Mem Cogn (2016) 44:390–402
DOI 10.3758/s13421-015-0575-6

Semantic congruency but not temporal synchrony enhances long-term memory performance for audio-visual scenes

Hauke S. Meyerhoff[1] · **Markus Huff**[2]

Published online: 30 November 2015

Abstract Human long-term memory for visual objects and scenes is tremendous. Here, we test how auditory information contributes to long-term memory performance for realistic scenes. In a total of six experiments, we manipulated the presentation modality (auditory, visual, audio-visual) as well as semantic congruency and temporal synchrony between auditory and visual information of brief filmic clips. Our results show that audio-visual clips generally elicit more accurate memory performance than unimodal clips. This advantage even increases with congruent visual and auditory information. However, violations of audio-visual synchrony hardly have any influence on memory performance. Memory performance remained intact even with a sequential presentation of auditory and visual information, but finally declined when the matching tracks of one scene were presented separately with intervening tracks during learning. With respect to memory performance, our results therefore show that audio-visual integration is sensitive to semantic congruency but remarkably robust against asymmetries between different modalities.

Keywords Audio-visual scenes · Long-term memory · Scene memory · Semantic congruency · Massive memory

Electronic supplementary material The online version of this article (doi:10.3758/s13421-015-0575-6) contains supplementary material, which is available to authorized users.

✉ Hauke S. Meyerhoff
h.meyerhoff@iwm-kmrc.de

[1] Leibniz-Institut für Wissensmedien, Tübingen, Germany

[2] Department of Psychology, University of Tübingen, Tübingen, Germany

Introduction

The human long-term memory for objects and visual details of pictures has a tremendous capacity. Humans remember thousands of pictures (Standing, 1973), objects (Brady, Konkle, Alvarez, & Oliva, 2008; Brady, Konkle, Gill, Oliva, & Alvarez, 2013), and static scenes (Hollingworth, 2005; Konkle, Brady, Alvarez, & Oliva, 2010). Even after attending to hundreds of intervening objects, human observers recognize objects in static scenes well above chance level (Hollingworth, 2004). Besides spatial and object-based information, long-term memory as well as working memory also includes spatiotemporal information such as motion. Consequently, adding motion information to the visual scene further improves memory performance in a subsequent recognition test (Matthews, Benjamin, & Osborne, 2007; Papenmeier, Huff, & Schwan, 2012; see also Buratto, Matthews, & Lamberts, 2009). Importantly, this dynamic superiority effect does not rely on the additional visual information of multiple static frames. Instead, the dynamic motion itself acts as a cue for retrieval. In this report, we test how auditory information contributes to long-term memory performance. When compared to visual stimuli, memory performance for auditory stimuli is inferior (Cohen, Horowitz, & Wolfe, 2009). However, no research has yet addressed whether auditory information enhances long-term memory representations of dynamic scenes.

Most of the studies that explored audio-visual interactions focused on perceptual and attentional processes. With regard to the perception of audio-visual stimuli, there is cumulating evidence for an early integration of audio-visual information (Falchier, Clavagnier, Barone, & Kennedy, 2002; Giard & Peronnet, 1999; van der Burg, Talsma, Olivers, Hickey, & Theeuwes, 2011). For instance, attending to audio-visual stimuli elicits more neural activity in multimodal areas of

the brain but less in the corresponding unimodal areas (Bushara et al., 2003). Such audio-visual interactions have been demonstrated to disambiguate perceptual information (Sekuler, Sekuler, & Lau, 1997; Shams, Kamitani, & Shimojo, 2000) or to guide spatial attention (e.g., Santangelo & Spence, 2007; van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008; see also Spence, 2010). In order to elicit these beneficial influences on perceptual processes, auditory and visual information needs to be presented in close temporal proximity (e.g., Meredith, Nemitz, & Stein, 1987; Stevenson, Zemtsov, & Wallace, 2012; van Wassenhove, Grant, & Poeppel, 2007; Zampini, Guest, Shore, & Spence, 2005). Beyond audio-visual synchrony, semantically congruent auditory and visual information (see Spence, 2011, for a review) further enhances the detection (e.g., Iordanescu, Guzman-Martinez, Grabowecky, & Suzuki, 2008; Iordanescu, Grabowecky, & Suzuki, 2011) as well as the recognition of the corresponding objects (Amedi, von Kriegstein, van Atteveldt, Beauchamp, & Naumer, 2005; Giard & Peronnet, 1999) or events (Grassi & Casco, 2010).

