Current Psychology
https://doi.org/10.1007/s12144-023-04997-0

# One perspective or two? Viewpoint dependency in visual events

Ayşe Candan Şimşek[1] · Tolgahan Aydın[1] · Zeynep Gunes Ozkan[2]

Accepted: 14 July 2023

## Abstract

Viewpoint dependency in dynamic events is still an open question. Movies present a unique case of complex visual stimuli where consecutive shots are filmed from multiple viewpoints. In the present study, we have examined whether people remember viewpoint-specific information in movie-like visual scenes. We have used naturalistic activities which involved two actors where a) the sequence is presented from one or both actors' viewpoints and b) the individual actions were in a natural order or they were scrambled. The results indicated that memory for individual shots decreased when the sequence was presented from both actors' perspectives. Also, people were mostly unaware of the order manipulation, and reordering the individual actions did not lead to a decrease in memory performance. The results favor the film-form model, which suggests that the spatial relations in movie scenes are represented by taking the first shot of a scene as the basis and incorporating the views in subsequent shots accordingly. This argues for an economical encoding in visual events, which favors narrative continuity over spatial relations.

**Keywords** Viewpoint · Spatial perception · Visual events · Movies · Order

> "*Space surrounds us, omnipresent. Yet narratives are linear: like attention, they take things one after another … Consistency of perspective is presumed to be necessary for the construction of a mental spatial framework in which to place each object or landmark.*" (Tversky, 2004, p.380).

Spatial perception in everyday life incorporates weaving together multiple viewpoints into a spatial map as the human observer moves in the environment. As Tversky (2004) argues, the viewpoint is omnipresent, which gives us access to a 360-degree perspective. However, space in visual narratives is constructed in a sequential nature by exposure to shots in a successive manner. To integrate each object and actor into the spatial map, movies make use of multiple editing techniques geared towards maintaining spatial continuity (Bordwell & Thompson, 1986; Cutting, 2021; Loschky et al., 2020; Smith et al., 2012). This alleviates the cognitive load of the viewer and allows them to focus on the narrative while keeping the spatial relations constant. Research so far does not answer the question of how much people encode the viewpoint of an agent in movie scenes, which may present a different type of computation compared to everyday spatial cognition. In this paper, we investigated *viewpoint dependency*, a concept that examines how much people use the viewpoint of the visual scene for encoding and remembering the position of objects and agents. We inquired whether the viewpoint of a movie shot is remembered when a film-like sequence is presented from either one or two actors' viewpoints. Another purpose was to examine viewpoint dependency when there are concurrent tasks present, which relates to following the action sequence. Event segmentation theories consider multiple dimensions when a viewer perceives a visual narrative. In addition to space, time, and motion, top-down features like actor goals and intentions are monitored in comprehending visual events (Radvansky & Zacks, 2010; Zacks et al., 2007). Action order may be an instrumental part of tracking the goals of the actors and the time course of action as it relates to predictive processing. As people track space and time concurrently while processing visual events, how much spatial perception of actor positions relates to the order of action requires a better understanding. Previous research showed that the time course of an action sequence is instrumental to comprehension and working memory (Claus & Kelter, 2006; Hymel et al., 2016; Raisig et al., 2010; Ruby et al., 2002). Following the action

Ayşe Candan Şimşek
ayse.simsek@yasar.edu.tr

[1] Department of Psychology, Yaşar University, Bornova, Izmir, Turkey

[2] Department of Psychology, Bournemouth University, Bournemouth, Dorset, UK

Published online: 29 July 2023

provides an additional load on cognitive resources, which may affect the accuracy of people's spatial memory. Therefore, spatial perception in movie scenes may be influenced by whether the action order is maintained or violated.

Studying visual narratives is a current interest in psychology due to an increase in studies investigating the perception of movies and other complex visual media. Interest in movies as a type of visual media (comics, movies, video games, etc.) grows substantially due to their similarities to real-life action (Levin & Simons, 2000; Smith et al., 2012). In movies, filmmakers benefit from editing rules to achieve coherent spatial representations (Magliano et al., 1996; Schwan & Ildirar, 2010), and this is how the viewer is not bothered by different camera viewpoints. Prominently used in films, *continuity editing rules* permit cuts to be "invisible" (Bordwell & Thompson, 1986; Cutting, 2021; Hutson et al., 2022; Smith et al., 2012). Those involve the usage of various editing techniques, such as the 180-degree rule, shot-reverse-shot (SRS) sequences, gaze matches, over-the-shoulder shots, and match-on action cuts. This, in turn, presents the audience with a seamless experience. Research into the physical structures underlying the perception of movies is still new, but there is still much to discover about how people comprehend and remember spatial relations (Bordwell, 1985; Smith, 2012; Smith et al., 2012). The investigation into the perception of spatial relations in movies can benefit the study of working memory. Research using complex visual stimuli such as movies enables the manipulation of visual scenes, thereby facilitating the acquisition of fresh insights on the constraints of working memory capacity.

Tracking spatial relations in movie scenes can burden cognitive resources as the viewer follows the narrative. Recently proposed Scene Perception & Event Comprehension Theory (SPECT) (Loschky et al., 2020) argues that people actively monitor the current information in an event in working memory and store the past information in long-term episodic memory to comprehend the entire narrative (Cohn-Sheehy et al., 2022; Hutson et al., 2017, 2022). Similarly, Cognitive Load Theory (Sweller et al., 1998) proposes that the limited nature of working memory is bound by a capacity, and if the cognitive load of a task is high, it comes with a cost for the processing of simultaneous information. This, in turn, has been shown to affect encoding and memory as well as hinder comprehension and learning (Brich et al., 2021; Forster & Lavie, 2008; Hitch et al., 2019; Mayer, 2014; Rey, 2014). Therefore, when the viewpoint changes with each successive shot, this may increase the cognitive load. Continuity editing rules are employed to provide consistent spatial relations across film shots. The proper usage of those rules alleviates the cognitive burden of the viewer and eliminates the need to monitor spatial details.

