Predicting episodic memories of movie events

Sarah M. Dowcett

Thesis submitted to the faculty of Emmanuel College in partial fulfillment of the requirements for the degree

Bachelors of Science
in
Biology

Dr. Gabriel Kreiman, Principle Investigator
Dr. Todd Williams, Chair
Dr. Joel Kowit, Course Advisor

Emmanuel College
April 2015
Boston, Massachusetts

Acknowledgements

I would like to thank Dr. Gabriel Kreiman and Hanlin Tang for their support over the course of this study. I have learned invaluable research skills that have helped prepare me for future endeavors. Thank you for being amazing mentors!

Abstract

Episodic memory refers to the recollection of autobiographical events. They are long lasting and allow us to travel back in time and recount specific details. Our choice of what memories to remember is the result of a complex filtering of sensory details. We sought to determine the factors that affect memory recall, and whether repeated exposure to related but different events enhanced memory. We asked 51 subjects to watch an episode of the TV show '24' and complete six sessions of memory recall testing. Subjects' recollections were insensitive to low-level stimulus manipulations, but sensitive to high-level manipulations. While subjects were tested on different segments in every session, the segments were related in content. To determine if this repeated testing affected subjects' recollection, we also asked 44 subjects to watch the same episode of the TV show '24' and complete two sessions of memory recall testing that did not contain either low or high-level stimulus manipulations. The results demonstrate that there are determining factors that affect memory recall, and that repeated testing improved subjects' recall performance.

Table of Contents

Introduction

Before the late 1900's, different forms of memory systems were not widely accepted amongst psychologists, who were the main group of people interested in studying memory (1). Within the last forty years, the notion that episodic memory is distinct from semantic memory has become more widely accepted. Episodic and semantic memory are two information processing systems that selectively receive information from perceptual systems or other cognitive systems, retain various aspects of the information, and can recount specific details that have been retained to other systems (3). Semantic and episodic memory systems, however, differ in terms of the nature of stored information, autobiographical versus cognitive reference, conditions and consequences of retrieval, and their vulnerability to interference, which results in transformation and erasure of stored information. Episodic memory is unique in that it receives and stores information about temporally dated episodes, or events, and temporal-spatial relations amount these events. An experience can be stored in the episodic memory system solely based on its terms of perceptible properties and it is always stored in terms of autobiographical reference to pre-existing contents within the episodic memory system store.

In comparison, semantic memory is the memory system necessary for the use of language. It serves as a mental organization of the knowledge a person possesses about words and other verbal symbols, their meaning and relationships, and rules (3). Semantic memory allows the retrieval of information that was not directly stores in it, and retrieval of information from the system leaves its contents unchanged. The semantic memory system is thought to be much less vulnerable to involuntary transformations and loss of information than episodic memory, and is relatively independent of the episodic memory system in terms of recording and maintaining information.

Episodic memory is a unique and complex system that allows us to mentally travel back in time and recall specific details about an event. The importance of understanding episodic memory has grown drastically over the last forty years, and has since become defined as a major neurocognitive memory system because while it parallels certain aspects of semantic memory, it also has its own special functions and properties. Episodic memories are long lasting which allows us to filter through dense amounts of detail and recall specific events, yet they are more vulnerable than other memory systems to neuronal dysfunction and likely unique to humans. Memories represent the output of a constructive process that selects, filters and interprets incoming inputs. To have episodic memories, our brain uses what is known as episodic "retrieval mode", by activating a combination of cortical and subcortical brain regions. Episodic memory makes mental time travel through subjunctive time possible, past to present, and is thought to embody "re-experiencing" events through autonoetic awareness (5).

The way in which visual information is processed in the brain is well understood, and plays a critical role in what we remember. Among the many functions of vision, object recognition is arguably one of the most important. Visual object recognition is essential for everyday tasks, including reading, navigation, and face recognition, all of which encompass a basic definition of memory. In as little as 150 ms, we can recognize complex shapes and categorize objects and scenes (1). Visual object recognition depends on combining two key properties, visual selectivity, and the robustness of recognition to object transformation (4). Visual selectivity refers to the ability to distinguish between similar objects despite manipulations, such as size, rotation and illumination. Given that an object can cast an infinite number of projections on the retina, our ability to recognize objects in a way that is robust to object transformation, is likely to have played a role in the evolution of the visual system (5).

Many studies have examined memory formation, inadvertently visual object recognition, for individual items such as words, faces, objects or scenes lacking temporal and/or spatial context. In 2007, Vogt et al. researched long-term memory for large numbers of color photographs of doors using a two-alternative forced-choice method. Through their investigation, they found that pictorial memory was very good and that information is maintained in long-term memory. In a study conducted by at the University of Washington, researchers proved that long-term memory is not permanent, therefore explaining why all past experiences cannot be recalled (1). We are proposing that there are certain factors that influence what we are able, and not able, to recall from past experiences, specifically past events.

To understand recall under natural conditions, it is critical to incorporate the temporal context, chronological arrangement of events, and spatial context that lead to episodic events and memories. One approach in this direction has focused on recollection of specific information within narratives. While several efforts have examined memory recollection for real-life events, it is often difficult to systematically study real-life events due to the challenges involved in establishing ground truth, reproducibility, appropriate controls, amount of practice or exposure, and other variables. A unique alternative that extends single item recollection measurements is the study of movies. Movies contain several important aspects of episodic information that are difficult to interpret from single item studies including temporal sequences, spatial and temporal context, affective components and an underlying narrative. Subjects can form vivid and detailed memories for movie events as assessed by cued recall, visual object recognition and meta-memory confidence estimates.

To study the formation of episodic memories under natural conditions, it is necessary to systematically define each event to be explored, and a mechanism to evaluate those

memories. The extent of memory recall versus failure depends on multiple factors including who is tested, e.g. subject's age, what contents are evaluated, e.g. single items versus episodes, the presence of meaning and context, how similar altered items are in comparison to the information to be remembered, when memory is probed, specifically the time since encoding, and how recollection is evaluated, e.g. free recall versus two-alternative forced choice. Free recall tests are commonly used to study memory and involve presenting subjects with a list of items that must be remembered. At the end of the list the subject is asked to recall the items in any order they desire, which gives this type of testing its name. On the contrary, two-alternative forced choice is a psychophysical method developed by Gustav Theodor Fechner for eliciting responses from a person about his or her experiences of a stimulus (3). Two-alternative forced choice is a more controlled method than free recall testing and is a good method when testing choice behaviors.

Our initial study sought to investigate whether the outputs of the complex cognitive selection and interpretation processes that lead to episodic memory formation under natural conditions can be predicted solely on the properties of the stimulus. We examined recollection of 200 short audiovisual segments from movies as a coarse proxy to real-life episodic memory in 51 subjects that performed a two-alternative forced choice task in six sessions. Our two-alternative forced choice task consisted of answering either yes or no. The six sessions took place from fifteen minutes up to one year post initial encoding (initial movie viewing). Subjects' recollections were reproducible within and across individuals, imperfect yet accurate, and insensitive to low-level stimulus manipulations but sensitive to high-level stimulus manipulations. Recognition was similarly high for single frames, even one year post initial encoding. We evaluated whether or not visual, auditory, and emotional content can predict what

subjects do and do not remember from a movie. First, we systematically quantified recollection accuracy for short movie events. Next, we used a semi-manual approach to extract low-level and high-level content properties from each movie event. We demonstrate that removing sound or color, flipping the frames horizontally, occluding 75% of each frame, or reversing the temporal order of the frames affected memorability within and across subjects. We demonstrate that the length of time of the shot viewed affected memorability.

Due to the above average recall performance of subjects, we asked whether or not there was an added effect of repeated testing throughout the six sessions. First, we asked 44 subjects to perform a two-alternative forced choice task in only one recall test, either one week or one month post initial encoding. Second, we increased the number of short audiovisual segments from 200 to 1000, and removed all low and high-level stimulus manipulations. We demonstrate that repeated testing increased memory recall performance by 5.5% for one-week subjects and by 9.2% for one-month subjects.

Materials and Methods

Subjects and Ethics Statement

One hundred and eleven subjects (54 female, 18 to 33 years old, 70 college students or recent graduates, normal vision, no reported color blindness) participated in this study. All tests were performed with the subjects' consent and followed the protocols approved by the Institutional Review Board.

Experiment 1:

Movie presentation and eye tracking

Fifty-seven subjects watched a 42-minute movie, TV series "24", Season 6, Episode 1, in the laboratory. None of the subjects had watched any episode from this TV series before. Subjects were instructed to "sit down, relax and enjoy the movie". During recruitment, subjects were told: "You will be asked questions about the movie in six evaluation sessions". There was no explicit mention about studying or testing memory but it can be assumed that subjects knew that memory was involved by virtue of the fact that they were going to be asked questions about the movie.