The aim of our study was to test how auditory and visual information interact in memory representations. Therefore, we used filmic stimuli to test how auditory information affects long-term memory for dynamic scenes. We measured recognition performance for brief auditory, visual, or audio-visual clips after retention intervals of 1 day or 1 week. In Experiment 1, we show that audio-visual clips elicit higher long-term memory performance than unimodal clips. In Experiment 2, we further show that this audio-visual superiority effect is larger for matching auditory and visual information than for mismatching stimulus combinations. The results of Experiments 3a and 3b show that the memory advantage of audio-visual scenes is robust against violations of audio-visual synchrony. Even further, Experiment 4a confirms that memory performance for audio-visual scenes remains intact when the auditory and the visual tracks of the same scene were presented sequentially (i.e., immediately following each other) in the learning session. Nevertheless, Experiment 4b shows that memory performance declines when auditory and visual track of the same clips were presented sequentially but with interleaving tracks from other clips. Taken together, these results indicate that memory for audio-visual scenes does not result from a pure summation of independent retrieval cues for auditory and visual information. Instead, our results suggest that matching combinations of auditory and visual information are integrated into a common memory representation. However, due to the remarkable robustness against violations of audio-visual synchrony, this integration process does not just reflect a transfer from audio-visual integration from perceptual processes into memory representations.

Experiment 1

Experiment 1 was designed to demonstrate that memory performance for audio-visual information is superior to the performance for their unimodal counterparts. We asked participants to learn brief clips in one session and to recognize these clips among distractors in a second session. Because we are not aware of any other similar study, this study also served for exploratory purposes. Most importantly, we manipulated the modality of the clips (auditory, visual, audio-visual). Although our main interest in this manuscript focuses on the distinction between visual and audio-visual scenes, we also included an auditory-only condition in order to replicate the previous finding that visual memory is superior to auditory memory (Cohen et al., 2009). Further we manipulated the length of the clips. The goals of this manipulation were twofold. First, we aimed to make the clip length unpredictable to the participants in order to avoid any possible effects that might arise from anticipating the stimulus duration. Second, this manipulation served as a check of plausibility for our results. Because longer scenes include more information than shorter scenes, they provide more retrieval cues for the recognition test. Thus, longer clips should elicit more accurate memory performance than shorter clips in all of our experiments. Finally, we also manipulated the retention interval as a between factor in this experiment for exploratory purposes.

Methods

Participants

Forty-eight students (47 female; 19–44 years old) from the University of Tübingen participated in Experiment 1 for course credit or payment. All experiments were approved by the institutional review board of the Knowledge Media Research Center and all participants provided informed consent prior to testing. Assuming correlations of $r = .5$ between the repeated measures factors,[1] 22 participants are necessary to reliably (power = .99) detect effects of $\eta^2_p = 0.15$ (Faul, Erdfelder, Lang, & Buchner, 2007). Our counterbalancing procedure for the within-subject factors required the number of participants to be a multiple of six. Because our main focus was on the within-subject manipulations, we tested 2 × 24 participants in Experiment 1 (the retention interval of 1 day vs. 1 week was implemented as a between-subject factor), but 24 participants in Experiments 2–4b.

[1] The assumed within-subject correlation of $r = .5$ was supposed to reflect a conservative estimate for recognition experiments based on our prior experiences. The actual correlations between the modality conditions within the same retention interval in Experiment 1 were higher, all $r(22) > .67$. Thus, the actually achieved statistical power was higher than suggested by the a priori calculation.

Apparatus, stimuli, and procedure

The stimuli were programmed in Matlab using the Psychophysics Toolbox extension (Brainard, 1997; Pelli, 1997). The experiments were presented on a 23-in. LCD monitor (60 Hz, 1,920 × 1,080 pixels) controlled by a MacMini at an unrestricted viewing distance of approximately 60 cm.

