International Journal of Cyber Behavior, Psychology and Learning
Volume 14 • Issue 1 • January-December 2024

Attracting Visual Attention in a Digital Age: Measuring the Determinants of Interestingness of Videos Using Biosensors

Quincy Conley
https://orcid.org/0000-0002-7526-6677
A.T. Still University, USA

ABSTRACT

The purpose of this study was to determine whether previously established visual attention patterns remained intact during video scenes designed to elicit specific emotions using a novel suite of biosensors. To examine the relationship between visual attention and emotion, data from eye tracking, facial expression recognition (FER), and galvanic skin response (GSR) combined with survey data were used to identify the bottom-up and top-down features of saliency in videos that contributed to their "interestingness." Using a mixed-methods design and convenience sampling, participants (N = 42) watched 60 video clips designed to evoke different emotional responses (positive, neutral, or negative). The results indicated that using a suite of biosensors to examine the impacts of bottom-up and top-down features of visual attention was effective.

KEYWORDS

Emotion, Eye Tracking, Facial Expression Recognition (FER), Galvanic Skin Response (GSR), Interestingness, Saliency, Visual Attention

INTRODUCTION

Attention is the process by which a person selects sights, sounds, or feelings to focus on from an array of stimuli in a given moment (Itti & Koch, 2000). Visual attention, a specific type of attention, is the subconscious process whereby a person devotes cognitive resources to process a visual input deeper than others in a manner that influences a behavioral response (Itti & Koch, 2000; Lamme, 2003). A concept related to visual attention is "interestingness" (Grabner et al., 2013; Gygli et al., 2013; Yoon & Pavlovic, 2014). Gygli et al. (2013) defined the phenomenon of interestingness as the characteristics that attract a person's visual attention. What makes something visually interesting are the qualities of a stimuli that they believe they will remember (Gygli et al., 2013). What people find interesting is often connected to how it makes them feel (Yoon & Pavlovic, 2014). Although people can express whether something is interesting to them, they are usually unable to fully explain why a particular object or scene captured their attention.

What makes something visually interesting is of relevance in the current digital era. Today, people are often presented with rich media content, such as text, images, and video, attempting to influence their attention in a way to motivate them to do something (e.g., click this ad or watch the video in this news article). Because people's senses are constantly bombarded through television, computers, and smartphones, visual content is the dominant way to gain a person's attention. As

DOI: 10.4018/IJCBPL.359336

such, visual attention is heavily studied in the psychology of consumer behavior (Clement, 2007; Simmonds et al., 2020), marketing (Karmarkar & Yoon, 2016), entertainment (Smith et al., 2012), neuroscience (Desimone & Duncan, 1995; Hickey & Peelen, 2015; Itti & Koch, 2000), computer science (Olshausen et al., 1993), and more recently, education (Alemdag & Cagiltay, 2018; Wang et al., 2014; Zu et al., 2020). Although what makes people click a video and keep watching has been studied, their reasons for doing so are still unclear. Thus, investigations to understand what captures people's attention, guides it, and keeps it are more critical than ever.

With the pervasiveness of visual information, there is renewed interest in understanding attention and human behavior by psychologists, marketers, journalists, computer scientists, and educators (Tunga & Cagiltay, 2023). Such knowledge will help people understand the fundamentals of the human visual system and design practical applications. For example, marketers could develop better advertisements that influence buyer behavior and increase sales. Computer scientists could use the most compelling graphics to enrich digital content, and even educators could identify which types of educational media motivate students the most. In all such situations, one theme is consistent: If you can attract a person's visual attention, you can potentially guide their behavior.

THEORETICAL FRAMEWORK

In addition to offering a definition of visual interestingness, Gygli et al. (2013) developed a framework for characterizing it. Interestingness is defined as the qualities of a visual stimuli that "people believe they will remember" that is attributed to why the give visual attention to (Gygli et al. 2013, p. 1633). To make their framework more usable, they categorized interestingness into three factors: (a) aesthetics, (b) unusualness, and (c) preferences (Gygli et al., 2013). Grabner et al. (2013) also attempted to define the elements contributing to interestingness, adding the factors of context and novelty. There have been many investigations of these categories of interestingness of visual information in the hopes of providing a more holistic perspective of what captures a person's visual attention.

Components of Interestingness

Two well-studied subcomponents of interestingness address bottom-up (exogenous) and top-down (endogenous) stimuli. Bottom-up stimuli, such as object color, shape, position, and movement, are considered low-level visual characteristics that capture people's attention. Top-down stimuli, such as the recognition of people, places, and emotions, are considered effective, yet higher level characteristics that require more in-depth physiological resources (Hutson et al., 2022; Smith et al., 2012). In this domain of literature, implications of interestingness vary based on reported differences between bottom-up and top-down features. For example, in a notable experimental study on the topic by Niu et al. (2012), eye movements of college-aged participants when viewing pictures were examined to identify the difference between positive, neutral, and negative scenes in terms of their influence on visual attention as measured by an eye tracker. Results suggested that indeed emotion derived from the scenes impacted eye movement.

Furthermore, Niu et al. (2012) reported that the participants were more likely to notice people, than neutral objects in a given scene. Thus, Niu et al. (2012) theorized that top-down stimuli can override bottom-up stimuli regarding visual attention. However, Niu et al. (2012) used static images instead of dynamic stimuli and conducted their study without concurrently comparing stimuli with top-down and bottom-up elements. Although Niu et al.'s (2012) study represents how researchers have learned how humans visually recognize and locate areas or objects, it focused on only bottom-up or top-down components, not both, using less complex images. Therefore, more research is needed to understand the impact of bottom-up or top-down components and emotion on the human visual system to better understand how human behavior works related to more complex digital stimuli.

Saliency

When grouped, these bottom-up or top-down components are often defined as having saliency, which describes the intrinsic characteristics of visual data that draw the viewer's attention (Itti & Koch, 2000; Liang & Hu, 2015; Treisman & Gelade, 1980; Yan et al., 2013). Specifically, saliency involves the qualities of objects or areas that stand out the most to a person regardless of their bottom-up or top-down features. Knowing the key qualities of objects or areas in a visual stimulus has important implications for predicting common human behavior, such as recognition of and reaction to objects or areas of a video segment. Research explorations of saliency for various stimuli have led to consistent results indicating that bottom-up or top-down features have limited influence individually over where people fix their visual attention.

However, evidence increasingly suggests these components are more powerful when applied together. In one computational study, Liang and Hu (2015) compared various models for saliency and found that the combination of bottom-up and top-down components outperformed either of these components when considered separately. The most prominent saliency models for attracting visual attention included features such as object animacy, position and motion, and emotion. Therefore, rather than ask whether bottom-up or top-down stimuli capture the most attention, a more suitable study question is, *what combination of saliency features increase the interestingness of a stimuli to gain and sustain human visual attention?*

Tools to Measure Interestingness

When researchers are conducting visual attention studies, the tools used to measure visual attention are crucial. Many researchers use a variety of measures to capture responses to visual and emotional states, including self-reported and physiological responses. Data collection in early studies investigating what makes visual information captivating was limited in scope and mostly restricted to self-reported data, even though such data are inherently subjective (Chan, 2009; Schellings & Van Hout-Wolters, 2011). Other researchers varied these methods slightly to minimize response bias and better determine what makes images and videos visually interesting to viewers (Berger et al., 2011; Shrivastava & Tyagi, 2014). Concerted efforts also were made to generate ground truth in an image set to examine multiple regions and levels of interest (Berger et al., 2011; Shrivastava & Tyagi, 2014). Similarly, some studies predicted interestingness from videos. Yoon and Pavlovic (2014) used affective analysis to search popular video hosting services using the word "interestingness" as the query term. The resulting 400 videos were analyzed using two computational frameworks to identify static and temporal features of each video and generate a relative ranking of saliency features (Yoon & Pavlovic, 2014). However, this approach is considered mostly subjective because of its dependence on crowdsourcing from a large group of people without a shared context. Therefore, computational predictions only partially explain visual attention when comparing interestingness of one video with another.

More contemporary studies incorporated eye-tracking technology to collect data to make inferences about the interestingness of visual stimuli (Espigares-Jurado et al., 2020; Susac et al., 2019; Wang et al., 2014; Zu et al., 2020). Eye tracking is a common physiological tool that involves biosensing technologies. Alemdag and Cagiltay (2018) conducted a systematic review of the use eye-tracking technology as a form of measurement during multimedia learning studies. The results of 58 studies, most of which involved college students, revealed that eye-tracking data may be used to draw conclusions about the cognitive processes involved in choosing, compiling, and integrating multimedia content. Moreover, user emotion was a major factor that guided eye movement while viewing multimedia (Alemdag & Cagiltay, 2018).