Six subjects were excluded from analyses, as 3 of them are researchers in this study and were not considered further to eliminate any potential biases. One subject had a low number of trials, <400, another subject showed significant biases in the responses with >75% "yes" answers, and one of them had low overall performance, <60% overall. None of the conclusions in the study would be altered if these 6 subjects were included in the analyses. All analyses in the text are based on 51 subjects.

The movie was presented on a Sony Multiscan G520 21-inch cathode-ray tube monitor, Sony Corporation, Tokyo, Japan. The movie presentation was controlled by an Apple MacBook Pro computer (Apple Computer, Cupertino, California) using MATLAB software (MathWorks, Natick, Massachusetts) with the Psychophysics Toolbox and Eyelink Toolbox extensions. The movie subtended approximately 7.5x12.5 degrees of visual angle and was presented in color at 30 frames/sec. The audio was delivered via headphones and subjects were allowed to adjust the volume at will. Eye movements were monitored throughout the movie using infrared corneal reflection and pupil location, with nine-point calibration (Eyelink D1000, SR Research, Mississauga, Ontario). There were no "recalibrations" during the movie presentation but accurate calibration was monitored at the end of the movie. Eye tracking data were synchronized to the movie presentation.

Definition of movie shots and content annotation

The sequence of frames during the movie was split into *shots* defined using a computational algorithm to detect sharp transitions, *cuts*, between two consecutive frames. The content of all the movie shots was described using a semi-supervised procedure that included computational annotations and manual annotation by 10 subjects. There was no overlap between the subjects performing the annotations and those who participated in the memory recall experiment. The annotations included "low-level" audio and visual properties: contrast, color content, sound level, and sound frequency spectrum. The annotations also included a series of "high-level" properties.

These "high-level" properties included whether the shot depicted emotional content, whether the shot elicited emotions in the viewer, whether the shot happened indoors or outdoors,

presence or absence of each one of 29 different characters, viewpoint for each character, presence or absence of 13 possible sounds, presence or absence of 20 possible emotions, and the presence or absence of 25 different objects. Although there was a small degree of variability in the annotations, particularly for the more subjective aspects of the shot content such as which emotion a character conveyed in a given shot, overall there was significant consistency in the annotations. We used the mode, majority vote, across different annotators when the annotations disagreed. We only considered properties that appeared in at least 10 shots for analyses.

Control shots

As described, the memory recall testing sessions included shots from Episode 1, the episode that subjects watched, and Episode 2, which had not been watched by the subjects. We chose shots from Episode 1 that had a corresponding shot in Episode 2 that was matched as close as possible in terms of the content annotations for characters and their viewpoints. For every shot shown from Episode 1, there was a trial with a matching shot from Episode 2 containing the same characters and viewpoints. This is an important aspect of the experimental design, because without this matching procedure, the shots in one episode could be completely distinct from those in the other episode and therefore easy to distinguish between.

Memory recall evaluation

In each trial, subjects were presented with a shot from the main movie, Episode 1, or a shot from the control movie, Episode 2. Shots from either episode were shown in pseudo-random order and with equal probability, chance performance equaling 50%. During recall testing, subjects were asked to indicate in a two-alternative forced choice manner whether they

recalled having seen the events in the shot during the movie presentation or not. Responses were provided using a computer mouse.

We refer to the presentation of unaltered shots as the default condition. Additionally, a series of modifications of each shot were introduced during the memory recall sessions: (i) presentation of single frames (randomly chosen from within the test shots); (ii) removal of sound; (iii) horizontal flip of each frame from left to right ; (iv) grayscale presentation; (v) occlusion, by presenting only one quadrant (randomly selected) and covering the other three quadrants with a black occluder; (v) temporal reversal of the frames within the shot. Subjects were instructed to indicate whether they remembered the events depicted in the shot regardless of such transformations. The order of presentation of shots and these manipulations was also pseudo-randomized.

Recall performance was evaluated in six recall testing sessions: Session 1, immediately after watching the movie (referred to as 0 days); Session 2, between 22 and 26 hours after watching the movie (referred to as 1 day); Session 3, between day 6 and day 8 after watching the movie (referred to as 7 days); Session 4, between day 27 and day 33 after watching the movie (referred to as 30 days); Session 5, between 85 and 95 days after watching the movie (referred to as 90 days); Session 6, between 335 and 395 days after watching the movie (referred to as 365 days). Subjects were offered a monetary incentive that grew with the number of sessions in which they participated. Still, not all subjects finished all 6 sessions (average 3.7±1.1 sessions/subject). Only 12 subjects participated in Session 6. Subjects were instructed not to watch any episode of this TV series during the entire testing period of 365 days. All subjects reported compliance with this rule.

In order to evaluate self-consistency, unknowingly to the subjects, a small fraction, 3%, of the shots were repeated at random times during the test. These repeat trials were equally distributed between the main movie and the control. None of the conclusions would be altered if these trials were excluded from the analyses, except of course that we would not be able to report self-consistency. There was no systematic trend in performance when comparing the first presentation of each shot and subsequent repetitions for this small set of 3% of repeated trials.

Data analyses

We computed the total number of "yes" and "no" responses for each subject. With the exception of one subject who was excluded from analyses, discussed above, the proportion of "yes" and "no" responses was close to 50% (50.5±4.9%, mean±SD across subjects). Throughout the manuscript, we summarized performance for each experimental condition by reporting the percentage of trials in which subjects were correct, "percentage correct". We only computed percentages for a given condition if we had a minimum of 20 trials. The first 5 trials in each experimental session were removed from analyses. Throughout the manuscript and unless otherwise stated, statistical analyses are based on a two-sided non-parametric permutation test with Bonferroni correction. When evaluating the degree of consistency, within and across subjects, we compared results against the null hypothesis according to which performance was independent across trials.

Experiment 2

Movie presentation and eye tracking

Forty-four subjects watched a 42-minute movie, TV series "24", Season 6, Episode 1, in the laboratory. None of the subjects had watched any episode from this TV series before. Subjects were instructed to "sit down, relax and enjoy the movie". During recruitment, subjects were told: "You will be asked questions about the movie in six evaluation sessions". There was no explicit mention about studying or testing memory but it can be assumed that subjects knew that memory was involved by virtue of the fact that they were going to be asked questions about the movie.

The movie was presented on a Sony Multiscan G520 21-inch cathode-ray tube monitor, Sony Corporation, Tokyo, Japan. The movie presentation was controlled by an Apple MacBook Pro computer (Apple Computer, Cupertino, California) using MATLAB software (MathWorks, Natick, Massachusetts) with the Psychophysics Toolbox and Eyelink Toolbox extensions. The movie subtended approximately 7.5x12.5 degrees of visual angle and was presented in color at 30 frames/sec. The audio was delivered via headphones and subjects were allowed to adjust the volume at will.

Definition of movie shots and content annotation

As described in **Experiment 1**.

Control shots

As described in **Experiment 1**.

Memory recall evaluation

In each trial, subjects were presented with a shot from the main movie, Episode 1, or a shot from the control movie, Episode 2. Shots from either episode were shown in pseudo-random order and with equal probability, chance performance equaling 50%. During recall testing, subjects were asked to indicate in a two-alternative forced choice manner whether they recalled having seen the events in the shot during the movie presentation or not. Responses were provided using a computer mouse. Unique to **Experiment 2**, there were no manipulations to any of the shots shown in recall testing session.

Recall performance was evaluated in only one recall testing sessions, either one week after initial encoding (referred to as 7 days), or one month after initial encoding (referred to as one 30 days). Subjects were offered a monetary incentive only for participating in both the initial encoding and recall testing session. In order to evaluate self-consistency, unknowingly to the subjects, the same method as described in **Experiment** 1 was used.

Data Analysis

As described in **Experiment 1**.