The experiment was divided into two sessions (learning and testing). In the learning session, the participants attended to their target items. Before this session, participants were informed about the different modality conditions and were told that they would need to perform a recognition test in the second session.[2] In the second session (1 day or 1 week later), participants performed an old/new recognition test. There were twice as many stimuli as in the learning session. After the full presentation of each clip, the stimuli were replaced by a brief question whether the recent clip was old or new. Only then were participants allowed to enter their response by pressing the corresponding button on a keyboard. After the second session, participants received a list of the 50 movies and were asked to mark which of them they had seen within the last 5 years. Across all reported experiments, this number varied from one to 27 movies. Because excluding familiar movies from the analysis did not affect the results of any experiment, we will not discuss this issue further.

The stimuli were 900 brief clips from 50 Hollywood movies (1935–2008). From each movie, we extracted 18 clips without any filmic cuts (cinemetrix database; www.cinemetrix.lv). From these clips, we presented the sound track only (audio), the video track only (visual), or the sound and video track together (audio-visual). In the audio condition, we additionally presented a loudspeaker icon. For each movie, the clips were distributed equally across lengths of 3, 3.5, or 4 s. Therefore, for each movie and length, we had six distinct clips. In order to eliminate any potential influence from the video clips, we counterbalanced the assignment to the modality condition as well as the target-distractor identity. Two of the six clips of each movie and length were assigned to each of the three modality conditions. This assignment was counterbalanced across groups for three participants. In other words, the two clips that were presented in the auditory condition of the first participant appeared in the visual condition of the second participant and in the audio-visual condition of the third participant. Orthogonally to this, we exchanged the target-distractor assignment for half of the participants. Therefore, our stimuli were completely counterbalanced within groups of six participants (3 modalities × 2 target/distractor assignments). To further exclude any systematic influences from the order of the different conditions, we randomly intermixed trials from all conditions. For each group of counterbalanced participants, we generated two random orders of the clips; one for the learning session and one for the testing session. Due to the counterbalancing procedure of the modality of the clip, this procedure automatically also counterbalances the presentation order of the different modalities. As an example, within a counterbalanced group of six participants, all testing sessions started with the same clip. For three of the six participants, this clip was a target (auditory for one of the participants, visual for another, and audio-visual for the remaining participant). For the remaining three participants, this clip was a matched distractor to a target which, in return, was the matched distractor for the first three participants.

Following this procedure, three were 100 clips for each modality and length (i.e., 300 for each modality condition). Half of them were presented once in the learning session whereas all stimuli were presented once in the testing session. The movies were presented in their original resolution (768 × 576 pixels or 1,024 × 576 pixels) in the center of the screen. Because there were matched target and distractor stimuli from the same movie, all aspects particular to movies such as varying resolutions are controlled for in our experiments. The clips were separated by an inter-stimulus-interval of 2 s and participants were allowed to pause after sequences of nine clips.

Results

Memory performance was more accurate for audio-visual scenes than for purely visual scenes (see Fig. S.1 in the Supplementary Material). Also, memory performance was less accurate for purely auditory scenes than for visual scenes thus replicating the previously observed superiority of visual memory above auditory memory (Cohen et al., 2009). For our analysis, we calculated d' as sensitivity measurement (see Table 1, which also includes values for the response bias c). An ANOVA with retention interval (1 day, 7 days) as between-subject factor and modality (audio, visual, audio-visual) as well as clip length (3 s, 3.5 s, 4 s) as within-subject factors revealed that memory performance was more accurate following a retention interval of 1 day than following a retention interval of 7 days, $F(1, 46) = 8.08$, $p = .007$, $\eta^2_p = 0.15$, 95 % CI 0.01–0.33. Memory performance also increased with an increasing length of the clip, $F(2, 92) = 16.61$, $p < .001$, $\eta^2_p = 0.27$, 95 % CI 0.11– 0.39. Most importantly, we observed a

[2] Our participants were aware that they were participating in a memory experiment, however, they were not aware of the specific objectives of the experiment. We decided to use intentional (rather than incidental, i.e., cover story) instructions to ensure that our participants attend to the clips without speculating about the objectives of the experiment. Further, without prior instructions, some of our experimental conditions such as mismatching sound (Experiment 2) or reversed visual playback (Experiments 3a and 3b) might appear as equipment failures. Finally, because the participants signed up for two sessions, they might infer the memory component even with incidental instructions. At least our intentional instructions thus ensured that this awareness was identical across all participants. Given that the integration of auditory and visual information arises automatically, it is very unlikely that the type of instruction interacts with experimental conditions that are of major interest for this research project.