In another study measuring emotional response to stimuli, Koelstra et al. (2010) found strong correlations between self-reported user data and physiological signal processing from electroencephalogram (EEG), galvanic skin response (GSR), electrocardiogram (ECG), electrooculogram (EOG), and other biosensors. Similarly, Wang and Ji (2015) used EEG, GSR, ECG,

and heart rate monitoring to complete a comprehensive review of visual attention in videos. Chanel et al. (2007) used multimedia content, physiological responses, and self-reported emotional valence to develop a model that estimates points in an arousal-valence space. When comparing results from physiological signals, Chanel et al. (2007) showed that biosensors are a viable method for determining the valence and arousal of emotional states while viewing multimedia content. More specifically, in a recursive feature elimination (RFE) study, Torres et al. (2013) found that GSR results were associated more with emotional arousal and less with emotional valence. To assess emotional valence, more contemporary biosensors for measuring emotion include facial expression recognition (FER) software. For example, Azevado and Gašević (2019) summarized how eye tracking and FER data can be used to complement traditional data collection instruments and statistical analyses to measure psycho-emotional processes in response to visual information and more. In an experimental study by Ahmad et al. (2022), multiple FER software packages were compared for accuracy measuring viewers' facial expressions while watching political advertisement videos. They found that the emotional responses differed by gender, age, and ethnicity, but FER tools such as FACET, AFFDEX, and Affectiva accurately detected emotional valence. However, in general, how much these measures correlate with each other and which emotions they accurately measure are debatable. Clearly, the accuracy of these tools requires additional investigation.

Another issue with visual attention research is whether lab-generated stimuli evoke real-world responses. To address this concern, Wang et al. (2014) compared physiological responses from lab experiments with results from a field study. They found that physiological responses from lab-generated stimuli were not easily generalized to real-world situations (Wang et al., 2014). A second issue is how to correctly associate low-level properties of visual stimuli (e.g., line orientation or color) with human emotion. Ellis et al. (2014) called this problem the affective gap. Therefore, researchers are also investigating novel ways to use biometric data from viewers' responses to improve affective content. For example, Yadati et al. (2013) developed a framework for analyzing affective video content to determine the best placement of personal advertisements for media. This type of data can also be used to predict viewers' responses to content. For instance, using their model and participant self-reported arousal-valence ratings, Ellis et al. (2014) predicted evoked emotion during movie trailers. In another study, Xu et al. (2008) developed a hierarchical structure for categorizing valence and arousal levels evoked by emotion in movies; their model was designed to help users more efficiently search and locate content by emotional tone.

HYPOTHESES DEVELOPMENT

Animacy, Position, and Motion

Animacy has been shown to influence the interestingness of stimuli when investigating visual attention. According to Tremoulet and Feldman (2006), animacy is the interpretation of an object as having characteristics of being alive or lifelike. This tendency of viewers to detect and gravitate toward things that appear alive, such as people or animals, can be used as an active visual attention-getting mechanism. In a study by New et al. (2007), participants more quickly and accurately detected scene changes of animals than inanimate objects. Similarly, Pratt et al. (2010) found that objects with animate motion (e.g., moving in a way that resembled a live creature) captured visual attention more quickly than objects without it. When researchers use eye-tracking tasks in studies, animate objects tend to be detected more (Yang et al., 2012); they also are viewed similarly and with more consistency than inanimate objects (Ković et al., 2009). More importantly, animate objects are associated with better recognition and memory than inanimate objects (Bonin et al., 2014).

Associated features of animacy are position and motion. Physiologically, humans are predisposed to detect position and movement (Gavrila, 1999; Mital et al., 2011; Smith et al., 2012), and the position of an object is important for interestingness because of the human ability to make relational decisions about two objects or areas, such as when viewing video content (Loschky et al., 2020).

Scholl and Tremoulet (2000) reported that people are sometimes cued to the animacy of an object based on its orientation to other objects. Furthermore, according to Michotte's (1991) perception of causality, behaviors can be inferred based on position; therefore, two objects (animate or inanimate) can be perceived to possess a lifelike presence in two-dimensional renderings. Unexpected changes in position can also be used to guide visual attention. Howard and Holcombe (2010) found that change in direction effectively directed a viewer's attention. Their study involved a task requiring participants to identify a target object's orientation, and participants performed worse when the target object collided with an irrelevant moving object (Howard & Holcombe, 2010). This unexpected collision likely caused an attention shift away from the task (Howard & Holcombe, 2010).

Motion has also been shown to have a marked influence on visual attention. Meyerhoff et al. (2014) measured the ability of college students to detect motion and found that animate objects were easier to discern during an interactive search activity. Similarly, Mahapatra et al. (2008) used video stimuli to investigate the connection between motion and visual attention and were able to generate visual attention maps using tagged frames labeled by participants. Osberger and Rohaly (2001) also investigated visual attention using videos and found that certain aspects, such as people, position, and movement, had substantial effects on attention. The authors used eye-tracking data to show that motion effectively guided the viewers' attention into predetermined regions of the videos (Osberger & Rohaly, 2001). Because animate objects in motion are considered more effective at attracting the viewer's visual attention than inanimate objects, I propose the following hypotheses:

- H_1: Videos featuring animate objects will evoke a faster time to the first fixation than videos featuring inanimate objects.
- H_2: Videos featuring animate objects will evoke more fixations than videos featuring inanimate objects.
- H_3: Videos featuring animate objects will evoke more positive emotional responses and higher positive emotional intensity than videos featuring inanimate objects.
- H_4: Videos featuring animate objects will evoke fewer correct responses and lower confidence ratings in the engagement questions than videos featuring inanimate objects.

Emotion

Emotional tone in images and videos can be created through people, position, and movement. Emotion is often measured using valence and intensity (Mohanty & Sussman, 2013). Valence refers to the direction of emotion (i.e., positive versus negative), whereas intensity refers to the strength of the emotion (i.e., high or low intensity). Researchers have previously investigated the links between valence and intensity and visual attention. Using biometric sensors, Eisenbarth and Alpers (2011) found distinct patterns in detailed eye-tracking data when participants were asked to view facial expressions. When participants viewed emotional facial expressions, initial fixations were toward the eye and mouth areas (Eisenbarth & Alpers, 2011).

However, fixation areas differed by expression type: Participants mainly concentrated on the eyes for sad expressions and on the mouth area for happy expressions (Eisenbarth & Alpers, 2011). The eye and mouth areas were also equally helpful for recognizing fearful and neutral expressions (Eisenbarth & Alpers, 2011). A study by Simola et al. (2013) found that unpleasant emotional pictures evoked more emotional responses than neutral pictures and, overall, emotionally laden pictures had higher emotional intensity than neutral pictures. Nummenmaa et al. (2006) found that the first fixation for emotional pictures took longer than fixations for neutral pictures, even when participants were instructed to not view the emotional pictures. Fernández-Martín and Calvo (2016) presented emotionally pleasant, emotionally unpleasant, and neutral scenes peripherally to participants and found that participants were more selectively oriented to the central task when emotional scenes were presented instead of neutral scenes. Jefferies et al. (2008) also investigated the relationship between

emotion and attention using valence and intensity. During an attentional blink task, scenes with neutral emotional tone led to higher accuracy, and those with high intensity led to lower accuracy, suggesting that high intensity may impair attentional control (Jefferies et al., 2008). Similarly, Vogt et al. (2008) found that participants were slower to disengage from stimuli with detectable emotional tone than from stimuli without it. Interestingly, this finding was not affected by direction of valence (positive or negative). Regardless of valence or intensity, emotional scenes presented in the peripheral vision have been shown to affect vision as well. Calvo et al. (2015) tested whether emotional scenes containing people were processed in peripheral vision and found that these scenes were associated with higher amounts of task interference than neutral scenes. Because positive emotional tone is also considered effective at guiding the viewer's visual attention, I propose the following hypotheses:

- H_5: Videos with positive emotional tone will evoke a faster time to the first fixation than videos with negative and neutral emotion tones.
- H_6: Videos with positive emotional tone will evoke more fixations than videos with negative and neutral emotion tones.
- H_7: Videos with positive emotional tone will evoke a more positive emotional responses and higher positive emotional intensity than videos with negative and neutral emotional tone.
- H_8: Videos with positive emotional tone will evoke lower correct responses and confidence ratings than videos with negative and neutral emotional tone.