Results

Experiment 1: Factors influencing memorability of episodic events: 6 sessions- 200 frames with low and high-level stimulus manipulations

Fifty-one subjects watched a 42-minute movie, a TV series named "24", Season 6, Episode 1 while we monitored their eye movements (**Figure 1A, S1A)**. We sought to quantitatively evaluate whether or not memorability of movie events could be inferred based on the movie content. Subjects' recall of specific episodic content was evaluated in 6 recall testing sessions, conducted 15 minutes to 365 days after they watched the movie, initial encoding. Memorability was evaluated by presenting brief movie shots lasting between 1 and 200 frames (**Figure 1B, Figure S2A-B**). These movie shots were defined as the intervals between cuts denoting large changes between consecutive frames (**Figure S2A-B**). During the recall tests, shots from the movie were intermixed with an equal proportion of control shots from the next episode in the same TV series, Episode 2, which subjects had not watched. These control shots from Episode 2 were matched to those in Episode 1 in terms of visual content (**Figure S2C**). During recall testing subjects had to indicate in a two-alternative forced-choice manner whether or not they had seen the events in each shot during the movie presentation (**Figure 1B-C**). In addition to showing the shots during the recall test exactly the same way in which they were presented during the movie, shots were also presented in a modified format during the recall tests (**Figure S3B-f**). To evaluate how visual, auditory and temporal characteristics of each shot influenced recall performance, shots were modified during recall tests by removing sound or color, flipping the frames horizontally, occluding 75% of each frame, or reversing the temporal order of the frames.

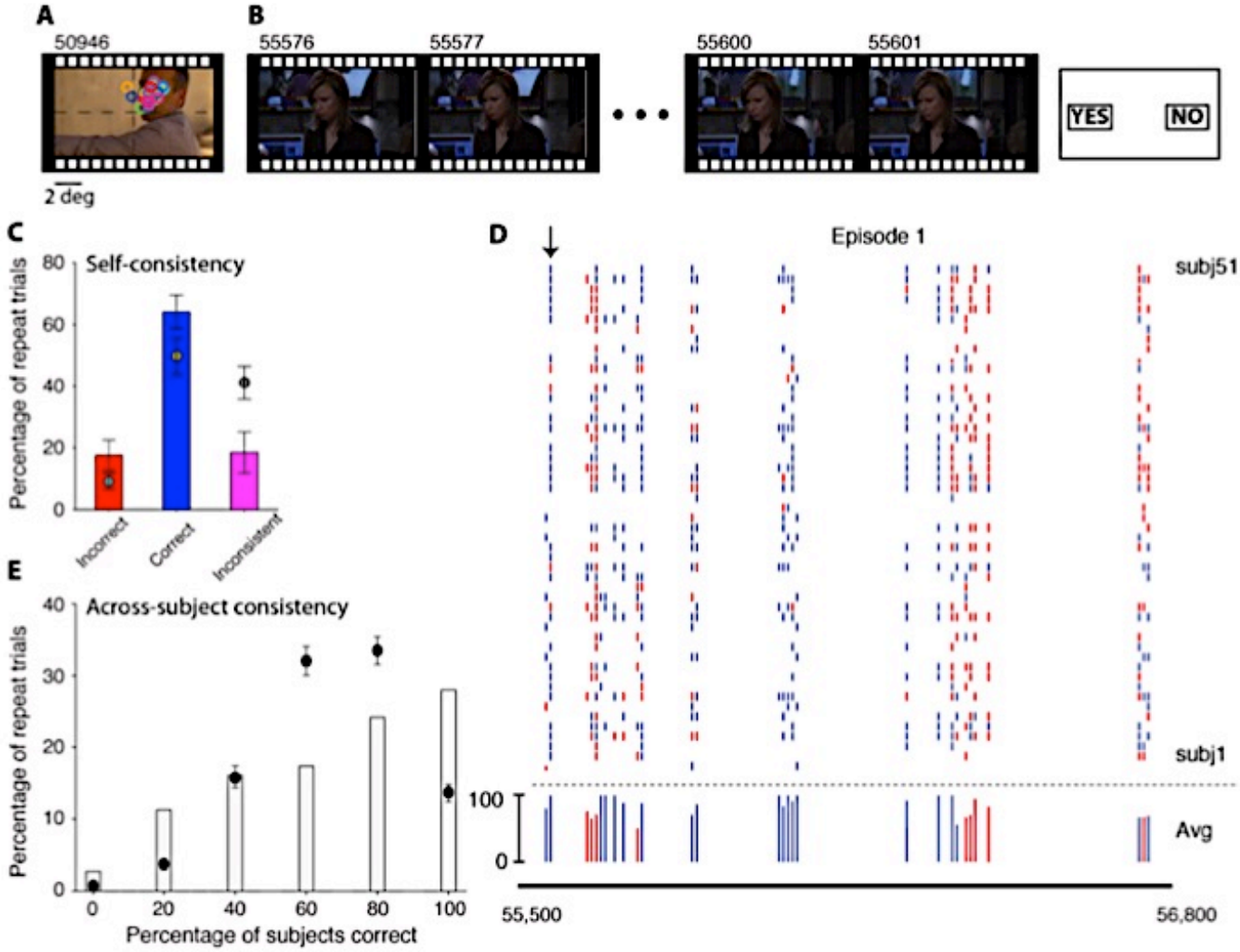

Figure 1. Experimental design and recall consistency

(A) Single frame (frame 50946) from the movie showing eye fixations from 25 subjects. Dashed lines and scale bar were not shown during the movie presentation.

(B) During recall testing, subjects were presented with a single shot (here from frame 55576 to frame 55601, duration = 0.833 seconds) and indicated whether or not they had seen the events in the shot during the movie.

(C) Degree of self-consistency averaged across subjects for movie shots. Error bars denote SD (n=51). The y-axis indicates the proportion of "repeat" trials where the subject was consistency incorrect (red), consistently correct (blue) or responded inconsistently (pink). The circles denote the expected proportions by chance, considering the overall performance of each subject. Self-consistency in repeat trials was significantly different from the values expected by chance in all three cases ($p<10^{-5}$, $p<10^{-8}$, $p<10^{-14}$, respectively, permutation test).

(D) Example raster indicating the performance of 51 subjects for multiple shots from frame 55,500 to frame 56,800. Each vertical mark indicates the subject's response (blue=correct, red=incorrect). Bottom: for each shot, if most subjects were correct, the height of a blue line indicates the percentage of subjects that were correct; if most subjects were incorrect, the height of a red line indicates the percentage of subjects that were incorrect.

(E) Considering random groups of 5 subjects, the y-axis reports the percentage of "repeat" trials where a given fraction of subjects is correct. Circles denote expected value under independence assumption considering performance. The distribution of across–subject consistency values were significantly different from that expected by chance ($p<10^{-5}$, KS test).

We summarize performance during the recall tests by reporting the percentage of trials when subjects were correct, chance level = 50%. On average, across all participants, shot manipulations, sessions, and conditions subjects correctly recalled 69.5±5.6% (mean±SD) of the trials. This performance was well above chance levels (50%) and well below ceiling levels (100%), providing sufficient range to investigate which variables contribute to recall performance. While overall recall of content in shots, lasting several tens to more than a hundred frames (30 frames/sec), could be expected based on everyday experience and previous studies, subjects also performed well above chance levels in trials containing only 1 frame, referred to as single frames, achieving 67.4±5.2% correct. The high performance in correctly recalling single frames is reminiscent of recent studies demonstrating a significant capacity to remember object details. These results extend earlier conclusions to the memorability of individual frames and shots that may be similar, but not identical, across two episodes in a situation where the frames are embedded in complex spatiotemporal context dictated by the movie (e.g. **Figure S2C**).

Recall responses in individual trials were consistent within and between subjects. Subjects responded self-consistently in repeat trials of the same shot (**Figure 1C**). Above chance levels of self-consistency would be expected merely from above chance overall recall performance, in the extreme case, a subject who was 100% correct would always be self-consistent. Subjects, however, were more self-consistent than expected under the null hypothesis of independence after considering the overall performance (**Figure 1C**). There was also strong consistency between subjects (**Figure 1D-E**). Examples of consistently correct and consistently incorrect answers in response to specific shots can be seen in **Figure S4**. Consistency between subjects was evident when comparing each subject to the mode response of all other subjects

(**Figure S6C-D**). There was stronger between-subject consistency than expected under the null hypothesis of independence after considering the overall performance (**Figure S6C-F**).

All subjects performed well above chance and below ceiling across all trials and manipulations, the range of recall performance was 60.9% to 84.2%. There was no significant difference in recall performance between trials from Episode 1 and Episode 2 (p=0.11, **Figure 2A**; in subsequent analyses and unless otherwise stated, data from Episode 1 and Episode 2 were pooled). We investigated whether several different types of shot manipulations during the recall tests (i.e. removal of sound or color, flipping the frames horizontally, occluding 75% of each frame, or reversing the temporal order of the frames) influenced memorability. As expected, recall performance was significantly higher for shots compared to single frames (**Figure 2B**). Removing sound during the recall test impaired performance, but visual information alone was sufficient to drive recall performance well above chance (**Figure 2C**). Reversing the temporal order of the frames in a shot also led to decreased performance, but subjects were still able to determine whether or not they had seen the shot during the movie (**Figure 2D**). In contrast with removing sound or reversing the temporal order of frames, there were two "low-level" manipulations that did not affect recall performance. Flipping the frames horizontally (**Figure 2E**, **2H**) and removing color (**Figure 2F**, **2I**), did not lead to changes in recall performance for either movie shots or single frames. Most notably, however, occluding 75% of the content of each frame led to a significant decrease in recall performance (**Figure 2G**, **2J**). Recall performance for occluded single frames was slightly, but significantly, above chance (58±5%, $p<10^{-4}$ two-sided t-test, **Figure 2J**). Conclusively, being able to see only one quarter of a single frame provided enough information to determine whether or not the corresponding event had been seen.

