Visual Recognition for Dynamic Scenes

by

Ryan Ferguson

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2014 by the
Graduate Supervisory Committee:
Donald Homa, Chair
Stephen Goldinger
Arthur Glenberg
Gene Brewer

ARIZONA STATE UNIVERSITY

May 2014

ABSTRACT

Recognition memory was investigated for naturalistic dynamic scenes. Although visual recognition for static objects and scenes has been investigated previously and found to be extremely robust in terms of fidelity and retention, visual recognition for dynamic scenes has received much less attention. In four experiments, participants view a number of clips from novel films and are then tasked to complete a recognition test containing frames from the previously viewed films and difficult foil frames. Recognition performance is good when foils are taken from other parts of the same film (Experiment 1), but degrades greatly when foils are taken from unseen gaps from within the viewed footage (Experiments 3 and 4). Removing all non-target frames had a serious effect on recognition performance (Experiment 2). Across all experiments, presenting the films as a random series of clips seemed to have no effect on recognition performance. Patterns of accuracy and response latency in Experiments 3 and 4 appear to be a result of a serial-search process. It is concluded that visual representations of dynamic scenes may be stored as units of events, and participant's old/new judgments of individual frames were better characterized by a cued-recall paradigm than traditional recognition judgments.

DEDICATION

This work is dedicated to my wife, Jennifer, who has encouraged, supported, and fed me throughout my time as a grad student.

ACKNOWLEDGMENTS

I would like to extend my thanks to my committee, Art Glenberg, Steve Goldinger, Gene Brewer, and Don Homa. Their support in the production of this project and throughout graduate school has had a profound effect on me as a person and as a scientist. I especially grateful to have Don Homa as an advisor and mentor. His kindness, humor, and passion has taught me to love and appreciate good data and good dogs.

TABLE OF CONTENTS

Page

LIST OF TABLES vi

LIST OF FIGURES vii

INTRODUCTION 1

Static Visual Memory 2

Memory for Dynamic Events 7

Visual Memory Errors 13

EXPERIMENT 1 19

Method 20

Results 22

Experiment 1 Disucussion 23

EXPERIMENT 2 25

Method 26

Results 26

Experiment 2 Disucussion 28

EXPERIMENT 3 31

Method 32

Results 34

EXPERIMENT 4 39

Method 40

Results 40

Experiment 3 & 4 Disucussion.....43

GENERAL DISCUSSION ..... ……. 48

GENERAL CONCLUSIONS…..……59

REFERENCES..... …… 60

LIST OF TABLES

Table Page

1. Overall and near edge recognition performance for Exp. 3 & 4.....47

LIST OF FIGURES

Figure Page

1. Recognition accuracy for Exp 1 23

2. Recognition accuracy for Exp 2 27

3. Hits and False Alarms for Exp 1&2 28

4. Recognition accuracy for Exp 3 34

5. Reaction Times for Exp 3 35

6. Accuracy by gap item quartile for Exp 3 36

7. Accuracy for seen items by gap distance 37

8. Recognition accuracy for Exp 4 41

9. Reaction Times for Exp 4 42

10. Accuracy by gap item quartile for Exp 4 43

INTRODUCTION

In a 2013 article in The New York Review of Books, famed American film director Martin Scorsese reflects on the origins of his love affair with motion pictures and the structures which contribute to movie magic. He describes the experience of cinema as an emergent one, "You take one shot, you put it together with another shot, and you experience a third image in your mind's eye that doesn't really exist in the two other images." This decoupling of physical stimuli and mental representation, the difference between what the eye and mind see – and further what the eye saw and what the mind recalls- has always fascinated cognitive scientists.

The current studies aim to investigate the visual component of our mental representations for dynamic scenes. Our memory for dynamic scenes must be able to produce both visual and temporal detail. For illustration purposes, consider the example of a televised trial. Visual memory could be stored like video footage, an eidetic and comprehensive representation of all that was observed. Although unlikely, visual memory for static objects has been called essentially unlimited (Standing, 1977) and is capable of containing a staggering amount of detail (Homa & Viera, 1988; Brady, Konkle, Alvarez, & Oliva, 2008). If this were the case, recall of a dynamic scene would simply be a processing of cueing up the correct footage. Alternatively, the temporal relations of the visual memory could be handled by a knowledge of the events and story of the trial (similar to a newspaper article about the trial) coupled with the visual knowledge of what the defendant, lawyers, and setting are. In this representation, recalling visual memories of the event is more of a reconstruction depending heavily on a higher level structure,

similar to story schema accounts of narrative comprehension (Thorndyke, 1977; Rumelhart, 1997). Both the eidetic and schema based representations lie on two ends of a spectrum which are broadly defined by bottom-up and top-down influences. Lying between the two is the possibility that the visual component of your memory for the event is mainly composed of the dramatic and surprising occurrences that happen within the trial, such as the defendant failing to fit a glove over his hand. This representation would be similar to the footage shown on the evening news which summarized the day's events. All visual information that fails to surprise falls to the wayside. If some occurrence does not surprise you during encoding, it would be easy to assume or fill in during recall. A visual representation could also be like that of a courtroom sketch artist, who collapses visual information across an event into a single illustration. The sketches are produced in order to be representative of a period of time, such as a cross examination, rather than to capture all of the visual detail.

While all of these possibilities seem plausible, the nature of visual memory for dynamic scenes is unlikely to be captured in a simple analogy. Visual memory is usually studied in by utilizing static images of either objects or scenes. The purpose of the current series of studies is to extend studies of visual recognition along a temporal dimension, and by doing so gain insight into the nature of the representations that perceiving dynamic scenes leave behind.

Static visual memory

Human's memory for still images has been shown to be both highly efficient and massive. Shepard (1967) demonstrated this by showing subjects 612 images for a short

duration (6s) and then employing a two-alternative choice test 2 hours later. His subjects correctly identified the studied photographs at a rate of 98% correct. Standing (1973) expanded on Shepard's finding by showing subjects 10,000 images for 15 seconds a piece over the course of 5 days. He found that subjects were able to recognize studied images 83% of the time, and images that were rated as "vivid" were recognized at an even higher rate. Some critics of these studies point out that subjects may be remembering a "gist" of the studied images rather than a rich visual representation (Chun, 2003; Simons & Levin, 1997). However, several studies have employed related foils in the recognition test and still found evidence for a robust memory for images (Konkle, Brady, Alvarez, & Oliva, 2010; Homa & Viera, 1988; Brady, Konkle, Alvarez, & Oliva, 2008). Homa and Viera (1988) varied the quality of foils by holding thematic detail constant but deleting extraneous physical detail, resulting in foils presented during recognition testing that had a very similar gist as those studied in a learning phase. They found that even when recognition testing was delayed 12 weeks, subjects were still able to reject the thematically related foils significantly above chance. Brady et al. (2008) had participants view thousands of unique objects in a single study session. Participants were then given a two-alternative forced choice test containing a studied object and a foil object that was either from a novel category, from the presented studied object's same category, or the exact same object as the studied object but in a different position. Even with the extremely difficult foils, participants achieved recognition accuracy of above 85% in every type of foil. These recognition accuracy rates suggest that representations

of static objects are not merely gist or basic-category level representations, but contain enough detail to enable accurate recognition in the face of highly similar foils.

While visual memory for images of individual objects is certainly large in terms of both capacity and fidelity, real world visual memories are for objects embedded within scenes, or scenes themselves. At the very least, scenes provide important context for expectations and identification of objects on which we plan to attend, especially natural scenes which tend to be predictable (Kersten, 1987). A great deal of information can be gathered about a scene from even a very short glance. Potter (1976) found that basic-level category knowledge (e.g., a birthday party scene) could be gathered by participants who were exposed to the images for less than 100ms. Even when images of scenes are distorted to the extent of disrupting individual object identification, the overall gist of a scene can still be obtained in very brief presentations. Schyns and Oliva (1994) presented images that contained the low special frequency information from a scene, thus destroying all fine detail and producing an image of a scene made up of fuzzy blobs. Participants were able to categorize the type of scene (highway, kitchen, hallway, etc.) even when presentation times were as short as 30ms. This quick identification of scenes, even in the absence of specific object information, may be due to a sort of 'global processing' which processes a scene as a whole and derives statistical regularities from it (Oliva & Torralba, 2001; Oliva, 2005). Using longer study durations, scenes can also be easily memorized and discriminated amongst similar foils at a performance level similar to that seen in static object visual recognition tasks. A study by Konkle et al (2010) had participants study almost 3000 images from 128 scene categories consisting of 1,4,16, or

64 exemplars each. Recognition memory was assessed using a two-alternative force choice test with either a novel or same-category scene. Mean performance across all of the exemplar sizes was at 80% correct, and the highly-homogeneous 64 exemplar categories were at a 76% correct recognition rate. Konkle concluded that the remarkable performance suggested that the representation of these scenes, which were intentionally studied for later recall, was highly unlikely to be at a gist or basic-level.

While some reports of vivid and accurate multimodal representations do exist, there is much more evidence that our perceptual system takes certain shortcuts in order to process important and salient information from our complex environment. Some of the most striking evidence for less than perfect visual representations of scenes come from the 'change blindness' literature (Simons & Levin, 1997; Rensink et al.,1997). Change blindness refers to the failure of observers to detect a large change in an environment such as actors being switched. While the extent to which observers are blind to changes is surprising, caution must be taken when taking change blindness as a condemnation of rich representations (Simons & Rensink, 2005). For instance, Hollingworth (2003) found excellent change detection when observers were cued to a potentially rotated object in a complex scene after the change took place. By accurately detecting the subtle changes in the complex scenes, the participants demonstrate that they still had access to a relatively rich pre-change representation, suggesting that some change blindness effects may be due to breakdowns in the comparison process rather than due to a sparse visual representation.

The majority of research on visual memory and recognition utilizes a design where single images are presented with study times ranging from 1 to 15 seconds. Rapid Serial Visual Presentation (RSVP) has also been used to investigate the temporal aspects of visual memory. Potter and Levy (1969) varied presentation times of thematically unrelated images from 113ms to 2,000ms. They found that the subsequent recognition test was greatly affected by presentation time, with images studied for 125ms correctly identified as old only 16% of the time, while images studied for 2000ms were correctly identified as old 93% of the time. However, when the task was changed to detection of a previously cued image within a sequence of rapidly presented images, subjects were able to correctly detect the cued image 70% of the time. This finding suggests that rapidly presented and unrelated images do not fall below a threshold for encoding; rather they are swiftly discarded if not cued as a target beforehand. Intraub (1980, 1984) obtained similar results by manipulating the interstimuli interval (ISI) between brief (110ms) presentations of images and found that recognition performance with a blank ISI was almost equivalent to recognition performance of a persistent stimulus presentation, indicating some iconic persistence. However, when the ISI was filled with to-be-ignored images, recognition performance for the target image fell sharply. Subramaniam, Biederman, and Madigan (2000) also investigated RSVP and the effect of repetition to prime target images. Once again, they found that identification of cued target images within RSVP was possible with presentation rates down to sub 100ms times, but subsequent recognition tests were at chance unless the presentation rate grew to above 200ms. Furthermore, they found no effect of repeating an image up to 31 times before it became a target on identification

tasks at short presentation intervals (76-126ms). This lack of priming, even with a large amount of repetition exposure, once again suggests that the presentation rate is falling below some established threshold that allows for meaningful encoding. They offer a speculative neurological explanation for their findings based on a theory from Potter (1976), who suggested that a fast enough RSVP would present "conceptual masking" that interfered with memory consolidation. Images below a threshold can be shallowly processed, and thus identified in line, but the thematic and conceptual translation is interrupted by the next stimulus presentation, preventing assimilation into longer term memory stores that are probed by recognition tests. Images that are rapidly presented that share a common theme, such as a video, may be able to overcome the RSVP recognition thresholds.

Memory for Dynamic Events

While static images are usually the topic of study when it comes to visual recognition research, our visual world is dynamic and real world recognition often involves recognizing moving objects or scenes. There is some evidence that our internal representations of dynamic visual stimuli are themselves dynamic, as illustrated by the phenomenon known as representational momentum (Freyd & Finke, 1984). Representational momentum is a systematic illusion which seems to suggest the visual system takes physical properties into account during perception and storage (Freyd & Johnson, 1987). In a typical representational momentum experiment, a short movie is shown depicting a shape moving across a static background. When the movie is abruptly stopped and observers are asked to indicate the last position of the shape, their judgments

are biased based on the delay after the offset of the shape and the acceleration of the shape, indicating that their representation of the shape continued to move as it had been.

While the representational momentum work may suggest dynamic representations, or at least representations that include predictions, it does not speak much of the fidelity and abstraction of the visual memory. A handful of studies have looked at visual memory using more standard study-test procedures such as those studies discussed in the previous section. Goldstein, Chance, Hoisington, and Buescher (1982) had participants either view film footage or view static images drawn from film footage and then had then complete a recognition test consisting of dynamic clips or static images. Those who studied static images only took a static recognition test. They found that there was a general advantage for dynamic encoding regardless of the type of recognition test taken, with those who were shown dynamic footage at learning and test performing the best. Matthews, Benjamin and Osborne (2007) also found evidence for the superior encoding of dynamic images. Using a better controlled design than Goldstein et al. in terms of stimulus equivalency, they presented participants with hundreds of static images or very short dynamic clips featuring a wide variety of subjects and had them return a week and four weeks later for a recognition test. In all conditions, items that were learned with dynamic representations were recognized with greater accuracy. This advantage was deemed the dynamic superiority effect (Matthews, Benjamin, & Osborne, 2007). It has since been demonstrated that there is a study-test congruence effect, where images are best recognized in their studied static or dynamic state (Burratto, Matthews & Lamberts, 2009).