Purpose of the Present Study

The purpose of the present study, which was built on the existing literature, was to determine whether previously established visual attention patterns produced by dynamic stimuli designed to elicit specific emotions remained intact measured by eye movement and emotion responses. In this study, I looked for the most important aspects of video that could potentially draw viewers in and keep them watching digital material. Therefore, I combined biosensor data from eye tracking, facial expression recognition (FER), and GSR with self-reported data to identify the bottom-up and top-down features of saliency in video stimuli that contribute to the interestingness of them.

The following research questions were used for this study:

- What is the effect of animacy on visual attention?
- What is the effect of emotional tone on visual attention?
- What is the effect of using a suite of biosensors to examine visual attention?

METHOD

Research Design and Participants

Using a mixed-methods design and convenience sampling, I recruited participants (N = 42) from a large university in the U.S. Pacific Northwest using a university-wide email announcement. The study was open to all undergraduate and graduate students. Because of the lack of demographic diversity, no identifying data besides gender, age, and grade level were collected to protect participants' anonymity. The local institutional review board approved the study, and self-selected participants completed an approved informed consent form before participation. Participants received $20 in U.S. currency as an incentive for their participation. The sample size was determined ahead of time based on an a priori power analysis, and data were not analyzed until after data collection was complete. Only complete participant datasets were included in the data analysis, resulting in four participants' datasets being removed from incompleteness. Thus, data from 42 college students representing

Table 1. Demographic characteristics of study participants (N = 42)

Characteristic	N	%
Gender		
Male	30	71.42
Female	12	28.57
Education level		
Freshman	11	26.19
Sophomore	6	14.29
Junior	11	26.19
Senior	8	19.04
Graduate student	6	14.29

Table 2. Number of videos for each object animacy and emotional tone condition

Animacy	Positive	Emotional Tone	Negative
		Neutral	
Animate	10	10	10
Inanimate	10	10	10

various majors (e.g., psychology, business, health sciences, biology, engineering, computer science) were included for this study. They ranged in age from 18 to 53 years (Median = 22). Additional demographic characteristics are presented in Table 1.

Visual Stimuli

To assess the effects of animacy and emotional tone on visual attention, I selected 60 video clips from a stock image and video subscription database (www.bigstockphoto.com). I chose dynamic videos without sound instead of static visual stimuli because of their ability to present objects and motion, both of which are important factors of visual attention (Li et al., 2009). Controls were applied for video position, size, and resolution. Video clip duration was approximately 10 seconds with the average frame rate of 30 frames per second. Video clips were categorized as having animate (i.e., humans, animals) or inanimate objects and by emotional tone (positive, neutral, and negative). Videos were manually sorted into six conditions of 10 videos each and served as the independent variables for the study (Table 2).

Before data collection, Amazon Mechanical Turk (MTurk) was used to confirm that chosen videos were accurately categorized by animacy and emotional tone. To verify, MTurk workers classified each of the 60 videos into one of the six conditions by clicking the appropriate option. Only videos with a 90% level of agreement or higher rated by the MTurk workers were used. None of the MTurk workers were participants in the study, and results from the sorting task were used only for external validation of video categorization.

Biosensors

The following biometric sensors, or biosensors, were used simultaneously to measure the visual attention of participants while watching the 60 videos on a computer monitor: eye tracking, FER, and GSR. The eye movements of participants while watching the videos were recorded using a Tobii Pro X2-30 eye tracker. The Tobii Pro X2-30 is a high-precision eye-tracking device manufactured by Tobii Pro, a subsidiary of Tobii Tech. It was designed for research and professional use in fields such as market research, usability testing, and human-computer interaction. It uses infrared technology

to track the user's gaze with a precision of 0.01 degrees and can track both eyes simultaneously at a sample rate of up to 120 hertz (Hz). The device is also portable and easy to set up and use, making it suitable for both laboratory and field settings. For this study, the eye-tracking gaze data were tracked at 30 Hz at an accuracy of 0.40° and a precision of 0.32°. The eye tracker was positioned at the base of the computer monitor using a magnetic bracket. To analyze the participant's facial expressions in reaction to the videos, Affectiva software was used to measure FER. A Universal Serial Bus camera was placed near the top of the monitor, and the software tracked changes in the participant's facial features, generating an interpretation of emotional valence (i.e., positive, neutral, and negative) in real time. The strength of emotional valence caused by the videos was recorded using a Shimmer 3 NeuroLynq GSR sensor device. A GSR sensor is a device that measures changes in the electrical conductance of the skin. These changes are thought to be related to changes in the sweat gland activity, which in turn is related to the level of emotional or physiological arousal. GSR sensors are often used in research to measure emotional responses to various stimuli, such as images or videos. The data are collected and can be analyzed to provide information about the level of emotional or physiological arousal (Benedek & Kaernbach, 2010). Through measurement of perspiration on the surface of the skin, the GSR equates electrical conductance as emotional arousal, where low conductance equals low emotional arousal and high conductance equals high emotional arousal. The GSR sensor was worn on the wrist and two middle fingers of participants, making it noninvasive and allowing for natural movements while using a keyboard and mouse. Data from all three tools were combined and time-stamped in the iMotions Attention Tool software during data analysis.

Study Procedure

Participants attended a single one-on-one lab session that lasted approximately 60 min. To start, they sat at a standard computer workstation in a stationary seat and were positioned approximately 18 in. from the computer monitor. After reviewing the consent form that the institutional review board approved, I placed the GSR sensor for the wrist on the participant's nondominant hand. Next, the eye tracker and FER were calibrated to ensure the participant's eye movements and facial expressions were accurately recorded during the experiment. After calibration, participants completed a brief pre-survey and were then instructed to carefully view the 60 video clips that were presented randomly at uniform intervals, while their eye movements and arousal activity levels were recorded. A static gray screen was presented as a transition between each video clip to standardize where participants looked. Participants were not explicitly told that their facial expressions and arousal activity levels were being monitored to avoid heightened reactions or potential bias. After each video, participants were instructed to answer a corresponding engagement question and provide an associated confidence rating. Once all video clips were viewed, participants completed a short post-survey and were dismissed from the study.

Outcome Measures and Data Analysis

For this study, I collected quantitative and qualitative data (Table 3). Quantitative outcomes included visual attention, emotional tone (valence and intensity), confidence, and engagement with the videos. Visual attention was assessed with the eye-tracking data and measured by time to first fix (TTFF) and fixation count. Emotional valence was measured with FER data and emotional tone with GSR data. Engagement questions with associated confidence ratings were used to measure engagement. Qualitative outcomes included visual attention assessed by heat maps created from the biometric sensors' data and open-ended questions from the post-survey.

Pre-Survey

For the pre-survey, participants completed an online questionnaire that included the consent form and collected demographic information (Table 1).

Table 3. Instrumentation and outcome measures

Instrumentation	Outcome Measure	Data Analysis
Pre-survey	Demographic data	Descriptive
Eye tracking	Time to first fixation	Quantitative
	Fixation count	Quantitative
	Heat map	Qualitative
Facial expression recognition	Emotional valence	Quantitative
Galvanic skin response	Emotional intensity	Quantitative
Engagement questions	Incorrect responses	Quantitative
Confidence rating	Confidence rating	Quantitative
Post-survey	Open-ended questions	Qualitative

Eye-Tracking Data

Eye-tracking data were collected and analyzed for patterns to determine which aspects of the videos participants paid attention to visually in terms of animacy. The data were measured by TTFF, which quantified how long it took participants to focus their attention on any area of the video (animate or inanimate). TTFF is a measure of how long it takes a user to first fixate on a specific point of interest (POI) on a video, webpage, image, or other visual stimuli. It is often used in user research and usability testing to evaluate the effectiveness of visual design and layout. A shorter TTFF generally indicates that the POI is more easily noticeable and relevant to the user, whereas a longer TTFF may indicate that the POI is less noticeable or less relevant. In eye-tracking studies, TTFF is typically measured in milliseconds and is calculated as the time elapsed between the presentation of the visual stimuli and the first fixation on the POI. TTFF can be used to evaluate the effectiveness of different design elements, such as color, contrast, location, and motion (Jacob & Karn, 2003).

The fixation count was used to quantify the number of times participants moved their eyes to focus on an area in the video. In eye-tracking studies, fixation count is typically measured as the number of times the user's gaze pauses on the POI, known as fixations, during a specific time or task. A higher fixation count generally indicates that the POI is more relevant or interesting to the user, whereas a lower fixation count may indicate that the POI is less relevant or interesting. Additionally, fixation count can be used to infer cognitive processes, such as attention and interest. Together, these outcomes measured visual attention, and a 3×2 analysis of variance (ANOVA) was used to evaluate the effects of animacy (animate or inanimate objects) and the three emotional tones (positive, neutral, and negative) on TTFF and fixation count separately. For fixation count, follow-up tests were conducted to evaluate pair-wise differences among the means for emotional tone. Because the variances were homogeneous and the sample sizes were similar, Tukey's honestly significant difference method was used. For qualitative inferences about visual attention patterns, heat maps were produced from the eye-tracking data to determine which objects or areas of the video participants viewed.