Another factor that can increase cognitive load and burden working memory is monitoring the order of action in a visual event. Event order is bound by a temporal prediction. According to predictive accounts of event cognition (Zacks et al., 2007), people engage in continuous comparisons when monitoring events, and when predictions fail due to an unexpected change, event boundaries emerge, leading to the formation of a new event model. Claus and Kelter (2006) state that "mental representation of the time course of a dynamic situation is a prerequisite for understanding" (p. 1042). This view would suggest that if the order in an action sequence is not maintained, it is reasonable to expect the cognitive resources to be strained. The literature so far provides limited and conflicting results about how people monitor order in action sequences. In such a study, Raisig et al. (2010) presented chronologically ordered or temporarily violated event sequences after providing a verbal title of an event to engage predictive processing. This research showed that the pupillary response of participants was larger for temporal violations, which was attributed to increased cognitive load. As temporal violations appear to burden resources, how this relates to awareness was also studied. Hymel et al. (2016) compared awareness in the case of regularly ordered actions versus reordered actions, in which subsections of activity were changed. They found that approximately half of the subjects were not aware of the order manipulation when not distracted, and this number decreased when they engaged in a concurrent task. Importantly, if critical action is the last event, people detect misordering better, possibly due to a thorough comparison of previous events. This indicates that concurrent tasks further decrease the cognitive capacity to monitor event order. Levin and Wang (2009) also examined whether the observer's gaze could be used in actions that use *canonical* (natural order) or *reversed* (misordered) sequences and showed that manipulating order was not as influential as expected for the memory of object locations. While the object was the focus of attention in canonical scenes, the actor was pursued in reversed scenes. The research so far is not conclusive about the possible role of action order in memory for visual events, and no study so far investigated the possible interaction between action order and spatial cognition. As people are mostly blind to changes in order, manipulations of order can affect working memory and therefore hinder the processing of concurrent information, such as memory for viewpoint.

We can say that the mapping of spatial relations in movies is costly as scenes show different viewpoints in successive shots. This may require the viewer to put themselves into the shoes of the actor facing the camera. How many people register the viewpoint of a given shot -viewpoint dependency- in movie scenes is still an open question. Tversky and Hard (2009) showed that when a visual scene included an actor, people tended to take the perspective of that actor, especially when the question was action-related. This is in line with previous studies that found longer reaction times when the scene included a person sitting at a table facing the camera (Cavallo et al., 2017) and showed a relationship between narrative engagement and perspective taking where the writer puts themselves in the shoes of a fictional character (Bientzle et al., 2021). The same was true even when there was an empty chair facing the viewer. An agent or even the possibility of one lead viewer to put themselves in the place of that person through mental rotation, thus explaining longer reaction times. Similarly, Del Sette et al. (2022) considered perspective taking judgments in complex real-world scenes and showed that taking another agent's perspective can be cued by specific prompts. This suggests that an egocentric perspective is not always inherent and that both top-down effects can bias a viewer to take another agent's perspective.

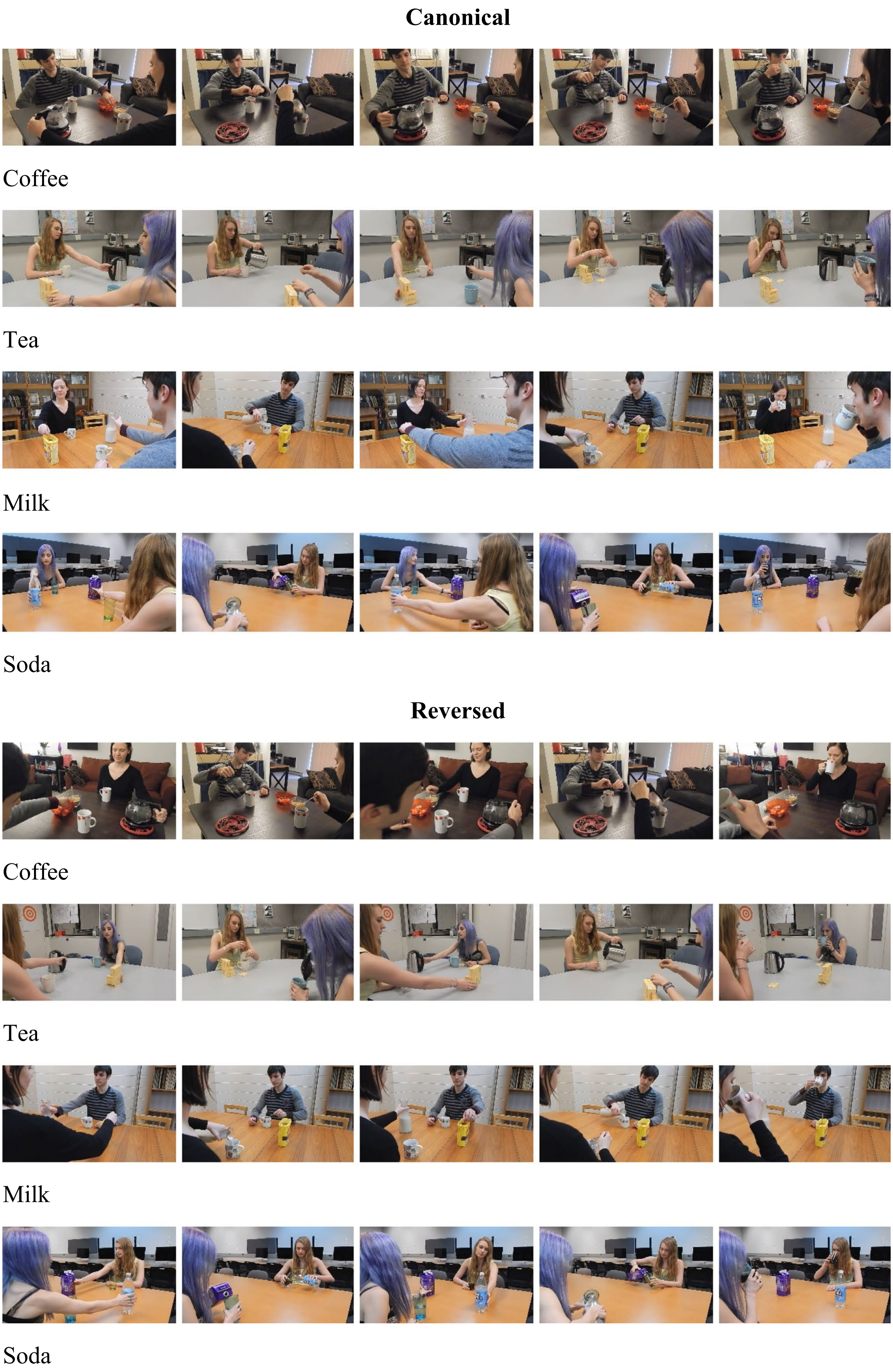
While viewpoint dependency was studied extensively in visual scenes, most of the literature comes from research that investigated static scenes (Jiang et al., 2013; Shelton & McNamara, 2004; Simons & Wang, 1998). Those studies heavily focused on the arrangement of objects, and the resulting mental representations were primarily viewpoint-dependent (David et al., 2006; Shelton & Mcnamara, 1997; Wolbers & Hegarty, 2010; Yu & Zacks, 2017). Research so far provides mixed results with respect to viewpoint dependency in dynamic visual scenes. Even though some studies support that dynamic scenes are represented in a viewpoint-dependent manner (Garsoffky et al., 2002; Sargent et al., 2019), literature also provides evidence in favor of viewpoint-independent representations (Garsoffky et al., 2007; Huff et al., 2009, 2011). Garsoffky et al. (2002) postulated three competing hypotheses for viewpoint dependency in dynamic visual scenes: the static-scene model, dynamic event model, and film-form model. The static-scene model posits that increasing viewpoints do not provide observers with a unified perspective, but rather, the observers remember scenes from specific viewpoints (Diwadkar & McNamara, 1997; Shelton & Mcnamara, 1997). Earlier studies indicated that reaction times increase linearly with the rotation angle (Shepard & Metzler, 1971). Therefore, according to the static model, an increase in the angle deviation from the original viewpoint results in a decreased performance due to mental rotation. The dynamic event model, on the other hand, states that dynamic scenes with multiple viewpoints lead to viewpoint-independent representation by helping the observer to establish a cognitive representation of the whole scene through abstraction (Allen et al., 1978; Franklin et al., 1992; Freyd, 1987; Huff et al., 2011). This abstraction allows viewers to unify different viewpoints, and novel viewpoints would not differ in terms of accuracy or speed. Additionally, the film-form model was proposed by Bordwell and Thompson (1986), which argues for the representation of a single viewpoint. This presents a compromise between the static-scene model and the dynamic-scene model. Spatial representation is still viewpoint-dependent, but the viewer uses an economic strategy to use the establishing shot at the beginning of a scene to build actor and object positions which are deemed consistent throughout the scene. While the static scene model posits that each additional viewpoint should allow for separate but equally strong representations, the film-form model predicts that only the first viewpoint of the establishing shot would be taken as the referential source of spatial layout. The motivation of this paper is to probe further into viewpoint dependency in film-like dynamic visual scenes.