. Dynamic presentation improves upon a visual memory system that has already been demonstrated to possess impressive capacity. Our own autobiographical memories can recall dynamic scenes with ease. Just how much are we encoding? One suggestion is that we are continuously sampling the world around us in an online fashion. An extreme example of such processing is found in the work of Penfield (1958). He used electrodes to probe the cortex of patients and discovered they recalled rich and fluid past memories when stimulated at certain locations. These memories were reported as being very detailed and specific, as if the patients had rewound a video tape to a particular place and hit play. In describing one patient's experience, Penfield (1958) wrote:

> When the electrode was applied in gray matter…, the patient observed: "I hear some music." Fifteen minutes later, the electrode was applied to the same spot again without her knowledge. "I hear music again," she said. "It is like radio." Again and again, then, the electrode tip was applied to this point. Each time, she heard an orchestra playing the same piece of music. (p.57)

Loftus and Loftus (1980) disputed Penfield's claims and dismissed his findings as unscientific, citing how easily memory is manipulated by using leading questions and different testing procedures. However, an investigation by Hamani, Stone, Laxton, and Lozano (2007) produced a report of similar autobiographical memories evoked during deep brain stimulation of the hypothalamic/fornix region. As with Penfield's findings, they reported very rich detailed memories that were consistently evoked when particular areas of the hypothalamic/fornix region were stimulated.

The structure of an event seems to play a role in its representation. Narrative comprehension and memory have been found to be strongly sensitive to narrative structure (Mandler & Johnson, 1977; Thorndike, 1977). Narrative structure is usually

presented as a hierarchical breakdown of the components of a story in which larger events can be broken down into sub component events. Mandler and Johnson (1977) analyzed folktales with the intent of identifying commonalities in structure among them and identified what could be referred to as a story grammar. This generic representation of narrative has been proposed in many different forms, but in general it contains a sort of scaffolding in which a story can rest. Thorndyke's (1997) grammar contained a hierarchical tree structure which at the highest level contained high level units of the story like setting, theme, plot, and resolution. Each of these units contain optional subunits that further elucidate their superordinate level, such as the setting level being modified by subordinate levels of time and place. Thorndyke found that participants' recall and comprehension of a story fell sharply if the story was reorganized in such a way that it could not have come from a story grammar. It has also been demonstrated that when participants were asked to recall stories presented in ungrammatical ways, their recall tended to reorder the units of the story into a more grammatical order –even when asked to recall the story verbatim (Mandler & Johnson, 1977, Stein & Nezworski, 1978; Thorndyke, 1977; Rumelhart, 1975). These results support the position that story grammars closely mirror our own hierarchical representations of narrative in memory.

Given the importance of internalized structure in recall and conceptualization, it may be equally important in perception. In order for a structured perception to occur there must be some segmentation of the dynamic environment. Stroud (1956) proposed that perception was continually segmented on a temporal basis in his perceptual moment hypothesis. His theory speculated that perception was processed in discrete chunks of

time, usually around 100ms, and that separate events that occurred within that time period would be perceived as simultaneous. Although generally discounted, more recent neurological evidence has found evidence for oscillatory brain frequencies that predict detection of separate events (VanRullen & Koch, 2003; Smith, Cottrell, Gosselin, & Schyns, 2005). Another theory is that perception is segmented into variable units that are context dependent. One such theory chunks perception into units defined by "breakpoints" (Newtson & Enquist, 1976). Newtson and Enquist suggest these breakpoints structure perception of dynamic events into action units (1977). These breakpoints are points at which a new and distinct action unit is created relative to a previous action unit. They found empirical support for this theory by having one set of subjects view a short film and press a button when they believe the action had shifted. With these breakpoints identified, the same film was shown to subjects with a small number of frames removed from breakpoint or non-breakpoint areas. Breakpoint areas showed higher levels of detection- up to 78% for a .5 second deletion- versus non-breakpoint deletions (a detection level of around 35%). In another experiment, they had one group of subjects watch a film and actively define breakpoints via button push, while another group passively watched the film. All subjects received a recognition test of both breakpoint and non-breakpoint frames. Breakpoint frames were recognized at a higher rate than non-breakpoint frames, and there were no significant differences between the passive and active groups when it came to recognition rate. Newtson and Enquist took this as evidence that subjects were automatically defining breakpoints, even when not explicitly told to do so.

Another theory of automatic and variable perceptual unit segmentation has been proposed by Zacks et al. (2007). In Newtson and Enquist's form of event segmentation, activity was segmented when there was a shift in action. Zacks and colleagues suggest that activity was segmented whenever activity deviated from a prediction model. This theory, known as the Event Segmentation Theory (EST), suggests that a predictive model is built upon initial exposure to some dynamic event. As long as the model continues to correctly account for ongoing action, it remains. However, when predictions begin to accrue error to exceed a threshold, the current predictive model is scrapped and a new one is constructed. When segmentation occurs, a process of perceptual reorientation is triggered in order to build a new predicative model. In everyday perception, these models can exist on a timescale that lasts a few seconds up to ten minutes (Kurby & Zacks, 2008). These events are said to be hierarchically structured, with smaller fine-grained events making up larger coarse-grained events. These fine-grained events are defined by boundaries that contain the "smallest natural and meaningful event," and coarse-grained events have boundaries which contain larger events (Zacks, Speer, & Swallow, 2007). For instance, when participants were asked to segment a video of a woman changing the sheets on a bed based on fine-grained events, they placed event boundaries after she had removed the pillow case from each individual pillow. Those who were asked to segment based on coarse-grained events placed boundaries after she had removed the cases from all of the pillows (Zacks, Tversky, & Iyer, 2001). These fine grained events are the building blocks of the coarse-grained events, with fine event boundaries lining up with the coarse event boundaries. Unlike the earlier work of Newtson and Engquist, Zacks has

focused on finding implicit neurological evidence for segmentation. Using functional magnetic resonance imaging (fMRI), synchronized brain activity was observed among subjects that lined up with event boundaries in movies of people performing everyday tasks (Zacks et al., 2001). Furthermore, brain responses were larger for boundaries of coarse events than for fine gained events. Similar synchronicity was observed using EEG (Kurby & Zacks, 2007) and eye movements (Swallow and Zacks, 2006). Zacks, Speer, Swallow and Maley (2010) had subjects watch an entire cinematic movie while in an fMRI machine and observed greater activity along event boundaries that coincided with new object interactions, changes in spatial locations, and goals. They theorize that these automatic responses along event boundaries form a structure for dynamic scenes that processes and stores ongoing action in terms of discrete units. A very complex dynamic scene that often defies the predictions of the observer is made up of many segments, while a simple and predictable scene is processed as relatively few segments. This sort of variable unit processing account of dynamic scenes allows a memory representation to be sparse when there is nothing of interest occurring or full of finely updated memory orientations when something complex or important occurs within the context of some ongoing action.

Visual Memory Errors

Visual illusions have long been used by psychologists to study perception. By isolating instances in which perception does not match objective reality, observations of the organizing principals of visual perception can be made. In the same way perceptual illusions inform us about perception, memory illusions or distortions inform us about

long term memory systems (Roediger, 1996). Despite the amazing visual memory discussed earlier, there are still situations where our memory for images falls short. Bartlett (1932) suggested that memory performs both reproductive and reconstructive tasks. Remembering a baseball glove seen in a learning task is highly reproductive, whereas remembering a highly contextualized item, such as the girl you sat next to in chemistry class in high school, is much more reconstructive. A classic example of reconstructive memory processes is demonstrated in the Deese-Roediger-McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995). In this task, subjects are usually read a list of words that semantically related to an unread target word. For instance, subjects may hear words like *bed, nap, drowsy, blanket, night* which are all semantically related to *sleep*. When later given a recognition test of the words heard earlier, subjects often false alarm to the target word (*sleep*) with a great deal of confidence. This paradigm has been repeated using visual stimuli, but to a lesser degree. Smith and Hunt (1998) either read a list of semantically related words or presented the written words visually. They found that subjects who auditorily encoded the word list recognized the un-heard target word at a similar rate to the rest of the words in the list, but those who visually saw the word list had an approximately 50% drop in false alarms. Israel and Schater (1997) found an even further decrease in false alarms when they presented line drawings instead of visual words.

While the visual modality seems to be somewhat more resilient to the types of false memory displayed in a DRM procedure, there are cases where memory for visual events appears vulnerable. Loftus, Miller and Burns (1978) performed a landmark study

involving visual memory, discovering what they called the "misinformation effect." In this study, participants saw a series of photographic slides of events leading up to a car accident. One group of subjects saw a car pulling up to a stop sign, and the other group saw a car pulling up to a yield sign. Subjects were later probed with questions that contained either consistent or inconsistent information regarding the traffic sign they observed in the slides. Finally, subjects were asked to identify slides that came from the original presentation of the car accident. Participants that were given inconsistent information in the probe questions later false alarmed to slides matching that inconsistent information. Miller and Gazzaniga (1998) also demonstrated fragile visual memory when they found that subjects often reported remembering the presence of items within a complex scene that were actually absent. They showed subjects a series of scenes that had a stereotypical item removed (for example, a beach scene with a beach ball removed). When subjects were later probed, they false alarmed to the absent but stereotypical items at similar rates to items that were actually present within the scene. Furthermore, subjects were often confident in their memory fabrications, as measured by remember/know judgments.

While it is difficult to reconcile an incredible visual memory with an episodic memory that is easily corrupted, there could be, as Bartlett suggested, two distinct memory processes involved. A recent study investigated detail retention and false memories by varying the context of studied items during test (Guerin, Robbins, Gilmore, & Schacter, 2012). Subjects studied everyday objects and then were given a recognition test that involved a choice between 3 items (or a choice of none). The triads consisted of

either a single related item and two novel items, two related items and one novel item, or a related item with a novel item and an item that was actually studied. They found that when either one or two related items were in the recognition triad, but an actually studied item was absent, they had false alarms between 35% and 50% of trials. However, when a studied item was presented alongside a related item, false alarms dropped to around 10%. This finding suggests that a detail-rich encoding of a studied item still persists, even when a relatively high amount of false alarms are occurring. While examples of memory distortions are usually presented as a breakdown within the memory system, some researchers have suggested that memory distortions are a result of a memory system that is constantly making predictions for future events (Schacter & Addis, 2007). Known as the constructive episodic simulation hypothesis (CESH), it is argued that episodic memories are retrieved and used to simulate possible outcomes in a person's immediate future. These simulations benefit from episodic memories that can be manipulated to fit different situations. CESH posits that memory distortions are a result of episodic memories that gain qualities from possible simulations and are replaced into their episodic memory store. Both CESH and Zach's Event Segmentation Theory propose that prediction plays a central role in processing of dynamic events. Strickland and Keil (2011) had participants view a video clip of a person approaching a soccer ball as if they were about to kick it. Before contact was made, the video cut to either an image of the ball bouncing down the field (implied contact) or to an empty field (non-implied contact). Participants who saw the ending with the implied contact had higher false alarms to images of the contact occurring than the non-implied contact group. Strickland and Kiel

suggested that the high false alarms occurred because the false alarmed to images fit in with a cohesive event model.

Other models of memory, proposed to account for false memory phenomena, fall into two broad categories, single and dual process models. Dual process models of memory propose two distinct types of memory play a part in recognition judgments: familiarity and recollection (see Yonelinas, 2002 for review). One of the dual process models often used to explain false memory is called Fuzzy Trace theory (Brainerd, Reyna, & Kneer, 1995). Fuzzy Trace theory (FTT) suggests that during encoding of some event, a number of traces are produced that vary on a continuum from gist to verbatim. Verbatim traces contain rich and specific detail. When a probed item is similar to a verbatim trace, it produces a strong feeling of recollection for that probed item within memory. When a probed item is compared to a gist trace, it is judged as being familiar. If this familiarity (the similarity to the gist trace) is sufficiently strong, it can cause a false recognition error (Brainerd et al, 1995). Single process models of memory suggest that subjective ratings of specific recall or general familiarity are not two distinct processes, but that strength of familiarity alone enables recognition judgments. If familiarity exceeds some threshold, an item is felt to be recognized. Global-matching models (Arndt & Hirshman, 1998; Hintzman, 1988; Hintzman, 1986) account for false recognition by once again suggesting that events are stored in traces that can be thought of as feature vectors. When an item is probed for recognition, the summed similarity of that item to all other relevant traces is calculated and it may or may not exceed a threshold for recognition- even if it is not an exact match of any stored probed item.

The current set of experiments utilize recognition judgments in order to assess the qualities of visual memory for dynamic scenes. The addition of a temporal dimension brings a host of new variables that studies utilizing pictures static objects or scenes have not needed to account for. The current experiments seek to build upon existing visual memory work by utilizing dynamic visual learning materials. Temporal structure and test-item similarity are manipulated in order to investigate the form which visual memories for dynamic scenes take.