A heat map is a color-coded visual representation of data in which red shows the most common parts of the video viewed. Still images of visual attention data were extracted from the midpoint of videos that contained similar stimuli (e.g., outdoors, cityscapes). The video midpoint was chosen for this analysis to ensure participant acclimation to the video stimuli.

Facial Expression Recognition Data

The FER data were used to compare the assigned emotional tone of the videos with the emotional valence of the participants' responses to better understand which elements of the videos contributed to those responses. FER is the process of identifying and interpreting human emotions by analyzing

facial features through various computer vision techniques. I used the software development kit (SDK) called Affectiva that employs a landmark recognition approach to detect and classify participants' emotions. FER is regularly used to study and understand emotions in human behavior (Sandbach et al., 2012). Valence values ranging from 0 (no expression) to 1 (a fully present expression) were recorded whenever facial muscle movements produced a pattern that was distinct enough to exceed an evidence threshold of 0.5 for a given emotion (positive, neutral, or negative). Every time a facial pattern exceeded this threshold, an emotion count was recorded. To account for differences in length for the 60 videos, emotional valence was measured as a percentage (emotion count/time in milliseconds). A 3 × 2 ANOVA was used to evaluate the effects of animacy (animate or inanimate objects) and the three emotional tones (positive, neutral, and negative) on the emotional valence.

GSR Data

The GSR sensor assessed the participants' emotional arousal or strength of the emotion. Emotional intensity values ranging from 0 (low intensity) to 10 (strong intensity) were used to examine how much emotion a participant experienced in response to the video stimuli (Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures, 2012). The average of the highest GSR signal per participant per video was calculated and compared among the six conditions. A 3 × 2 ANOVA was used to evaluate the effects of animacy (animate or inanimate objects) and the three emotional tones (positive, neutral, and negative) on emotional intensity.

Engagement Questions

A catch trial was used in the present study to keep participants engaged while watching the videos by making them answer a short yes or no question after each video. Common in the field of experimental psychology, catch trials are used to keep participants actively engaged (Nugent, 2013). Instead of passively watching the videos, the participants were asked to answer a basic comprehension question (i.e., Did a television appear in the video?) immediately after each video to ensure that they were paying attention. The mean number of incorrect answers per video was calculated and compared among the six video stimuli conditions. A 3 × 2 ANOVA was then used to evaluate the effects of animacy (animate or inanimate objects) and the three emotional tones (positive, neutral, and negative) on engagement responses.

Confidence Ratings

After each engagement question, participants indicated their level of confidence with their response on a Likert-type scale rating ranging from 1 (not confident at all) to 10 (extremely confident). To reduce guessing, this outcome was designed to assess the certainty of their answer to the engagement question and subsequently used to estimate which video stimuli condition was most effective at guiding the participants' attention (Sandberg et al., 2010; Seth et al., 2008). These data were analyzed by the number of correct answers per video stimulus condition using an ANOVA. Specifically, a 3 x 2 ANOVA was used to evaluate the effects of animacy (animate or inanimate objects) and the three emotional tones (positive, neutral, and negative) on the confidence level of engagement question responses.

Post-Survey

After viewing all the video stimuli, participants completed an electronic survey in which they answered three open-ended questions:

- Was there a particular video that stood out to you?
- Why did the particular video stand out to you?
- Did you find any of the materials presented during the experiment sensitive in nature?

Table 4. Means and standard deviations of time to first fixation (in ms) by video stimuli condition

Video Stimuli Condition	M	SD	N
Animate			
Positive	295.76	674.54	10
Neutral	350.60	1205.18	10
Negative	306.43	751.36	10
Total	317.42	904.82	30
Inanimate			
Positive	343.45	1157.19	10
Neutral	336.50	875.21	10
Negative	333.84	886.10	10
Total	338.00	982.53	30
Total			
Positive	321.15	961.70	20
Neutral	343.14	1042.98	20
Negative	320.37	822.13	20
Total	328.19	946.09	60

The questions were developed specifically for the present study to look for trends and anomalies in the participants' performance and to verify consistency of the data collection. Answers were qualitative in nature, but evaluated quantitatively by identifying themes and coding them into categories as recommended by Graneheim and Lundman (2004) and Patton (2005). The number of times a response occurred per category was then counted and reported as a percentage to preserve the essence of the content while identifying patterns.

RESULTS

TTFF

The means and standard deviations for TTFF as a function of the two factors of visual attention are presented in Table 4. No interaction was found between animacy and emotional tone, $F(2, 1{,}886) = 0.17$, $p = 0.84$, partial $\eta^2 < 0.01$. No main effects were found for animacy, $F(1, 1{,}886) = 0.22$, $p = 0.64$, partial $\eta^2 < 0.01$, or emotional tone, $F(2, 1{,}886) = 0.13$, $p = 0.88$, partial $\eta^2 < 0.01$. Neither animacy nor emotional tone contributed to how fast participants focused on an area of the video. However, consistent with hypothesis H_1, participants focused on areas of videos with animate objects the fastest. Converse to hypothesis H_5, participants focused on positive emotional tone the fastest.

Fixation Count

The means and standard deviations for fixation count as a function of the two factors of visual attention are presented in Table 5. No interaction was found between animacy and emotional tone, $F(2, 1{,}886) = 1.38$, $p = 0.25$, partial $\eta^2 = 0.01$. No main effect was found for animacy, $F(1, 1{,}886) = 0.68$, $p = 0.41$, partial $\eta^2 < 0.01$. There was a main effect for emotional tone, $F(2, 1{,}886) = 5.97$, $p < 0.01$, partial $\eta^2 = 0.01$. Emotional tone contributed to how much participants looked around the screen area of the video.

For the follow-up tests, there were differences in the fixation count between video stimuli designed to elicit negative versus positive responses and negative versus neutral responses, but not between positive and neutral responses. The 95% confidence intervals for the pair-wise differences, as well as the means and standard deviations for the three emotional tones, are presented in Table 6. Consistent with hypotheses H_2 and H_6, participants fixated more on videos with animate objects than on videos with inanimate objects. Participants also fixated more on videos with positive emotional tone than on videos with negative or neutral emotional tones.

Table 5. Means and standard deviations of fixation count by video stimuli condition

Video Stimuli Condition	M	SD	N
Animate			
Positive	37.80	19.10	10
Neutral	35.11	13.67	10
Negative	34.09	14.59	10
Total	35.64	15.99	30
Inanimate			
Positive	35.88	20.17	10
Neutral	36.37	19.44	10
Negative	32.68	20.20	10
Total	35.01	20.00	30
Total			
Positive	36.78	19.68	20
Neutral	35.78	17.78	20
Negative	33.37	16.86	20
Total	35.31	18.19	60

Table 6. Means, standard deviations, and 95% confidence intervals of pair-wise differences for emotional tone from fixation count

Emotional Tone	M	SD	Positive	Neutral
Positive	36.78	19.68	[−1.39, 3.40]	[0.01, 4.80*]
Neutral	35.78	17.78	[1.01, 5.80*]	
Negative	33.37	16.86		

Note. An asterisk indicates that the 95% confidence interval did not contain zero; therefore, the difference in means was significant at the 0.05 significance level using Tukey's honestly significant difference method.

Heat Map by Video Stimuli Condition

In Table 7, select segments of the video clips are presented with the aggregate eye-tracking data in detail. For the animate positive scene, people with pleasant looks on their faces are playing music on a beach. In the animate negative scene, two people run into each other while walking on a sidewalk with angry looks on their faces. Similarly in the animate negative scene, despite the opposite disposition, the heat map shows how participants concentrated on the more emotionally expressive person in the scene. Conversely, in the animate neutral scene with multiple people without emotional tone, participants' views were more widespread. In the video scenes featuring inanimate objects, regardless of emotional tone, participants had wide-ranging views. Both positive and negative animate object eye-tracking data indicated a greater visual focus on specific targets, such as faces, whereas videos featuring inanimate objects caused broader patterns of nonspecific visual attention.

