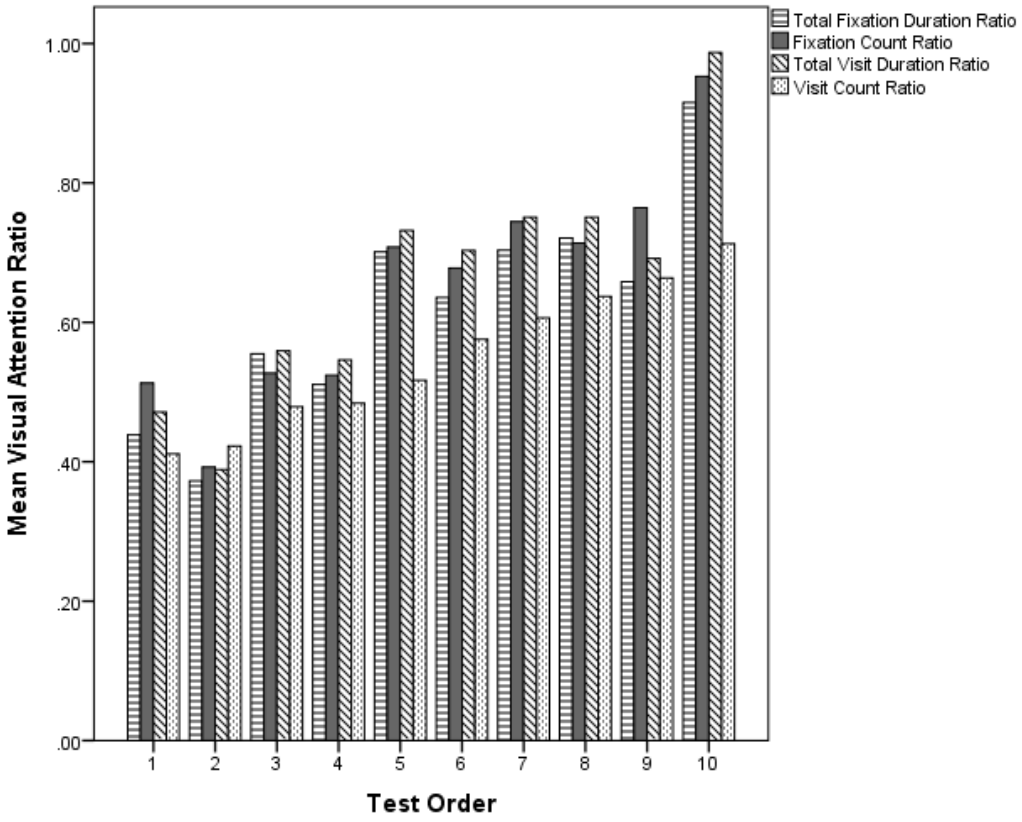

Figure 50: Mean Visual Attention by Test Order

Additional Findings: Order of Fixation in Areas of Interest

Figures 51 and 52 provide different views of in which Area of Interest subjects had their first fixations. Figure 51: Area of Interest for First Fixation by Subject shows that most subjects predominantly focused their first fixations on the Thumbnail AOI, which I would expect because the video player did not display the video in the Video AOI until a subject had clicked on a thumbnail to select it. Only Subject P18 had a strong inclination to have his first fixation in the Video Area of Interest, but because of focusing on the vertical bar that divides the thumbnail viewing area from the video playback area to move it. Subject P20 was the only subject who had all her first fixations in the Thumbnail AOI. The rest of the subjects did have some variance as to where they had their first fixation. A review of the test recordings reveal that when subjects focused first on the Video AOI, it was a result of either moving the vertical bar to change the size of

the thumbnail viewing area, or they intuitively clicked on a thumbnail image to have the larger still image appear in the video playback area.

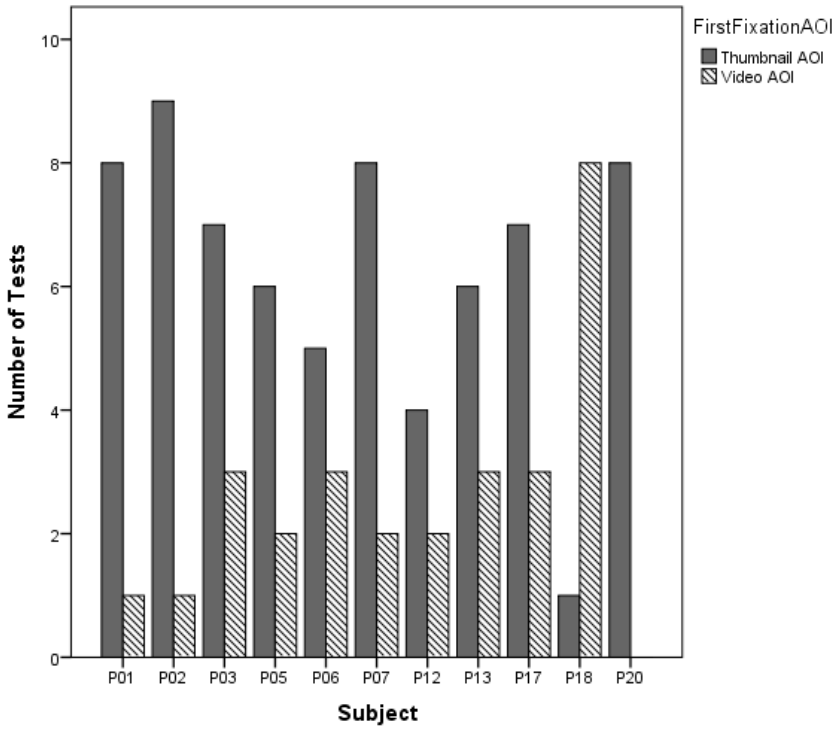

Figure 51: Area of Interest for First Fixation by Subject

Figure 52 shows the Area of Interest for the first fixation by Test Order. All subjects had their first fixation in the Thumbnail AOI for their first test, which I would expect to see, based on the default setting of the video player. Some variance occurred over the remaining tests. Although the pattern is not strong, Figure 52 suggests that subjects tended to vary their first fixations more as the test sequence progressed. The small dip at Test 6 may indicate that, with the shift to the Rank Ordered tests, subjects may have focused on the Thumbnail AOI first to assess whether the Rank Ordered thumbnail images might be easier to process. All subjects, except for Subject P18, transitioned to the Rank Ordered set of videos at Test 6. The test moderator informed the subjects that they were shifting to a pre-filtered, or rank-ordered set of videos, so they knew the video test sets were pre-filtered by a machine. Some subjects acknowledged in the post-test interview that the shift caused them to assess the video sets more closely at

first. Only Test 3 and Test 9 have a higher frequency for subjects fixating first on the Video AOI; the remaining tests all have a higher frequency for the Thumbnail AOI being the area of first fixation.

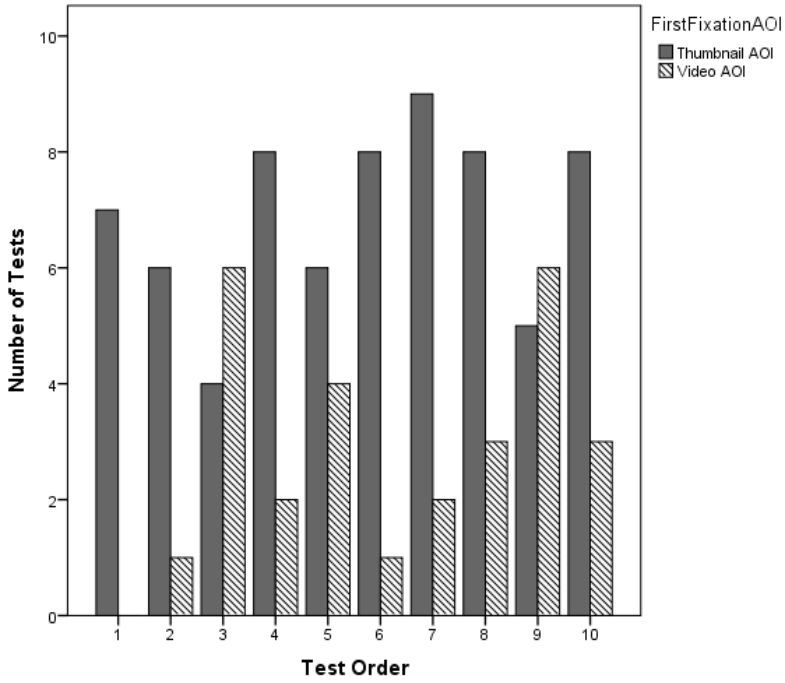

Figure 52: Area of Interest for First Fixation by Test Order

Figure 53: Area of Interest for First Fixation by Event Type does not reveal any noticeable pattern between Event Type and area of first fixation. Although four tests have a noticeably higher rate of the Video AOI for the first fixation, these events do not share similar attributes, such as Objects versus NoObjects, or perceived Difficulty. The four events with noticeably higher numbers of tests where subjects first fixated on the Video AOI include the following events:

- E007 – Changing a Vehicle Tire
- E008 – Flash Mob Gathering
- E014 – Repairing an Appliance
- E015 – Sewing Project

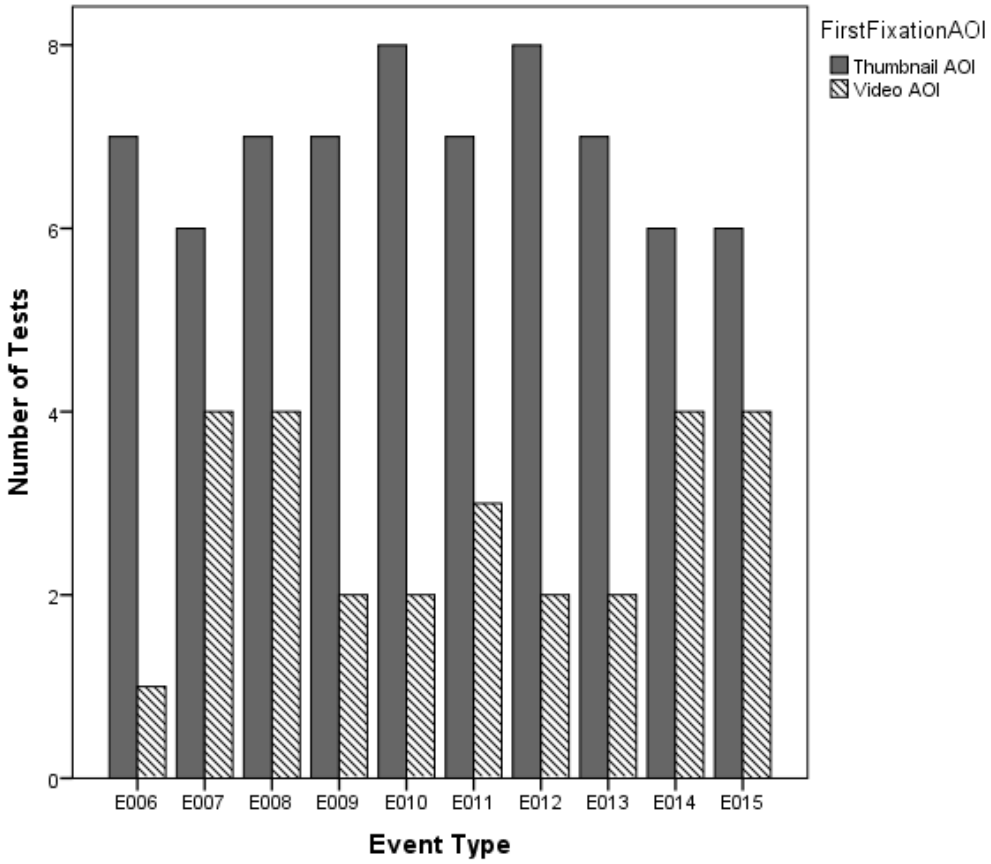

Figure 53: Area of interest for first fixation by event type

The positive correlation between Thumbnail Image size and AOI of First Fixation is not significant, $r(97) = .06$, $p = .57$. I would not expect a significant correlation pattern because both variables are discrete; the Thumbnail Image variable has three values (Small = 1, Medium = 2, and Large = 3) and the AOI of First Fixation has only two values (Thumbnail AOI = 1, Video AOI = 2). Figure 54 shows the scatterplot diagram of this correlation.