## Experimental overview

The present study examined the role of perspective and action order on viewers' spatial representations. We measured recognition memory to examine whether a single shot is remembered from the viewpoint of a specific actor. We used videos where the entire sequence was either presented from one actor's perspective (single perspective) or both actors' perspectives (SRS sequences). Also, the action sequences in the videos were either in natural order (canonical) or misordered (scrambled). Reliance on perspective was examined to compare three theories suggested by Garsoffky et al. (2002). Similar to Garsoffky et al. (2002), we used naturalistic action of daily events, but we edited videos using one basic continuity editing tool to make perspective change salient.

The hypotheses for the role of perspective on visual recognition were twofold. If the dynamic event model applies to visual narratives, we would see similar performance for single perspective and SRS sequences as the spatial model would be abstracted and remembered from every viewpoint. If the film-form model's predictions apply, we would see better performance for a single perspective where the perspective of the first shot, which would be considered an alternative to an establishing shot, is the same as the other shots in the sequence. The viewpoints in the SRS sequences, on the other hand, alternate between each shot, which in turn may lead to low memory performance for the SRS sequences. Research also suggests that, after a cut, the attention of the viewer goes to what's central in the scene, such as the actor's faces and other prominent objects (Cutting et al., 2012; Smith, 2012; Smith & Henderson, 2008; Smith et al., 2012). It is reasonable to expect binding action from multiple viewpoints would be prioritized, and the presentation of multiple viewpoints would be overlooked. This would in part, be due to the correct employment of the continuity editing rules, which make sure that the spatial continuity is preserved across shots and the viewer is not disoriented.

In addition to memory for viewpoint, we also measured memory for spatial orientation, where we asked whether people encoded the location of objects in relation to the actors in the scene. As memory for viewpoint measured whether people remembered a specific action component


Canonical
Coffee
Tea
Milk
Soda
Reversed
Coffee
Tea
Milk
Soda

◂**Fig. 1** Order of presentation in each activity of canonical and scrambled videos. In scrambled videos, the reaching shots are switched with one another to create scrambled video sequences. In canonical conditions, coffee and tea are single perspectives, and in scrambled conditions, milk and soda are single perspectives. The other half of each condition consists of shot reverse shot sequences

from a specific actor's perspective, memory for spatial orientation measured whether people registered the location of objects with respect to the actors regardless of the perspective. We expected to find similar results for the memory of perspective and spatial orientation such that the object relations would be remembered better in single-view narratives, which gives more opportunities to encode the spatial relations in multiple repetitions from a single actor's perspective.

Additionally, we intended to examine the possible interaction between order and perspective. Our manipulation was adopted from Levin and Wang's (2009) study, where researchers used either naturally ordered or misordered action sequences. If the order is disrupted, this may lead to increased cognitive load and therefore hinder memory further in SRS sequences. For single-perspective scenes, however, because the viewpoint is constant, the memory of separate viewpoints may not be affected by order manipulation. Therefore, we hypothesized that order would not affect memory for single perspective scenes, but scrambled order would lead to lower memory in SRS sequences.

## Method

### Participants

A sensitivity analysis with G*power (Faul et al., 2009) for 2*2 mixed design indicated that the lowest effect size our minimum sample ($N=100$) was able to capture with an alpha level error probability of 0.05, and power of 0.8 was 0.12 in terms of effect size *f*.

A total of 118 participants took part in the experiment. One participant was excluded from the data due to having more than one response time above a 3.92 standard deviation. Therefore 117 participants were included (94 females and 24 males, $M_{age}=21.9$, $SD_{age}=3.3$). The majority of the participants (84%) were undergraduate students. 47% of the participants were in the canonical group ($N=55$), while the remaining 53% were in the scrambled group ($N=62$). The experiment was approved by the Institutional Ethical Board and all participants provided informed consent before the study. Participants were compensated with course credit in predetermined classes.

None of the participants reported being color blind and all the participants had normal or corrected-to-normal vision.

### Stimuli

We filmed 4 activity videos (coffee, tea, milk, and soda) in the laboratory, where two actors performed everyday activities while seated face-to-face across a table. The actors reached, poured, and drank the necessary ingredients for their beverages in all activities. Each video included 5 shots taken from a 45-degree angle positioned over the shoulder of each actor. Each video was edited so that all the shots either came from a single actor's viewpoint (single perspective) or alternated between each actor's viewpoint with a technique called shot-reverse shot editing (shot-reverse shot). This manipulation was implemented as a within-subjects variable since half of the videos were assigned to the single perspective condition, and the other half were assigned to the shot-reverse-shot condition. In addition, the order manipulation was done similarly to Levin and Wang's (2009) study, in which the steps of each activity either followed a natural order (canonical) or the steps were reordered (scrambled). Canonical and scrambled conditions refer to the edited format of the presentation videos. In the canonical condition, all the shots in the video were presented in the natural order for each activity. In the scrambled condition, the position of shot 2 and shot 4 were exchanged, which created misordered action sequences. Order manipulation was administered between subjects as the first group watched all 4 videos in the canonical order, and the second group watched all 4 videos in the scrambled order. The sequences ended with a coherent conclusion (actors drinking the beverage).