Emotional Valence

The means and standard deviations for emotional valence as a function of the two factors of visual attention are presented in Table 8. No interaction was found between animacy and emotional tone, $F(2, 174) = 0.01$, $p = 0.99$, partial $\eta^2 < 0.01$. No main effects were found for animacy, $F(1, 174) = 0.05$, $p = 0.83$, partial $\eta^2 < 0.01$, or emotional tone, $F(2, 174) = 0.02$, $p = 0.98$, partial $\eta^2 < 0.01$. Neither animacy nor emotional tone contributed to a change in emotional valence. However, consistent with hypotheses H_3 and H_7, videos with animate objects produced a more emotional valence than videos with inanimate objects regardless of emotional tone; videos with positive emotional tone had a stronger emotional valence than videos with negative and neutral emotional tone.

Table 7. Still heat map images of video midpoint by video stimuli condition

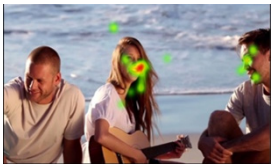
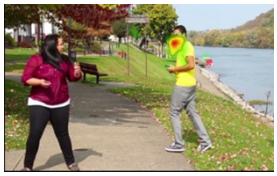
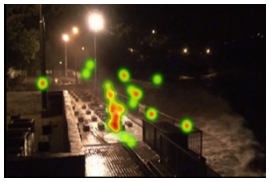
Animacy	Emotional Tone		
	Positive	Neutral	Negative
Animate			
Inanimate			

Table 8. Means and standard deviations of emotional valence by video stimuli condition

Video Stimuli Condition	M	SD	N
Animate			
Positive	9.06	11.79	10
Neutral	8.56	11.40	10
Negative	8.72	11.68	10
Total	8.78	11.50	30
Inanimate			
Positive	8.43	11.35	10
Neutral	8.19	11.09	10
Negative	8.59	11.69	10
Total	8.41	11.25	30
Total			
Positive	8.74	11.48	20
Neutral	8.38	11.15	20
Negative	8.66	11.59	20
Total	8.59	11.34	60

Emotional Intensity

The means and standard deviations for emotional intensity as a function of the two factors of visual attention are presented in Table 9. No interaction was found between animacy and emotional tone, $F(2, 54) = 0.96$, $p = 0.39$, partial $\eta^2 = 0.03$. No main effects were found for animacy, $F(1, 54) = 3.57$, $p = 0.06$, partial $\eta^2 = 0.06$, or emotional tone, $F(2, 54) = 2.19$, $p = 0.12$, partial $\eta^2 = 0.08$. Neither animacy nor emotional tone contributed to a change in emotional intensity. However, consistent with hypotheses H_3 and H_7, videos with animate objects produced a stronger emotional intensity than videos with inanimate objects; videos with positive emotional tone had a stronger emotional valence than videos with negative and neutral emotional tone.

Engagement Questions

The means and standard deviations for engagement questions as a function of the two factors of visual attention are presented in Table 10. No interaction was found between animacy and emotional response, $F(2, 54) = 1.40$, $p = 0.27$, partial $\eta^2 = 0.05$. No main effects were found for animacy, $F(1, 54) = 0.26$, $p = 0.61$, partial $\eta^2 < 0.01$, or emotional tone, $F(2, 54) = 0.02$, $p = 0.98$, partial $\eta^2 < 0.01$.

Table 9. Means and standard deviations of emotional intensity by video stimuli condition

Video Stimuli Condition	M	SD	N
Animate			
Positive	0.18	0.04	10
Neutral	0.15	0.05	10
Negative	0.15	0.03	10
Total	0.16	0.04	30
Inanimate			
Positive	0.16	0.02	10
Neutral	0.14	0.02	10
Negative	0.14	0.03	10
Total	0.14	0.03	30
Total			
Positive	0.16	0.04	20
Neutral	0.14	0.03	20
Negative	0.14	0.03	20
Total	0.15	0.04	60

Table 10. Means and standard deviations of percentage correct of engagement questions by video stimuli condition

Video Stimuli Condition	M	SD	N
Animate			
Positive	0.92	0.12	10
Neutral	0.86	0.31	10
Negative	0.96	0.09	10
Total	0.91	0.20	30
Inanimate			
Positive	0.86	0.20	10
Neutral	0.95	0.08	10
Negative	0.84	0.29	10
Total	0.88	0.21	30
Total			
Positive	0.89	0.16	20
Neutral	0.90	0.23	20
Negative	0.90	0.22	20
Total	0.90	0.20	60

Neither animacy nor emotional tone contributed to the number of incorrect answers. Inconsistent with hypotheses H_4, videos with animate objects resulted in a higher number of correct answers than videos with inanimate objects; yet, consistent with $H_{8,}$ videos with positive emotional tone resulted in a lower number of correct answers than videos with negative and neutral emotional tones.

Confidence Ratings

The means and standard deviations for confidence ratings as a function of the two factors of visual attention are presented in Table 11. No interaction was found between animacy and emotional tone, $F(2, 54) = 0.67$, $p = 0.52$, partial $\eta^2 = 0.02$. No main effects were found for animacy, $F(1, 54) = 1.55$, $p = 0.22$, partial $\eta^2 = 0.03$, or emotional tone, $F(2, 54) = 0.38$, $p = 0.69$, partial $\eta^2 = 0.01$. Neither animacy nor emotional tone contributed to confidence ratings. However, contrary to hypotheses H_4 and H_8, videos with animate objects resulted in a higher confidence rating than videos with inanimate objects; videos with positive emotional tone resulted in a higher confidence rating than videos with negative and neutral emotional tones.

Table 11. Means and standard deviations of confidence rating by video stimuli condition

Video Stimuli Condition	M	SD	N
Animate			
Positive	9.52	0.43	10
Neutral	9.25	0.81	10
Negative	9.27	0.79	10
Total	9.35	0.69	30
Inanimate			
Positive	9.08	0.54	10
Neutral	9.31	0.71	10
Negative	8.98	0.82	10
Total	9.12	0.69	30
Total			
Positive	9.30	0.53	20
Neutral	9.28	0.74	20
Negative	9.13	0.80	20
Total	9.24	0.69	60

Table 12. Responses to the post-survey question asking about a particular video that stood out (N = 34)

Video Stimuli Condition	N	%
Animate		
Positive	6	17.65
Neutral	2	5.88
Negative	4	11.76
Inanimate		
Positive	7	20.59
Neutral	0	0
Negative	15	44.12

Post-Survey Open-Ended Questions

As shown in Table 12, at least one video stood out to 34 participants. Inanimate videos with negative emotional tone stood out to participants the most, and inanimate videos with neutral emotional tone stood out to participants the least. As shown in Table 13, most of the responses landed in one of five saliency categories (colors, novelty, expectation, brightness, and motion). The colors included in the video were the most common reason why participants reported the video stood out to them. Of the 34 participants, most participants 82%, $n = 27$) did not find the materials in the videos sensitive in nature. The other participants (18%, $n = 6$) reported at least one video was sensitive in nature because of the emotional tone, dark color, or subject matter connotation.

Table 13. Thematic categories and responses for the post-survey question asking why the particular video stood out (N = 19)

Thematic Category	N	%
Colors	5	26.32
Novelty	4	21.05
Expectation Brightness	4	21.05
Motion	3	15.79

DISCUSSION

When comparing animate and inanimate videos, I discovered that the TTFF was noticeably quicker for animate videos, but no statistically significant differences were found. Additionally, positive animate videos produced the fastest TTFF, whereas neutral animate videos produced the slowest TTFF. Furthermore, although there were no statistically significant differences found for emotional tone, positive animate videos had the strongest saliency for directing or guiding the user's visual attention.

Similarly, positive animate videos produced the highest fixation counts, and negative inanimate videos produced the lowest. Videos with positive emotional tone produced higher fixation counts than those with negative emotional tone. These results suggested an aversion effect caused by the negative nature of some videos; participants were cautious or hesitant to look elsewhere in the scene. This finding was supported by the responses to the open-ended question asking participants whether any of the videos were sensitive in nature. Some of the participants expressed feeling anxiety after watching the negative videos. Participants not only expressed negative emotions about negative animate videos but also reported positive emotions about positive animate videos. The statistically significant main effect found for animacy supports both these findings.

The heat map results supported the hypotheses that suggested videos with animate objects would produce more fixations than inanimate objects. Consistent with previous work by Meyerhoff et al. (2014), this finding reinforces existing knowledge that viewers can efficiently and quickly distinguish between animate and inanimate objects. In addition to animacy, emotional tone may have contributed to the heat map results. For example, in Table 7 for the inanimate negative video heat map, lighting seemed to be a factor. In that darker scene, participants seemed to concentrate more on the lighter areas of the scene. Regarding the other heat maps in Table 7, depth (foreground versus background) and contrast of motion (i.e., static building versus motion of cars) may also have contributed to observed results.