EXPERIMENT 1

Experiment 1 examines the detail of visual memory for dynamic events. In many ways, it is similar to other studies that have found that our visual memory for static events is so massive (Standing, 1973; Shepard, 1967, Konkle et al., 2019; Brady et al., 2008). Because our visual system evolved in a dynamic environment, it is tempting to assume that recognition for dynamic scenes would also be impressive. However, research has shown that memories for dynamic events can be manipulated by post-event misinformation (Wright, Loftus, & Hall, 2001; Loftus, Miller, & Burns, 1978; Clifasefi, Garry & Loftus, 2007). This malleability suggests that top down information plays an important part in recall. Given the effects of structure and its impact on recall of simple narratives (Thorndyke, 1977), how well a participant understands an event should have some effect on how well they encode and can later recall visual memories.

Participants watched three videos, each followed by a recognition task that has them make old/new judgments for frames that either come from the video they just watched or came from a highly similar scene containing the same characters in the same settings. There are two between subject conditions: a linear condition, in which participants watch the film clips as they were originally produced, and a jumbled condition, in which the film clips have been cut into many pieces and rearranged in order to cause some deficit in the perception of global narrative structure. The three videos chosen (12 Angry Men, Dr. Who, and Looney Tunes) were chosen because they had highly consistent settings, which aided in identifying foils. They also represented variation in terms of color versus black and white (Looney Tunes and Dr. Who versus 12 Angry Men), and a continuum of the number of human faces (12 Angry men has many

faces in every scene, Dr. Who has half human faces and half alien masks, and Looney Tunes has no human faces). Both color (Wichmann, Sharpe, & Gegenfurtner, 2003; Matthews, Benjamin, & Osborne, 2007) and faces (Lander & Bruce, 2003) have been investigated in terms of their effects on memory for dynamic stimuli.

Method

Participants

A total of 54 Arizona State University undergrads from introductory psychology classes participated in Experiment 1 in order to fulfill a course requirement, 26 in the linear condition (no jumbling) and 28 in the jumbled condition. Subjects were randomly assigned to a condition. At the conclusion, participants were queried if they had previously seen any or all the video clips.

Materials and Apparatus

Three video clips were chosen to reflect wide sampling of content and various levels of reality. One of the clips was a 5 minute section of the 1957 movie "12 Angry Men." This black and white film was selected because of the low amount of variation in scenes and characters throughout the movie (the movie takes place in the same room with the same 12 men). Another 5 minute clip came from the BBC television series "Dr. Who" from 1984. This clip contained actors either in futuristic dress or in full makeup that completely hides facial features. The last 5 minute clip came from the cartoon Looney Tunes featuring a coyote chasing after a road runner. This clip was fully animated throughout and also contained no voices. For each of the three 5 minute clips chosen, a matching 5 minute clip was chosen from another point in the video from which to draw

the foils. The clips from which the foils were drawn were chosen because they contained the same characters in the same general setting. This was a non-issue for 12 Angry Men and Looney Tunes because the setting is consistent. The clips and foils from Dr. Who were chosen because they had the same characters in the same locations. None of the foil clips immediately preceded or followed the presented videos. Subjects in both the linear and jumbled conditions received the same foils. The "Dr. Who" and Looney Tunes clips had a frame rate of 25 frames per second (fps) and the "12 Angry Men" had an fps of 23.97.

In the jumbled condition, the video clips were cut into 27 pieces and randomly rearranged to produce clips that contained the same visual information as the original. The clips ranged from 7 to 12 seconds.

For the subsequent recognition test, 50 individual frames were selected from both the experimental and foil video clips yielding 100 still images.

All clips and still images were presented on a 19 inch LCD monitor using a 640x480 resolution.

Procedure

Participants were randomly assigned to one of two conditions, linear and jumbled. In the linear condition the video clips were shown in the temporal order in which they were originally produced. In the jumbled condition, the video clips were sliced into 27 pieces and ordered randomly. In each condition, participants were shown a 5 minute video clip immediately followed by 100 still images. Participants viewed the clip, with the included audio track, in a darkened room. The order of the images was randomized

for each subject. Participants gave old/new judgments for each still image by pressing the "o" or "n" key on a keyboard. They did not receive any feedback. Images remained on the screen until a judgment was given. Following the recognition test, the procedure was repeated for the second and third video clip in an identical procedure. The order of the video clips and their respective recognition tests were randomized. After the subjects completed the recognition test, they filled out a short survey assessing familiarity with the films shown.

Results

Figure 1 shows the accuracy for each video, both in the linear and the jumbled condition. Overall, the mean accuracy (the average of hits and correct rejections) for the linear condition was .791 and .775 for the jumbled condition, $F(1, 52) = 0.77$, MSe = .015, $p > .05$. No subject in either condition performed worse than chance (.500). The effect of movie was significant, $F(2, 104) = 34.13$, MSe = .007, $\eta^2 = .396$; the interaction between condition (linear, jumbled) and movie was not significant, $F(2, 104) = 1.17$, $p > .05$. Overall, subjects were most accurate in identifying frames from the Dr. Who clip (.854), worst on the 12 Angry Men movie (.723), and intermediate on the Looney Tunes clip (. 771).

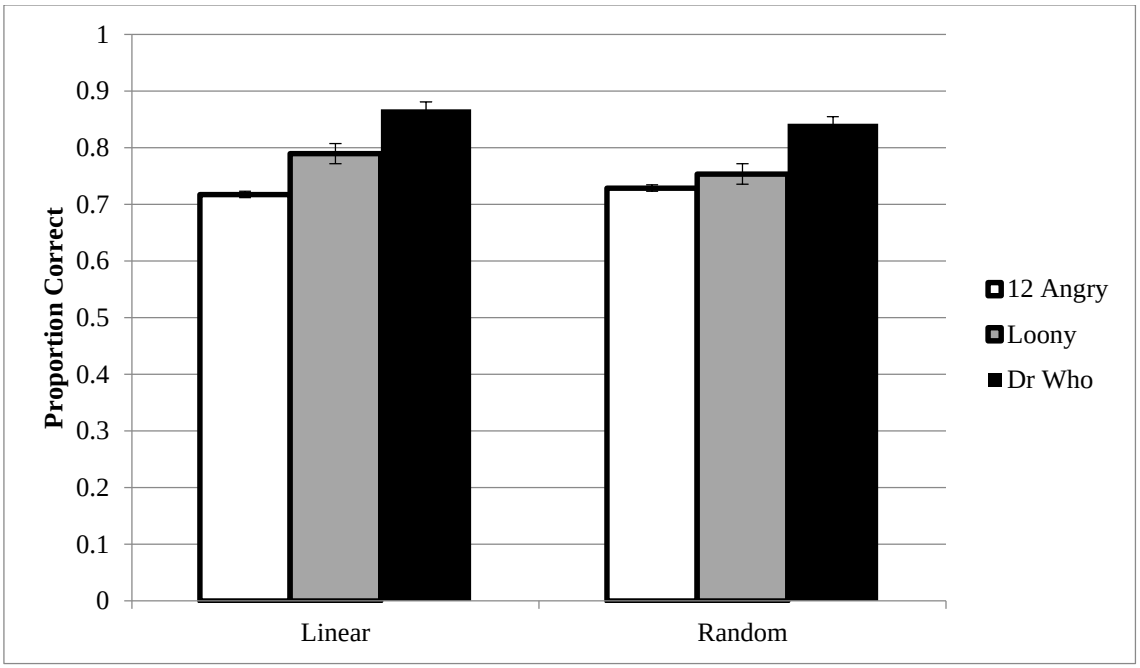


Figure 1. Accuracy for the old/new recognition test in Experiment 1.

When queried at the conclusion of the study, only a few students (< 20%) claimed to have viewed the movie '12 Angry Men', with only slightly greater numbers recognizing the other two clips (20-30%). A number of students were uncertain if they had viewed the particular clips. Regardless, analysis of those who professed to have seen or may have seen the previous clips did not alter accuracy on the later recognition test, $p > .20$.

Experiment 1 Discussion

In general, subjects correctly recognized 78% of the test frames, with minimal effect due to jumbling and only slightly modified by type of movie clip. This level of recognition performance is impressive and is consistent with previously discussed research examining recognition performance for static scenes (Konkle et al. 2010). A similar result was obtained by Zacks Speer, Vettel, and Jacoby (2006) when they had

normal controls and Alzheimer patients watch a movie of a mundane action and perform a recognition task. Zacks et al. had subjects watch a movie of an actor performing a mundane action such as watering a plant or washing a car. In the recognition task, subjects had to differentiate seen images from images from foil videos that depicted the same actor performing the same action on a different item (e.g. a different flower pot or a different car). They found that their normal control group was able to make this distinction at around 86%, which is in line with our own results.

Zacks et al. (2006) also found that participants who had more agreement with the group about where to segment the videos (in the same process as Newtson & Enquist, 1976) had higher recognition scores. However, this result is complicated by the fact that there did not seem to be any relation to the correct recognition of any one item and that items distance from a participant's event segmentation. Our own participants did not seem to be affected by the temporal disruption brought on by clip reordering, which had to have at least some detrimental effect on understanding the narrative within each clip. This result appears to discount theories of encoding that relied on a strong schema or structure in order to make judgments at recognition as randomly reordering events within a movie should have at least some disruptive effect on the formation of any sort of framework for what is actually occurring in the movie (Thorndyke, 1977; Rumelhart, 1975).

EXPERIMENT 2

The high recognition accuracy in Experiment 1 could be explained by simply saying that subjects were able to correctly recollect the target frames in the recognition task. An alternative explanation could be that because of the amount of exposure participants had to the scenes from the clips they saw in the learning portion, they were able to easily build a representation that allowed them to easily reject the lure frames. We choose the stimuli and the lure images in an attempt to minimize this, but there is the possibility that the lures were too distinct to actually get a measure of recognition. Nonetheless, we opted to address this concern by removing all of the surrounding video from the frames that appear in the recognition test. In effect, these subjects first viewed the 50 'old' frames in the study phase, and then saw these same 50 frames intermixed with the same foils used previously in Experiment 1 in the test phase. If the test frames functioned as a distinct group, then subjects should be able to discriminate these old from the new test frames.

The 50 test frames for each movie that functioned as test items in experiment 1 functioned as the study set in experiment 2. Again, half the subjects viewed the frames in a linear format (continuous in time) and half viewed the frames in a jumbled order. As was the case in experiment 1, the three movies were presented in a random order.

Method

Participants.

The subjects were 55 Arizona State University undergraduates, selected from the same introductory classes as in Experiment 1. A total of 27 subjects viewed the frames in a linear (correct temporal) order and 28 viewed the same frames in a jumbled order.

Materials and Procedure.

The target images from the recognition test in Experiment 1 were used as learning stimuli in this experiment. Each image was displayed in its native resolution for 83ms (5 refreshes of an LCD running at 60Hz). A black and white static mask was then displayed. An interstimulus interval of approximately 5 seconds separated each frame. At the conclusion of the study phase, the recognition test phase was conducted in a manner identical to that of Experiment 1.

Results

Figure 2 shows the accuracy for the linear and random conditions, separately for each movie. Overall, the mean accuracy for the linear and jumbled conditions was .587 and .600, respectively, a difference that failed to reach significance, $F < 1$, $p > .20$. However, the effect of movie was significant, $F(2, 106) = 21.85$, MSe = .005, $\eta^2 = .292$, $p < .001$, with accuracy lowest on the movie '12 Angry Men' (.539) and higher on the Looney Tunes (.619) and Dr. Who (.622) movie clips. Since chance was .50, significance on any movie would require that the subject correctly recognize (mean of hits and correct rejections) 60 of the 100 test items ($z = 2.00$, z-approximation to a

binomial). For the linear condition, 12 of the 27 subjects had a recognition score of .60 or higher; for the jumbled condition, 14 of 28 did so.

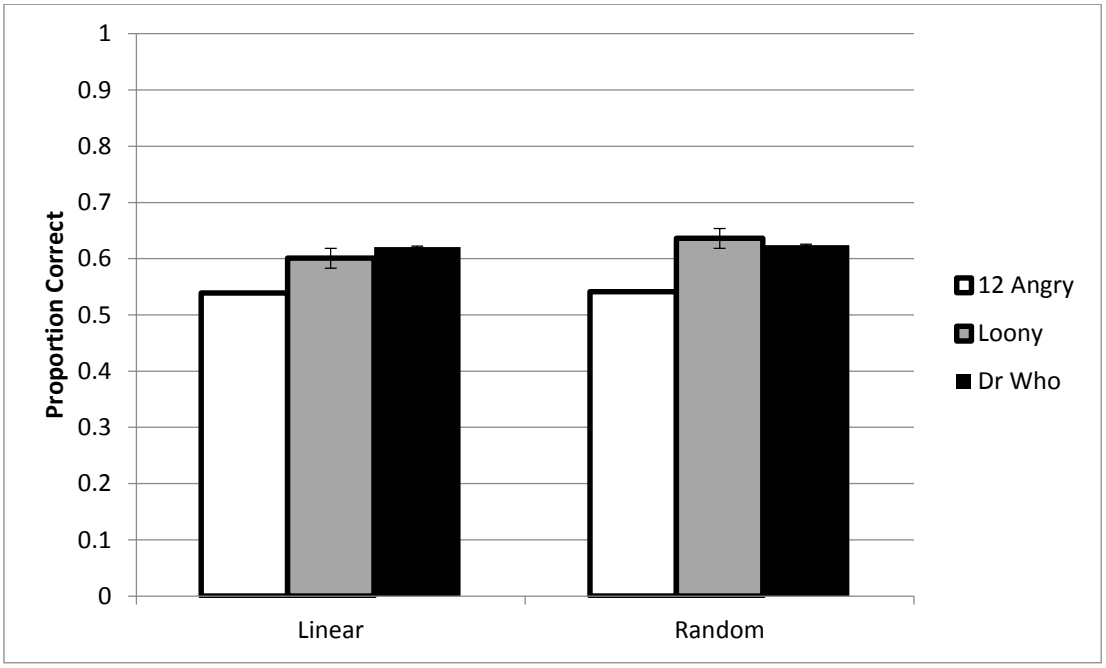


Figure 2. Accuracy for old/new recognition test for Experiment 2.