Table 1 Results of Experiments 1–4 separately for different lengths of the clips

	Clip length					
	3 s		3.5 s		4 s	
	d' *M (SD)*	c *M (SD)*	d' *M (SD)*	c *M (SD)*	d' *M (SD)*	c *M (SD)*
Experiment 1 (1 day)						
Audio	0.90 (0.40)	0.43 (0.39)	1.00 (0.46)	0.42 (0.37)	1.23 (0.50)	0.37 (0.39)
Visual	1.36 (0.56)	0.26 (0.29)	1.54 (0.52)	0.24 (0.23)	1.46 (0.55)	0.26 (0.30)
Audio-visual	1.70 (0.64)	0.16 (0.29)	1.81 (0.79)	0.17 (0.27)	1.98 (0.76)	0.06 (0.30)
Experiment 1 (7 days)						
Audio	0.62 (0.39)	0.47 (0.46)	0.66 (0.50)	0.35 (0.45)	0.83 (0.57)	0.32 (0.40)
Visual	1.01 (0.37)	0.17 (0.32)	1.10 (0.53)	0.17 (0.40)	1.24 (0.54)	0.19 (0.35)
Audio-visual	1.26 (0.44)	0.15 (0.39)	1.45 (0.50)	0.08 (0.35)	1.49 (0.51)	0.03 (0.39)
Experiment 2						
Congruent	1.32 (0.52)	0.06 (0.32)	1.59 (0.52)	0.07 (0.33)	1.67 (0.55)	0.02 (0.37)
Incongruent	1.17 (0.49)	0.29 (0.35)	1.42 (0.62)	0.30 (0.32)	1.52 (0.65)	0.28 (0.35)
Visual	1.08 (0.44)	0.23 (0.32)	1.12 (0.61)	0.18 (0.30)	1.15 (0.46)	0.20 (0.34)
Experiment 3a						
Visual forward	0.86 (0.43)	0.48 (0.33)	1.08 (0.38)	0.40 (0.39)	1.12 (0.45)	0.45 (0.36)
Visual backward	0.95 (0.36)	0.46 (0.26)	1.04 (0.48)	0.44 (0.37)	1.06 (0.36)	0.43 (0.28)
Audio/visual forward	1.33 (0.47)	0.29 (0.35)	1.46 (0.55)	0.24 (0.30)	1.59 (0.63)	0.29 (0.37)
Audio/visual backward	1.28 (0.51)	0.39 (0.30)	1.37 (0.57)	0.23 (0.36)	1.55 (0.50)	0.33 (0.28)
Experiment 3b						
Audio/visual forward	1.07 (0.65)	0.30 (0.43)	1.14 (0.58)	0.23 (0.38)	1.24 (0.63)	0.24 (0.40)
Audio/visual backward	1.04 (0.65)	0.31 (0.41)	1.23 (0.72)	0.29 (0.47)	1.22 (0.70)	0.26 (0.44)
Experiment 4a						
Audio first	1.69 (0.44)	0.23 (0.38)	1.86 (0.65)	0.24 (0.37)	2.07 (0.81)	0.12 (0.40)
Video first	1.53 (0.57)	0.27 (0.38)	1.71 (0.73)	0.32 (0.42)	1.97 (0.69)	0.29 (0.48)
Audio-visual	1.50 (0.53)	0.35 (0.36)	1.79 (0.56)	0.32 (0.44)	1.97 (0.76)	0.29 (0.43)
Experiment 4b						
Separate	1.33 (0.49)	0.15 (0.42)	1.60 (0.56)	0.04 (0.42)	1.68 (0.48)	0.03 (0.36)
Audio-visual	1.60 (0.57)	0.05 (0.39)	1.78 (0.77)	-0.12 (0.39)	2.00 (0.68)	-0.15 (0.32)