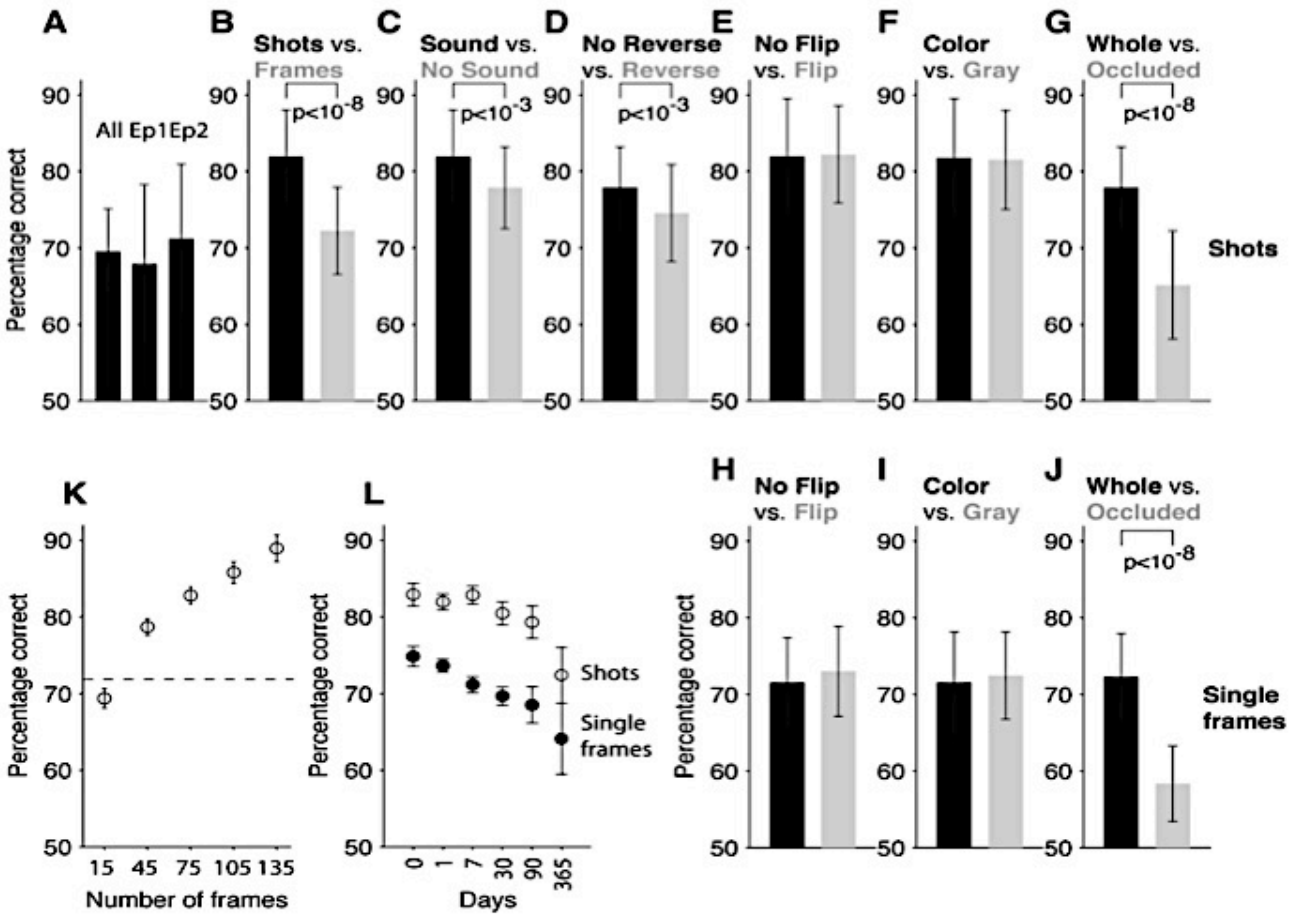

Figure 2. Recall performance was insensitive to low-level stimulus manipulations and sensitive to disruption of the spatiotemporal events

(A) Overall performance including *all* manipulations. There was no significant difference when comparing all trials, Episode 1 trials and Episode 2 trials. In this and subsequent subplots, N=51 subjects, error bars denote SEM across subjects, *p* values indicate permutation test (Materials and Methods).

(B) Recall performance was higher for shots (black, no occlusion, no temporal reversal) versus single frames (gray).

(C) Recall performance was higher for shots including sounds (black) versus shots where sound was removed (gray).

(D) Reversing the temporal order of the frames in a shot (gray) led to decreased recall performance (here shots did not include sounds).

(E) Horizontally flipping the frames in a shot (gray) did not change recall performance.

(F) Removing color from the frames in a shot (gray) did not change recall performance.

(G) Occluding 75% of the frames in a shot (gray) led to decreased recall performance (here shots do not include sounds in both bars).

(H) (Single frames) Horizontally flipping a single frame (gray) did not change recall performance.

(I) Removing color from single frames (gray) did not change recall performance.

(J) Occluding 75% of a single frame (gray) led to decreased recall performance.

(K) Recall performance (shots including sounds, no occlusion, no temporal reversal) increased with the number of frames in the shot. The dashed line shows recall performance in single frames. Bin size = 30 frames; results are shown in the center of each bin.

(L) Recall performance for shots (empty circles) and single frames (filled circles) decreased with time after watching the movie. Note that the scale on the x-axis is not linear in time (test points are shown at equidistant intervals along the x-axis).

We also found that recall performance increased with the length of each shot (**Figure 2K**). Recall performance for very brief shots was slightly below performance for single frames, likely due to the fact that some of those shots contained brief blurry images and camera movements that were difficult to interpret. Performance reached approximately 90% for shots lasting greater than four seconds. Recall performance also showed a significant decrease with the amount of time elapsed between initial encoding and the recall test (**Figure 2L**, $p<0.01$ for both shots and single frames, permutation test). These results are consistent with previous studies of the retention function based on single images, narratives or autobiographical information. Recall performance was above chance for single frames, even when tested one year post initial encoding (64±11%), but this value was only slightly significant, likely due to the small number of subjects who agreed to participate ($p<0.03$, n=12 subjects, two-sided t-test). In summary, the variables that led to an increase in the number of errors that subjects made when recalling specific content from brief shots included distortion of temporal sequences, removal of audio-visual content cues and the amount of time between encoding and testing. The consistency, accuracy and flexibility of recall performance shown here extends previous studies to the realm of spatiotemporal sequences present in movies, and establishes memorability of movie shots as a robust variable that must be explained from the events occurring during encoding.

We next sought to determine which factors of the content in each shot corresponded with successful recall. To do this, we used a semi-supervised procedure to annotate each shot in terms of low-level audio-visual properties (contrast, color content, sound volume, and sound frequency spectrum), high-level audio-visual properties (specific objects, characters, actions, and sounds) and other high-level cognitive properties (e.g. emotional content). An example of these

annotations showing the presence, and viewpoint, of each character across the entire first episode is shown in **Figure S6.** For these analyses, we only considered the default trials without any of the shot manipulations during the recall test (**Figure S3A**) and we also restricted the analyses to the first three recall sessions, up to one week post-encoding. Several of the annotated content properties showed a correlation with successful recall. For example, subjects demonstrated heightened recall of shots containing "action" (92±28% correct) versus shots without action (84±37% correct) (**Figure S7A**, permutation test $p<10^{-4}$). Shot properties that correlated with recall performance included the scene location (**Figure S7D**), the emotional content (**Figure S7G**), shot duration (**Figure S7H**, see also **Figure 2K**) and the presence of specific characters (**Figure S7K**), sounds (**Figure S7L**), emotions (**Figure S7M**) or objects (**Figure S7N**). By contrast, other variables, such as the number of objects, number of characters, or camera movement did not correlate with recall performance (**Figure S7**).