A comparison of the results from Experiment 1 and 2 is shown in Figure 3 broken down by hits and false alarms. A combined analysis of the two experiments revealed that performance was significantly higher in Experiment 1, $F(1, 105) = 215.52$, MSe = .019, $\eta^2 = .672$, $p < .001$.

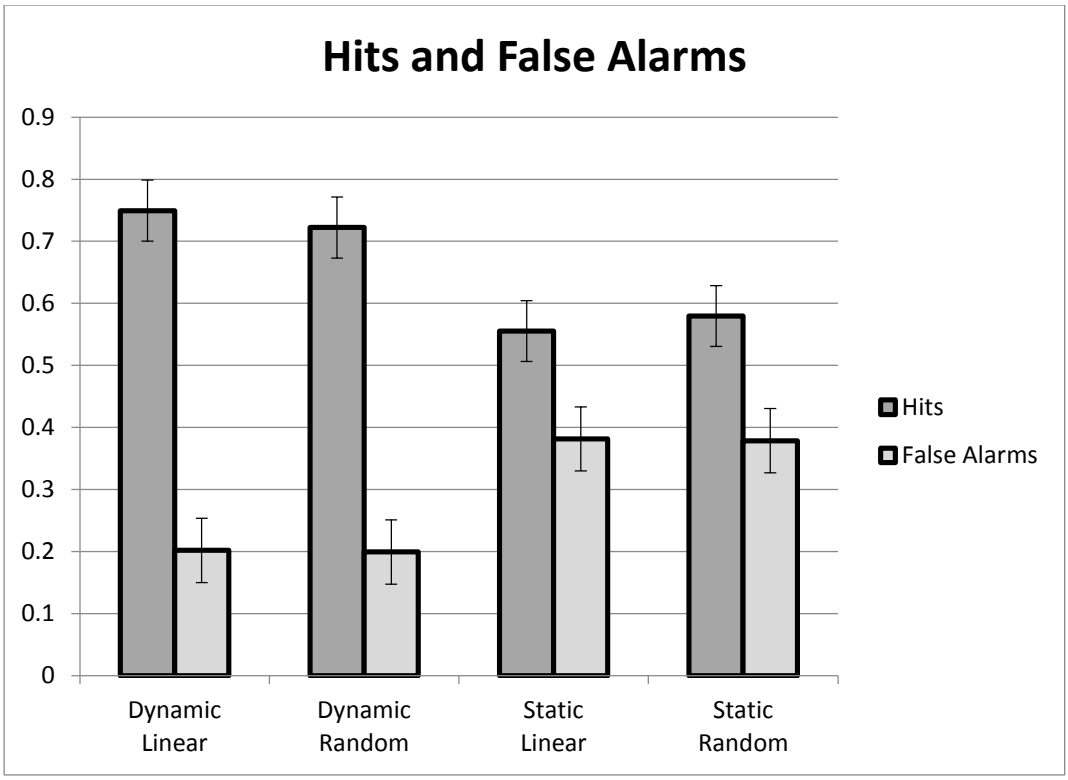


Figure 3. Hits and False alarms for Experiments 1 and 2 collapsed across different videos.

Experiment 2 Discussion

The removal of all non-test frames from the videos in Experiment 2 had a detrimental effect on recognition ability. Participants correctly recognized old frame at a level slightly above chance, which contrasts starkly with the recognition rates we saw in Experiment 1 (around 78% correct). The only difference between the two studies was the presence of the surrounding frames in experiment one. Dynamic presentation appears to be vital to the participant's ability to recognize viewed frames from frames that come from later parts in the film.

The clips that were in color had better recognition rates than those without color. This is consistent with several studies that have shown that color gives an advantage in brief scene categorization (Wichmann, Sharpe, & Gegenfurtner, 2002; Oliva & Schyns,

1997). Other studies have demonstrated accurate scene classification and recognition at exposure durations that that were shorter than our own, however it is difficult to directly compare results because all of our stimuli were highly related where other studies used unrelated scenes (Greene & Oliva, 2009; Grill-Spector & Kanwisher, 2005; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007). Most of these studies describe the level of identification as gist-level or some understanding of global image properties. The constant setting of our stimuli should have helped guide attention towards salient objects or people within the scenes (Itti, Koch, & Niebur, 1998; Chun & Jiang, 1998), but the poor recognition rates suggest that the learning representations did not contain sufficient detail to differentiate from the foils.

It seems paradoxical that recognition for target frames is aided by the inclusion of extraneous visual information (the non-target frames in Experiment 1), and discrimination drops when the extra information is removed. The results from experiment 2 suggests that our visual memory for events relies on a sort of perceptual structure in which a certain amount or duration of exposure is necessary to build a lasting representation. Studies utilizing static scenes have suggested that different qualities about a scene are processed along different timelines. Scene category or gist is one of the quickest qualities derived from brief presentations of scenes (Hollingworth, 2003; Greene & Oliva, 2009). Layout or position of objects is processed later, and may provide a sort of map that can direct attention to areas within the scene that are likely to contain important visual information (Tatler, Gilchrist, & Rusted, 2003; Rensink, 2000). Scene category and layout information is largely invariant in natural scenes, and once they are

established attention can shift to objects within the scene that are harder to predict (Rensink, 2000, but see Hollingworth & Henderson, 2002 for an alternative theory). Dynamic scenes, such as the clips used in Experiment 1, would make establishing layouts of scenes even easier than that in static scenes because movement of an object within a scene provides excellent cues for depth (Rogers & Graham, 1979). The brief exposure times used in Experiment 2 may have created representations which contained little more than the gist of the scene. Because the lures used in the recognition phase were highly similar, the gist representation was unable to discriminate. The lack of discrimination between target and lure items in the recognition test in Experiment 2 suggests that in order to achieve recognition performance as high as was observed in Experiment 1, more detailed representations derived from dynamic scenes were needed. Experiments 3 and 4 investigate the dynamic representation's fidelity by using foils that systematically vary in similarity compared to viewed dynamic events.

EXPERIMENT 3

Experiment 1 showed that participants were able to identify seen frames and new frames from continuous movies with high accuracy regardless of if the movie was presented in a linear or random fashion. Experiment 2 showed that isolated exposure to individual frames was not enough to obtain high recognition performance and that old and new frames did not appear categorically different at test. The question remains, what is the mechanism in which visual recognition is aided by dynamic presentation? There are two general explanations that could be driving the superior performance for dynamic presentation. First, participants could be using the longer exposure to build a detailed representation of the scenes and the objects and actors contained within (Rensink, 2000; Hollingowrth & Henderson, 2002). This bottom-up approach would allow participants to make recognition judgments by comparing probe frames to stored rich representations. Another possibility is that participants are forming more general representations which are more akin to scripts, focusing on sparse event descriptions or goal-oriented actions of characters (Mandler & Johnson, 1977; Thorndike, 1977). The lack of any real difference between the linear and jumbled conditions makes it difficult to argue for a sort of visual representation which is based on a hierarchical narrative structure. However, it could be that the fine-grained events, which are the building blocks of the coarse-grained events, are preserved (Newtson & Engquist, 1976; Zacks, Tversky, & Iyer, 2001). These small-unit events could be a theoretical unit of dynamic visual memory even if they are outside a hierarchical framework.

Although the foils from Experiment 1 were from different parts of the same movie, they may not have been visually distinct enough from the viewed footage to

present much of a challenge. Alternatively, the process of presenting the film in a random order may have hindered the formation of an overall narrative or formation of coarse-grained events, but it may have preserved fine-grained events. These fine event representations may enable a sort of simulation at the recognition test where the participants can ask, "Did I see this happen?/Could this have been a part of what I saw happen?" Instead of pulling foils from before or after the viewed footage, Experiments 3 and 4 pull foils from within the footage. A clip from the movie A Touch of Evil (1956) was chosen because it utilized a series of long continuous shots. Several small gaps (ranging from .5 seconds to 30 seconds) were removed and the film was stitched back together. If the viewer notices the deletion, it appears that the film just skips ahead. Foils in the recognition test were then taken from the removed gaps. These foils are highly similar to the viewed footage, and that similarity approximately varies with the gap size. Frames from the .5 second gaps were similar to the viewed footage, whereas frames from the 30 second gaps were, on average, less similar.

Methods

Participants

One hundred and fifteen Arizona State University students participated in order to fulfil a course requirement. None of the participants had seen A Touch of Evil within the past 3 years.

Materials and Apparatus

A 26 minute clip from the beginning of the Orson Well's cut of A Touch of Evil (1956) was used. Twenty-four gaps were removed consisting of four repetitions of six

different gap sizes:.5 seconds, 1 second, 2.5 seconds, 5 seconds, 15 seconds, and 30 seconds. The gaps were distributed pseudo-randomly ensuring that each gap size was represented within each quadrant of the movie. Gaps were also not allowed to occur over any hard cut (a change in scene location or cut away to a different event). Frames for the recognition test were sampled randomly from any point of the seen footage or any point from the removed gaps.

The movie was presented on 16:10 monitors with black bars along the top and bottom to preserve the original aspect ratio. The original frame rate (23.976 frames per second) was also preserved. Frames from the recognition test were presented as the same size and aspect ratio as they would have appeared in the movie.

Procedure

The experiment proceeded in a similar way to that in Experiments 1&2. The footage was cut in half in order to maintain the attention of participants and prevent too much decay from frames appearing at the beginning of the clips. Participants were told they were to watch a clip from a movie with several small gaps removed from them. Whenever they detected that the film skipped forward, they were asked to press the spacebar as soon as possible. As soon as the clip was finished, they were presented with a recognition task consisting of 72 items: 36 seen frames and 36 frames from the gaps. They frames were presented individually, and the subjects were asked to identify if the frames were seen in the footage they just watched, or if the frame was new. The exact procedure repeated with the second half of the footage. In total, 144 recognition judgments, along with reaction times, were obtained from each subject.

Results

Recognition Accuracy

Subject's overall recognition accuracy for test frames was near chance (M = .55, SD = .04). There is a significant difference in recognition accuracy for frames that were present in the displayed video (seen frames-M = .76, SD = .098) and for items that were pulled from the removed gaps of various sizes (new frames- M = .34, SD = .098). As seen in Figure 4, subject's recognition performance seemed to be a function of the gap size form which the test frame was taken. A repeated measures ANOVA shows that accuracy is significantly different among the different gap sizes, $F(6,684) = 270.5$, $p < .001$, $\eta^2 = .7$. Pairwise comparisons between the six different gap sizes and seen items found that all item sources differed significantly from one another ($p < .05$) with the exception of the items from .5 second gaps and 1 second gaps, $p = .48$.

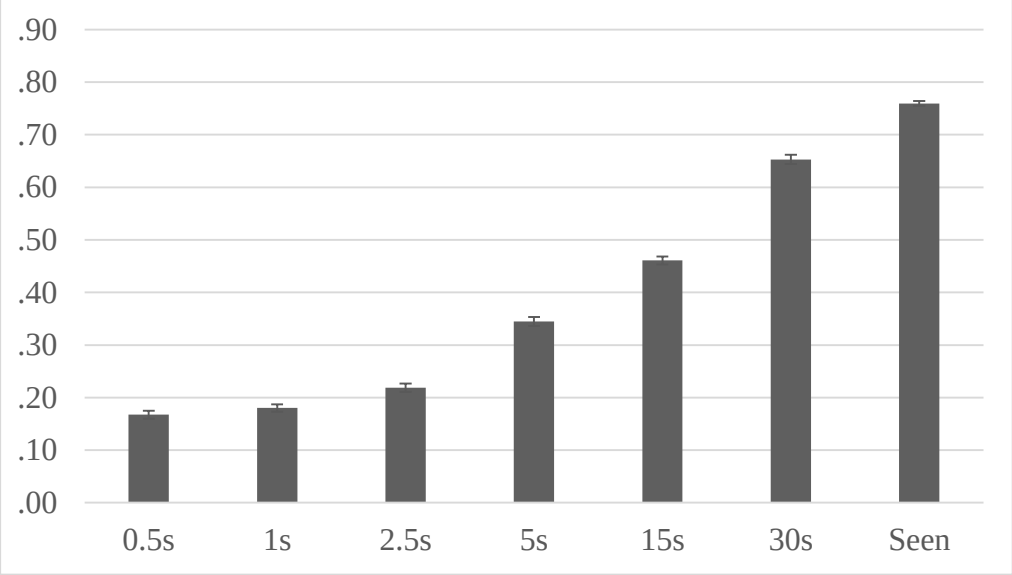


Figure 4. Recognition accuracy for Experiment 3. Performance is broken down by item source, with foil items labeled by their gap size.