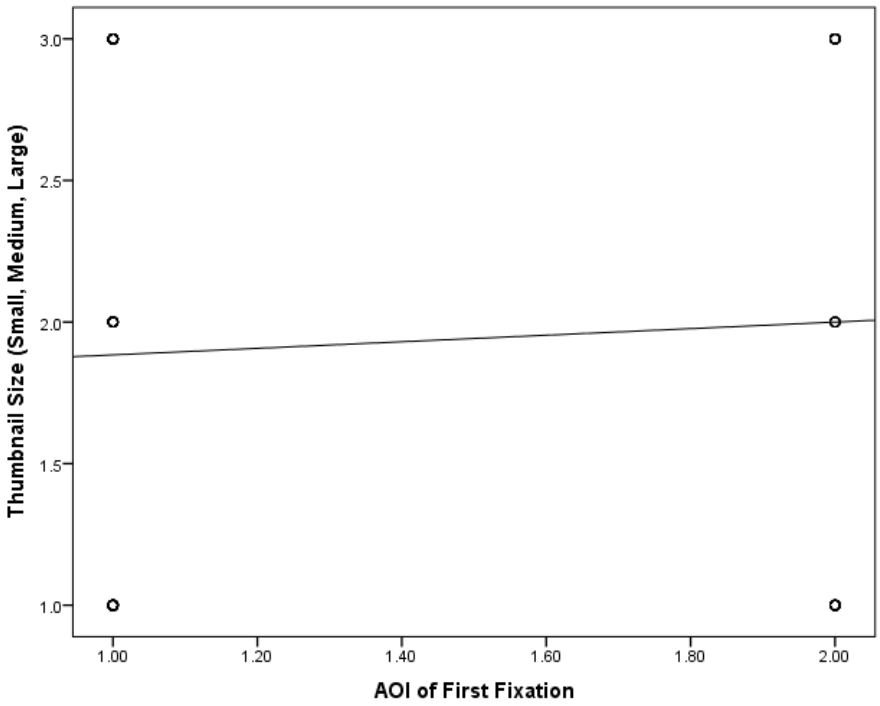

Figure 54: Correlation of AOI of First Fixation to Thumbnail Size

The negative correlation, however, between the Thumbnail AOI size and AOI of First Fixation is significant, $r(96) = -.35$, $p = .00$. Figure 55 shows the scatterplot diagram of this correlation.

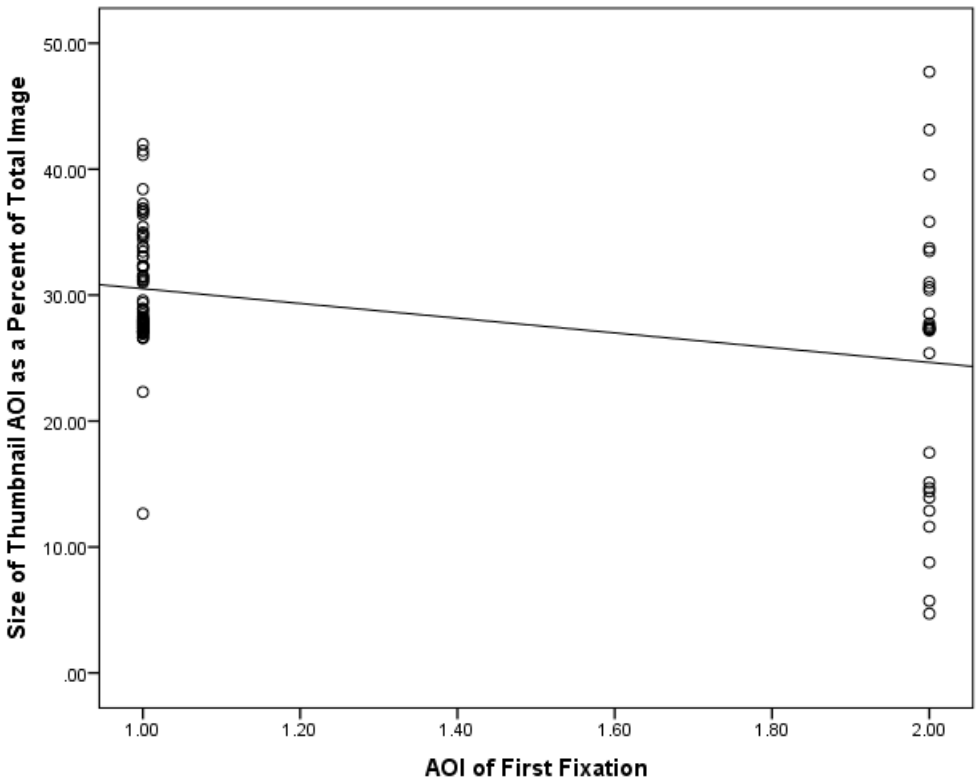

Figure 55: Correlation of AOI of First Fixation to Size of Thumbnail AOI

This significant correlation suggests that the larger the Thumbnail AOI is, the more likely the subject is going to focus on the Video AOI first. This finding corresponds to the earlier finding that large Thumbnail AOIs do not receive as much focused visual attention as smaller Thumbnail AOI areas. If a large Thumbnail AOI has many small thumbnail images, the subject may be more apt to quickly click on a thumbnail to display the still image in the Video Playback area, and then focus visual attention as the Video Playback still as a surrogate to decide on the video clip. The qualitative analysis obtained from viewing the eye-tracking recordings confirm this pattern of behavior. The next section presents more detail on the qualitative findings of the subjects' behavior during the tests, based on a review of the eye-tracking recordings and the post-test interviews.

Qualitative Results

This section presents qualitative results summarized from observations during the testing process and in summarizing the post-test interviews. In general, the observation during the baseline testing process was that most subjects appeared to be motivated by a goal of speed or throughput, as measured by processing all videos in a test data set before reaching the two-hour time limit. Most of the subjects adjusted their approach to processing videos, including their allocation of visual attention, to support or improve their speed or throughput in completing the task. Only one participant did not display behavior associated with the goal of processing all the videos during each test, and had a consistently lower level of throughput than the other participants. Some participants seemed to set a goal of finishing the processing of all videos as quickly as possible, completing the task for some tests in a little more than one hour.

Although the adjustments that subjects made to both the video player interface and their visual attention patterns were diverse, a few patterns of behavior emerged. Most of the subjects processed the videos in a linear order; that is, they reviewed the thumbnails in the order in which they appeared in the thumbnail view, and decided on the thumbnail if possible, based on the scene or the appearance of objects relevant to the event type. Many subjects felt confident in deciding based on the thumbnail image alone. This behavior is reflected in the overall low percentage of videos played (53% for all tests; 45% for tests with 99% of the videos processed). Figure 56 shows the mean number of videos played, as a percentage of the total number videos, and the amount of each video played, as a percentage of the total length of the video, for each Event Type.

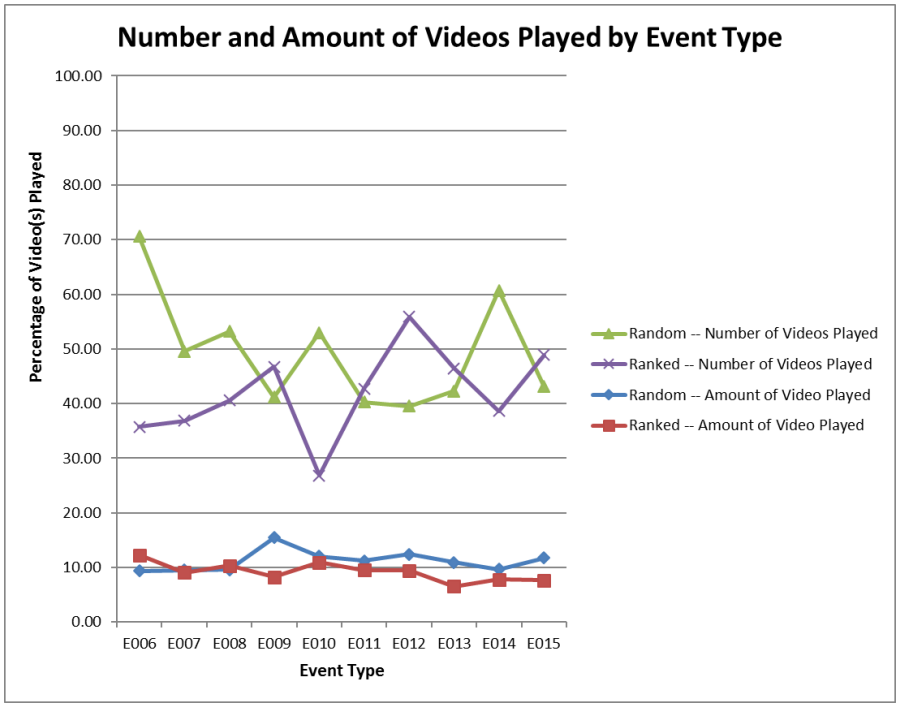

Figure 56: Number of videos played, and amount of each video played, by event type

Triage Strategy: Most of the subjects developed a strategy of triage for processing the thumbnail images based on scene characteristics. For example, most subjects reported in the post-test interview that if the general scene of the thumbnail image did not match the event description, they would eliminate the video clip based on the thumbnail, or still image, alone. If the Event was Making a Sandwich or Repairing an Appliance, and they saw outdoor scenes, they would eliminate it. Likewise, if they saw indoor scenes and the Event was Parkour, Parade, or Getting a Vehicle Unstuck, they would vote No based on the still image. For Grooming an Animal, subjects reported if they did not see an animal in the image, they would eliminate it. Some subjects would set questionable thumbnails aside to play as a batch, often saving several to review at the end of the test, when they felt more comfortable about having enough time left to view a greater percentage of the videos.