For emotion outcomes, animate videos produced more frequent emotional responses than inanimate videos as measured by the FER data. More specifically, positive animate videos produced the most frequent emotional responses. As expected, neutral inanimate videos produced the least frequent emotional responses. Similarly, animate videos produced stronger emotional intensity than inanimate videos as measured by the GSR data. Positive animate videos produced the strongest emotional intensity, and negative and neutral inanimate videos produced the weakest. These findings are supported by the previous GSR studies (Ellis et al., 2014; Koelstra et al., 2010; Torres et al., 2013; Wang et al., 2014; Xu et al., 2008; Yadati et al., 2013).

The results from engagement questions also supported the GSR data. When participants were asked an engagement question about whether they had seen a particular object in each video clip, they were more likely to respond incorrectly while watching animate videos than inanimate videos. Similarly, participants watching positive animate videos were more likely to get the engagement questions incorrect, whereas participants watching negative inanimate videos were more likely to get the question correct. One explanation for this result is that participants were more focused on the animate object in positive videos and less focused on other aspects of the video to which the question may have pertained.

Contradictory to findings for the engagement questions, the confidence ratings suggested that participants were more confident in their answers for animate videos than inanimate videos. Even though the participants were more likely to get the question incorrect while watching the positive animate videos, they reported the highest level of confidence in their responses to the engagement questions for them. Similarly, participants reported the lowest confidence in their answers while watching negative inanimate videos despite being more likely to get the question correct. A possible explanation for this result may be that the questions were not balanced across object animacy. For instance, in the positive animate videos, the questions sometimes required participants to recall something that was outside the focus of the video, such as asking them whether they saw the red ball

when the video was about someone walking on a street. Participants may have had fewer incorrect answers and higher confidence ratings if the question asked about the person walking on a street.

When participants were asked whether a particular video stood out to them as part of the post-survey questionnaire, there was no clear patterns in terms of animacy and emotional tone. However, participants identified negative inanimate videos more than the others. This finding suggests that videos with a negative tone may be more memorable. Thematic analysis of participant responses to the question asking why a particular video stood out suggested that colors, novelty, expectation, brightness, and motion were factors contributing to more memorable videos. When asked whether they found any material in the video clips to be sensitive, the few participants who indicated videos were sensitive referred most to ones with negative emotional tone.

Limitations and Future Directions

Besides a relatively small, homogenous sample, which limits the generalizability of the results, a concern was that the emotional tones in the 60 videos were not strong enough. Although categorizing each of the videos into positive, neutral, or negative tones was easy, it was not guaranteed that the strength of the participant's emotional valence would be discernible to the FER and GSR software tools. The consistently low negative emotional valence and intensity observed in the study supports this notion. Therefore, an important question to address in future studies is why participants had such a low emotional response to negative videos. A possible reason could be that the participants did not see the object (animate or inanimate) that elicited an emotional response. In a future study, this could be verified by gaze-over-object analysis to confirm how often users fixated on the regions (or objects) of interest. Other reasons could be that perhaps the videos were not categorized correctly or that the threshold settings of the biometric sensors were set too high or too low to detect emotional signals. Furthermore, because emotions are complex (Grossmann et al., 2016), quantifying these physiological responses is an inherent limitation of this type of data. Carefully considering the best-rated videos for positive or negative emotional tones, researchers may find results from this study useful for designing future studies that are more replicable.

Another limitation of this study was related to the engagement questions. When considering the engagement questions by video stimuli during data analysis, I discovered that one question (Did the trees have leaves on them?) produced an unusually high number of errors compared with the other questions. The video that corresponded to that question involved a forest trail with coniferous trees. It seems likely that the incorrect responses were caused by the term "leaves." Most people think of a "tree" as "an object with leaves," so when asked whether the trees had leaves, participants probably assumed the trees they saw most likely had leaves because they saw objects that looked like trees (representativeness) and trees often have leaves (availability). Thus, this result may have resulted from an error of logic rather than pure observation. Another possible explanation is that participants answered no because evergreen trees have needle-shaped leaves and most people incorrectly believe that needles are not leaves. When this question was removed, positive animate objects focused the participants' visual attention to the point that they stopped scanning and failed to see other objects in the video. Although two participants did not reply to two engagement questions for unknown reasons, the missing data did not change results.

In this study, video was the sole focus of visual attention, and audio was not included. Once we have a better understanding of the visual components of videos and what makes them interesting, the next logical step is to investigate the impact of audio on visual attention. Such studies may also address one of the limitations of the present study: the lack of consideration of accessibility needs (Moray, 2017). When considering how users experience digital information, researchers definitely need to have more compassion to better support users with sensory exceptions (Conley et al., 2018). Thus, developing an algorithm that better describes cross-media is necessary so that images and video can be converted into text descriptions, improving accessibility. Furthermore, future studies could investigate whether artificial intelligence allows people to *hear* and *see* better online. Current

methods rely on human efforts, such as MTurk, to manually evaluate results and create models based on observed patterns, but artificial intelligence may reduce that reliance on human input (Rehman & Saba, 2014).

CONCLUSION

The purpose of this study was to investigate the saliency of visual elements in videos by measuring visual attention with a novel combination of eye tracking, FER, GSR, and surveys. The first research question for the study involved the effect of animacy on visual attention. Data from the biometric sensors and the survey supported the hypotheses that animate objects in videos more often captured attention than inanimate objects. The second research question for the study involved the effect of emotional tone on visual attention. Findings also supported the hypotheses that emotional tone in videos, specifically positive emotional tone, more often captured attention than negative or neutral emotional tones. For the third research question, the results indicated the novel approach of using a suite of biosensors to examine the impacts of bottom-up and top-down features of visual attention, together, was effective. Besides eye-tracking data, the FER and GSR biosensor data were useful for understanding not just emotional valence, but also emotional intensity. Overall, this study suggested that animacy and emotion, especially positive emotional tone, effectively guided users' visual attention while watching videos. Considering how ubiquitous video and digital content are today, I believe these results are important for improving visual content. Furthermore, outcomes from this area of research may be useful for training technological tools, such as video repository searches and artificial intelligence vision sensors. Whether applied for marketing, usability, or educational purposes, the need for understanding human perception and behavior is increasingly necessary to design and deliver intuitive and interesting viewing experiences.

CONFLICTS OF INTEREST

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. All data and study materials are available upon request. To address bias in citation practices, the author sought to choose references that reflect the diversity of the field in discipline, gender, ethnicity, geography, and other factors. Unfortunately, the current methodology is limited to considering gender as a binary variable.

STATEMENT OF FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The project was carried out with support from the department of Organizational Performance & Workplace Learning (OPWL) at Boise State University, which provided access to necessary facilities, resources, and equipment, including the biosensors used for data collection. All research activities were conducted independently by the author without direct financial support.

AUTHOR NOTE

Quincy Conley https://orcid.org/0000-0002-7526-6677

Address any correspondence concerning this article to Quincy Conley, A.T. Still University, 5835 Still Circle, Mesa, Arizona, 85206. Email: quincyconley@gmail.com

PROCESSING DATES

This manuscript was initially received for consideration for the journal on 10/12/2023, revisions were received for the manuscript following the double-anonymized peer review on 10/23/2023, the manuscript was formally accepted on 05/12/2024, and the manuscript was finalized for publication on 10/15/2024.

ACKNOWLEDGMENTS

Thank you to contributors associated with the iPerform Research Lab and generous colleagues from the Organizational Performance & Workplace Learning (OPWL) department at Boise State University. Thank you to Dr. Steven Cutchin and friends at Dimensional Mechanics, who guided the direction of this study. Also, extraordinary thanks are owed to Dr. Dominque Davis; this study could not have happened without her direct support. Furthermore, I'd like to thank Ms. Jamie Capawana, who was instrumental in this study. Lastly, thank you to Ms. Deborah Goggin for her keen eye and quality copy editing and my mentor, Dr. Mark Grabe for his relentless support.