Driven by these correlations, we then asked whether or not it was possible to predict recall performance based on the content of properties. We use a multivariate linear regression that encompassed the average recall performance as a linear combination of the content properties. On average, this multivariate linear regression model accounted for 44% of the variability (**Figure 3A-B**) and was able to account for the degree of memorability in each shot (**Figure 3A**) and single frame (**Figure 3B**) for both Episode 1 and Episode 2.

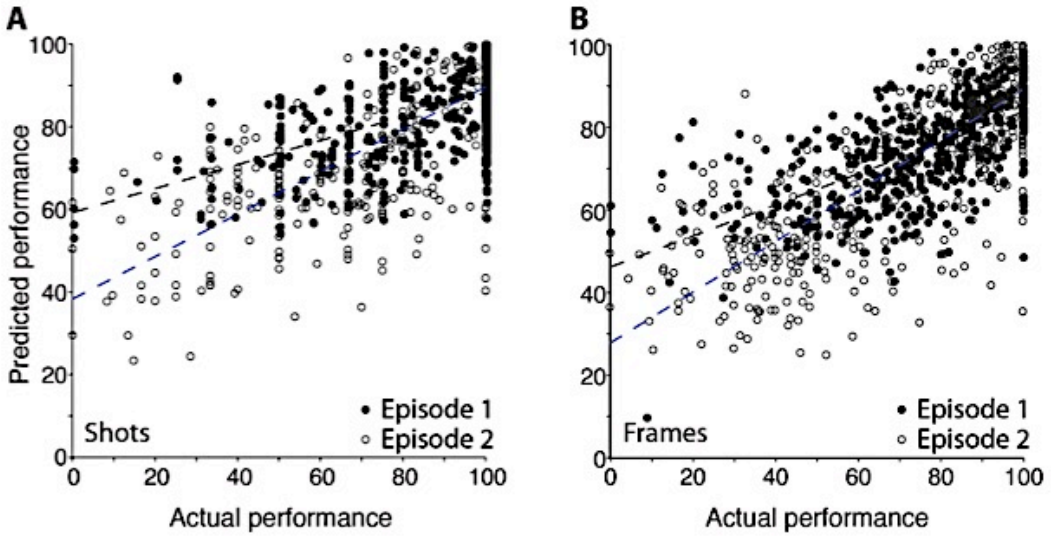

Figure 3. A multivariate linear regression model accounts for a significant fraction of the variance in recall performance. Multivariate linear regression prediction of recall performance (y-axis) against actual recall performance (x-axis, the percentage of subjects that was correct for a given shot) for shots (A) or single frames (B) for Episode 1 (filled circles) or Episode 2 (empty circles). The dashed lines denote the best linear fits: A, ρ=0.54 (Ep1) and 0.72 (Ep2); B, ρ = 0.61 (Ep1) and 0.78 (Ep2).

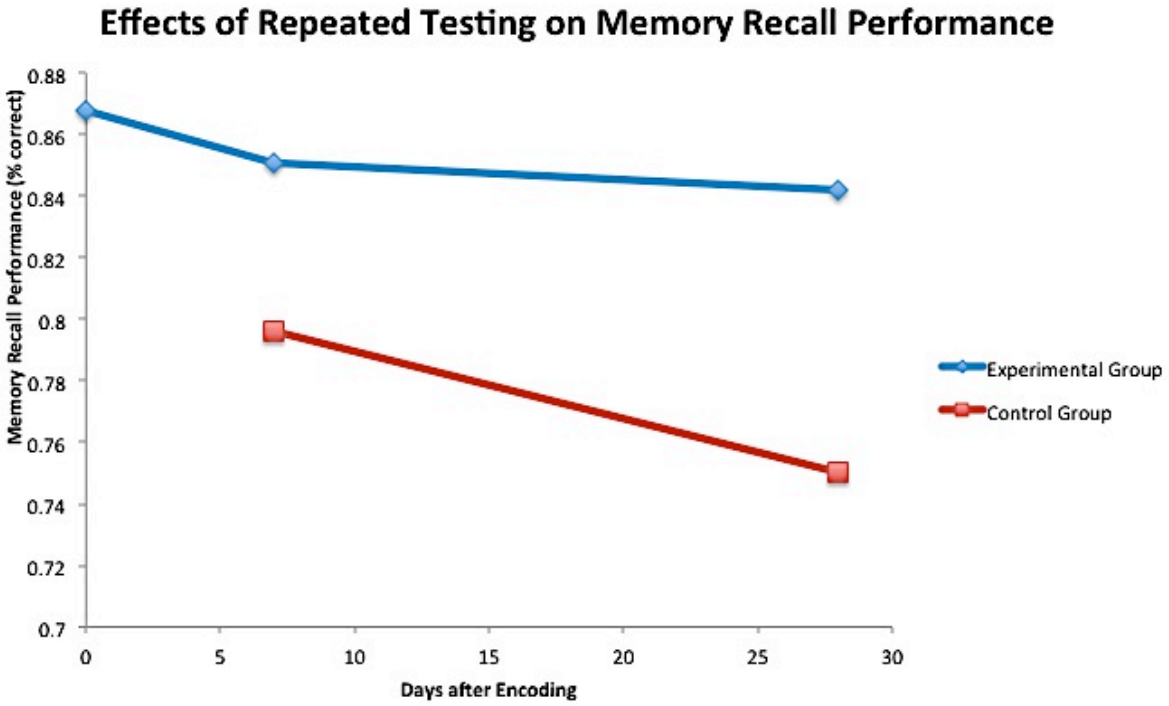

Figure 5. Repeated testing had a positive effect on memory recall performance. Subjects from Experiment 1 and Experiment 2 memory recall performance at one week post initial encoding and one month post initial encoding.

Experiment 2: Effects of Repeated Testing of Recall Performance- 1000 frames with low or high-level stimulus manipulations

The results presented from Experiment 1 describe how memorability of episodic events can be predicted based on content and how manipulations affected successful recall performance. We extended those observations by asking whether or not, when quantifying how much a subject recalls correctly, results are affected by repeated testing. We determined that repeated testing had a significant effect on memory recall performance, and likely explains the high recall success seen in **Experiment 1**.

44 subjects watched a 42-minute movie, a TV series named "24", Season 6, Episode 1. Through the use of two control groups, we sought to quantitatively evaluate whether or not memory recall performance was enhanced due to repeated testing. Subjects' recall of specific episodic content was evaluated in only 1 session, conducted either one week or one month post initial encoding. Memorability was again evaluated by presenting brief movie shots lasting 200 frames, but control groups were presented with 1000 movie shots, rather than 800. These movie shots were defined as the intervals between cuts denoting large changes between consecutive frames. During the recall tests, shots from the movie were again intermixed with an equal proportion of control shots from the next episode in the same TV series, Episode 2, which the subjects had not watched. These control shots from Episode 2 were matched to those in Episode 1 in terms of visual content. Subjects had to indicate in a two-alternative forced-choice manner whether or not they had seen the events in each shot during the movie presentation. Subjects were not evaluated on how visual, auditory, and temporal characteristics of each shot influenced recall performance, rather **Experiment 2** was solely used to determine if the high successful recall performance observed in **Experiment 1** was influenced by repeated testing sessions.

We summarize performance during the recall tests by reporting the percentage of trails when subjects were correct, chance level at 50%. On average, across all participants, subjects recall performance was well above chance, 77.3%±5.7% (mean±SD), for all trails. The performance was well above change levels (50%) and below ceiling levels (100%), providing ample range to investigate whether or not there is an added effect of multiple testing.

Recall response in individual trials were consistent both within and between subjects. Subjects responded self-consistently in single trials. All subjects performed well above chance and below ceiling levels. Across all trials, the range of recall performance was 66.0% to 89.0%. We investigated whether or not there was an effect of repeated testing on recall performance, which influenced memory. As expected, recall performance was significantly lower in subjects who were not exposed to repeated testing than those subjects who completed six sessions of recall testing. Subjects who were only tested one week after initial encoding had overall decreased performance, 79.6%±5.6% ($p<10^{-4}$ two-sided t-test), in contrast to subjects who had also completed an additional recall testing prior to the one week recall testing, 85.1%+/-5.03% (**Figure 5)**. Similarly, recall performance was also significantly lower in subjects who were only tested at one month post-initial encoding, 74.9%±5.7%, in comparison to those subjects who had been tested twice prior to one month recall testing, 84.2%±5.7%. We determined that repeated testing in **Experiment** 1 improved subjects' recall performance by 5.5% during one week testing, and by 9.2% during one month testing.