Ergonomic Adjustments: Some subjects established a strategy to maximize the ergonomic effectiveness of the custom interface to accommodate the speed in processing. Thumbnails would disappear from the thumbnail view after the subject decided, causing the thumbnails to move up and realign. Some subjects minimized the effects of this constant shifting of thumbnails by using the largest size of the thumbnail and reducing the viewing window to a single column. Then they would focus on one constant thumbnail position, allowing a new thumbnail to populate the location of their focus. Often this position was the thumbnail slot closest to the video controls and determination buttons to minimize the mouse movement that they would need to make to select a thumbnail, play it, and decide on it. Subjects P05, P18, and P20 specifically described using this approach in their post-test interviews.

Not all subjects who chose the largest thumbnail size went to these efforts; they would use the large thumbnail primarily to see more detail in the image before they selected it. Other subjects, however, preferred the smaller thumbnail sizes, and displayed as many as possible in the thumbnail viewing pane for apparent rapid scanning of multiple thumbnails. Subject P20 took this approach. Subject P20 left the video player in the default configuration of small thumbnail images arranged five to a row, and then would rapidly click on the thumbnail to display the static image in the video area, and make most of her determinations on the image in the Video AOI.

We also had subjects comment about the lack of spacing between the voting buttons, and their proximity to the video control buttons. Some subjects commented that because of the voting buttons being so close, they would accidentally hit a “No” button when they intend to hit a “Yes” button. This usability issue would also affect the accuracy of the accuracy measures of the study, given the small sample size and the small number of target videos within each set.

Playing Videos: When playing videos, most subjects would skip to specific locations in the video instead of using the Play or Fast Forward features to view most the video. Subjects often described their approach as playing a few seconds, then skipping to the

middle and then skipping again to the end of the video, playing a few seconds at each location to determine if the scene was consistent throughout the video or to validate the occurrence of an event. Several subjects either specified the 25% - 50% - 75% spots while others used a more generic "beginning, middle, and end" estimation. One subject even reported using the Voting Buttons below the Video area as visual aids to gauge the spots.

Use of Audio: Subjects reported making very little use of audio clues to identify or confirm events, although some subjects used it more than others. Subjects who used audio tended to use it both to verify targets and to eliminate non-target videos. For example, several subjects knew that parkour usually has a certain type of music associated with it. If they heard that type of music, they would pay closer attention to the video. If they did not hear the music, they would quickly eliminate the video. Birthday party, Parkour, Flash Mob, and Parade are some of the event types for which several subjects reported using or relying on audio clues to confirm the event. Some subjects would listen for a voice track or narrative. If they heard a voice track, they would listen more closely to determine if the narrative provided clues on the event types. But very few subjects reported a consistent strategy for using audio clues across all tests, and several subjects reported that in many tests they did not listen for any audio clues.

Learning/Familiarity Effect: Several subjects commented that they detected certain categories or types of video not associated with the test event types that would appear frequently in the test data sets. Examples of these types of video that subjects identified included cooking shows, highland dance competitions and performances, girls putting on makeup, and people receiving haircuts. Subjects commented that these videos were often obvious from the thumbnails, and could easily be determined as negatives without playing the video. They felt that this developed familiarity of the general categories of videos that are available on the Internet enhanced their speed (throughput) for the tests that contained many of these types of videos. Subjects also reported "learning" that with the ranked video sets the targets would occur early in the set of videos, so they could move more quickly in processing the videos toward the end of the test.

The above summaries represent patterns that apply to several or most of the subjects, as either observed through the eye-tracking recordings or reported by the participant during the post-test interview. But another interesting comparison is between two subjects – P18 and P20 – who demonstrated very different behaviors to achieve the same goal of processing the videos as quickly as possible. Both subjects had average throughput scores above 990: P18 averaged 998; P20 averaged 990. Both subjects also worked toward finishing the tests in a relatively short amount of time as compared to the other subjects, in some cases finishing their tests around the one-hour mark. Some of their visual search behaviors were similar, but in aspects of visual search and their use of the interface, they took very different approaches to the process.

Their areas of similarity included a heavy reliance on the thumbnail or static image to decide about the video. They both described using scene characteristics to establish a set of filters for each event to quickly eliminate videos based on the thumbnail only that they had high confidence would not contain the target. They also both described viewing the video at three points – beginning, middle, and end – when they felt they needed to watch the video to decide. Their differences occurred in two main areas: 1) how they configured the video player interface, and 2) their approach for adapting their visual search as they progressed through the sequence of tests.

After his first test, Subject P18 configured the video player to narrow the thumbnail viewing area to the width of one thumbnail image. For the second and third tests, he still used the small thumbnail size in the narrow view. Subject P18's explanation for narrowing the thumbnail viewing area to the width of one video was to focus his attention more on the thumbnail. In the post-test interview for his second test, Subject P18 said that "with the narrower band of thumbnails his eyes did not wander as much; he found focusing on a smaller region was more helpful. Since Subject P18 did not have to focus on where to look, it was easier to focus on what he was looking at. The change also made it easier for him to click on the No button because he would focus on the thumbnail closest to the buttons, instead of at the top of the column. But in the middle of his fourth test, he changed the size of the large thumbnail image from small to large. Subject P18

said he changed the size to improve his focus. He also played more of the video for his fourth test, which was E009 – Getting a Vehicle Unstuck. Subject P18 was the one subject who had his first set of rank-ordered videos for the fifth test. For this test, he noted that he developed a practice of watching a larger number of videos for the first 100 tests to get a sense of the filters, and then could more confidently filter out videos based on the thumbnails only.

For his Test 6, which was E006 – Birthday Party, P18 said he changed his overall approach from a sequential scan of the thumbnails to ones that obviously had the event. He wanted to identify the obvious target videos first, and then spend more time reviewing the remaining videos. He made this change to his approach because he was getting frustrated with trying to identify the target events and felt he would run out of time. Subject P18 noted during the first test that he could visually process the thumbnail images faster than the interface loaded them – at times he wanted to move faster but the interface was holding him back. Subject P18 made this same comment on a few of his post-test interviews, at one point stating that he wished he had the option to reduce the resolution of the images to speed up the process of loading them.

Subject P20 did not configure the video player interface, but instead left it at the default settings of small thumbnail images in a thumbnail viewing area that was five thumbnails wide. Instead of focusing on the thumbnail images, she would click on them to view the still image in the larger video player viewing area. Subject P20 would move quickly between the Thumbnail AOI and the Video AOI and the Voting Buttons, usually selecting thumbnail images that were closest to the Voting Buttons. On one test, Subject P20 reported in the post-test interview that she was aware that she had accidentally clicked the "No" button on one or two thumbnail images when she meant to click on "Yes." She reported changing the sequence or order of how she processed the videos, instead of following a straight sequential order. But, for the most part, aside from developing a sense of the filters based on scene characteristics, or using audio clues to support some tests where the audio added value in distinguishing, such as on E008 –

Flash Mobs and E006 – Birthday Party, Subject P20 did not vary her approach as she progressed through the tests.

These two different approaches taken by subjects P18 and P20 resulted in similar rates of throughput, but different rates in accuracy. Subject P18 had an accuracy rate of 90.41 percent, and a false alarm rate of only 0.92 percent. Subject P20 had an accuracy rate of 81.06 and a false alarm rate of 8.11 percent. Clicking on the wrong voting button a couple of times, as Subject P20 reported, alone should not account for a nine percent different in accuracy, and would not influence seven percent higher false alarm, unless she made similar errors in the other direction. Although comparing two subjects to each other is not statistically valid, this comparison serves to identify further areas of experimentation with video player interfaces to validate what features and multimedia search behaviors promote both speed and accuracy at the same time.

The comparison also highlights another potential factor that influenced the subjects' performance, which is their motivation, and the need to control for that factor in future studies. Although the intent of the test was to measure both speed and accuracy, the conditions of the test environment did not reward the subjects for either variable. Subjects were paid to complete in the tests, defined as either reaching the two-hour time limit or by making a decision on all 1,000 thumbnails and videos in the test set. Subjects received their payment regardless of whether they took the full two hours or significantly less time, and whether they had a high or low rate of accuracy. Some subjects may have been motivated to finish as quickly as possible without a significant concern for accuracy because they knew that they would receive the same payment regardless of the amount of time spent. Subject P20's behavior may be indicative of a lack of concern with accuracy, as evidenced by her lack of experimenting with different configurations of the interface to improve performance, as subject P18 did. Subject P18 made specific comments in the post-test interviews about how he altered the configuration to reduce eye movement and improve focus on the thumbnails as approaches to improving his performance. Subject P18 also asked whether he would receive the results of his accuracy, indicating that he was concerned about that aspect of his performance.

Chapter 5: Conclusions

The results of this study failed to identify any consistent patterns of how subjects allocate their visual attention between the Thumbnail AOI and the Video AOI that correlate to improved accuracy or how they approach searching for target events with objects included in the semantic definition. The results, however, did indicate a preference for the use of Thumbnail AOI and Thumbnail Images, especially when subjects were motivated by speed, or throughput. The first result that indicates a preference for using the Thumbnail AOI is the difference in the allocation of visual attention by perceived difficulty of the test Event. The ANOVA results, and the chart that shows each subject's mean ratios for each of the four metrics of visual attention, indicate that the subjects had a higher ratio of Thumbnail AOI use for LessDifficult Events than for Difficult Events, indicating a preference for relying on obtaining the "gist" of a video through the surrogates of static images. Another indicator of the preference for the Thumbnail AOI is that the allocation of visual attention by Test Order indicates that, as subjects progressed through the sequence of tests, they came to rely less on the Video playback feature. The feedback that subjects provided during the post-test interviews reinforce the data, as most of them reported that, as they learned more about each event type and the set of videos, they gained confidence in using only the thumbnail images to make decisions – especially with the rank-ordered set of videos that had machine-learning filters applied to them.

The positive correlation of the ratios of the allocation of visual attention to throughput also indicates that, to increase their throughput, subjects increased their allocation of visual attention to the Thumbnail AOI, or decreased their allocation of visual attention to the Video AOI. This finding makes sense – playing videos takes time, and when throughput is a motivator, subjects will reduce the amount of time they spend viewing videos. While the findings indicate that size matters, the size of the thumbnail image, and not the overall amount of the interface allocated to the thumbnails, seems to matter more. Subjects who increased the size of their thumbnail images had more focused

visual attention, but less overall time in the Thumbnail AOI area, indicating that they made more focused use of their visual attention. As noted in the Qualitative Results section, one subject commented specifically about changing to the larger Thumbnail Image size, but reducing the Thumbnail AOI size, to increase his focus on the content of the thumbnails instead of moving between the two areas.