The videos were counterbalanced based on activity between the two groups to avoid an additional effect of content. Activities were divided in half to be placed between conditions. Coffee and tea activities were assigned to single perspective conditions in canonical videos and shot reverse shot conditions in scrambled videos. Similarly, milk and soda were assigned to single-perspective conditions in canonical videos and shot reverse shot conditions in scrambled videos. Manipulation of order was aimed to be further tested with a funneling type of post-questionnaire at the end of the experiment. Figure 1 provides screenshots taken from all activities showing the single perspective or shot reverse shot sequences in either canonical or scrambled conditions.

The average duration of the activity videos was 33.25 s. The average duration of the test shots was 6.68 s, and 6.4 s for distractor shots. The videos were filmed and later edited by the authors using iMovie software from the MacOS platform in an amateur capacity. The actors were graduate students who did not have any professional acting experience. Care and attention are given to make sure that the lighting of the room and the positioning of the camera angle and height is consistent across all videos. The videos were filmed and presented without audio. The stimuli videos can be found at (http://osf.io/qx5hd).

## Procedure

The experiment was implemented using PsychoPy (Peirce et al., 2019). Pavlovia.org was used as the platform for the experiment.

In the first part of the experiment, participants watched 4 activity videos. Each trial block had one activity video, then a visual recognition test with 5 tests and 5 distractor shots. The experiment included 4 trial blocks. Participants were in one of the two order groups (canonical vs. scrambled). In both groups, the presentation order of the videos was randomized. The test shots and distractor shots that follow the activity videos were also given in random order.

In the second part of the experiment, the screenshots taken from the second and third shots of the main videos were presented to the participants in their original and mirror form in a 2AFC (alternative-forced choice) format. The first and last shots were not used because the actors were not interacting with the target objects in those as they provided an introduction and end to the activity. We also wanted to use two test trials to present the objects from both actors' perspectives. Mirror images were generated by rotating the original image 180 degrees on the horizontal axis. After being tested with videos, participants were asked to indicate which image was part of the original video. The target picture's position was also randomized. Figure 2 shows an illustration of a trial with the canonical order from the shot-reverse-shot perspective.

The third and last part of the experiment consisted of three open-ended, funneling-type post-experiment questionnaires like the ones used in previous studies on *change blindness* (Angelone et al., 2003; Levin & Simons, 1997; Simons & Levin, 1998). This part was presented only for the group who received the scrambled videos since the order was administered between-subjects. The purpose of this questionnaire was to test the awareness of order manipulation. The questions were as follows:

1. *Did you detect any difference in the videos you watched?*
2. *Do you think that the events in the videos follow naturally?*
3. *Did you notice that the order of the events changed in some videos?*

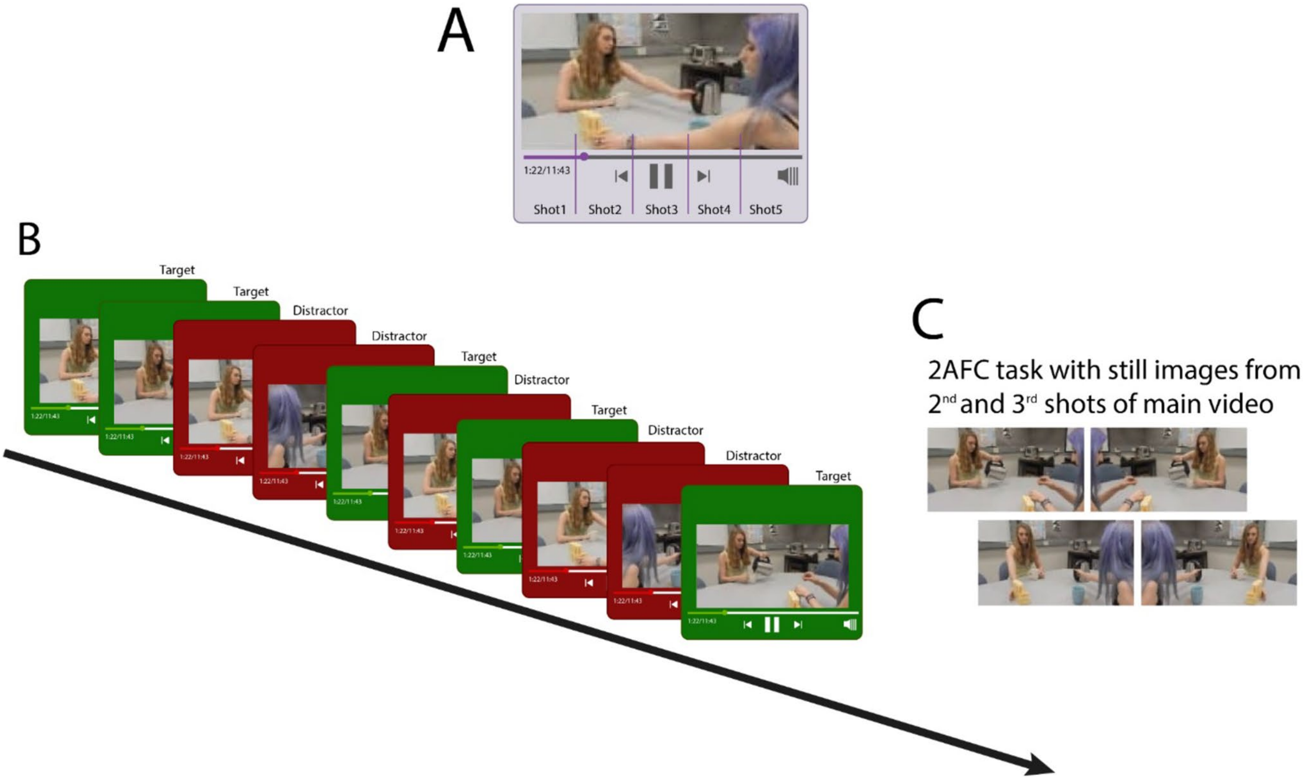


**Fig. 2** An illustration of an example trial block. **A**. Presentation video consisting of 5 shots that the participants watched before test part. **B**. The test part consists of the test and distractor shots that the participant saw in random order after the presentation video. **C**. Two alternative forced choice task including images taken from the presentation video

Open-ended answers were taken and later coded for content by two researchers. Assistants categorized answers and discussed answers to reach a complete agreement. The answers were categorized to indicate the percentage of participants who noticed anything different in the video (*general awareness*) and the percentage of people who noticed the order manipulation (*order awareness*) in the respective three questions.