Supplemental Figures

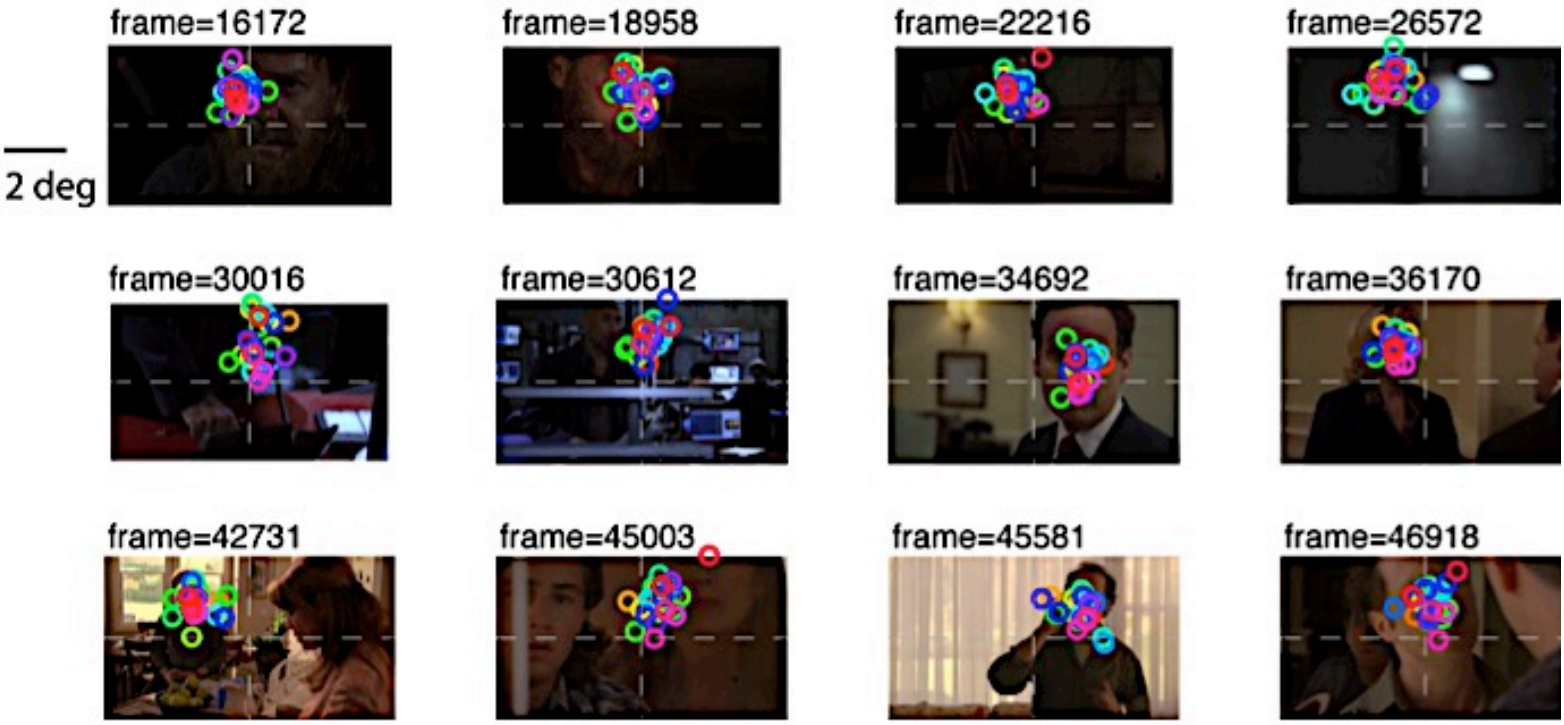

Figure S1. Eye movements of plot-interesting cases.

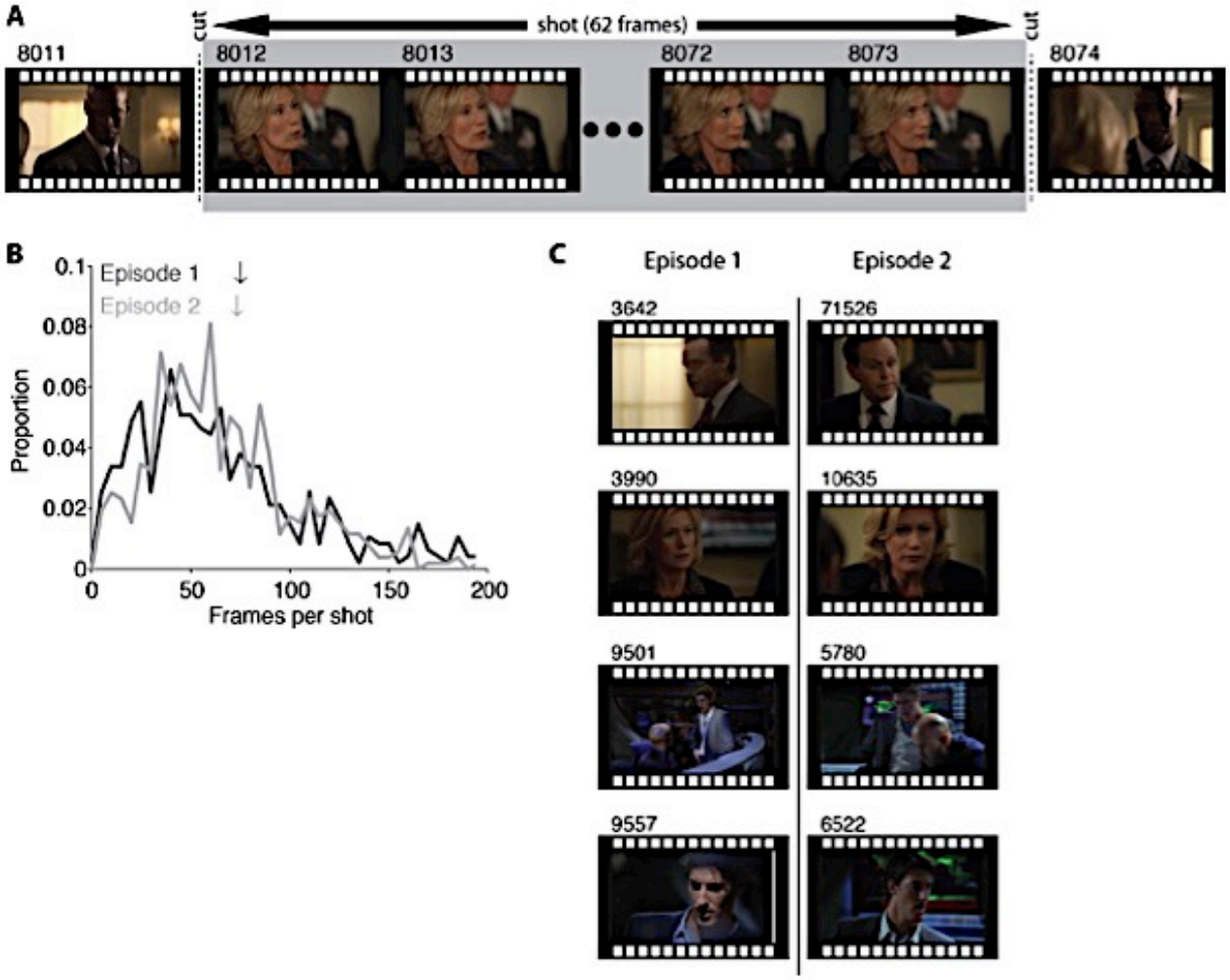

Figure S2. Cut definitions.
(A-B) Intervals between cuts denoting large changes between consecutive frames.
(C) Control shots from Episode 1 and Episode 2 in terms of visual content.