These results reinforce that humans have a strong ability to quickly discern "gist" from observing visual scenes, or images, and prefer to use that ability to expedite the multimedia event detection process. The overall high rate of accuracy of the subjects, combined with indicators of a preference for allocating their visual attention to thumbnail surrogates over videos, suggests that humans rely on that ability when faced with decision-making situations under time-critical circumstances. The results of this study provide the following insights to the design of interfaces for multimedia information retrieval systems:

1. Ensure that users can easily identify how to change both the size of the thumbnail image and the size of the thumbnail viewing area to support individual preferences for viewing the surrogates
2. Provide markers or controls for the video playback feature to enable users to either skip to different segments of the video, either by quantifiable measures (i.e., quarter marks) or by a machine-detected change in the visual scene, or to enable playback at increased speeds, to reduce the amount of playback time required.
3. Allow users to select the resolution of images displayed to enhance the speed at which the images are rendered in the video player. Allow users to change the resolution on demand if they need a higher resolution to decide based on the image.
4. Allocate adequate size and spacing for video controls and other buttons or controls that determine the disposition of a video clip to minimize mistakes made caused by moving the mouse quickly between locations to increase throughput.

The findings from this study may also inform training or guidance to new multimedia analysts who need to identify semantically-defined events. This guidance could address visual search behaviors that reinforce accuracy over speed, such as configuring the interface to enable focus on the image. The guidance or training could also use examples of different types of semantically defined events to help the analysts quickly identify the characteristics that make events easier to discern from images and the characteristics that make it more difficult to identify targets so that they can consciously assess how to adjust their visual search approach to the context of the event. The qualitative results also suggest that multimedia analysts should take time to assess the impact of how applied machine-learning filters rank video sets so the analysts can adjust their approach to video search and review to maximize the human contributions of accurate detection to the process. Any training and guidance should also caution multimedia analysts against relying too heavily on machine-applied filters to a point of sacrificing accuracy for the sake of throughput.

Recommendations for Further Study

The overall high accuracy and high throughput results of the initial baseline study suggest that perhaps the NIST MED'11 corpus was not robust enough to provide a challenge that is representative of the challenge that multimedia event detection analysts face in their daily work. Future studies on human performance in multimedia event detection should use a more challenging corpus that includes events that

- cannot easily be determined by setting/environment or objects alone,
- are defined primarily by actions or activities
- use the presence of objects to help define the event, but the lack of an object does not necessarily negate the presence of a target event, or
- are embedded in the video as a minor portion of, or background to, the overall theme or subject of the video.

The combination of the small sample size and the tight clustering of accuracy scores and overall throughput limits the insights provided by the statistical analysis in this study. Future studies should increase the number of participants and use a more challenging test

corpus that produces a wider range of accuracy and throughput scores to validate not only the correlations that this study identified as significant, but to also test the correlations on the hypotheses for which this study failed to reject the null hypothesis statement.

Another consideration for a study that is more representative of the work that multimedia event detection analysts face would be to conduct this experiment in a more realistic environment of how multimedia analysts currently perform their work instead of a laboratory environment with time constraints on the tests. A more realistic setting without time-bound tasks would result in estimates of throughput as a ratio of processing time.

Further analysis on this study could include taking the time to create an area of interest for each individual thumbnail in the thumbnail viewing area to calculate the count and duration of fixations on individual thumbnail images. This analysis would provide more detailed insight into how each subject allocated visual attention to individual thumbnail images prior to deciding whether a video contained the target. This information would be especially useful for comparing visual attention behavior between thumbnail images of different sizes.

In this study, subjects knew that the goal was to achieve high accuracy in a limited amount of time. Both speed and accuracy were important, but with only a clock for feedback, subjects had no knowledge of how accurately they were performing the task. Also, subjects were paid to complete the test, regardless of how long they took to complete the test or how accurate they were. Therefore, speed naturally became the priority consideration, or perhaps the only consideration, for most subjects. While keeping each test double-blind, moderators could provide feedback on the subject's performance after a test session as an indirect incentive for accuracy. Or perhaps the test moderators would reward the subjects based on accuracy and speed, instead of only for completing the test. Measuring the allocation of visual attention when accuracy is a driver of performance may yield different patterns, or a wider range of user behaviors.

To understand better the influence of the features of the multimedia player on user behavior, future studies should control for the configuration of specific features to assess

the extent to which they impact user performance. However, this recommendation is not as critical for general usability improvements, as usability issues and opportunities for improving interface design can be identified through less formal means of usability testing and experimentation.

References

Ahmad, G. (2016). Associative relevance based eye fixations enhance decision making processes in scene perception. *Journal of Management Research*, 8(2), 1-17.

Amir, A., Berg, M., & Permuter, H. (2005). Mutual relevance feedback for multimodal query formulation in video retrieval. *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval* (pp. 17–24). New York: ACM.

Belke, E., Humphreys, G. W., Watson, D. G., Meyer, A. S., & Telling, A. L. (2008). Top-down effects of semantic knowledge in visual search are modulated by cognitive but not perceptual load. *Attention, perception, & psychophysics*, *70*(8), 1444-1458.

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177*, 77-80.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 753-763.

Castelhano, M. S., & Witherspoon, R. L. (2016). How You Use It Matters Object Function Guides Attention During Visual Search in Scenes. *Psychological science*, 0956797616629130.

Coutrot, A., Guyader, N., Ionescu, G., & Caplier, A. (2012). Influence of soundtrack on eye movements during video exploration. *Journal of Eye Movement Research*, *5*(4), 2.

Cunningham, S. J., & Nichols, D. M. (2008). How people find videos. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 201-210). Pittsburgh, PA: ACM.

De Groot, F., Huettig, F., & Olivers, C. N. (2016). When meaning matters: The temporal dynamics of semantic influences on visual attention. *Journal of experimental psychology: human perception and performance*, *42*(2), 180.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, *11*(5), 14-14.

Fine, M. S., & Minnery, B. S. (2009). Visual salience affects performance in a working Memory Task. *The Journal of Neuroscience*, *29* (25), 8016 -8021.

Gravetter, F.J. & Wallnau, L.B. (2014). *Essentials of statistics for the behavioral sciences*, 8th ed. Belmont, CA: Wadsworth, pp. 232-236.

Hart, S. G. (2006). Nasa-Task Load Index (Nasa-TLX): 20 Years Later. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, *50*(9), 904-908.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload*, *1*, 139–183.

Hauptmann, A. G., Lin, W.-H., Yan, R., Yang, J., & Chen, M.-Y. (2006). Extreme video retrieval: joint maximization of human and computer performance. *Proceedings of the 14th annual ACM international conference on Multimedia*, (pp. 385–394). New York, NY: ACM.

Hearst, M. (2011). User interfaces for search. In R. Baeza-Yates & B. Ribeiro-Neto (Eds.) *Modern Information Retrieval: The Concepts and Technology Behind Search* (2nd ed.), p. 21-55.

Henderson, J.M., & Ferreira, F. (2004) Scene perception for psycholinguists. In *The interface of language, vision, and action: Eye movements and the visual world.* New York: Psychology Press, pp. 1-58.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, 50(1), 243-271.

Henderson, J. M., Williams, C. C., Castelhano, M. S., & Falk, R. J. (2003). Eye movements and picture processing during recognition. *Perception & Psychophysics*, *65*, 725-734.

Hughes, A., Wilkens, T., Wildemuth, B. M., & Marchionini, G. (2003). Text or pictures? An eyetracking study of how people view digital video surrogates. In E. M. Bakker, M. S. Lew, T. S. Huang, N. Sebe, & X. S. Zhou (Eds.), *Image and Video Retrieval* (Vol. 2728, pp. 271-280). Berlin, Heidelberg: Springer . Retrieved August 8, 2011 from http://www.springerlink.com/content/v1qjmbkhudtna59x/.

Hwang, A. D., Wang, H. C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision research*, *51*(10), 1192-1205.

IBM developerWorks: IBM Multimedia Analysis and Retrieval System Communities. Retrieved October 11, 2011 from https://www.ibm.com/developerworks/mydeveloperworks/groups/service/html/communityview?communityUuid=7dc62548-8bc8-42c4-b2e9-150dde7c649a.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10-12), 1489-1506.

Järvelin, K., & Ingwerson, P. (2004). Information seeking research needs extension towards tasks and technology, *Information Research* 10(1) paper 212. Retrieved August 12, 2011, from http://informationr.net/ir/10-1/paper212.html.

Lee, H., & Smeaton, A. (2002). Designing the user interface for the Físchlár Digital Video Library. *Journal of Digital Video Processing*, *2*(4), pp. 1-20.

Malcolm, G. L., Groen, I. I., & Baker, C. I. (2016). Making sense of real-world scenes. *Trends in Cognitive Sciences*, *20*(11), 843-856.

Malcolm, G. L., Rattinger, M., & Shomstein, S. (2016). Intrusive effects of semantic information on visual selective attention. *Attention, Perception, & Psychophysics*, *78*(7), 2066-2078.

Mei, Yang, Hua, and Li. (2011) Contextual video recommendation by multimodal relevance and user feedback. *ACM Transactions on Information Systems*, 29(2), Article 10, 24 pages.

National Institute for Standards and Technology (NIST). (2011). 2011 TRECVID Multimedia Event Detection Track (updated August 7, 2011), NIST Information Technology Laboratory, Retrieved August 9, 2011, from http://www.nist.gov/itl/iad/mig/med11.cfm.

Neider, M. & Zelinksy, G. (2006). Scene context guides eye movements during visual search. *Vision Research, 46*, 614-621.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Empowering People*, (pp. 249–256). Seattle, WA: ACM.

Nelson, W. W., & Loftus, G. R. (1980). The functional visual field during picture viewing. *Journal of Experimental Psychology: Human Learning & Memory*, *6*, 391-399.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3), 145-175.

Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003, September). Top-down control of visual attention in object detection. In *Proceedings of the 2003 International Conference on Image Processing*, (Vol. 1, pp. I-253). IEEE.

Over, P., Awad, G., Michel, M.., Fiscus, J., Kraaij, W. & Smeaton, A. (2011). TRECVID 2011 -- An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics, *Proceedings of TRECVID 2011*, National Institute of Standards and Technology.

Ponceléon, D. & Slaney, M. (2011). Multimedia information retrieval. In R. Baeza-Yates & B. Ribeiro-Neto (Eds.) *Modern Information Retrieval: The Concepts and Technology Behind Search* (2nd ed.), p. 587-639.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, 62, 1457-1506.

Sauro, J. & Lewis, J.R. (2012). *Quantifying the user experience.* Waltham, MA: Morgan Kaufmann, pp. 117-119.

Schoeffmann, K., Hopfgartner, F., Marques, O., Boeszoermenyi, L., & Jose, J. M. (2010). Video browsing interfaces and applications: a review. *SPIE Reviews*, *1*(1), 018004.

Schoeffmann, K., Ahlström, D., Bailer, W., Cobârzan, C., Hopfgartner, F., McGuinness, K., & Bai, H. (2014). The video browser showdown: a live evaluation of interactive video search tools. *International Journal of Multimedia Information Retrieval*, *3*(2), 113-127.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages* (pp. 336-343).

Song, Y., & Marchionini, G. (2007). Effects of audio and visual surrogates for making sense of digital video. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (p. 867). Presented at the SIGCHI conference, San Jose, California, USA.

Tobii Technology AB. (September 2010). *Tobii Studio™ 2.X User Manual*, Manual release 1.0, 130-135.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766-786.