### Data analysis

The study used a 2*2 mixed design with *perspective* as the within-subjects variable and *order* as the between-subjects variable. To investigate whether different perspectives and orders of videos had effects on recognition performance, dprime (d') scores were calculated. D' scores were measured as the difference between standardized scores of hit rates and false alarm rates ($z$(H)—$z$(F)). The hit rate is the proportion of test trials answered correctly (hits / (hits + misses)), while the false alarm rate is the proportion of distractor trials answered incorrectly (false alarms / (false alarms + correct rejections)). Since infinite d' scores were obtained for H or F scores of 0 or 1, a correction called *log-linear* was implemented. We added 0.5 to both hits and false alarms and added 1 to both signal and noise trials, as suggested by Stanislaw and Todorov (1999). Following the calculations of d' for each participant, accuracy and all reaction times were done by the 2*2 analysis of variance (ANOVA). The Mann–Whitney U test and Pearson's correlation test were used in order to investigate demographic information (See OSF file for detailed analysis).

To compute all analyses, the ezANOVA function from the ez package v.4.4–0 (Lawrence, 2016) and to create all plots, the ggplot function from the ggplot2 package v.3.3.6 (Wickham, 2016) was used within the R Environment for Statistical Computing (R Development Core Team, 2011).

## Results

Prior to the hypothesis testing the whole sample was included in the exploratory analysis and was not separated according to perspective and order conditions. We observed no major differences for demographic factors of gender and age, and we did not find any significant effect of the activity type.

### Shot memory

We observed a main effect for *perspective* ($F$ (1, 115) = 53.19, $p < 0.001$, $\eta_p^2 = 0.32$), however, neither the main effect of order ($F$ (1, 115) = 1.76, $p = 0.19$, $\eta_p^2 = 0.015$), nor the interaction of perspective and order ($F$ (1, 115) = 0.54, $p = 0.46$, $\eta_p^2 = 0.005$) were statistically significant. Also, no significant main effect of order ($F$ (1, 115) = 0.003, $p = 0.96$, $\eta_p^2 < 0.001$), perspective ($F$ (1, 115) = 1.91, $p = 0.66$, $\eta_p^2 = 0.002$), or interaction ($F$ (1, 115) = 1.68, $p = 0.20$, $\eta_p^2 = 0.014$) were detected for reaction times. Figure 3A shows d' for perspective and order.

When mean d` were examined, we observed that d` of single perspective were higher than d' of SRS sequences. Also, participants were not more accurate in canonical videos compared to scrambled videos and they responded with a similar pace to all video conditions. The mean and standard deviation of d` and RT scores can be found in Table 1 below.

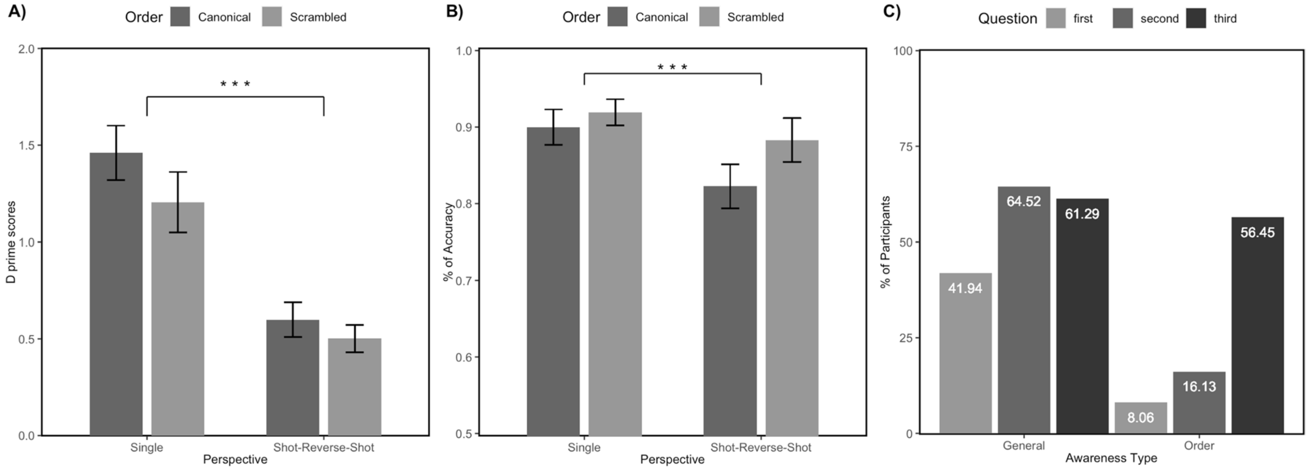


**Fig. 3** The d' scores for the shot memory, accuracy percentages for the spatial orientation memory for perspective and order, and general awareness and order awareness percentages for each question. **A**. The d' scores for the shot memory for perspective and order. **B**. The accuracy percentages for the spatial orientation memory for perspective and order. **C**. Scores for general awareness and order awareness percentages for each question

**Table 1** D' scores and reaction times in shot memory for different orders and perspectives

| *Perspective* | *D Prime Scores* | | *Reaction Times* | |
|---|---|---|---|---|
| | *Order* | | | |
| | *Canonical* | *Scrambled* | *Canonical* | *Scrambled* |
| Single Perspective | 1.46 ± 1.05 | 1.2 ± 1.23 | 1.43 ± 0.90 | 1.60 ± 1.09 |
| Shot Reverse Shot | 0.60 ± 0.67 | 0.50 ± 0.56 | 1.67 ± 1.88 | 1.48 ± 0.95 |

Mean (and standard deviation) d' scores and reaction times in shot memory for different order and perspectives

**Table 2** Accuracy percentages and reaction times in spatial orientation memory

| *Perspective* | *Accuracy Percentage* | | *Reaction Times* | |
|---|---|---|---|---|
| | *Order* | | | |
| | *Canonical* | *Scrambled* | *Canonical* | *Scrambled* |
| Single Perspective | 0.90 ± 0.17 | 0.92 ± 0.13 | 3.73 ± 1.87 | 3.33 ± 2.49 |
| Shot Reverse Shot | 0.82 ± 0.21 | 0.88 ± 0.23 | 4.10 ± 2.17 | 3.44 ± 1.77 |

## Spatial orientation memory

Similar results were observed between shot memory and spatial orientation memory. For accuracy, only the main effect of perspective was statistically significant, $F$ (1, 115) = 7.56, $p$ = 0.007, $\eta_p^{2} = 0.062$. The main effect of order ($F$ (1, 115) = 1.89, $p$ = 0.16, $\eta_p^2 = 0.017$) and interaction effect ($F$ (1, 115) = 0.99, $p$ = 0.32, $\eta_p^2 = 0.008$) were not statistically significant. Figure 3B shows accuracy percentages for perspective and order.