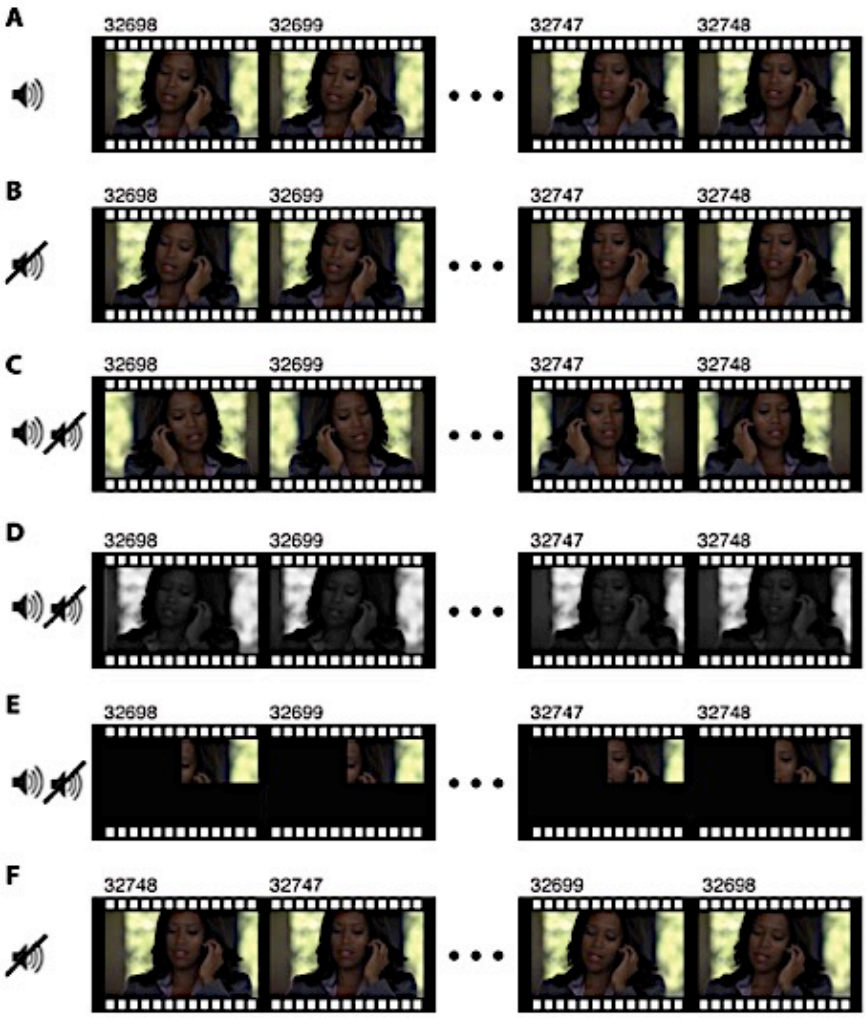

Figure S3. Manipulations.

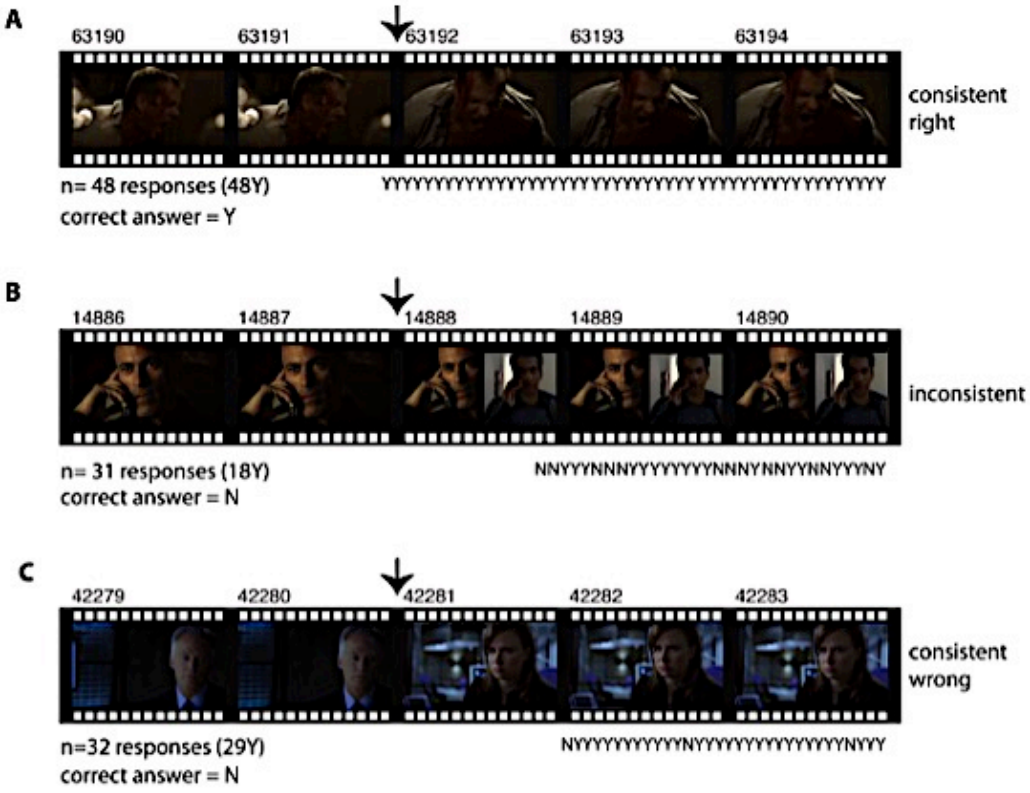

Figure S4. Examples of answers consistently correct and consistently incorrect in response to specific shots.

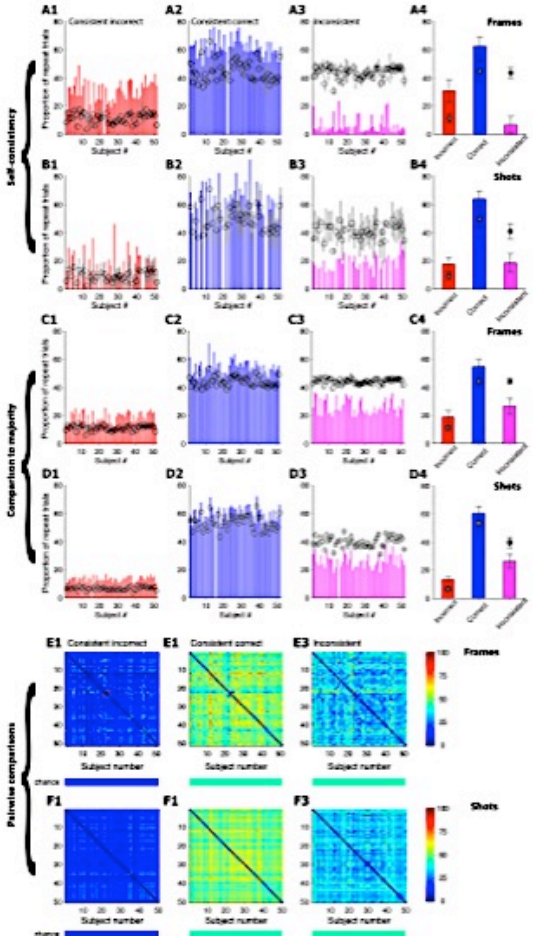

Figure S5. Consistency between subjects.
(A) Call consistency across subjects- **Figure 1**- frames 1
(B) Call consistency across subjects- **Figure 1**
(C) Call consistency across subjects- **Figure 3**- frames 1
(D) Call consistency across subjects- **Figure 3**
(E) Call consistency across subjects- **Figure 4**- frames 1