M mean, *SD* standard deviation, *d'* sensitivity, *c* response criterion

main effect of the modality, $F(2, 92) = 117.36$, $p < .001$, $\eta^2_p = 0.72$, 95 % CI 0.61– 0.78. Performance was most accurate in the audio-visual condition and least accurate in the audio-only condition. Here, Bonferoni-Holm corrected *t*-tests confirmed that all modality conditions differed from each other, all *t*s > 8.11, all *p*s < .001. Neither the three-way interaction nor any two-way interaction reached significance, all *F*s < 1.12, all *p*s > .35. Because the retention interval did not interact with the other factors, we dropped this manipulation for the remaining experiments.

Overall, the results of this experiment show that memory performance for audio-visual scenes is superior to memory performance for purely visual scenes. With regard to the number of potential retrieval cues, this audio-visual memory advantage is not too surprising. In this terminology, auditory information might provide an additional retrieval cue thus enhancing memory performance for audio-visual scenes. Nevertheless, this experiments serves as a basic demonstration of the audio-visual memory advantage, which we will be exploring in more detail in the subsequent experiments.

Experiment 2

The results of Experiment 1 show that memory performance for audio-visual scenes is more accurate than for purely visual scenes. A simple explanation for this finding is that auditory information provided a redundant retrieval cue, which enhances memory performance by a probabilistic summation. An alternative explanation for the audio-visual memory

advantage is that auditory and visual information are integrated into a more elaborated memory representation of the dynamic scene, which elicits memory performance that is higher than expected by a pure summation of distinct auditory and visual retrieval cues. In order to test these explanations, we manipulated the semantic congruency between the auditory and the visual track in Experiment 2. Auditory and visual track were either from the same scene (i.e., matching) or from a different scenes (i.e., mismatching). Additionally, we included a control condition without auditory track. Hereby, audio-visual integration during encoding predicts better memory performance for matching than mismatching combinations of auditory and visual information. This is because semantically matching auditory information has been demonstrated to improve audio-visual integration (e.g., Amedi et al., 2005; Chen & Spence, 2010; Grassi & Casco, 2010). In contrast, such a result pattern cannot be explained with a pure summation of independent retrieval cues.

Methods

Participants

Twenty-four new students (18 female; 19–56 years old) participated in Experiment 2.

Apparatus, stimuli, and procedure

Apparatus, stimuli, and procedure were identical to Experiment 1 with the following exceptions: The retention interval was 1 day. In both the learning and the testing session, the audio track either matched or mismatched the video track, or was absent (i.e., video only). We generated the clips with mismatching audio-visual information (counterbalanced across clips and participants) by combining the visual and auditory track from different clips. No visual or auditory track was presented twice within the learning or testing session. Importantly, we maintained study-test congruency; matching and mismatching combinations of auditory and visual information were the same for each participant during learning and testing.

Results

Memory performance was more accurate for audio-visually matching clips than for mismatching clips, while both audio-visual conditions revealed a higher memory performance than the purely visual condition (see Fig. S.2 in the Supplementary Material). We analyzed the observed sensitivity d' with a repeated measures ANOVA with audio-visual match (matching, mismatching, visual) as well as clip length (3 s, 3.5 s, 4 s) as independent variables. As in the previous experiments, memory performance increased with an increasing duration of the clip, $F(2, 46) = 15.02$, $p < .001$, $\eta^2_p = 0.40$, 95 % CI .16–.54. Most importantly, we observed an effect of the audio-visual match, $F(2, 46) = 30.76$, $p < .001$, $\eta^2_p = 0.57$, 95 % CI .36–.68, with the highest memory performance for audio-visually matching clips, while memory performance was lowest in the visual-only condition. Here, Bonferoni-Holm corrected *t*-tests confirmed that all audio-visual match conditions differed from each other, all ts > 2.91, all ps < .008. We also observed an interaction between audio-visual match and clip length, $F(4, 92) = 2.66$, $p = .038$, $\eta^2_p = 0.10$, 95 % CI 0–.20. This interaction (if reliable since the confidence interval includes zero) indicates that memory performance increased with clip length only in the matching and mismatching conditions but not in the purely visual condition (see Table 1).