For reaction times, we did not find any significant effects of order, ($F$ (1, 115) = 2.38, $p$ = 0.13, $\eta_p^2 = 0.020$.) and perspective ($F$ (1, 115) = 1.79, $p$ = 0.18, $\eta_p^2 = 0.015$) and there was no significant interaction between the factors ($F$ (1, 115) = 0.57, $p$ = 0.45, $\eta_p^2 = 0.005$).

Participants were more accurate in single perspective than in SRS sequences and no effect of perspective and order was observed on reaction times. No main or interaction effect was observed for order suggesting that viewpoint information was coded independently from action order. The mean and standard deviation for accuracy percentages and reaction times can be seen in Table 2.

## Post-experiment questionnaire

The answers were categorized to indicate the percentage of participants who noticed anything different in the video (*general awareness*) and the percentage of people who noticed the order manipulation (*order awareness*) in the respective three questions. To give an example of general awareness, the following comment to the second question, "No, for example, water suddenly passes to the left hand while it was in the right hand", shows that the participant realized that the videos were not natural, but they were not aware of the order manipulation. Similarly, another answer to the same question clearly stated that the participant was aware of order manipulation: "No, the order of some sections was changed." Fig. 3C shows general and order awareness percentages for each question.

# Discussion

First of all, the results of the current study indicated that memory for individual shots was better for single-perspective sequences than SRS sequences, which employed two actors' perspectives consecutively. This result suggests a conditional conclusion that depends on the cognitive load of the task. Since single-view sequences do not require the viewer to keep track of the changing viewpoints, the consistency of the perspective leads to a viewpoint-dependent representation where the actor's viewpoint is incorporated into the narrative. However, when the sequence is in the form of the SRS sequence, which is the most conventional form in conversation scenes in movies, viewers are faced with a decision to use their cognitive resources more efficiently.

When a visual narrative includes more than one actor's perspective, the resulting mental representation favors the film-form model (Bordwell & Thompson, 1986), which suggests that the spatial relations are incorporated into the first shot, generally called the establishing shot. The film-form model proposes viewpoint dependent representation, in which memory would suffer with increasing viewpoints. Better detection in single-perspective scenes compared to SRS sequences suggests a viewpoint dependent representation as each alternate perspective burdens the viewer, and leads to a more economical approach favoring narrative over spatial details. In our videos, we can argue that the participants may have treated the first shot as the establishing shot. The film-form model can further help explain our results since viewpoints different from establishing shots could have led to a decrease in memory for viewpoint specific information, which is observed with SRS sequences.

One other explanation for reduced memory for viewpoint in SRS sequences can be related to the Cognitive Load Theory, which suggests that tracking simultaneous information puts a burden on cognitive resources (Forster & Lavie, 2008; Hitch et al., 2019; Mayer, 2014; Rey, 2014). Increasing the number of viewpoints to be tracked can strain working memory and lead to insufficient encoding of viewpoint-related information. New research shows that structured strategies can offload portions of the demand in high cognitive load tasks focused on visual-spatial techniques (Brich et al., 2021). This tells us that working memory load is not an all or none phenomenon, and additional factors can help alleviate the burden. Tracking the concurrent spatial details while engaging in narrative comprehension in a visual narrative can be considered a simultaneous task with following the narrative.

Secondly, the order of the action had no main effect on memory, and no interaction was found between perspective and memory. This suggests that reordering the steps of the action did not lead to a decrease in performance. The explanation for this comes from the literature, which shows that people do not often realize order manipulations as long as the narrative flow is continuous (Hymel et al., 2016). The open-ended answers given to the questions on manipulation awareness also support the idea that people did not realize that the steps of the action were reordered in the sequence unless specifically directed. This result may be due to several reasons, one being the fact that the depicted activity, which started and ended with a coherent direction, such as someone starting to prepare tea and drinking the prepared tea, provides a coherent sequence where interchanging the middle steps might not have disrupted the narrative flow. Also, the narrative order might be coded independently of the perspective information. Besides narrative flow, there is one more possible explanation for why the order did not yield any effect. The findings of Levin and Wang (2009) suggested that people follow the actor in scrambled videos, while the actor's interaction with the object is what people focus on in canonical videos. Participants might have missed the scrambled order of events since it required attending to interactions with objects, which is only possible in canonical videos. Future studies could implement such a design by asking participants to specifically focus on objects, to examine whether scrambled videos can be detected in a narrative flow. We cannot eliminate this explanation since we did not specifically ask participants to focus on objects.

As Baker and Levin (2015) demonstrated, people notice changes less often in an ongoing event. Because the event in our videos was coherent overall, people did not notice any order violations. Support for this argument also comes from Hymel et al. (2016), where researchers observed that when reordering the action components at the end of the event where no other action follows, people notice order manipulation more. However, when the reordering is introduced in the middle of the event, people notice the change less. Because in our videos, the order manipulation was in the middle, and the event ended coherently, this may have led to lower awareness of the order reversal. This may be one of the reasons why the order did not interact with perspective memory.

Lastly, we found similarities between memory for dynamic shots and still video images. Our intention for using two measures for the same stimuli was to test whether participants would correctly remember the positions of the objects if they were presented simultaneously with a mirror image of the scene. Dynamic scenes provide more information than still images as actors engage with objects, which could represent information on perspective besides memory for relative actor and object orientation. However, still images only test for knowledge of spatial mapping. Results show that spatial orientation is stored, and it is not mere recollection of action.

## Limitations and future implications

We used simple activity videos filmed in the laboratory, which involved amateur actors. Those may lack the artistic qualities of professionally produced movies that incorporate aesthetic concerns. Videos filmed in laboratories may sometimes be more monotonic or robotic, which may decrease the motivation and engagement of the viewer. Using cinematic sequences from existing movies may be considered for future research to avoid any potential issues related to the enjoyment of the videos. Also, the present videos were not formed to include establishing shots; therefore, future research can focus on the potential role of establishing shots in viewpoint dependency.

In addition, as proposed above, the activities we used started with logical beginnings and ended with a resolution. The order manipulation was administered to the middle shots that may be interchangeable. Also, a more narrative-specific question to investigate the comprehension of the activity can give us more information about whether the action continuity is disrupted. Future research can use a more extensive order manipulation to specifically compare the role of each shot in an SRS sequence. For example, if the order manipulation was done to the first and last event so that the person drinks the tea and then starts preparing it, then the order might interact with viewpoint information. Why perspective information is discarded when more than one actor is present in a narrative is still an open question.