Tzoukerman, E. Ambwani, A. Chipman, L., Davis, A., Farrell, R., Houghton, D., Jojic, O., Neumann, J., Rubinoff, R., Shevade, B. & Zhou, H. (2012). Semantic multimedia extraction using audio and video. In Maybury, M., (ed.), *Multimedia information extraction: Advances in video, audio, and imagery analysis for search, data mining, surveillance, and authoring.* Hoboken, NJ: John Wiley & Sons, pp. 159-174.

Ware, C. (2008). *Visual thinking for design*. Amsterdam: Morgan Kaufmann.

Wilson, M.L., Kules, B., Schraefel, M.C., Schneiderman, B. (2010) From keyword search to exploration: designing future search interfaces for the web. *Foundations and Trends in Web Science*, 2(1), p. 1-97.

Wilson, M. L., Schraefel, M.C., & White, R. W. (2009). Evaluating advanced search interfaces using established information-seeking models. *Journal of the American Society for Information Science and Technology, 60*(7), 1407-1422.

Xu, Q. (2012). Interface design and usability. In Costello, V. (2016). *Multimedia foundations: Core concepts for digital design*. Oxford, UK: CRC Press.

Yang, M., & Marchionini, G. (2005). Deciphering visual gist and its implications for video retrieval and interface design. *CHI '05 extended abstracts on Human factors in computing systems* (pp. 1877–1880). Portland, OR, USA: ACM.

Yarbus, A. (1967). *Eye movements and vision.* New York: Plenum Press.

Appendix A: Acronyms and Definitions

Acronym	Term	Definition
AOI	Area of Interest	A static portion of the stimuli (viewing area) of an eye-tracking study used to measure the frequency and time spent looking within the area by subjects during a test.
FC	Fixation Count	This metric measures the number of times the participant fixates on an AOI or an AOI group.
IMARS	IBM Multimedia Analysis and Retrieval System	A system that can be used to automatically index, classify, and search large collections of digital images and videos.
MED	Multimedia Event Detection	The detection of a target event type within multimedia (i.e., videos). The Multimedia Event Detection (MED) evaluation track is part of the TRECVid Evaluation. The goal of MED is to assemble core detection technologies into a system that can search multimedia recordings for user-defined events based on pre-computed metadata.
MIR	Multimedia Information Retrieval	Systems that support the search and selection of multimedia files for review.
NIST	National Institute for Standards and Technology	Government agency that prepared the TRECVid video corpus used for the Multimedia Event Detection Human Performance Baseline test.
TFD	Total Fixation Duration	This metric measures the sum of the duration for all fixations within an AOI, thus the N

		value used to calculate the descriptive statistic is based on the number of recordings.
TRECVid	Text Retrieval Conference Video Retrieval Evaluation	A conference series sponsored by NIST devoted to research in automatic segmentation, indexing, and content-based retrieval of digital video.
TVD	Total Visit Duration	This metric measures the duration of all the visits within an AOI or AOI group. In this case the N value used to calculate descriptive statistics is based on the number of recordings. A visit is defined as the interval of time between the first fixation on the AOI and the next fixation outside the AOI.
VC	Visit Count	This metric measures the number of visits within an AOI or AOI group. Each individual visit is defined as the interval of time between the first fixation on the AOI and the next fixation outside of the AOI.

Appendix B: NIST Multimedia Event Detection Event Descriptions

Event Name: Birthday Party (E006)

Definition: An individual celebrates a birthday with other people

Explication: A birthday in this context is the anniversary of a person's birth. Less commonly, the term "birthday" can be used to refer to the anniversary of an organization's establishment, but a celebration for an organization does not satisfy the event definition.

A birthday celebration is a gathering of people who have been invited by the host or hosts to come to a set location (often a private home, sometimes a restaurant, bar, nightclub, park, or other public venue) to socialize in honor of the person(s) whose birthday it is (the birthday celebrant(s)).

Birthday parties, as with other parties/celebrations, will typically feature an assortment of food and beverages. Birthday parties are often accompanied by colorful decorations, such as balloons and streamers, and some people may wear cone-shaped "birthday hats". The decorations may include signs or banners displaying a message for the birthday celebrant. Often, especially for children's parties, guests will bring cards and/or gifts wrapped in shiny/colorful paper or bags, which will be opened by the birthday celebrant(s), or by their parents/siblings if the birthday celebrants are too young to open the gifts by themselves. A cake (or sometimes cupcake or other food item) with lit candles, called the "birthday cake", is often served. The song "Happy Birthday to You" may sung by the guests while the birthday cake with lit candles is carried to a table or counter where the birthday celebrant(s) are seated. The birthday celebrant(s) then blow out the candles, usually after the song is finished, and the guests then clap and cheer. Birthday parties may also involve games or other organized group activities.

Evidential description:

- scene: indoors (a home, a restaurant) or outdoors (backyard, park); day or night

- objects/people: decorations (balloons, streamers, conical hats, etc), birthday cake (often with candles), birthday celebrant, guests, gifts
- activities: singing, blowing out candles on cake, playing games, eating, opening gifts
- audio: singing "Happy Birthday to You"; saying happy birthday; laughing; sounds of games being played

Event name: Changing a vehicle tire (E007)

Definition: One or more people work to replace a tire on a vehicle

Explication: A vehicle is any device, motorized or not, used to transport people and/or other items. Tires are ring-shaped inflated objects, usually made of rubber, that fit over the wheel of a vehicle. The process for replacing a tire includes removing the existing tire and installing the new tire onto the wheel of the vehicle. Tires typically are replaced because they are damaged or worn down. If a tire is damaged and loses air pressure as a result, it is called a "flat tire". Generally the driver of the vehicle with a flat tire will stop the vehicle as soon as possible and replace the affected tire with a temporary tire called a "spare tire", which may be stored elsewhere on/in the vehicle. In other cases, the tire may be changed not by the vehicle operator, but by a professional (e.g. a mechanic) who may use dedicated tools and work in a repair shop or similar setting.

Evidential description:

- scene: garage, outdoors, street, parking lot
- objects/people: tire, lug wrench, hubcap, vehicle (car, bike, lawn mower, etc.), tire jack
- activities: removing hubcap, turning lug wrench, unscrewing bolts, pulling rim out of tire
- audio: narration of the process; sounds of tools being used; street/traffic noise; background noises from repair shop

Event name: Flash mob gathering (E008)

Definition: A coordinated large group of people assemble suddenly in a public place, perform a predetermined act to a surprised public, then disperse quickly
Explication: A flash mob is a group of people in a public place surprising the public by doing something unusual in a coordinated fashion. Flash mobs usually consist of people either suddenly starting to perform a coordinated activity (e.g. dancing, freezing in place, pretending to fight each other, re-enacting a movie scene), or performing normal activities but with coordinated appearance/clothing or holding a coordinated prop object. In coordinated activity flash mobs, participants (known as "flashmobbers") start out by engaging in routine activities in a public place, and are initially indistinguishable from people who are not part of the flash mob -- until they begin performing the pre-determined coordinated activity. A flash mob typically ends with all of the flash mobbers suddenly going back to normal activities at the same time. Often flash mobs begin with one or two people and then more and more flash mobbers join. Sometimes flash mobs (particularly ones where dancing is the activity) are accompanied by background music. A previously-designated sound such as a clock tower chiming, a whistle being blown, or music beginning/ending may be used to signal the start or end of the flash mob. In some cases a clip may show the flash mob group assembling prior to the flash mob activity starting; this may consist the flash mob leader/leaders speaking to the group using a megaphone.

Evidential description:

- scene: indoor or outdoor, public place
- objects/people: a very large group of people, typically no objects involved
- activities: a wide range of activities can be performed, including dancing or singing in unison, moving in a coordinated fashion, or simply milling around
- audio: background music; sound that designates start/end of the flash mob activity; leader speaking to group of assembled flash mobbers

Event name: Getting a vehicle unstuck (E009)

Definition: One or more people work to free a motorized or unmotorized vehicle that is stuck.

Explication: A stuck vehicle is one that either cannot move, or can only move in a limited area (e.g. a parking space). Usually movement is restricted either because the vehicle is located on/in a substance that prevents the wheels of the vehicle from having enough friction to propel the vehicle forward (e.g. mud, mud covered in water, snow, etc.), or because the angle of the vehicle and/or surrounding obstacles prevent it from moving normally. Getting a vehicle unstuck can be done by maneuvering the vehicle itself or by using another device (another vehicle, an object, etc.) with the intention of getting the vehicle to a location or angle where it can again move normally.

Evidential description:

- scene: typically outdoors, day or night, any weather
- objects/people: vehicle (car, boat, bicycle, construction vehicle), person operating the vehicle, possibly other people assisting in or outside the vehicle, possibly additional vehicle(s) assisting
- activities: steering, hitching, chaining, driving, sliding, skidding, pulling
- audio: vehicle tires spinning against surface; engine noise from vehicle; narration or commentary on process from participants

Event name: Grooming an animal (E010)

Definition: One or more people groom an animal

Explication: Grooming refers to caring for the hygiene/cleanliness and appearance of the animal. A very common form of grooming is bathing the animal, usually accomplished by either immersing the animal in water or spraying the animal with water, often followed by application of soap/shampoo and then additional rinsing with water. Other grooming activities include trimming of hair and nails, cleaning of teeth, eyes, and ears, and brushing, combing, and styling the fur of the animal. Procedures performed to address a medical ailment, such as application of topical ointment to the skin, are not generally considered to be examples of grooming.

Evidential description:

- scene: outdoors, in a yard or corral, indoors in bathroom, grooming salon,
- exhibition center
- objects/people: sink, bathtub, hose, shower, soap, shampoo, scissors, clippers
- activities: spraying hose, putting animal on table, rinsing, blow-drying fur, cutting fur, clipping nails
- audio: human talking to the animal; animal noises (e.g. purring);
- narration of the process

Event name: Making a sandwich (E011)

Definition: Constructing an edible food item from ingredients, often including one or more slices of bread plus fillings

Explication: Sandwiches are generally made by placing food items on top of a piece of bread, roll or similar item, and placing another piece of bread on top of the food items. Sandwiches with only one slice of bread are less common and are called "open face sandwiches". The food items inserted within the slices of bread are known as "fillings" and often include sliced meat, vegetables (commonly used vegetables include lettuce, tomatoes, onions, bell peppers, bean sprouts, cucumbers, and olives), and sliced or grated cheese. Often, a liquid or semi-liquid "condiment" or "spread" such as oil, mayonnaise, mustard, and/or flavored sauce, is drizzled onto the sandwich or spread with a knife on the bread or top of the sandwich fillers. The sandwich or bread used in the sandwich may also be heated in some way by placing it in a toaster, oven, frying pan, countertop grilling machine, microwave or grill. Sandwiches are a popular meal to make at home and are available for purchase in many cafés, convenience stores, and as part of the lunch menu at many restaurants.