REFERENCES

Ahmad, K., Wang, S., Vogel, C., Jain, P., O'Neill, O., & Sufi, B. H. (2022). Comparing the performance of facial emotion recognition systems on real-life videos: Gender, ethnicity and age. In K. Arai (Ed.), *Proceedings of the Future Technologies Conference (FTC) 2021* (vol. 1, pp.193–210). Springer. DOI: 10.1007/978-3-030-89906-6_14

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education*, *125*, 413–428. DOI: 10.1016/j.compedu.2018.06.023

Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: Issues and challenges. *Computers in Human Behavior*, *96*, 207–210. DOI: 10.1016/j.chb.2019.03.025

Benedek, M., & Kaernbach, C. (2010). Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology*, *47*(4), 647–658. DOI: 10.1111/j.1469-8986.2009.00972.x PMID: 20230512

Berger, A. L., Schohn, G. C., & Wener, M. J. (2011). *Regions of interest in video frames*. (U.S. Patent No. 7876978 B2). United States Patent and Trademark Office. https://patentimages.storage.googleapis.com/31/ce/68/e03e4b74c7142a/US7876978.pdf

Bonin, P., Gelin, M., & Bugaiska, A. (2014). Animates are better remembered than inanimates: Further evidence from word and picture stimuli. *Memory & Cognition*, *42*(3), 370–382. DOI: 10.3758/s13421-013-0368-8 PMID: 24078605

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207. DOI: 10.1109/TPAMI.2012.89 PMID: 22487985

Calvo, M. G., Rodríguez-Chinea, S., & Fernández-Martín, A. (2015). Lateralized discrimination of emotional scenes in peripheral vision. *Experimental Brain Research*, *233*(3), 997–1006. DOI: 10.1007/s00221-014-4174-8 PMID: 25511169

Chan, D. (2009). So why ask me? Are self-report data really that bad? In Lance, C. E., & Vandenberg, R. J. (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 309–336). Routledge.

Chanel, G., Ansari-Asl, K., & Pun, T. (2007). Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In *Proceedings of the 2007 IEEE International Conference on Systems, Man and Cybernetics* (pp. 2662–2667). Institute of Electical and Electronics Engineers (IEEE). DOI: 10.1109/ICSMC.2007.4413638

Clement, J. (2007). Visual influence on in-store buying decisions: An eye-track experiment on the visual influence of packaging design. *Journal of Marketing Management*, *23*(9–10), 917–928. DOI: 10.1362/026725707X250395

Conley, Q., Scheufler, J., Persichini, G., Lowenthal, P. R., & Humphrey, M. (2018). Digital citizenship for all: Empowering young learners with disabilities to become digitally literate. *International Journal of Digital Literacy and Digital Competence*, *9*(1), 1–20. DOI: 10.4018/IJDLDC.2018010101

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222. DOI: 10.1146/annurev.ne.18.030195.001205 PMID: 7605061

Eisenbarth, H., & Alpers, G. W. (2011). Happy mouth and sad eyes: Scanning emotional facial expressions. *Emotion (Washington, D.C.)*, *11*(4), 860–865. DOI: 10.1037/a0022758 PMID: 21859204

Ellis, J. G., Lin, W. S., Lin, C.-Y., & Chang, S.-F. (2014). Predicting evoked emotions in video. In *Proceedings of the 2014 IEEE International Symposium on Multimedia* (pp. 287–294). IEEE. DOI: 10.1109/ISM.2014.69

Espigares-Jurado, F., Muñoz-Leiva, F., Correia, M. B., Sousa, C. M. R., Ramos, C. M. Q., & Faísca, L. (2020). Visual attention to the main image of a hotel website based on its position, type of navigation and belonging to Millennial generation: An eye tracking study. *Journal of Retailing and Consumer Services*, *52*, 101906. DOI: 10.1016/j.jretconser.2019.101906

Fernández-Martín, A., & Calvo, M. G. (2016). Selective orienting to pleasant versus unpleasant visual scenes. *Cognition*, *155*, 108–112. DOI: 10.1016/j.cognition.2016.06.010 PMID: 27371766

Gavrila, D. M. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, *73*(1), 82–98. DOI: 10.1006/cviu.1998.0716

Grabner, H., Nater, F., Druey, M., & Van Gool, L. (2013). Visual interestingness in image sequences. In *MM '13: Proceedings of the 21st ACM International Conference on Multimedia* (pp. 1017–1026). Association for Computing Machinery (ACM). DOI: 10.1145/2502081.2502109

Graneheim, U. H., & Lundman, B. (2004). Qualitative content analysis in nursing research: Concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today*, *24*(2), 105–112. DOI: 10.1016/j.nedt.2003.10.001 PMID: 14769454

Grossmann, I., Gerlach, T. M., & Denissen, J. J. A. (2016). Wise reasoning in the face of everyday life challenges. *Social Psychological & Personality Science*, *7*(7), 611–622. DOI: 10.1177/1948550616652206

Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., & Van Gool, L. (2013). The interestingness of images. In *Proceedings of the 2013 IEEE International Conference on Computer Vision* (pp. 1633–1640). IEEE. DOI: 10.1109/ICCV.2013.205

Hickey, C., & Peelen, M. V. (2015). Neural mechanisms of incentive salience in naturalistic human vision. *Neuron*, *85*(3), 512–518. DOI: 10.1016/j.neuron.2014.12.049 PMID: 25654257

Howard, C. J., & Holcombe, A. O. (2010). Unexpected changes in direction of motion attract attention. *Attention, Perception & Psychophysics*, *72*(8), 2087–2095. DOI: 10.3758/BF03196685 PMID: 21097853

Hutson, J. P., Chandran, P., Magliano, J. P., Smith, T. J., & Loschky, L. C. (2022). Narrative comprehension guides eye movements in the absence of motion. *Cognitive Science*, *46*(5), e13131. DOI: 10.1111/cogs.13131 PMID: 35579883

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. DOI: 10.1016/S0042-6989(99)00163-7 PMID: 10788654

Jacob, R. J. K., & Karn, K. S. (2003). Commentary on section 4. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 573–605). Elsevier B.V.

Jefferies, L. N., Smilek, D., Eich, E., & Enns, J. T. (2008). Emotional valence and arousal interact in attentional control. *Psychological Science*, *19*(3), 290–295. DOI: 10.1111/j.1467-9280.2008.02082.x PMID: 18315803

Karmarkar, U. R., & Yoon, C. (2016). Consumer neuroscience: Advances in understanding consumer psychology. *Current Opinion in Psychology*, *10*, 160–165. DOI: 10.1016/j.copsyc.2016.01.010

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *The American Psychologist*, *73*(1), 1–2. DOI: 10.1037/amp0000263 PMID: 29345483

Koelstra, S., Yazdani, A., Soleymani, M., Mühl, C., Lee, J.-S., Nijholt, A., Pun, T., Ebrahimi, T., & Patras, I. (2010). Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos. In Yao, Y., Sun, R., Poggio, T., Liu, J., Zhong, N., & Huang, J. (Eds.), *Brain informatics* (pp. 89–100). Springer. DOI: 10.1007/978-3-642-15314-3_9

Ković, V., Plunkett, K., & Westermann, G. (2009). Eye-tracking study of inanimate objects. *Psihologija (Beograd)*, *42*(4), 417–436. DOI: 10.2298/PSI0904417K

Lamme, V. A. F. (2003). Why visual attention and awareness are different. *Trends in Cognitive Sciences*, *7*(1), 12–18. DOI: 10.1016/S1364-6613(02)00013-X PMID: 12517353

Li, J., Tian, Y., Huang, T., & Gao, W. (2009). A dataset and evaluation methodology for visual saliency in video. In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo*, pp. 442–445. IEEE. DOI: 10.1109/ICME.2009.5202529

Liang, M., & Hu, X. (2015). Feature selection in supervised saliency prediction. *IEEE Transactions on Cybernetics*, *45*(5), 914–926. DOI: 10.1109/TCYB.2014.2338893 PMID: 25122849

Loschky, L. C., Larson, A. M., Smith, T. J., & Magliano, J. P. (2020). The Scene Perception & Event Comprehension Theory (SPECT) applied to visual narratives. *Topics in Cognitive Science*, *12*(1), 311–351. DOI: 10.1111/tops.12455 PMID: 31486277

Mahapatra, D., Winkler, S., & Yen, S.-C. (2008). Motion saliency outweighs other low-level features while watching videos. In B. E. Rogowitz & T. N. Pappas (Eds.), *Proceedings of SPIE, 6806, Human vision and electronic imaging XIII*. SPIE, the international society for optics and photonics. DOI: 10.1117/12.766243

Mayer, R. E. (2002). Cognitive theory and the design of multimedia instruction: An example of the two-way street between cognition and instruction. *New Directions for Teaching and Learning*, *2002*(89), 55–71. DOI: 10.1002/tl.47

Meyerhoff, H. S., Huff, M., & Schwan, S. (2013). Linking perceptual animacy to attention: Evidence from the chasing detection paradigm. *Journal of Experimental Psychology. Human Perception and Performance*, *39*(4), 1003–1015. DOI: 10.1037/a0030839 PMID: 23181685