Finally, we did not ask any questions about the content of the visual events in our videos, so we cannot draw any direct links to the viewer's motivation to follow the narrative. As the SPECT theory (Loschky et al., 2020) suggests, people hold active information in working memory to follow the entire narrative in a movie. Coherence of a narrative leads to better recall of previously experienced events (Cohn-Sheehy et al., 2022). Future research may consider how much following the narrative may interact with tracking the actors' positions as a secondary task. As more than one actor's perspective is depicted, this may lead to discarding the spatial details in favor of the narrative to optimize cognitive load better. Because the narrative is often the viewer's objective, this may come at the expense of encoding the viewpoint information.

## Conclusions

Studying complex visual events can broaden our understanding of the cognitive strategies viewers adopt when dealing with multiple viewpoints. Film-like visual events contain rich and complex spatial cues. Continuity editing rules are used to provide consistent spatial relations to free the viewer's cognitive resources. The study of continuity editing rules is still new to cognitive research, but it provides a promising tool that can help us understand everyday interpersonal interactions where considering another agent's viewpoint is crucial. As SPECT theory suggests, people use top-down and bottom-up information to process and comprehend a narrative (Loschky et al., 2020). The viewpoint of an actor can be considered among the bottom-up information that forms the spatial map of a visual event. Our study suggests that in visual narratives that involve more than one actor, people do not register viewpoint-specific information and that order of action may be overlooked if the action starts and ends with a coherent progression. Event Segmentation Theory indicates that people monitor visual events along multiple dimensions in a predictive manner (Zacks et al., 2007). The present study adds to understanding of how people process spatial relations together with action order.

**Acknowledgements** We thank Melissa Elston, Steven Armstrong, Kacie Armstrong and Erin İşbilir for their help in stimulus creation and James Eric Cutting for their shared comments and ideas.

**Author contributions** ACŞ developed the study concept and the study design. TA and ZGO programmed the experiment. Data collection, data analysis, drafting the manuscript, and providing critical revisions for drafts were combined work of three authors. All authors approved the final version of the manuscript for submission.

**Funding** This work was supported within the scope of the scientific research project which was accepted by the Project Evaluation Commission of Yaşar University under the project number and title of "BAP107_ Spatial Perspective Taking in Dynamic Visual Scenes".

**Data availability** Data and analyses scripts have been made publicly available via the Open Science Framework and can be accessed at http://osf.io/qx5hd

## Declarations

**Ethics approval** Ethics approval was granted by the University Ethics Committee of the corresponding author.

**Consent to participate** Informed consent was obtained from all individual participants included in the study.

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

Allen, G. L., Siegel, A. W., & Rosinski, R. R. (1978). The role of perceptual context in structuring spatial knowledge. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 617–630. https://doi.org/10.1037/0278-7393.4.6.617

Angelone, B. L., Levin, D. T., & Simons, D. J. (2003). The Relationship between change detection and recognition of centrally attended objects in motion pictures. *Perception, 32*(8), 947–962. https://doi.org/10.1068/p5079

Baker, L. J., & Levin, D. T. (2015). The role of relational triggers in event perception. *Cognition, 136*, 14–29. https://doi.org/10.1016/j.cognition.2014.11.030

Bientzle, M., Eggeling, M., Kanzleiter, M., Thieme, K., & Kimmerle, J. (2021). The impact of narrative writing on empathy, perspective-taking, and attitude: Two randomized controlled experiments on violations of Covid-19 protection regulations. *PLoS One, 16*(7), e0254501.

Bordwell, D. (1985). *Narration in the fiction film*. Routledge.

Bordwell, D., & Thompson, K. (1986). *Film art: An introduction* (2nd ed.). Knopf.

Brich, I. R., Bause, I. M., Hesse, F. W., & Wesslein, A.-K. (2021). How spatial information structuring in an interactive technological environment affects decision performance under working memory load. *Computers in Human Behavior, 123*, 106860. https://doi.org/10.1016/j.chb.2021.106860

Cavallo, A., Ansuini, C., Capozzi, F., Tversky, B., & Becchio, C. (2017). When far becomes near: Perspective taking induces social remapping of spatial relations. *Psychological Science, 28*(1), 69–79. https://doi.org/10.1177/0956797616672464

Claus, B., & Kelter, S. (2006). Comprehending narratives containing flashbacks: Evidence for temporally organized representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(5), 1031–1044. https://doi.org/10.1037/0278-7393.32.5.1031

Cohn-Sheehy, B. I., Delarazan, A. I., Crivelli-Decker, J. E., Reagh, Z. M., Mundada, N. S., Yonelinas, A. P., Zacks, J. M., & Ranganath, C. (2022). Narratives bridge the divide between distant events in episodic memory. *Memory & Cognition, 50*(3), 478–494. https://doi.org/10.3758/s13421-021-01178-x

Cutting, J. E. (2021). *Movies on our minds: The evolution of cinematic engagement*. Oxford University Press. https://doi.org/10.1093/oso/9780197567777.001.0001

Cutting, J. E., Brunick, K. L., & Candan, A. (2012). Perceiving Event Dynamics and parsing Hollywood films. *Journal of Experimental Psychology: Human Perception and Performance, 38*(6), 1476–1490. https://doi.org/10.1037/a0027737

David, N., Bewernick, B., Cohen, M., Newen, A., Lux, S., Fink, G., ... Vogeley, K. (2006). Neural representations of self versus other: Visual-spatial perspective taking and agency in a virtual ball-tossing game. *Journal of Cognitive Neuroscience*, *18*(6), 898–910

Del Sette, P., Bindemann, M., & Ferguson, H. J. (2022). Visual perspective-taking in complex natural scenes. *Quarterly Journal of Experimental Psychology (2016), 75*(8), 1541–1551. https://doi.org/10.1177/17470218211054474

Diwadkar, V. A., & McNamara, T. P. (1997). Viewpoint dependence in scene recognition. *Psychological Science, 8*(4), 302–307. https://doi.org/10.1111/j.1467-9280.1997.tb00442.x

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160.

Forster, S., & Lavie, N. (2008). Failures to ignore entirely irrelevant distractors: The role of load. *Journal of Experimental Psychology: Applied, 14*, 73–83.