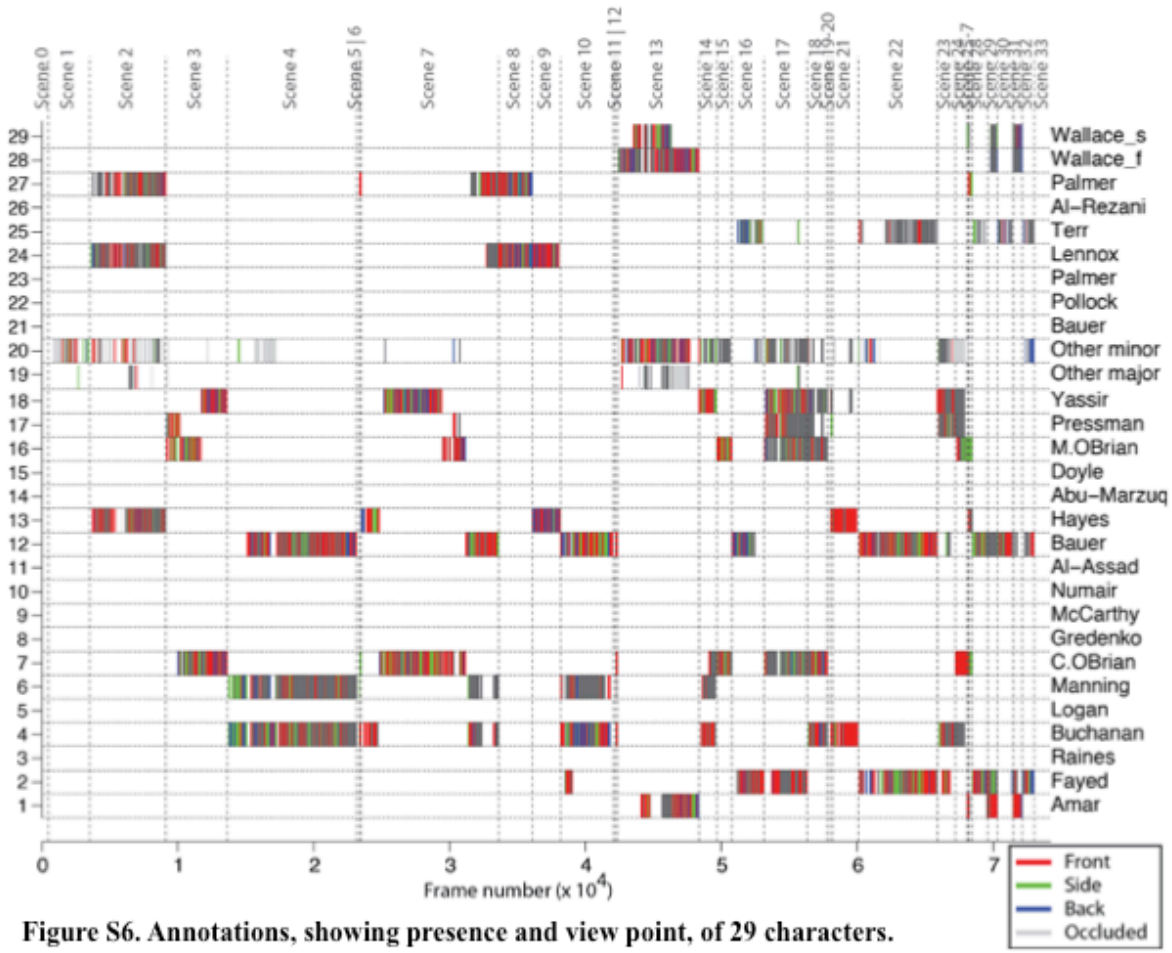

Figure S6. Annotations, showing presence and view point, of 29 characters.

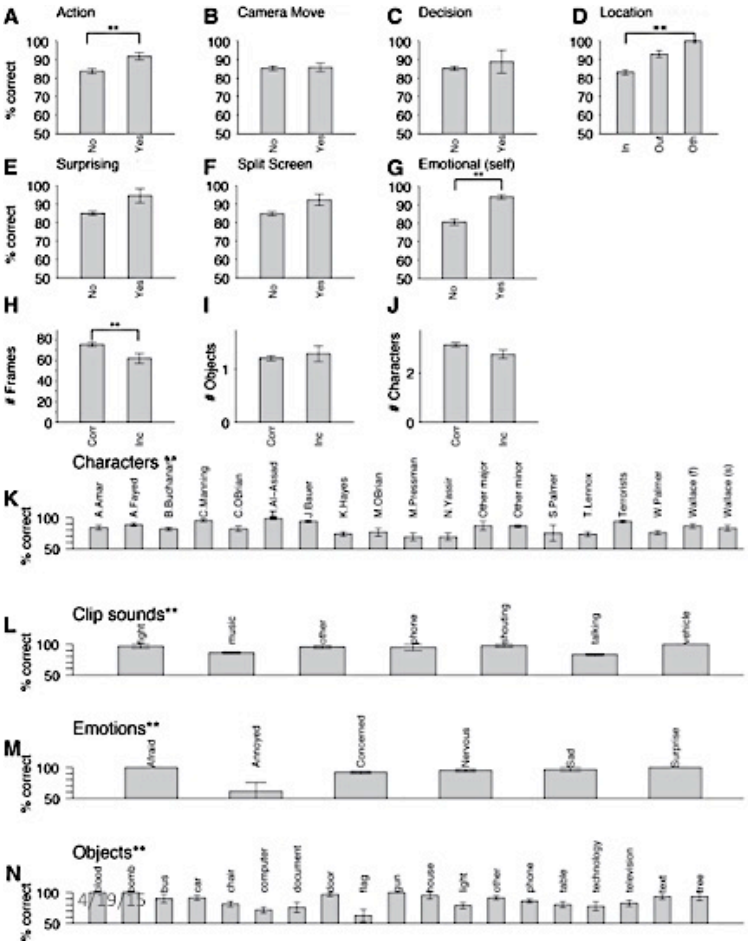

Figure 7. Correlate annotations and recall aggregates of subjects.

Discussion

Movie trials and memory recall testing enabled us to characterize factors that influence episodic memory in human subjects. Due to the brain's mechanism of episodic memory recall it is difficult to detect the pathways that allow for episodic memory. However, our data revealed many factors that influence episodic memory recall. We tested memory recall performance of 51 subjects who completed six memory recall performance assessments. To determine if there is an effect of repeated testing, we tested memory recall performance of 44 subjects who completed one memory recall performance assessment. By applying low and high level stimuli during testing, we characterized the factors that influence memory recall. Additionally, by eliminating repeated testing, we determined there was a positive effect on memory recall.

Subjects might have made educated guesses on the content of individual questions, or combined guesses independently across multiple fragments. Instead, we found that performance was affected by many factors. Due to the high recall performance correct scores, we decided to test a new group of subjects only once after initial encoding to determine if there were any effects of repeated testing. We found that subjects who were exposed to repeated testing did 5.5% and 9.2% better that the subjects who were only tested once, one week and one month after initial encoding.

Our senses are continuously bombarded with information, only a small fraction of which is remembered. The information flow is very large even considering eye movements as a coarse proxy for attention and assuming that only a small fraction of what is seen is processed. Extensive previous research has demonstrated that retained information represents the output of selective and constructive filtering to extract meaning based on prior knowledge, goals, associations, and abstraction. Here, we demonstrated that we can capture a glimpse of the output

of these complex cognitive operations by using only visual, auditory, and emotional cues and making relatively accurate predictions about what subjects remember.

Movies offer the unique opportunity to examine memory formation for even sequences that are close to the basic elements of everyday episodic recollections. Subjects can form memories for specific movie events that are accurate and sufficiently robust to be reproducible across repeated testing, and yet constantly imperfect, thus following the basic properties of episodic memory formation demonstrating in other domains. The observation that recollections are consistent across subjects corroborates that there are specific aspects of the content of each shot that contribute to the remembering and forgetting. This study provides characterization of how some aspects of the audio, visual, and cognitive contents of brief shots contribute to memory recall and demonstrates that those content properties can be used to provide reasonably accurate predictions of what people will or will not remember from specific events embedded within a movie narrative.

Previous research has found similar findings to ours, except they sought to determine what makes an individual image memorable. Vogt et al. used a two-alternative forced choice method to test subjects' memory using 400 original and edited images 0.5 hours after initial encoding and 9 days after initial encoding. This study found that subjects' performance was relatively high when asked to recall an original photograph, 85% correct rate, and they concluded that there were specific details of visual senses that contribute to long-term memory of those senses. Although the specific implementation is distinct from our studies, the results from those single image studies also depict memory as accurate yet error-prone, with a high degree of consistency within and across subjects. The degree of memorability across subjects in those studies could also be predicted from variables describing contents of each picture. Specific

contents, such as the presence of a face, showed a strong correlation with image memorability. The results presented in our study extends those findings by examining time scales of weeks to months, by considering alternative items that are extremely similar to the test items in terms of basic properties, and by making predictions about memory recall for episodic events that include spatiotemporal context and emotional valence embedded in a narrative.

Even though using movies as stimuli provides a rich arena to quantitatively examine the formation of episodic memories, commercial movies, such as the ones used here, constitute artificial stimuli where the director attempts to guide the observers' viewpoints, attention, feelings, and even recollections. Therefore, the extent to which the conclusions about factors influencing the predictability of episodic memory formation from audio, visual, and cognitive content can be extrapolated to real life memories remains to be determined and will require further investigation. There are future analyses of **Experiment 2** data that would expand our understanding of the factors that influence memorability. Analysis at the subject level, in addition to group analysis that was conducted in this study, would allow for a more in depth understanding of any differences in factors influencing memory recall than what we previous found. The initial steps presented in our study provide a methodological approach that opens the doors to building more complex quantitative models to capture the output of the selective filtering and subjunctive processes that forms the essence of episodic memories.

References

1. Blumberg J & Kreiman G (2010) How Cortical neurons help us see: visual recognition in the human brain. *The Journal of Clinical Investigation* 120(9) 3054-3063.
2. Vogt S & Magnussen S (2007) Long-term memory for 400 pictures on a common theme. *Exp Psychol* 54(4):298-303.
3. Loftus EF (2005) Planting misinformation in the human mind: a 30-year investigation of the malleability of memory. *Learning & memory* 12(4):361-366.
4. Tulving E (2002) Episodic memory: from mind to brain. *Annual review of psychology* 53:1-25.
5. Schacter DL, Norman KA, & Koutstaal W (1998) The cognitive neuroscience of constructive memory. *Annual review of psychology* 49:289-318.
6. Loftus EF, Loftus EF (1980) On the Permanence of Stored Information in the Human Brain. *American Psychologist* 35(5): 409-420.