Evidential description:

- scene: indoors (kitchen or restaurant or cafeteria) or outdoors (a park or backyard)
- objects/people: bread of various types; fillings (meat, cheese, vegetables), condiments, knives, plates, other utensils
- activities: slicing, toasting bread, spreading condiments on bread, placing fillings on bread, cutting or dishing up fillings
- audio: noises from equipment hitting the work surface; narration of or commentary on the process; noises emanating from equipment (e.g. microwave or griddle)

Event name: Parade (E012)

Definition: A large group of people process for a celebration/commemoration of some event

Explication: A parade is a group of people processing either in celebration or commemoration of some event. Parades typically involve one or more groups of people proceeding down a route between two lines of spectators. Most often parades process down a street and the spectators are lined up on either side of the street. People in the parade may be driving cars, riding horses, and/or walking, dancing as a group in coordinated special dress or costumes, or riding on a parade float. Parade floats are decorated platforms that sit on top of a vehicle or are pulled by a vehicle or by people as part the procession. Parades are generally accompanied by music and by cheering or clapping from the spectators. Military groups may participate in parades, but not all military demonstrations constitute a parade.

Evidential description:

- scene: typically outdoors, any season, usually on a street
- objects/people: a very large group of people, with floats, costumes, props, vehicles, horses, megaphones
- activities: marching, walking, singing, dancing, clapping, yelling
- audio: music from bands; crowd cheering or clapping; announcers describing the goings-on; horns or other vehicle noises

Event name: Parkour (E013)

Definition: A person travels by foot from one point to another while performing various gymnastic maneuvers over/on/through pre-existing man-made obstacles, typically outdoors.

Explication: In parkour a person travels by foot as quickly as possible from one point to another while performing various acrobatic maneuvers (e.g. flips, swinging body over an object) and climbing over/on/through pre-existing obstacles such as trees, street lamps, fences, walls, children's playground equipment. Parkour is typically performed outdoors and often in an obstacle-dense urban environment. Free-running is a related sport in which people interact with pre-existing obstacles via acrobatic tricks. For purposes of this event kit, free-running is considered as a form of parkour. Though typically performed outdoors, parkour may be practiced or demonstrated indoors, usually in a gym. Although acrobatic tricks occur in other sports/activities (skateboarding, cheerleading, gymnastics, dancing, snowboarding, etc.), these sports are distinct from parkour. In parkour, the acrobatic tricks are performed as part of the interaction with the pre-existing obstacles and/or to enable the athlete to maintain as much efficient forward movement as possible.

Evidential description:

- scene: usually outdoors in an urban setting
- objects/people: natural environment, stairways, buildings, rooftops, fire escapes, playground equipment
- activities: running, scaling walls, jumping over obstacles, rolling
- audio: sounds of human hitting the various obstacles (e.g. feet hitting a picnic table); background music; occasional crowd reactions

Event name: Repairing an appliance (E014)

Definition: One or more people make repairs to a household appliance

Explication: Appliances are machines that are used for functions related to the maintenance or care of a home or office (e.g. cleaning, cooling/heating), or to assist people in performing a household task (e.g. cooking, washing). Major household

appliances are typically large and are often metallic, black or white in color, and may include: air conditioners, dishwashers, clothes dryers, drying cabinet, freezer, refrigerator, kitchen stove, water heater, washing machine, trash compactor, microwave ovens and induction cookers. Non-major household appliances are typically small appliances that perform a more specialized task, and include: coffee makers, toasters, stand mixers, food processors. Repairs to these items may include removing and replacing a part, or adjusting a part without replacing it.

Small machines that are held and moved with the hand while operating them, such as hand mixers, hair dryers, or electric toothbrushes, may also be considered appliances. Large, permanently-installed machines that perform household tasks (e.g. a garage door opener, a central heating/cooling unit) are also not usually classified as appliances. Finally, machines primarily used for entertainment rather than household tasks, such as televisions, CD and DVD players, cameras, clocks, video game consoles, home cinema systems, telephones, answering machines, etc. are not generally considered Appliances.

Evidential description:

- scene: typically indoors, in a home (kitchen, garage, basement)
- objects/people: appliance (dishwasher, refrigerator, toaster oven, washing machine etc); tools, rags, machine parts
- activities: unscrewing/screwing parts, lifting machine parts, squatting down, bending over, holding objects
- audio: sounds of tools making contact with object being repaired; sounds from power tools; narration of the process; sounds emanating from the appliance itself (e.g. garbage disposal being operated)

Event name: Working on a sewing project (E015)

Definition: One or more people work on constructing a garment or other object from fabric, sewn by hand or by machine.

Explication: Sewing consists of passing a needle attached to thread or other type of filament through the sewable material known as "fabric", which is typically a textile

but may also include leather, plastic or other materials. Sewing may be performed by hand or with the aid of a motorized or non-motorized machine. Sewing may involve constructing, repairing, or altering a garment or other object by sewing pieces of material together; it may also consist of sewing objects (e.g buttons, decorative trim) onto the material. Related activities involving needle and thread, known as "needle crafts" include knitting and embroidery but these are not generally considered sewing.

Evidential description:

- scene: typically indoors, in a workroom or craft room
- objects/people: sewing machine, needles, threads, fabric, patterns, scissors, pins, tape measure, ironing board & iron
- activities: measuring, cutting, sewing, ripping, tracing, bending over, pressing
- audio: sounds of sewing machine being operated; sound of fabric ripping; narration of process

Appendix C: Post-Test Questionnaire and Results

Post-Test Questionnaire

I based the post-test survey questions on the NASA Task Load Index (Hart, 2006; Hart & Staveland, 1998), which is an instrument designed to measure the physical and mental effort and levels of frustration that subjects perceive in performing tasks. The survey asked subjects to use a five-point Likert scale, with descriptors for each point in the scale, to respond to the following questions:

1. How difficult was it for you to identify videos that contain the event?
2. How successful do you think you were in identifying all the videos that contained this specific type of event?
3. How difficult was it to decide whether an event matched the event description you were given?
4. How difficult was it for you to visually search for this event type?
5. How difficult was it for you to maintain your focus on the task during the test?
6. Please indicate the amount of mental and perceptual activity required for the task (e.g., thinking, deciding, remembering, looking, searching, etc.)?
7. Please indicate the amount of time pressure you felt in completing the task.
8. Please rate the amount of frustration you felt during the task.

For all questions except number 2, the 1-5 Likert scale had 1 as being very easy and 5 as being very difficult. For question 2, the scale ranged from 1 as Found None (i.e., not successful) to 5 Found Most to All (i.e., very successful). Although the ordered Likert scales are not truly continuous interval values, they are averaged in the figures in this section to facilitate data visualization.

Post-Test Questionnaire Results

Figure 57 shows the average scores for the question on how each subject perceived the difficulty in identifying events by event type and test. This chart shows

similar patterns for event types that had high numbers of missed detections or false alarms. Specifically, the following events had higher scores:

- Random set E006 – Birthday party
- Random set E008 – Flash mob
- Random set E009 – Getting a vehicle unstuck
- Ranked set E012 – Parade

These event and test types also had more Medium and Low confidence ratings associated with their determinations, so these survey responses correspond both to the accuracy and the confidence ratings for these tests.

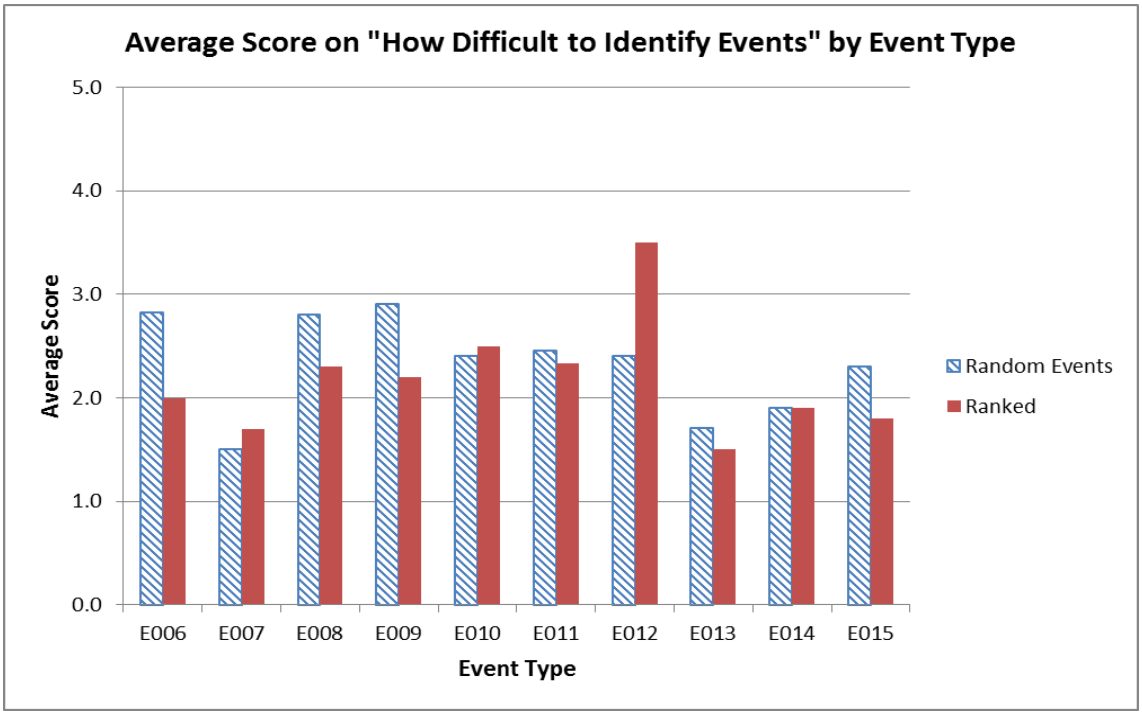

Figure 57: Post-test survey scores on "How Difficult to Identify Events" by event type

Figure 58 shows the average scores for the question "How successful do you think you were in identifying all the videos that contained this specific type of event?" The scale for this question ranged from 1 being "Found None" to 5 being "Found most to all," so the higher the number the greater the subject's perceived success. The average scores across all subjects are consistently high, with the all but one test having an average

score above 4.0. This overall high degree of confidence corresponds with the high P_{DET} scores across all subjects and all event types and to the confidence ratings associated with Correct Detections. The one test that has an average score below 4.0 is E010 – Grooming an Animal from the Ranked Order data set. This test did not have any targets, which may have caused the lower perceived success.