Recognition Reaction Time

Reaction time was broken down by response in order to maintain an equivalent number of responses across all types of recognition test items. If we only analyzed reaction times on correct trials, almost half of the trials would be discarded, with some item sources losing over 80% of their responses. Subjects were slower to call recognition test items new (M = 2260ms, SE = 86.1ms) than seen (M = 1844ms, SE = 64.2ms), $F(1, 58) = 47.5$, $p < .001$, $\eta^2 = .45$. The main effect of item source was also significant, $F(6,348) = 3.192$, $p < .01$, $\eta^2 = .05$. There was also a significant interaction between new/old response and recognition item source, $F(6,348) = 3.4$, $p < .01$, $\eta^2 = .06$. This interaction is a result of all item sources being slower identified as new except for those from the 30 second gap, which were called new at statistically similar speeds.

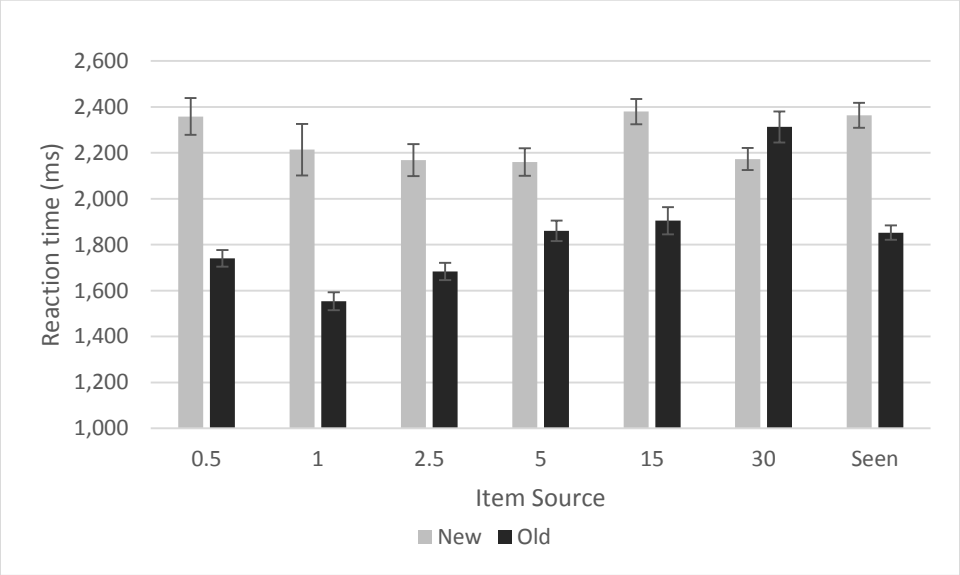


Figure 5. Reaction time for the old/new recognition test in Experiment 3. Recognition test items are broken down by source and response, with foils labeled by their gap size.

Recognition Test item position and accuracy

It is also possible that recognition accuracy is mediated by a test item's position within the gap. Items at the beginning and end of each gap have a high amount of visual

similarity to seen items that also border the gap. Each test item was assigned a quartile in which it appeared within each gap. A repeated measures ANOVA of accuracy data found a significant main effect of quartile, $F(3,340) = 5.1$, $p < .01$, $\eta^2 = .07$. Gap length was also a significant main effect, $F(5,340) = 90.8$, $p < .001$, $\eta^2 = .57$. The quartile x gap length interaction was also significant, $F(15, 1020) = 5.18$, $p < .001$, $\eta^2 = .071$. Figure 6 illustrates the interaction. The smaller gap sizes (.5s, 1s, 2.5s, and 5s) are relatively linear across all quartiles. For the 15s gap size, quartiles 2 and 3 are significantly more accurate than quartiles 1 and 4 ($p < .05$). The 30s gap seems to be driving the interaction, with the last two quartiles being significantly different than the first two ($P < .05$).

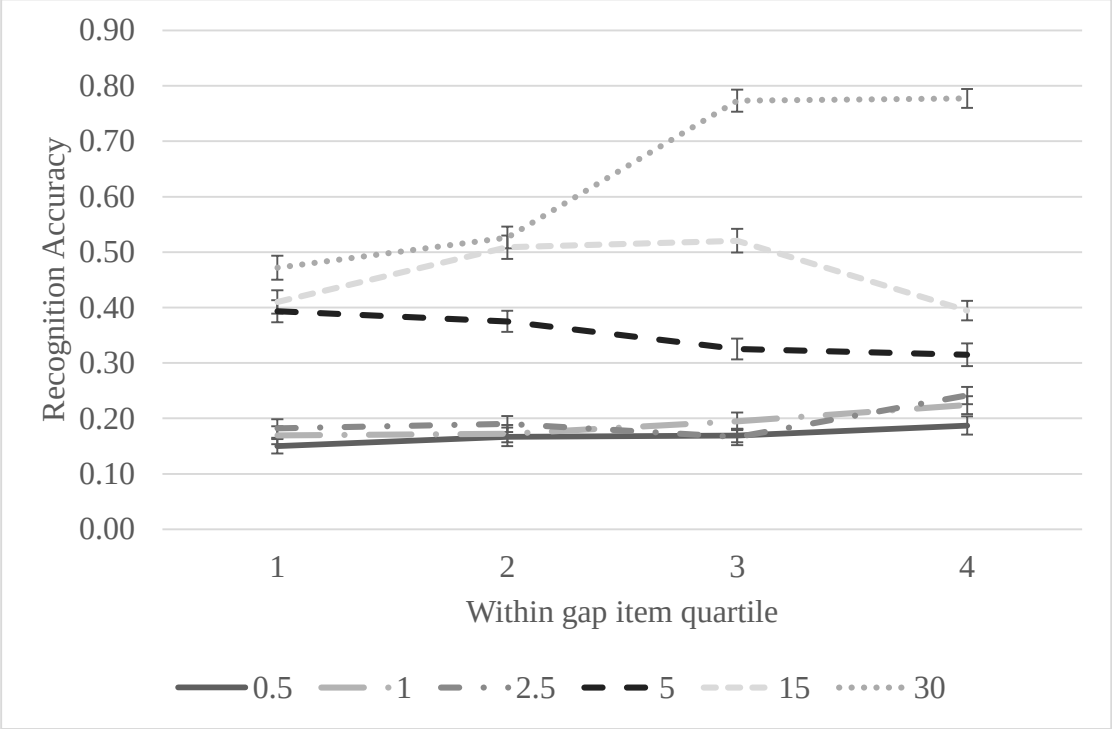


Figure 6. Recognition accuracy for gap items by quartile in Experiment 3. Accuracy displayed for each gap duration broken down by quartile.

Items from the recognition test that were present in the video (seen items) were also analyzed. A binary logistic regression of the distance of a frame from any gap on

recognition accuracy was performed, but frame distance was found to be an insignificant predictor (*B*=0, *p*=ns). Figure 7 shows the accuracy for frames as a function of distance from any previous gap.

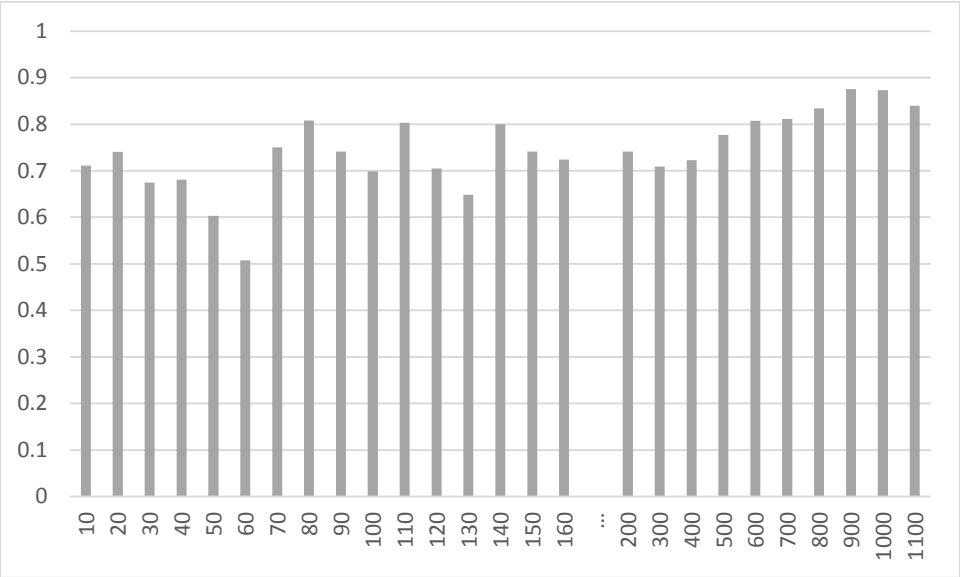


Figure 7. Recognition accuracy for seen frames as a function of distance from gap in Experiment 3. Gap distance is represented as number of frames presented between nearest gap and the seen frame.

Gap Detection

Subjects were asked to press the space bar when they noticed that a part of the movie they were watching had "jumped ahead," identifying when a gap was present. In order to identify a window in which a particular response was coded as a hit or false alarm, a scree plot methodology was adopted where hits and false alarms were plotted for various window sizes. Hits increased and false alarms decreased as the window was increased from 1000ms to 3000ms, where they remained relatively stable. So, 3000ms was chosen as our gap window in order to identify hits and false alarms. Each movie contained 24 removed sections (gaps), so the maximum amount of hits would be 24. There is not an upper limit to the number of false alarms, which proves problematic for

calculating proportions necessary to carrying out signal detection analysis. Regardless, using the 3000ms window, there were an average of 16.3 hits (SD = 4.4) and 10.12 false alarms (SD = 9.65). Gap detection probability differed significantly among gap sizes, $F(1,114) = 35.1$, $p < .01$. The probability that the half second and one second gaps were detected was significantly lower than the rest of the gap sizes. Gap detection percentages for each subject were calculated for each gap size, and the correlation among each of these percentages were significant among one another. This suggests that for each subject, the ability to detect a gap is relatively stable regardless of how large that gap was. However, gap detection of a certain gap size did not correlate with recognition performance for that particular gap size with the exception of the fifteen-second gap, suggesting that gap detection and recognition performance are not coupled.

EXPERIMENT 4

Experiment 3 had strikingly different results from Experiment 1 in terms of recognition performance. Breaking down recognition performance by gap size shows some sort of lawful pattern relating image similarity to recognition probability. Interestingly, images taken from the very short gaps (<2.5s) are called "old' more often than images that were actually present in the viewed footage.

Experiment 3 was different than Experiment 1 in two key ways. First, the foils in Experiment 3 were more similar to the target images within the viewed footage. Second, the foils used in Experiment 3 were heavily implied by the surrounding sequence of events. Although the gaps were noticeable, they did not totally disrupt the narrative the movie portrayed.

By jumbling the movie in the same way it was done in Experiment 1, the effect of perceptual and narrative implication is reduced. Although the same visual information is present in the jumbled version of the movie in Experiment 4 and the linear version in Experiment 3, the experience for the observer is quite different. The jumbled presentation will make developing a precise narrative structure more difficult, if not impossible. This should reduce the narrative implication which may be responsible for the high false alarms in Experiment 3. Also, the detection of the gaps will be rendered moot as the entire movie is full of rough cuts which jump from place to place. The reorganization of the movie preserves the visual detail while damaging potential top-down influences on visual recognition.

Method

Participants

One hundred and thirty-nine Arizona State University students participated in order to fulfil a course requirement. None of the participants had seen A Touch of Evil within the past 3 years.

Materials and Procedure

The movie from Experiment 3 was randomly cut into 138 clips with durations ranging from 6.2-12.7 seconds and an average duration of 9.95s. These clips were then ordered pseudo-randomly as to ensure that no two clips would have a consistent temporal order before and after randomization. The same target and foil frames from Experiment 3 were employed.

The experiment was carried out exactly as it was in Experiment 3 with the exception of the randomized movie. Participants received all of the same instructions as Experiment 3, with an additional statement explaining the randomization procedure. The recognition test portion of the experiment was identical in Experiments 3 and 4.

Results

Recognition Accuracy

Recognition accuracy was very similar to that in Experiment 3. Overall accuracy was poor and close to chance (M = .55, SD = .05). Seen frames (M = .75, SD = .13) were identified with significantly more accuracy than gap frames (M = .35, SD = .12), $t(138) = 20.48$, $p < .05$. Repeated measures ANOVA showed that accuracy was significantly different among the different gap sizes, $F(6,828) = 261.1$, $p < .001$, $\eta^2 = .65$. Pairwise comparisons revealed that all gap sizes were different from one another with the

exception of the .5 second and 1 second gaps, p=.79. All of these results mirror those found in the previous experiment.