In line with Experiment 1, these results show that presenting auditory and visual tracks simultaneously enhances memory performance relative to a purely visual presentation because both audio-visual conditions outperformed the purely visual condition. This memory advantage can be interpreted in terms of a probability summation of multiple retrieval cues. In other words, in the audio-visual conditions, participants miss target clips only when they fail to retrieve the visual track as well as the auditory track. However, probabilistic summation cannot explain why semantically congruent clips elicit more accurate memory performance than mismatching combinations of auditory and visual information. A plausible explanation for this benefit is that auditory and visual information are integrated into a more elaborate memory representation which elicits recognition performance that is higher than expected by a pure summation of independent retrieval cues. During perception, temporal synchrony between auditory and visual signals typically is the key factor triggering the integration process (e.g., Sekuler et al., 1997; van der Burg et al., 2008). If these perceptual processes transfer to long-term memory, semantically matching scenes would result in more accurate memory performance due to their higher synchrony of auditory and visual transients. Alternatively, the semantic match itself might trigger the integration process irrespective of temporal synchrony. In this case, a rapid comparison of the gist of the auditory and visual tracks (Potter, 1976; Potter & Levy, 1969) would precede the integration process. Only when this comparison confirms the actual match between the different sensory information they would be integrated. In the remaining experiments, we will address this question by investigating the influence of audio-visual synchrony on memory performance.

Experiments 3a and 3b

The results of Experiment 2 provided evidence that matching combinations of auditory and visual information elicit higher memory performance than mismatching combinations. This

observation agrees with the findings of studies on audio-visual integration during perceptual processes showing that semantically matching combinations of auditory and visual information increase the probability of audio-visual integration (see Chen & Spence, 2010, Grassi & Casco, 2010). A parsimonious explanation for this observation is that effects of audio-visual integration during the perception of the brief audio-visual scenes in our experiments transfer to long-term memory. Audio-visual integration during perceptual processes typically is highly sensitive to temporal offsets between auditory and visual information (e.g., Sekuler et al., 1997; van der Burg et al., 2008). In order to test whether the integration effect that we observed in Experiment 2 follows the same principles as audio-visual integration during perception, we maintained semantic congruency in Experiments 3a and 3b, but manipulated the temporal match between auditory and visual track by presenting half of the visual clips in reverse (backward), whereas auditory information was presented as before (forward). If audio-visual integration in memory representations is as sensitive to temporal offsets as audio-visual integration during perception, memory performance should decline with a reversed playback of the visual track. Alternatively, there might be qualitative differences between audio-visual integration for perception and memory. Here, it might be that semantically matching information from one modality can be integrated into the memory representation of another modality irrespective of temporal synchrony (i.e., a later integration in working memory). In this case, reversing the visual component of a brief audio-visual clip would leave memory performance unaffected.

Methods

Participants

Twenty-four new students participated in Experiment 3a (22 female; 18–27 years old). The final sample of Experiment 3b consisted of 21 new students (14 female; 18–33 years old). Data from three additional participants were excluded from the analysis due to chance level performance.

Apparatus, stimuli, and procedure

Apparatus, stimuli, and procedure were identical to Experiment 1 with the following exceptions: The retention interval was 1 day. We increased the number of clips to 1, 200 (in order to maintain 150 targets and 150 distractors for each condition). In Experiment 3a, these clips were allocated to four conditions that resulted from the combinations of the factors modality (visual vs. audio-visual) and the playback mode of the visual track (forward vs. reverse; auditory information was always forward). In Experiment 3b, all clips were presented audio-visually and we manipulated only the playback mode of the visual track (forward vs. reverse). We conducted Experiment 3b because we observed a slight numerical advantage for audio-visual scenes presented forwardly. Although this advantage was far from statistical significance, we wanted to eliminate any doubts by directly contrasting these two conditions with an increased amount of stimuli (i.e., 300 targets and 300 distractors per condition).