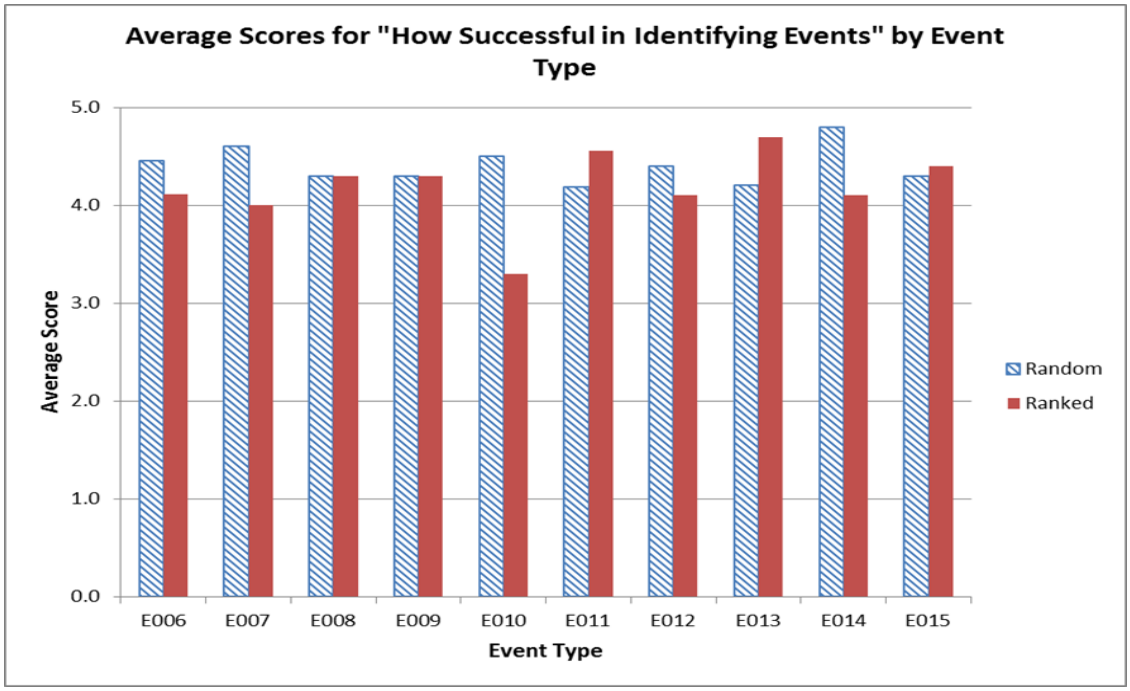

Figure 58: Post-test survey results for "How Successful in Identifying Events" by Event Type

Figure 59 shows the post-test survey scores for the question "How difficult was it to match events to their description," by event type. These scores are very similar to the scores for the question on difficulty to identify events, with the same four events and test types emerging as being more difficult for subjects. These events include

- Random set E006 – Birthday party
- Random set E008 – Flash mob
- Random set E009 – Getting a vehicle unstuck
- Ranked set E012 – Parade

These events also had higher false alarm rates than the other events, reinforcing the finding that many of the false alarms were a result of ambiguity of the event definitions and annotations.

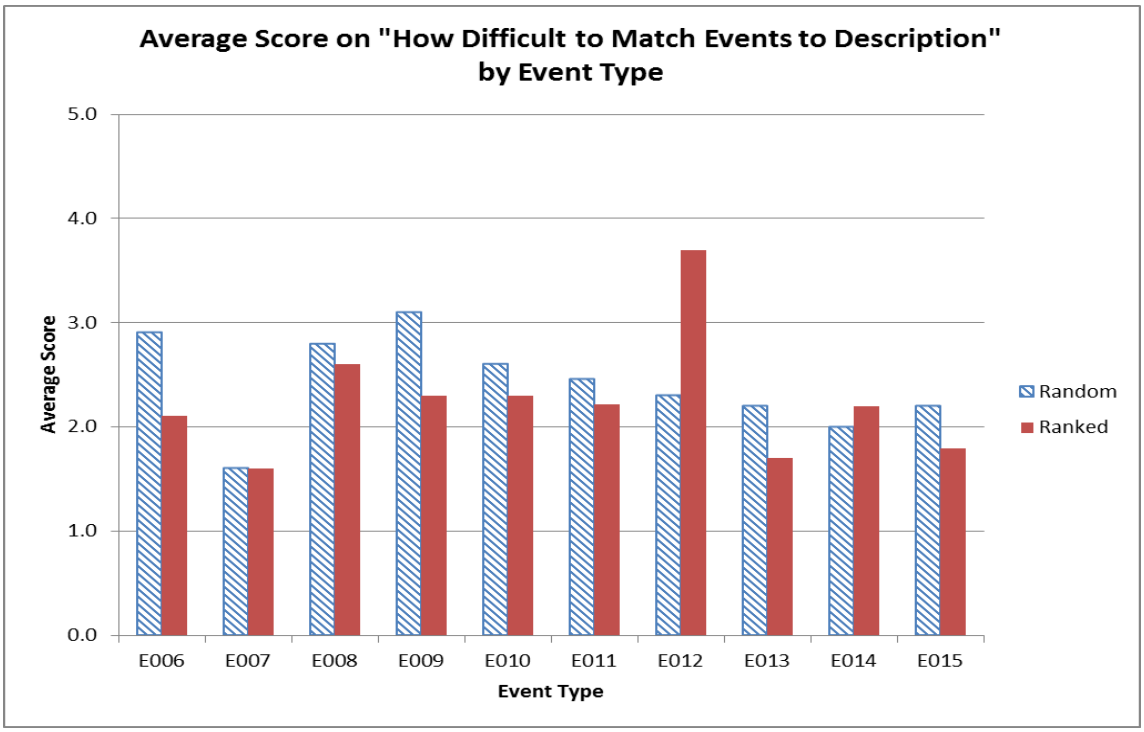

Figure 59: Post-test survey scores for "How difficult to match events to description" by Event Type

Figure 60 shows the post-test survey scores for the question "How difficult was it to visually search for events," by event type. The scores for this question show three of the same four event types emerge as being more difficult:

- Random set E006 – Birthday party
- Random set E009 – Getting a vehicle unstuck
- Ranked set E012 – Parade

Although Random E008 – Flash Mob does not appear as difficult to visually search for, another event and test type emerged as being more difficult: Random E015 – Working on a sewing project, which is the test on which subjects had 0 missed detections.

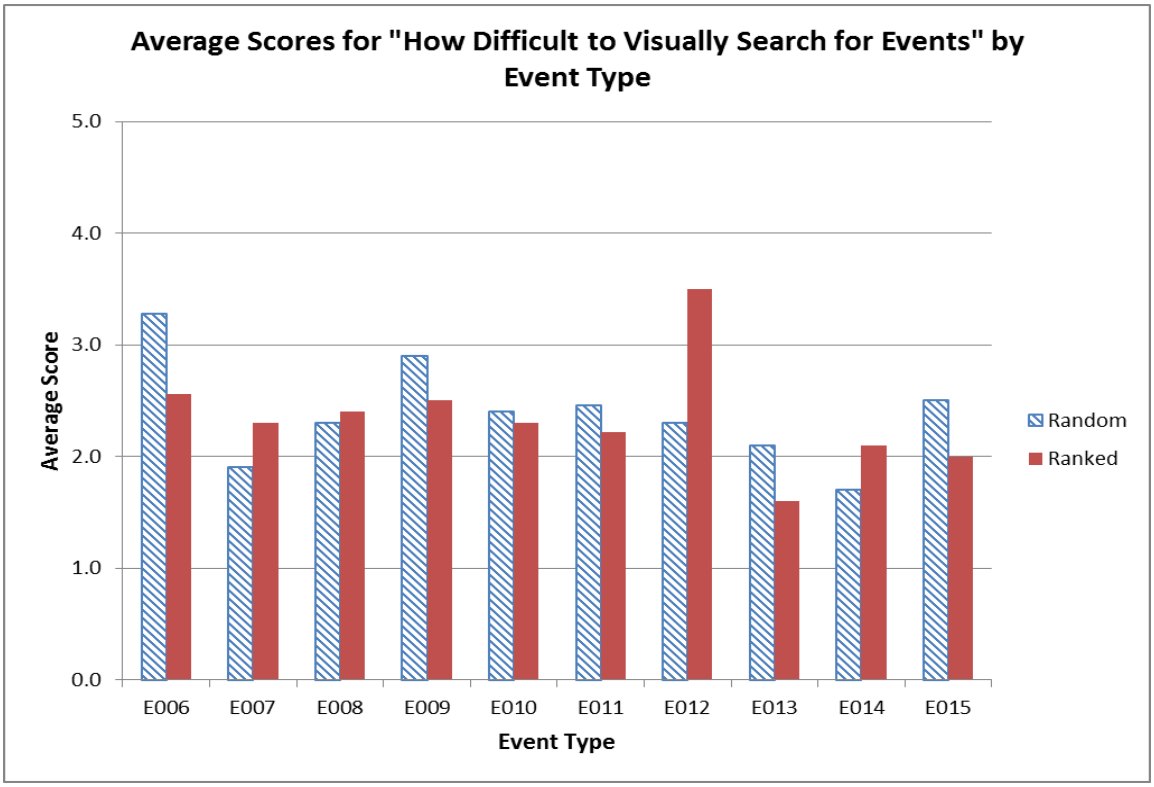

Figure 60: Post-test survey scores for "How difficult to visually search for events"

Figure 61 shows the results of the scores for the question of "How difficult was it to stay focused," by event type. Only two tests emerge as having higher than average scores compared to the other tests: Random E009 – Getting a Vehicle Unstuck and E015 – Working on a Sewing Project.

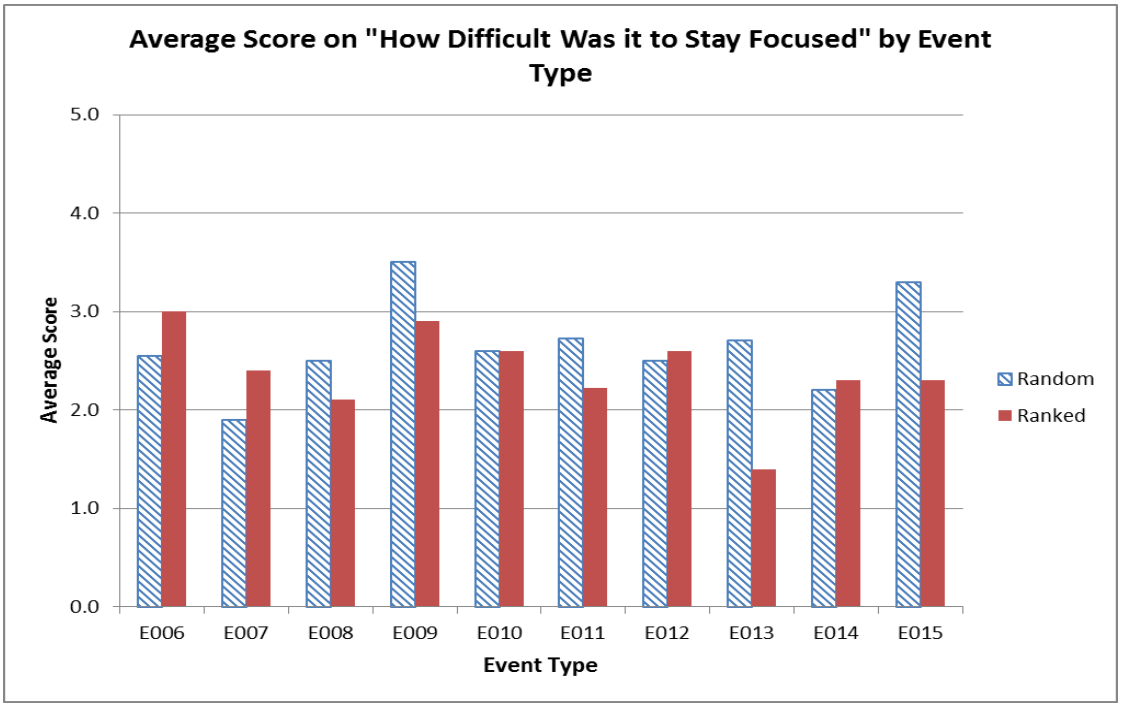

Figure 61: Post-test survey scores for "How difficult was it to stay focused"

Figure 62 shows the post-test survey scores for the question on the "Amount of mental and perceptual activity required for the task?" by event and test type. These results are similar to the results for the questions on Identifying events and Matching events. Three of the four same event and test types emerged as being more difficult. These events include:

- Random set E006 – Birthday party
- Random set E009 – Getting a vehicle unstuck
- Ranked set E012 – Parade

In addition, Random E015—Working on a Sewing Project emerges as being one of the more difficult tests for mental and perceptual effort and for Ability to Stay Focused. These results may result from the very low number of target events in the data set. The Random set E015 had only two targets; the Ranked set had only three targets to find. Other data sets, however, had equally low numbers and had lower scores for difficulty. Although subjects may have found Random E015 more difficult, it is the one test on which all subjects identified all the target videos.