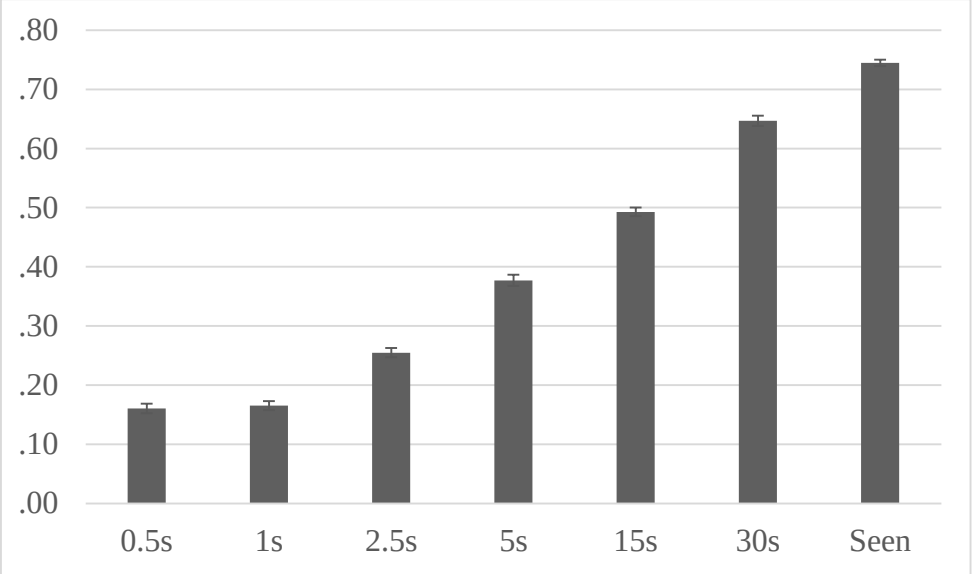


Figure 8. Recognition accuracy for Experiment 4. Performance is broken down by item source, with foil items labeled by their gap size.

Recognition Reaction time

Reaction time was once again broken down by response. Subjects were slower to call recognition test items new (M = 2475ms, SE = 128ms) than seen (M = 1855ms,SE = 80ms), $F(1,52)=47.4$, $p < .001$, $\eta^2 = .48$. The main effect of item gap size was not significant, $F(6,312) = .920$,$p = .34$, $\eta^2 = .017$. There was a significant interaction between item gap size and response, $F(1,52) = 10.67$, $p < .005$, $\eta^2 = .17$. Pairwise comparisons found that seen responses were significantly faster for all gap sources except for 15s and 30s gap items ($p < .05$).

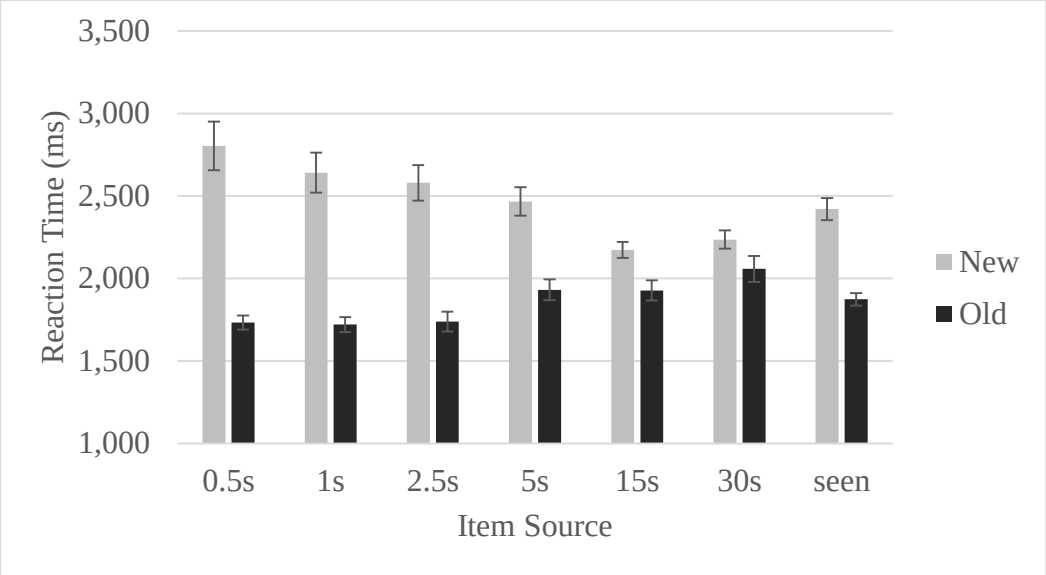
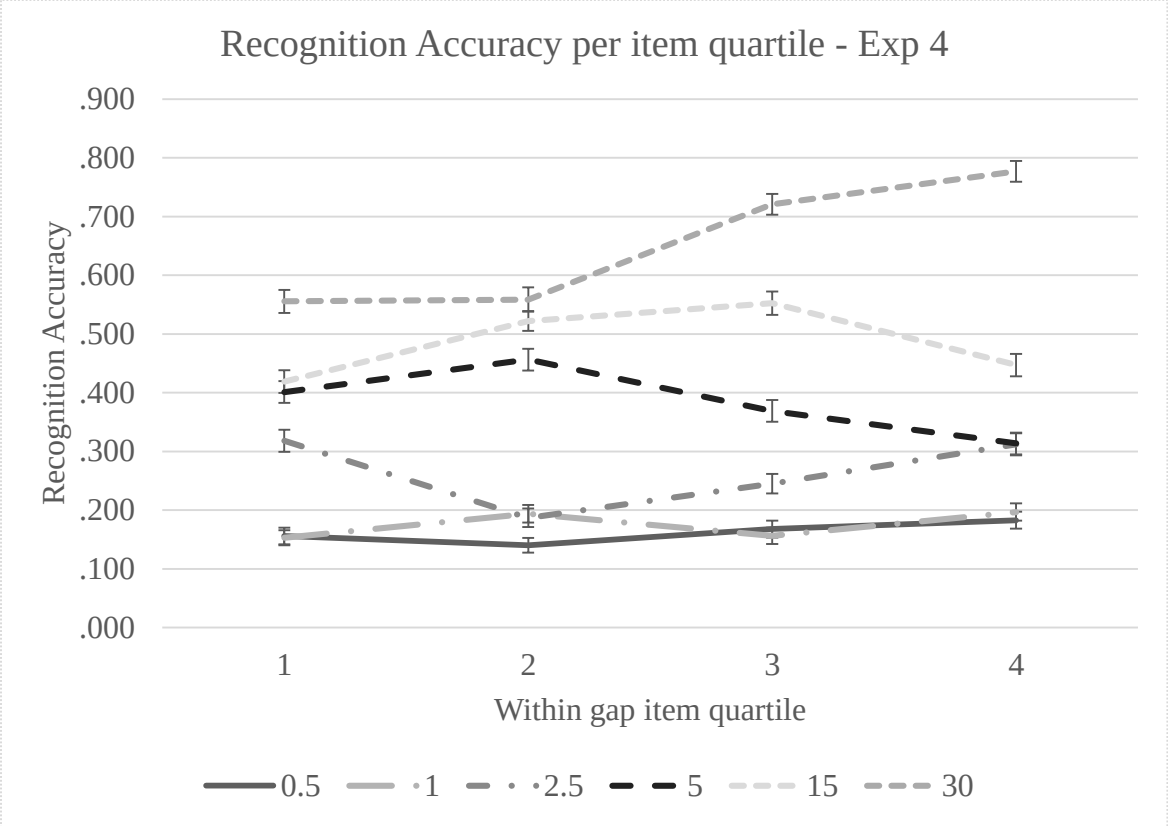


Figure 9. Reaction time for the old/new recognition test in Experiment 4. Recognition test items are broken down by source, with foils labeled by their gap size.

Recognition Accuracy and gap frame position

Recognition accuracy was also examined according to a recognition test item's position within the gap it was pulled from. Each gap was broken up into quartiles and recognition accuracy was assessed using a repeated measures design with item quartile and item gap source as main effects. Trends were much less lawful than they were in the previous experiment. Item quartile was a significant main effect, $F(3,231) = 19$, $p < .001$, $\eta^2 = .197$. Items from the 4th quartile were identified with the highest accuracy ($M = .42$, $SE = .02$) and items from the 1st quartile were identified with the lowest accuracy ($M = .3$, $SE = .02$). The effect of item gap was also significant, $F(5,385) = 18.6$, $p < .001$, $\eta^2 = .34$, as was the interaction, $F(15,1155) = 36.3$, $p < .001$, $\eta^2 = .32$. Figure 10 illustrates the complicated interaction. Figure 10 shows that, much like in Experiment 3, the 30 second gaps drive the overall differences among the quartiles.



Experiment 3 & 4 Discussion

Overall accuracy for Experiment 3 was at 54.8% and for Experiment 4 it was at 54.7%. Performance in terms of accuracy in the recognition tests for both experiments were both strikingly similar and strikingly poor. Experiment 3 and 4 replicate the pattern seen in Experiments 1 and 2 in which the linear or jumbled presentation does not seem to have any effect on the recognition task. The lack of effect for the ordering of the movie goes against many findings from the text narrative comprehension literature that coherent and canonical narrative structure is best learned and recalled (Schwarz & Flammer, 1981; Stein & Nezworski, 1978; Mandler & Johnson,1977). One issue with comparing the current results to those found in the story grammar research is that many of these studies

assess narrative understanding through free recall of events contained within a story, which may be a different processes than that employed by recognition tests (Brown, 1976). Yarkoni et al. (2008) had subjects read either a normal or scrambled version of a story followed by a recognition test containing sentences that were previously read or similar foils. Although their methodology is more analogous to that in the current studies, they found a significant decrease in recognition performance in the scrambled groups where we observed none. They concluded that sentence recognition was aided by structure, but that does not appear to be the case for the visual recognition task in the current study.

The addition of visual feedback is obviously the difference between the current task and similar narrative comprehension studies that found a difference for jumbled presentation. There are two possibilities for these results. The first is that the visual information in the presented visual narrative somehow bootstraps the process of hierarchical representation so well that participants in the jumbled condition are essentially able to reconstruct what is happening in the films. This process would have to be so efficient that the representations for the normal and jumbled conditions are equivalent at the recognition test. This is unlikely because there should have been some performance cost observed for the jumbled conditions if they were actively reconstructing at recognition or encoding. Instead, we see performance levels on recognition in Experiment 3 and 4 that are essentially equal. The second possibility is that participants are not constructing any sort of higher order representation that takes into

account narrative structure, or are at least not utilizing that structure at all when making recognition judgments.

While overall accuracy was at around chance, there seems to be a lawful pattern that emerges whenever the foil items are broken down into the gap sizes from which they originate. In both experiments, accuracy correlates strongly with gap size. In both Experiments 3 & 4, accuracy for foils from the 30 second gaps was at 65%, while accuracy for the .5 and 1 second gaps is at less than 20% in both studies. In fact, items from the .5 second and 1 second gaps are called 'old' or seen more often than items that actually were seen in the viewing portion of the task. This false recognition effect has appeared in both of the reported studies here and in two separate pilots which employed different videos. While the gaps are referred to by their duration in comparison to an uncut version of the film, to the participants they are all just blips in the visual stream-if they are even detected at all. What the gaps really represent are different bins of variation from viewed footage. All of the frames within the smaller gaps are more similar to both themselves and to viewed footage. The within gap variance and the average similarity to seen images decreases as the gap sizes get larger. One potential explanation of the accuracy data is that recognition judgments are based on visual similarity alone. On average, the recognition test items from the 30 second gaps are less similar to seen images, so it is easier to recognize them as new. Although this explanation certainly accounts for some of patterns in the data, it fails to account for others. A pure similarity explanation would predict that new items that are near the edges of the gaps would be recognized with more accuracy than items in the middle of the gaps. As shown in Figure

6 and Figure 10, this is not the case. Any trend in item quadrant's predictive ability for accuracy is driven by the 30 second gap. The rest of the trends remain flat across all quadrants. Table 1 shows the accuracy for items within 20 frames of the beginning or end of gap, collapsed across all subjects. Because gap items were randomly sampled across the entirety of the gap, there are a greatly reduced number of observations in the larger gap sizes. Additionally, because the smaller gaps have less than 40 frames, all of them are included. Large amounts of alpha inflation as a result of the number of observations for the smaller gap sizes, coupled with the wildly uneven number of observations between gap sizes suggests these numbers should be approached with caution. Regardless, the general trends of the whole gap accuracy are reflected in the accuracy for the items within 20 frames. If recognition accuracy was only a function of visual similarity, there would be no large difference between the accuracy for the 5-second-within-20-frames items versus the 30-second-within-20-frames items.

Table 1. Overall and near edge recognition performance for Exp. 3 & 4.Overall accuracy and accuracy for foil frames within 20 frames of a gap in Experiment 3 and 4.

		Less than 20 frames from an edge			Whole Gap		
	Gap Size	Accuracy	SE	N	Accuracy	SE	N
Linear	0.5	.167	.011	1380	.167	.012	1380
	1	.180	.011	1380	.180	.012	1380
	2.5	.230	.013	912	.219	.012	1380
	5	.327	.019	437	.345	.012	1380
	15	.462	.034	145	.461	.012	1380
	30	.544	.049	68	.653	.012	1380
Random	0.5	.158	.010	1662	.158	.011	1662
	1	.162	.010	1662	.162	.011	1662
	2.5	.269	.012	1063	.253	.011	1662
	5	.347	.017	559	.376	.011	1662
	15	.425	.030	186	.491	.011	1662
	30	.557	.043	88	.646	.011	1662

Distance from a gap also did not seem to affect judgments for seen items, as seen in Figure 7. This is somewhat surprising, especially considering how dynamic events are said to be segmented. Event Segmentation Theory (EST) suggests that ongoing events are monitored and a prediction model is formed (Zacks et al, 2007). As the events deviate from the model, errors begin to accrue and eventually the model is reset, resulting in an event boundary. When a new model is formed, the scene is examined resulting in increased processing of scene details. The sudden termination of action that would occur when a gap occurred should have been surprising enough to trigger a new boundary, but we saw no evidence of increased processing near a gap. It could be the case that EST's prediction of increased processing could be focusing on more semantic forms of information within the scene, and not reflected in our visual recognition test.