Franklin, N., Tversky, B., & Coon, V. (1992). Switching points of view in spatial mental models. *Memory & Cognition, 20*(5), 507–518. https://doi.org/10.3758/BF03199583

Freyd, J. J. (1987). Dynamic mental representations. *Psychological Review, 94*(4), 427–438. https://doi.org/10.1037/0033-295X.94.4.427

Garsoffky, B., Huff, M., & Schwan, S. (2007). Changing viewpoints during dynamic events. *Perception, 36*(3), 366–374. https://doi.org/10.1068/p5645

Garsoffky, B., Schwan, S., & Hesse, F. W. (2002). Viewpoint dependency in the recognition of dynamic scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(6), 1035–1050. https://doi.org/10.1037/0278-7393.28.6.1035

Hitch, G., Allen, R. J., & Baddeley, A. D. (2019). Attention and binding in visual working memory: Two forms of attention and two kinds of buffer storage. *Attention, Perception & Psychophysics, 82*, 280–293.

Huff, M., Jahn, G., & Schwan, S. (2009). Tracking multiple objects across abrupt viewpoint changes. *Visual Cognition, 17*(3), 297–306. https://doi.org/10.1080/13506280802061838

Huff, M., Schwan, S., & Garsoffky, B. (2011). When movement patterns turn into events: Implications for the recognition of spatial configurations from different viewpoints. *Journal of Cognitive Psychology, 23*(4), 476–484. https://doi.org/10.1080/20445911.2011.541152

Hutson, J. P., Smith, Tim J., Magliano, J. P., & Loschky, L. C. (2017). What is the role of the film viewer? The effects of narrative comprehension and viewing task on gaze control in film. *Cognitive Research: Principles and Implications*, *2*(1).

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

Hymel, A., Levin, D. T., & Baker, L. J. (2016). Default processing of event sequences. *Journal of Experimental Psychology: Human Perception and Performance, 42*(2), 235–246. https://doi.org/10.1037/xhp0000082

Jiang, Y. V., Swallow, K. M., & Capistrano, C. G. (2013). Visual search and location probability learning from variable perspectives. *Journal of Vision, 13*(6), 1–13. https://doi.org/10.1167/13.6.13

Lawrence, M. A. (2016). ez: Easy analysis and visualization of factorial experiments. Retrieved June 13, 2022 from https://CRAN.R-project.org/package=ez

Levin, D. T., & Simons, D. J. (1997). Failure to detect changes to attended objects in motion pictures. *Psychonomic Bulletin & Review, 4*(4), 501–506. https://doi.org/10.3758/BF03214339

Levin, D. T., & Simons, D. J. (2000). Fragmentation and continuity in motion pictures and the real world. *Media Psychology, 2*(4), 357–380.

Levin, D. T., & Wang, C. (2009). *Spatial Representation in Cognitive Science and Film Projections, 3*, 24–52. https://doi.org/10.3167/proj.2009.030103

Loschky, L. C., Larson, A. M., Smith, T. J., & Magliano, J. P. (2020). The scene perception & event comprehension theory (SPECT) applied to visual narratives. *Topics in Cognitive Science, 12*(1), 311–351.

Magliano, J. P., Dijkstra, K., & Zwaan, R. A. (1996). Generating predictive inferences while viewing a movie. *Discourse Processes, 22*(3), 199–224. https://doi.org/10.1080/01638539609544973

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 43–71). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.005

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Radvansky, G. A., & Zacks, J. M. (2010). Event perception. *WIREs Cognitive Science, 2*(6), 608–620.

Raisig, S., Welke, T., Hagendorf, H., & van der Meer, E. (2010). I spy with my little eye: Detection of temporal violations in event sequences and the pupillary response. *International Journal of Psychophysiology, 76*, 1–8. https://doi.org/10.1016/j.ijpsycho.2010.01.006

R Development Core Team. (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Retrieved June 13, 2022 from http://www.R-project.or

Rey, G. D. (2014). Seductive details and attention distraction – An eye tracker experiment. *Computers in Human Behavior, 32*, 133–144. https://doi.org/10.1016/j.chb.2013.11.017

Ruby, P., Sirigu, A., & Decety, J. (2002). Distinct areas in parietal cortex involved in long-term and short-term action planning: A PET investigation. *Cortex, 38*, 321–339. https://doi.org/10.1016/S0010-9452(08)70663-4

Sargent, J. Q., Zacks, J. M., Hambrick, D. Z., & Lin, N. (2019). Event memory uniquely predicts memory for large-scale space. *Memory & Cognition, 47*(2), 212–228. https://doi.org/10.3758/s13421-018-0860-2

Schwan, S., & Ildirar, S. (2010). Watching film for the first time: How adult viewers interpret perceptual discontinuities in film. *Psychological Science, 21*(7), 970–976. https://doi.org/10.1777/0956797610372632

Shelton, A. L., & Mcnamara, T. P. (1997). Multiple views of spatial memory. *Psychonomic Bulletin & Review, 4*(1), 102–106. https://doi.org/10.3758/bf03210780

Shelton, A. L., & McNamara, T. P. (2004). Spatial memory and perspective taking. *Memory & Cognition, 32*(3), 416–426. https://doi.org/10.3758/BF03195835

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science (New York, N.Y.), 171*(3972), 701–703. https://doi.org/10.1126/science.171.3972.701

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review, 5*(4), 644–649. https://doi.org/10.3758/bf03208840

Simons, D. J., & Wang, R. F. (1998). Perceiving real-world viewpoint changes. *Psychological Science, 9*(4), 315–320. https://doi.org/10.1111/1467-9280.00062

Smith, T. J. (2012). The attentional theory of cinematic continuity. *Projections, 6*(1), 1–27. https://doi.org/10.3167/proj.2012.060102

Smith, T. J., & Henderson, J. M. (2008). Edit Blindness: The relationship between attention and global change blindness in dynamic scenes. *Journal of Eye Movement Research*, *2*(2). https://doi.org/10.16910/jemr.2.2.6

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

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*, 137–149.

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review, 10*, 251–296. https://doi.org/10.1023/A:1022193728205

Tversky, B. (2004). Narratives of Space, Time, and Life. *Mind and Language, 19*(4), 380–392. https://doi.org/10.1111/j.0268-1064.2004.00264.x

Tversky, B., & Hard, B. M. (2009). Embodied and disembodied cognition: Spatial perspective-taking. *Cognition, 110*(1), 124–129. https://doi.org/10.1016/j.cognition.2008.10.008

Wickham, H., (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978–3–319–24277–4. Retrieved June 13, 2022 from https://ggplot2.tidyverse.org

Wolbers, T., & Hegarty, M. (2010). What determines our navigational abilities? *Trends in Cognitive Sciences, 14*(3), 138–146. https://doi.org/10.1016/j.tics.2010.01.001

Yu, A. B., & Zacks, J. M. (2017). Transformations and representations supporting spatial perspective taking. *Spatial Cognition and Computation, 17*(4), 304–337. https://doi.org/10.1080/13875868.2017.1322596

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. https://doi.org/10.1037/0033-2909.133.2.273

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.