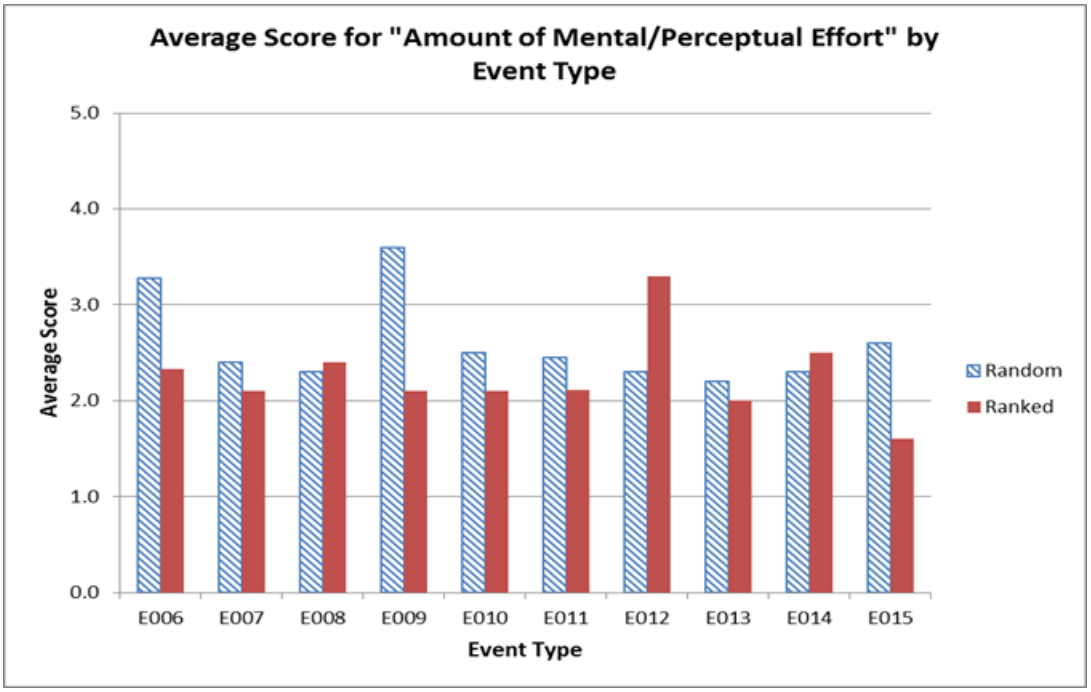

Figure 62: Post-test survey results for "Rate the amount of mental/perceptual effort"

Figure 63 shows the post-test survey scores for the question on “Amount of time pressure felt” during the test. Although all tests had the same two-hour time limit, the scores did vary slightly by event type. Also, almost all the Ranked tests had slightly lower scores (indicating less time pressure felt) than the Random tests. The one exception is Ranked E012 – Parade. Subjects felt slightly more pressure on this event in the Ranked set than in the Random set. Ranked E012 is the test that had the unusually high number of false alarms, and a lower percentage of videos played for those false alarm determinations. Subjects may have felt that they needed more time to play more of the videos to make accurate determinations.

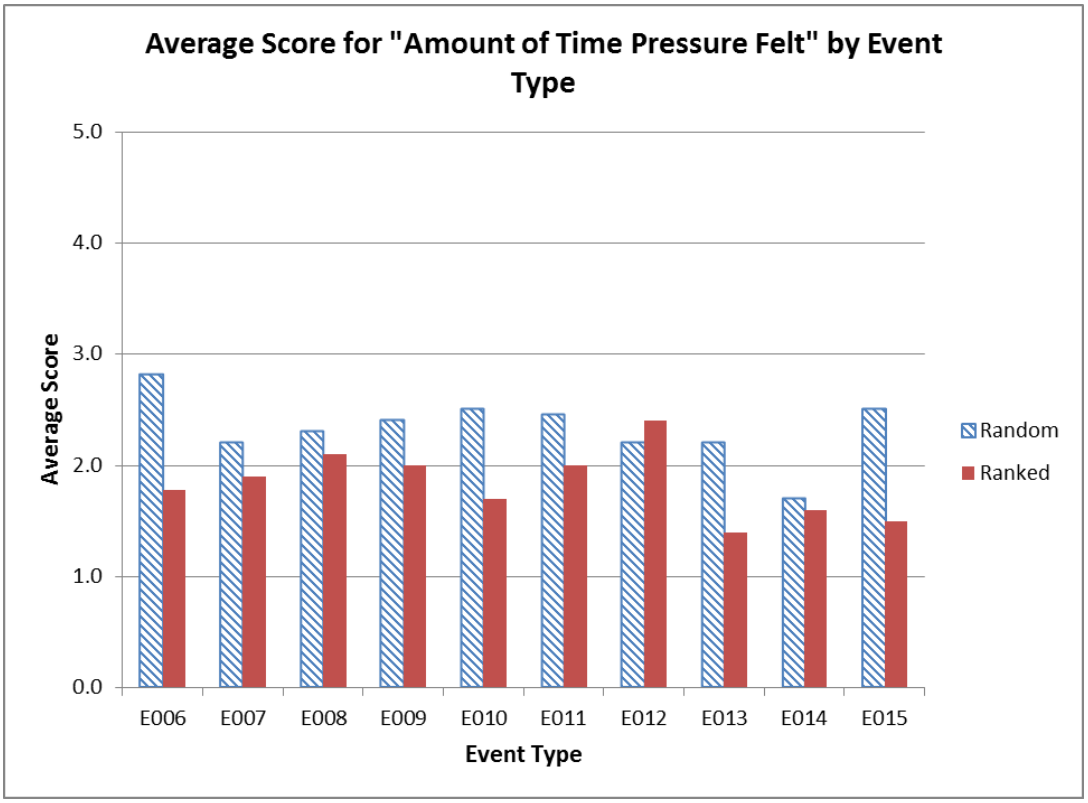

Figure 63: Post-test survey scores for "Amount of time pressure felt"

Figure 64 shows the post-test survey scores for the question on "Amount of frustration felt" by Event Type. The notably higher scores were the scores that reached or exceeded the 2.0 mark. For this question, the random order Event 009 – Getting a Vehicle Unstuck – had the highest score, perhaps because it was the event for which subjects were not clear on whether the vehicle had to become unstuck in the video for the video to match the target definition. Other events with higher frustration scores were:

- E006 – Birthday Party
- E010 – Grooming an Animal (Ranked)
- E011 – Making a Sandwich
- E012 – Parade
- E015 – Sewing Project

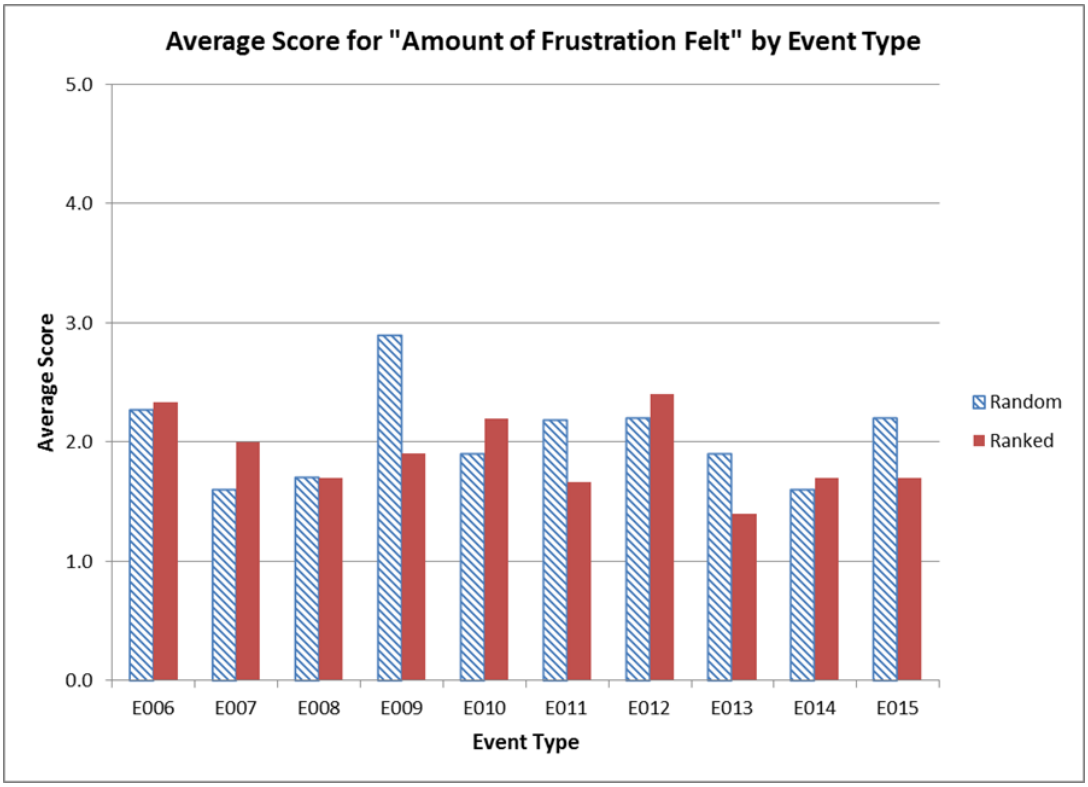

Figure 64: Post-test survey score for "Amount of frustration felt"

Appendix D: Post-Test Interview Questions

Post-Test Interview Questions

After each test, the test moderator would interview the subject to collect additional information on self-reported behaviors for their approach to the event type of the test. The specific questions asked during the interview included the following:

- Please describe the approach you took to review the thumbnails and videos. [*If you had the opportunity to observe their approach, provide your observations as examples to prompt the discussion.*]
- Did you change your approach during the test? If yes, please describe how you changed your approach, and what prompted you to change?
- To what extent did you use or rely on audio clues to confirm the event type?

[Ask the following questions only if this is NOT his/her first test]

- Did you use the same approach for this event type as you did for previous event types?
- How would you compare looking for this event type to the previous event type(s) that you have already searched for?