Participants were faster to judge items as seen regardless of the recognition item's origin. In general (although not significant in Experiment 4), participants were fastest to incorrectly judge items from the .5 second gaps as seen. Reaction times for incorrect judgments for the gap items increased with the gap size, which may suggest that there is a search function which is sensitive to item variance. New judgments were reliably slower regardless of item origin. In both Experiment 3 and Experiment 4, the smallest difference between reaction times for old and new judgments was for items from the 30 second gaps, and the largest difference was for items from the smaller gaps (.5 and 1 second). There are two main general trends that emerge from examining the reaction time in Experiments 3 and 4. First, reaction time was usually quicker to call an item old rather than new. This is unsurprising given that the majority of memory models that can predict reaction times give the 'negative' response (this frame was NOT seen before) as the default (Sternberg, 1966; Ratcliff, 1978; Hintzman, 1988). The second general trend present in the reaction time data is that reaction times for 'old' responses were shorter for the smaller gap sizes than they were for the longer gap sizes. Keeping in mind that for all of the gap items an 'old' response is incorrect, it is interesting to note that as gap size (and dissimilarity and foil variance) increase, so does reaction time.

General Discussion

It is difficult to gaze within our minds and examine the structure and ingredients of our own memories. To most people, the process of determining if a television show is a rerun or not is a quick and automatic judgment. To a psychologist, this act is much more complicated and most likely involves a complex interaction of visual memory and

various top down factors. The experiments reported here sought to examine the form (which itself is an echo of function) that visual memories for events take. Experiment 1 showed that recognition memory for images from movies was good (around 78%), but most interestingly, recognition memory for these movies seems unaffected by jumbling up the presentation of the movies. Experiment 2 showed that visual recognition declines sharply when non-target frames are dropped, which suggests that the encoding or retrieval process need some sort of visual scaffolding in order to form a good memory. Experiments 3 and 4 employed the use of much more difficult foils, and found that recognition performance like that seen in Experiments 1 and 2 quickly broke down for foils taken from very small gaps within the viewed movie. This performance degradation seemed to be a function of the size of the gap from which the foils were taken, and in both Experiments 3 and 4, items taken from the very smallest gap sizes were called 'old' more often than the items the participant actually observed.

Considering the wide range of performance we observed, it is questionable that any one model of recognition memory can describe our results. Broadly, most models of recognition assume that in a recognition test a probe item is compared to items in memory and a decision through the contributions of familiarity and recollection, or that of familiarity alone. A recognition judgment based on recollection arises from a match between a test probe and an item in memory. This comparison is a serial process because items in the memory set are compared one at a time to the probe item. While recollection based recognition is slow, it is also characterized by high confidence and high accuracy (Yonelinas, 2001, 1999). Given the terrible accuracy in Experiments 3 and 4, recollection

is probably not a strong contributor to the recognition judgment. This of course does not discount recollection as a contributor to normal recognition judgments or that dual process models are incorrect, only that the recognition judgments here are unlikely to involve the recollection process for the majority of the decisions. If it did, we would see higher accuracy that was more stable across gap sizes. Reduced reliance on recollection has been found when attentional resources or other task demands are manipulated, and it could be the case that our task did not foster memory representations that promoted recollection (Yonelinas, 2002).

On the other hand, recognition judgments based on familiarity are characterized by being a parallel process where activation to a probed item is summed over set of traces/vectors/representations in memory. Single and dual process models of memory utilize this parallel process of familiarity in order to account for quick recognition decisions that arrive early in the decision making process, even if there is a large memory set (Gronlund & Ratcliff, 1989; Hintzman & Curran,1994; Yonelinas, 2002). Familiarity based judgments are often more inaccurate and could offer an explanation for why participants performed better in Experiment 1 than they did in Experiments 3 and 4. However, the lack of relation between the distance from seen items and accuracy demonstrated in Experiment 3 casts doubt on the theory that judgments are made via familiarity via visual similarity. Familiarity is summed over all stored traces in order to account for its general advantage in speed over recollection, and to be able to model results which find similar latencies in old/new decisions (Gillund & Shiffrin, 1984). Familiarity produces a single score that represents the activation of the probe item on the

memory set. If this score exceeds a criterion, the probe item is called ‘old,’ otherwise it is called ‘new.’ For example, Nobel and Shiffrin (2001) examined reaction times in a simple word list recognition paradigm and found equivalent latency distributions for hits/correct rejections (‘old’/’new’) and false alarms/misses (‘old’/’new’). This was consistent with performance based on familiarity, which would output a recognition decision following the summation of activations to stored memory items. However, in our own data, reaction times for correct and incorrect judgments in Experiments 3 and 4 (Figure 5 and Figure 9) display a different pattern. Keeping in mind the terrible accuracy, response latencies of the positive responses (‘old,’ meaning they recognize it) appear to be a function of the gap size from which the lures were taken. Also, negative responses generally had longer response latency than positive responses, which is a feature that is normally found in models involving a serial search.

In order for recognition memory to account for our findings in terms of accuracy and response latency, we would need to borrow and rearrange the predictions of the processes which underlie recognition models. We would need the serial search of recollection to account for response latencies, without any of its accuracy. We would need the more abstract familiarity that could produce poor accuracy given highly similar foils, but we would need to ignore its parallel nature. Also, by applying a recognition model to the data from these studies, we are assuming there is a parity in structure between the probe item (a single frame) and the memory representation. Formalized models of recognition which depend on a direct comparison or an activation score assume that the probe item could just as easily be a memory set item. If recognition

memory is unable to describe our current results, the reason might be that the parity assumption has been violated and the performance patterns observed are more reflective of a different memory process.

Performance on a recognition task is not the only measure of how well items presented in a study session are retained in memory. Nobel and Shiffrin (2001) demonstrated this by having participants study a list of word pairs and using different procedures to measure performance. When they presented single words or word pairs at test and had participants indicate if they had seen them before within any of the studied word pairs, they found a pattern of recognition that indicated decisions were made using a measure of familiarity. Specifically, they found responses were fast and did not differ as a function of positive or negative response. However, when they presented participants with word pairs and asked them to judge if they were identical to study or if they were rearranged (associative recognition) or they presented participants with one of half of a studied word pair and had them produce the other (cued-recall), they found a much slower pattern of response. They concluded that while recognition decisions may be based on a single strength of signal response, the associative recognition and cued recall performance indicated that there was a slower sequential memory search occurring. In order to find a match in the cued-recall condition, participants were comparing a single word to the word pairs in memory. Because the single probe word did not have parity with the items in the memory set, items from the memory set were having to be considered one at a time in order to determine a match. The lack of parity between

recognition probe and the memory representation of the movie could indicate that our own study is more like a cued-recall task than a recognition task.

Cued recall is usually modeled as a serial process because response times are often shown to be much longer and more dependent on participant response or set size (Raaijmakers & Shiffrin, 1980, 1981; MacLeod & Nelson, 1984). The Search of Associative Memory (SAM) model illustrates how one such serial process might occur (Raaijmakers & Shiffrin, 1981). Let us consider a cued-recall task similar to the one described above from Nobel and Shiffrin (2001). SAM conceptualizes representations in memory as "images," which contain item and associative information. During the study phase, participants are shown a list of word pairs and each of the word pairs is then encoded and represented as a separate image. When a single word is presented at test, it acts as a cue which activates a number of images that the cue may be a part of. This activation strength is a measure of global activation, and behaves much like familiarity in the context of a recognition test. Because cued-recall is not assuming there to be a direct match in the search set, this measure of activation is not selecting an exact match to the cued word, but rather a set of items to serve as the search set that have the highest probability of containing a match. Individual images within this search set are then randomly sampled. If the cued word is found to exist within a selected image, then that image is recalled. If the selected image does not contain the cued word, then a different image is selected from the set. The process terminates after a certain amount of time or after a certain number of failed retrievals. SAM and other similar models that account for performance on cued-recall would predict a response latency for the negative response

that was consistently higher than that for the positive response, much like we observe in our own study.

If our own results suggest that a direct comparison between test item and memory representation is not occurring, then we must ask what the nature of the memory representation is in our task. We know our cue was a single frame, but what is it cueing in memory? Returning back to the analogy of the televised courtroom drama, we are able to eliminate some possibilities of representation based on our results. Memory for dynamic scenes is not like playing back recorded footage from a video camera. Although performance was good in Experiment 1, poor accuracy in later experiments suggest that visual memory for dynamic scenes cannot be explained solely by what visual information was observed. Likewise, a representation that is more like a script or a newspaper article does not explain our results either. One of the clearest trends with the current data is that jumbling the presentation order of the movies did not have any effect for performance on any of our measures. These data suggests that performance on the recognition test did not rely on a visual event representation that necessitated or utilized a structure such as a story schema in order to encode the visual event. This does not necessarily mean that structure is irrelevant to visual event memory. Effects of the jumbling would probably emerge if the experimental task probed broad plot points or required free recall of events which occurred within the movie. If that were the case, then the lack of effect in the current study could be explained as a levels of processing effect. But it would not change the fact that our task was possible, even with a jumbled presentation, in Experiment 1.

The establishment of a broad story schema does not appear to be a prerequisite to preforming the frame identification task.

We cannot, however, ignore all structure and explain the current results. The dissociation between performance in Experiment 1 and that of Experiments 3 and 4 may be a result of structural violations rather than solely a product of using more similar foils. The results of Experiment 3 showing the lack of relation between where a lure item was drawn and its ability to elicit a false alarm, both in terms of quadrant and absolute location, suggest that more than just similarity is contributing to false alarms. Experiment 2 also suggested that the target and lure frames in Experiment 1 were perceptually similar, yet they were easily identified. Confusability and temporal proximity data used as an indicator of similarity does bring along an implicit suggestion that the representation that the recognition probe is being compared to within memory is itself static or could produce a static representation. This is not necessarily the case. Whatever form the mental representations from the movies are taking, they were well formed in Experiment 1 but not in Experiment 2. Likewise, these representations caused high false alarms in some cases in Experiments 3 and 4, but did not in Experiment 1. If our recognition task was in actuality a cued-recall task, then the pattern of results could be explained by mental representations that are characterized by local, but not global structure. These representations would need to be sensitive enough to the episodic nature of the dynamic scene that they are encoding that they would be able to detect if a cue did not occur within their bounds, such as the lures from outside the movie in Experiment 1.

However, they cannot be so rich of a representation that they are not fooled by the lures from the short gaps in Experiments 3 and 4.

The most likely candidate for such a unit of representation is what Newton and Enquist (1976) termed a fine-grained event. As previously discussed, event segmentation is thought to be an automatic process during perception (Zacks et al, 2007; Zacks et al., 2010). Fine grained events emerge when people are asked to segment dynamic scenes in terms of the smallest natural and meaningful unit of activity, and neuroimaging tests show increased activity at event boundaries even if a participant is not consciously told to segment ongoing dynamic events (Zacks et al., 2010). Fine-grained events are nested within coarse-grained events as measured by the alignment of their event boundaries. These larger coarse-grained events are theorized to be nested within higher-order structures like story grammars and scripts (Zacks, Tversky, & Iyer, 2010). However, event segmentation is usually found using videos with a very simple structure, such as a man washing a car or making a sandwich. By jumbling the footage in our own experiment, we certainly have disrupted the canonical narrative structure that had been found to aid comprehension and recall in written story experiments, but our chunks of approximately 10 seconds most likely still contained many fine-grained events. The temporal length of fine or coarse events depends on their context, with fine-grained events having lengths of between 5-10 seconds with a great deal of variability (Zacks, Tversky, & Iyer, 2001; Zacks et al. 2006).

The findings of all 4 experiments are reconciled if the visual information obtained from encoding a dynamic scene is composed of fine-grained events. For the experiments

that feature dynamic scenes (1,3, and 4), the recognition test behaved more like a cued-recall test because of the lack of parity between the static frame test item and the memory representations of the movie. The static frame acted as a cue which activated a memory set consisting of similar events. Events from the memory set were selected and compared until a match was found based on probability of association, or a termination limit was reached. This resulted in negative responses ('new') having consistently higher response latencies than positive responses. In Experiment 1, none of the events in memory were directly associated with the test probe, leading to high performance in frame classification and low false alarms. In Experiments 3 and 4 there was a greater degree of total activation between the static frames and the events in memory compared to Experiment 1. Any frame, seen or not, that sufficiently activated the stored event from which it was pulled would be called old. If a frame cue came from the middle of a fine grain event, it would result in extremely high levels of false alarms because the representations are at the fine event level. In fact, these false alarms could only really be classified as errors by a privileged observer who can deconstruct an event as a series of frames. To the participant who encoded the fine grained event, a frame coming from a small gap of half a second within the event is equivalent to a seen frame, because the representations are referenced at event level. This explains the lawful (and awful) pattern of recognition accuracy in Experiments 3 and 4. Smaller gap items are more likely to be completely inside fine-grained events that are perfectly well represented. Frames from the longer gaps may be from fine-grained events that exist completely within the gap, and so they are consistently called 'new.' Reaction time variations are explained in a similar

way. A static frame coming from a small gap would activate a smaller memory set consisting of higher activated events compared to an item from a large gap. Positive responses for the smaller gap items would come quicker because of their smaller memory set, and the opposite would be true for the larger gap items.