Meyerhoff, H. S., Schwan, S., & Huff, M. (2014). Perceptual animacy: Visual search for chasing objects among distractors. *Journal of Experimental Psychology. Human Perception and Performance*, *40*(2), 702–717. DOI: 10.1037/a0034846 PMID: 24294872

Michotte, A. (1991). The emotions regarded as functional connections. In G. Thinés, A. Costall, & G. Butterworth (Eds.), *Michotte's experimental phenomenology of perception* (pp. 103–116). Erlbaum. (Original publication 1950)

Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, *3*(1), 5–24. DOI: 10.1007/s12559-010-9074-z

Mohanty, A., & Sussman, T. J. (2013). Top-down modulation of attention by emotion. *Frontiers in Human Neuroscience*, *7*, 102. DOI: 10.3389/fnhum.2013.00102 PMID: 23554590

Moray, N. (2017). *Attention: Selective processes in vision and hearing*. Routledge. DOI: 10.4324/9781315514611

New, J., Cosmides, L., & Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(42), 16598–16603. DOI: 10.1073/pnas.0703913104 PMID: 17909181

Niu, Y., Todd, R. M., & Anderson, A. K. (2012). Affective salience can reverse the effects of stimulus-driven salience on eye movements in complex scenes. *Frontiers in Psychology*, *3*, 336. DOI: 10.3389/fpsyg.2012.00336 PMID: 23055990

Nugent, S. (2013). Catch trial. *Psychology dictionary professional reference*. Retrieved on September 15, 2022, from https://psychologydictionary.org/catch-trial/

Nummenmaa, L., Hyönä, J., & Calvo, M. G. (2006). Eye movement assessment of selective attentional capture by emotional pictures. *Emotion (Washington, D.C.)*, *6*(2), 257–268. DOI: 10.1037/1528-3542.6.2.257 PMID: 16768558

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *13*(11), 4700–4719. DOI: 10.1523/JNEUROSCI.13-11-04700.1993 PMID: 8229193

Osberger, W. M., & Rohaly, A. M. (2001). Automatic detection of regions of interest in complex video sequences. In Rogowitz, B. E., & Pappas, T. N. (Eds.), *Proceedings of SPIE, 4299, Human vision and electronic imaging VI* (pp. 361–373). SPIE. DOI: 10.1117/12.429506

Patton, M. Q. (2005). Qualitative research. In Everitt, B. S., & Howell, D. C. (Eds.), Encyclopedia of statistics in behavioral science (pp. 1633–1636). John Wiley & Sons. DOI: 10.1002/0470013192.bsa514

Pratt, J., Radulescu, P. V., Guo, R. M., & Abrams, R. A. (2010). It's alive! Animate motion captures visual attention. *Psychological Science*, *21*(11), 1724–1730. DOI: 10.1177/0956797610387440 PMID: 20974713

Rehman, A., & Saba, T. (2014). Evaluation of artificial intelligent techniques to secure information in enterprises. *Artificial Intelligence Review*, *42*(4), 1029–1044. DOI: 10.1007/s10462-012-9372-9

Sandbach, G., Zafeiriou, S., Pantic, M., & Yin, L. (2012). Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image and Vision Computing*, *30*(10), 683–697. DOI: 10.1016/j.imavis.2012.06.005

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition*, *19*(4), 1069–1078. DOI: 10.1016/j.concog.2009.12.013 PMID: 20133167

Schellings, G., & Van Hout-Wolters, B. (2011). Measuring strategy use with self-report instruments: Theoretical and empirical considerations. *Metacognition and Learning*, *6*(2), 83–90. DOI: 10.1007/s11409-011-9081-9

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299–309. DOI: 10.1016/S1364-6613(00)01506-0 PMID: 10904254

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: Relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences*, *12*(8), 314–321. DOI: 10.1016/j.tics.2008.04.008 PMID: 18606562

Shrivastava, N., & Tyagi, V. (2014). Content based image retrieval based on relative locations of multiple regions of interest using selective regions matching. *Information Sciences*, *259*, 212–224. DOI: 10.1016/j.ins.2013.08.043

Simmonds, L., Bellman, S., Kennedy, R., Nenycz-Thiel, M., & Bogomolova, S. (2020). Moderating effects of prior brand usage on visual attention to video advertising and recall: An eye-tracking investigation. *Journal of Business Research*, *111*, 241–248. DOI: 10.1016/j.jbusres.2019.02.062

Simola, J., Torniainen, J., Moisala, M., Kivikangas, M., & Krause, C. M. (2013). Eye movement related brain responses to emotional scenes during free viewing. *Frontiers in Systems Neuroscience*, *7*, 41. DOI: 10.3389/fnsys.2013.00041 PMID: 23970856

Smith, T. J., Levin, D., & Cutting, J. E. (2012). A window on reality: Perceiving edited moving images. *Current Directions in Psychological Science*, *21*(2), 107–113. DOI: 10.1177/0963721412437407

Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, *49*(8), 1017–1034. DOI: 10.1111/j.1469-8986.2012.01384.x PMID: 22680988

Torres, C. A., Orozco, Á. A., & Álvarez, M. A. (2013). Feature selection for multimodal emotion recognition in the arousal-valence space. In *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 4330–4333). IEEE. DOI: 10.1109/EMBC.2013.6610504

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. DOI: 10.1016/0010-0285(80)90005-5 PMID: 7351125

Tremoulet, P. D., & Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Perception & Psychophysics*, *68*(6), 1047–1058. DOI: 10.3758/BF03193364 PMID: 17153197

Tunga, Y., & Cagiltay, K. (2023). Looking through the model's eye: A systematic review of eye movement modeling example studies. *Education and Information Technologies*, *28*(8), 9607–9633. Advance online publication. DOI: 10.1007/s10639-022-11569-5

Vogt, J., De Houwer, J., Koster, E. H. W., Van Damme, S., & Crombez, G. (2008). Allocation of spatial attention to emotional stimuli depends upon arousal and not valence. *Emotion (Washington, D.C.)*, *8*(6), 880–885. DOI: 10.1037/a0013981 PMID: 19102600

Wang, C., Geelhoed, E. N., Stenton, P. P., & Cesar, P. (2014). Sensing a live audience. In *CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1909–1912). ACM. DOI: 10.1145/2556288.2557154

Wang, S., & Ji, Q. (2015). Video affective content analysis: A survey of state-of-the-art methods. *IEEE Transactions on Affective Computing*, *6*(4), 410–430. DOI: 10.1109/TAFFC.2015.2432791

Xu, M., Jin, J. S., Luo, S., & Duan, L. (2008). Hierarchical movie affective content analysis based on arousal and valence features. In *MM '08: Proceedings of the 16th ACM International Conference on Multimedia* (pp. 677–680). ACM. DOI: 10.1145/1459359.1459457

Yadati, K., Katti, H., & Kankanhalli, M. (2013). Interactive video advertising: A multimodal affective approach. In S. Li, A. El Saddik, M. Wang, T. Mei, N. Sebe, S. Yan, R. Hong, & C. Gurrin (Eds.), *Advances in multimedia modeling* (pp. 106–117). Springer. DOI: 10.1007/978-3-642-35725-1_10

Yan, Q., Xu, L., Shi, J., & Jia, J. (2013). Hierarchical saliency detection. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1155–1162). IEEE. DOI: 10.1109/CVPR.2013.153

Yang, Y., Nie, F., Xu, D., Luo, J., Zhuang, Y., & Pan, Y. (2012). A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 723–742. DOI: 10.1109/TPAMI.2011.170 PMID: 21844624

Yoon, S., & Pavlovic, V. (2014). Sentiment flow for video interestingness prediction. In *HuEvent '14: Proceedings of the 1st ACM International Workshop on Human Centered Event Understanding from Multimedia* (pp. 29–34). ACM. DOI: 10.1145/2660505.2660513

Zu, T., Hutson, J., Loschky, L. C., & Rebello, N. S. (2020). Using eye movements to measure intrinsic, extraneous, and germane load in a multimedia learning environment. *Journal of Educational Psychology*, *112*(7), 1338–1352. DOI: 10.1037/edu0000441

Quincy Conley is a researcher specializing in the intersection of educational technology, human-computer interaction, and cognitive psychology. With a focus on using innovative technologies like eye-tracking, facial expression recognition, galvanic skin conductance, and other biometric sensors to understand cognitive processes, Dr. Conley aims to enhance user engagement and optimize learning experiences in digital environments. Dr. Conley holds a doctorate from Arizona State University and has published several articles exploring the application of biometric technology in educational contexts. His work bridges the gap between cutting-edge biometric technology and practical applications in education, contributing to the advancement of interactive and efficacious learning systems.