By conceptualizing the participant's visual memory for dynamic scenes as a set of fine-grained events we are able to unify all of the results from the current study. These results do not demonstrate that the fine-grained events cannot be further reduced or nested within any higher order structure. Representations may be structured within a folk taxonomy in a similar way to cognitive categories. Individual objects can be identified at a superordinate level (e.g., tool), a basic level (hammer), or a subordinate level (ball-peen). Within this taxonomy, the level that is used most often in naming tasks and is also shown to be the level that is first used to categorize an object is the basic level (Rosch et al, 1976). This is because the basic level representation of an object contains the highest cue validity, leading it to be more differentiated from other categories (Rosch, 1978). While this is generally true, depending on the task or expertise, the privileged level of representation can shift to higher or lower levels on the taxonomy. For example, Tanaka and Taylor (1991) found that experts' subordinate-level categories were as differentiated as their basic level categories, and so the subordinate level was used in naming and categorization tasks. In the current study, the fine-grained event was the preferred level of representation.

General Conclusions

The present study investigated the representations for dynamic scenes within memory. These representations were found to contain enough visual detail to discriminate between images that occurred within or outside of a particular viewed movie, suggesting that the representations are visually rich. There did not seem to be any effect of the disruption of the movie's sequences within any of our experiments. While this does not discount canonical structure for episodic memories, it does suggest that such a structure was not necessary in order to encode the visual information. Old/new judgments of individual frames that came from within a viewed movie exhibited a pattern of accuracy and response latency that was not consistent with recognition performance that assumes parity in form between a test cue and an item within memory. Instead, the data suggests that visual information from this task was represented at the event level.

REFERENCES

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79*(2), 97-123. doi:10.1037/h0033773

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language, 39*(3), 371-391. doi:10.1006/jmla.1998.2581

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. New York, NY, US: Cambridge University Press, New York, NY.

Bradfield, A., & McQuiston, D. (2004). When does evidence of eyewitness confidence inflation affect judgments in a criminal trial?. *Law and human behavior, 28*(4), 369-387. doi:10.1023/B:LAHU.0000039331.54147.ff

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 105*(38), 14325-14329. doi:10.1073/pnas.0803390105

Brainerd, C. J., & Reyna, V. F. (1998). When things that were never experienced are easier to "remember" than things that were. *Psychological Science, 9*(6), 484-489. doi:10.1111/1467-9280.00089

Brainerd, C. J., Reyna, V. F., & Kneer, R. (1995). False-recognition reversal: When similarity is distinctive. *Journal of Memory and Language, 34*(2), 157-185. doi:10.1006/jmla.1995.1008

Brown, A. S. (1976). Spontaneous recovery in human learning. *Psychological Bulletin, 83*(2), 321-338. doi:10.1037/0033-2909.83.2.321

Buratto, L. G., Matthews, W. J., & Lamberts, K. (2009). When are moving images remembered better? Study–test congruence and the dynamic superiority effect. *The Quarterly Journal of Experimental Psychology, 62*(10), 1896-1903.

Chun, M. M. (2003). Scene perception and memory. (pp. 79-108). San Diego, CA, US: Academic Press, San Diego, CA. doi:10.1016/S0079-7421(03)01003-X

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36*(1), 28-71.

Clifasefi, S. L., Garry, M., & Loftus, E. (2007). Setting the record (or video camera) straight on memory: The video camera model of memory and other memory myths. (pp. 60-75) Oxford University Press, New York, NY.

Deese, J. (1959). Influence of inter-item associative strength upon immediate free recall. *Psychological Reports, 5,* 305-312. doi:10.2466/PR0.5.3.305-312

Freyd, J. J., & Finke, R. A. (1984). Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 126.

Freyd, J. J., & Johnson, J. Q. (1987). Probing the time course of representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(2), 259.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*(1), 1-67. doi:10.1037/0033-295X.91.1.1

Goldstein, A. G., Chance, J. E., Hoisington, M., & Buescher, K. (1982). Recognition memory for pictures: Dynamic vs. static stimuli. *Bulletin of the Psychonomic Society, 20*(1), 37-40.

Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(5), 846-858. doi:10.1037/0278-7393.15.5.846

Greene, M. R., & Oliva, A. (2009). The briefest of glances The time course of natural scene understanding. *Psychological Science, 20*(4), 464-472.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science, 16*(2), 152-160. doi:10.1111/j.0956-7976.2005.00796.x

Guerin, S. A., Robbins, C. A., Gilmore, A. W., & Schacter, D. L. (2012). Retrieval failure contributes to gist-based false recognition. *Journal of Memory and Language, 66*(1), 68-78. doi:10.1016/j.jml.2011.07.002

Hamani, C., Stone, S., Laxton, A., & Lozano, A. M. (2007). The pedunculopontine nucleus and movement disorders: Anatomy and the role for deep brain stimulation. *Parkinsonism & Related Disorders, 13*, S276-S280. doi:10.1016/S1353-8020(08)70016-6

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*(4), 411-428. doi:10.1037/0033-295X.93.4.411

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528-551. doi:10.1037/0033-295X.95.4.528

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language, 33*(1), 1-18.

Homa, D., & Viera, C. (1988). Long-term memory for pictures under conditions of thematically related foils. *Memory & Cognition, 16*(5), 411-421. doi:10.3758/BF03214221

Hollingworth, A. (2003). Failures of retrieval and comparison constrain change detection in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 29*(2), 388.

Hollingworth, A. (2005). The relationship between online visual representation of a scene and long-term scene memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(3), 396-411. doi:10.1037/0278-7393.31.3.396

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 28*(1), 113.

Intraub, H. (1980). Presentation rate and the representation of briefly glimpsed pictures in memory. *Journal of Experimental Psychology: Human Learning and Memory, 6*(1), 1-12. doi:10.1037/0278-7393.6.1.1

Intraub, H. (1984). Conceptual masking: The effects of subsequent visual events on memory for pictures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 115-125. doi:10.1037/0278-7393.10.1.115

Israel, L., & Schacter, D. L. (1997). Pictorial encoding reduces false recognition of semantic associates. *Psychonomic Bulletin & Review, 4*(4), 577-581.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence, 20*(11), 1254-1259.

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research, 47*(26), 3286-3297. doi:10.1016/j.visres.2007.09.013

Kersten, D. (1987). Predictability and redundancy of natural images. *Journal of the Optical Society of America, A, Optics, Image & Science, 4*(12), 2395-2400

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological Science, 21*(11), 1551-1556. doi:10.1177/0956797610385359

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in cognitive sciences*, *12*(2), 72-79.

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive Science, 4*(2), 195-208. doi:10.1207/s15516709cog0402_4

Loftus, E. F., & Loftus, G. R. (1980). On the permanence of stored information in the human brain. *American Psychologist,35*(5), 409-420. doi:10.1037/0003-066X.35.5.409

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4*(1), 19-31. doi:10.1037/0278-7393.4.1.19

MacLeod, C. M., & Nelson, T. O. (1984). Response latency and response accuracy as measures of memory. *Acta Psychologica,57*(3), 215-235.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive psychology*, *9*(1), 111-151

Matthews, W. J., Benjamin, C., & Osborne, C. (2007). Memory for moving and static images. *Psychonomic Bulletin & Review,14*(5), 989-993. doi:10.3758/BF03194133

Miller, M. B., & Gazzaniga, M. S. (1998). Creating false memories for visual scenes. *Neuropsychologia, 36*(6), 513-520. doi:10.1016/S0028-3932(97)00148-6

Mudd, K., & Govern, J. M. (2004). Conformity to misinformation and time delay negatively affect eyewitness confidence and accuracy. *North American Journal of Psychology, 6*(2), 227-238.

Newtson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology*, *12*(5), 436-450. doi:10.1016/0022-1031(76)90076-7

Newtson, D., Engquist, G. A., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology, 35*(12), 847-862. doi:10.1037/0022-3514.35.12.847

Nobel, P. A., & Shiffrin, R. M. (2001). Retrieval processes in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(2), 384-413. doi:10.1037/0278-7393.27.2.384

Oliva, A. (2005). Gist of the scene. *Neurobiology of attention*, *696*, 64.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34*(1), 72-107.

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.

Penfeild, W (1955). The twenty-ninth Maudsley lecture: the role of the temporal cortex in certain psychical phenomena. *Journal of Mental Science. 101*(424),451–465.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2*(5), 509-522. doi:10.1037/0278-7393.2.5.509

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology,81*(1), 10-15. doi:10.1037/h0027470

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 14, pp. 207–262). New York: Academic Press.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. Psychological Review, *88*(2), 93-134. doi:10.1037/0033-295X.88.2.93

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59-108. doi:10.1037/0033-295X.85.2.59

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual cognition,7*(1-3), 17-42.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological science*, *8*(5), 368-373.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803-814. doi:10.1037/0278-7393.21.4.803

Rogers, B., & Graham, M. (1979). Motion parallax as an independent cue for depth perception. *Perception*, *8*(2), 125-134.

Rumelhart, D.E., 1975. Notes on a Schema for Stories. In: Representation and Understanding: Studies in Cognitive Science, Bobrow, D.G. and A.M. Collins (Eds.), Academic Press, New York, ISBN-10: 0121085503, pp: 211-236.

Schacter, D. L., & Addis, D. R. (2007). The ghosts of past and future: A memory that works by piecing together bits of the past may be better suited to simulating future events than one that is a store of perfect records. *Nature, 445*(7123), 27-27. doi:10.1038/445027a

Schwarz, M. N., & Flammer, A. (1981). Text structure and title—effects on comprehension and recall. *Journal of Verbal Learning & Verbal Behavior, 20*(1), 61-66.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science, 5*(4), 195-200.

Shepard R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning & Verbal Behavior,6*(1), 156-163. doi:10.1016/S0022-5371(67)80067-7

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences, 1*(7), 261-267. doi:10.1016/S1364-6613(97)01080-2

Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in cognitive sciences, 9*(1), 16-20.

Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychological Science*, *16*(3), 184-189.

Smith, R. E., & Hunt, R. R. (1998). Presentation modality affects false memory. *Psychonomic Bulletin & Review, 5*(4), 710-715. doi:10.3758/BF03208850

Standing, L. (1973). Learning 10,000 pictures. *The Quarterly Journal of Experimental Psychology, 25*(2), 207-222. doi:10.1080/14640747308400340

Stein, N. L., & Nezworski, T. (1978). The effects of organization and instructional set on story memory∗. *Discourse Processes*, *1*(2), 177-193.

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153*(3736), 652-654.

Strickland, B., & Keil, F. (2011). Event completion: Event based inferences distort memory in a matter of seconds. *Cognition,121*(3), 409-415. doi:10.1016/j.cognition.2011.04.007

Stroud, John M. (1956). The fine structure of psychological time. (pp. 174-207). New York, NY, US: Free Press, New York, NY.

Subramaniam, S., Biederman, I., & Madigan, S. (2000). Accurate identification but no priming and chance recognition memory for pictures in RSVP sequences. *Visual Cognition, 7*(4), 511-535. doi:10.1080/135062800394630

Tatler, B. W., Gilchrist, I. D., & Rusted, J. (2003). The time course of abstract visual representation. *Perception, 32*(5), 579-592. doi:10.1068/p3396

Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology, 9*(1), 77-110.

VanRullen, R., & Koch, C. (2003). Is perception discrete or continuous? *Trends in Cognitive Sciences, 7*(5), 207-213. doi:10.1016/S1364-6613(03)00095-0

White, C. T., & Harter, M. R. (1969). Intermittency in reaction time and perception, and evoked response correlates of image quality. *Acta Psychologica, Amsterdam, 30*, 368-377. doi:10.1016/0001-6918(69)90060-2

Wichmann, F. A., Sharpe, L. T., & Gegenfurtner, K. R. (2002). The contributions of color to recognition memory for natural scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(3), 509-520. doi:10.1037/0278-7393.28.3.509

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review, 11*(4), 616-641.

Yarkoni, T., Speer, N. K., and Zacks, J. M. (2008). Neural substrates of narrative comprehension and memory.*Neuroimage* 41, 1408–25. doi:10.1016/j.neuroimage.2008.03.062

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*(3), 441-517. doi:10.1006/jmla.2002.2864

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133*(5), 800-832. doi:10.1037/0033-2909.133.5.800

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133*(5), 800-832. doi:/10.1037/0033-2909.133.5.800

Yonelinas, A. P., Dobbins, I., Szymanski, M. D., Dhaliwal, H. S., & King, L. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness and Cognition: An International Journal, 5*(4), 418-441. doi:10.1006/ccog.1996.0026

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., . . . Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience, 4*(6), 651-655. doi:10.1038/88486

Zacks, J. M., Speer, N. K., Swallow, K. M., & Maley, C. J. (2010). The brain's cutting-room floor: Segmentation of narrative cinema. *Frontiers in Human Neuroscience, 4* doi:10.3389/fnhum.2010.00168

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin, 133*(2), 273-293. doi:10.1037/0033-2909.133.2.273

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General, 130*(1), 29-58. doi:10.1037/0096-3445.130.1.29