Results

The results of Experiment 3a replicated the superiority of audio-visual clips; however, we observed no influence of the reversal of the visual track (see Fig. S.3 in the Supplementary Material). A repeated measures ANOVA with sensitivity d' as the dependent variable and modality (visual, audio-visual), direction of the visual track (forward, reversed), and clip length (3 s, 3.5 s, 4 s) as independent variables confirmed that memory performance is more accurate for audio-visual than visual clips, $F(1, 23) = 62.18$, $p < .001$, $\eta^2_p = 0.72$, 95 % CI 0.49–0.82, and that memory performance increases with the length of the clip, $F(2, 46) = 17.87$, $p < .001$, $\eta^2_p = 0.43$, 95 % CI 0.20–0.58. With regard to the question of whether reversing the visual track of an audio-visual clip affects performance, we observed neither a main effect of the reversal of the visual tracks, $F(1, 23) = 1.35$, $p < .001$, $\eta^2_p = 0.06$, 95 % CI 0– 0.28, nor an interaction between the direction of the visual track and the learning condition, $F < 1$. Also, none of the other possible two- and three-way interactions reached significance, all Fs < 1.

The results of Experiment 3b confirmed the absence of any influence of the reversal of the visual track (see Fig. S.4 in the Supplementary Material). A repeated measures ANOVA with sensitivity d' as the dependent variable and direction of the visual track (forward, reversed) as well as clip length (3 s, 3.5 s, 4 s) as independent variables show no effect of the reversal of the visual track, $F < 1$, nor an interaction between the direction of the visual track and clip length, $F < 1$. However, as in the previous experiments, memory performance increased with an increasing duration of the clips, $F(2, 40) = 7.87$, $p = .001$, $\eta^2_p = 0.28$, 95 % CI 0.06– 0.45.

The results of these two experiments reveal a remarkable contrast between audio-visual integration with regard to long-term memory and audio-visual integration during perception. Whereas audio-visual integration during perception is highly sensitive to violations of audio-visual synchrony, memory performance seems to be unaffected by such manipulations (at least in the temporal range of our stimuli). A possibility for this pattern of results is that matching gist information of auditory and visual information rather than temporal synchrony triggers audio-visual integration with respect to memory representations. Such a mechanism also would predict more accurate memory performance for matching than mismatching scenes as observed in Experiment 2. In return, the auditory

track also would be integrated with the visual track when the visual track is presented in reversed order thus causing the results of the present experiments. A candidate for audio-visual integration that overcomes asynchronies between the distinct modalities is working memory (we will discuss other alternatives in the General Discussion), which might maintain information from one modality in order to integrate the delayed information from the other modality. We will further address this possibility in Experiments 4a and 4b.

Experiments 4a and 4b

The results of Experiments 2 and 3 provided an interesting overall pattern. Whereas mismatching auditory information resulted in less accurate memory performance for audio-visual scenes than matching auditory information, reversing the visual track of an audio-visual clip had no effect on memory performance. A potential explanation for this observation is that working memory processes might compensate for the delay between auditory and visual information. In Experiments 4a and 4b, we tested this possibility by further extending the temporal offset between auditory and visual information. In the testing session, all clips were presented audio-visually and simultaneously. In the learning session, however, auditory and visual track were either presented simultaneously or separately. In Experiment 4a, the separate presentation of matching auditory and visual tracks was operationalized in a sequential manner (i.e., immediately following each other). If working memory processes cannot compensate for temporal offsets between auditory and visual information, memory performance should decline with a sequential presentation of auditory and visual information. Because we did not observe an effect in this experiment, Experiment 4b was designed to demonstrate that there are limitations in the integration of asynchronous auditory and visual stimuli. In this experiment, matching auditory and visual tracks were randomly intermixed among all other clips of the learning session. Due to capacity limitations of the working memory, interference or decay (see Unsworth, Heitz, & Parks, 2008, for evidence suggesting interference) might prevent the integration of matching auditory and visual tracks that are intermixed among other tracks. Therefore, this manipulation should impair memory performance if working memory processes contribute to the observed robustness of memory representations against violations of audio-visual synchrony.

Methods

Participants

Twenty-four new students (15 female; 18–35 years old) participated in Experiment 4a. The final sample of Experiment 4a consisted of 23 new students (21 female; 18–28 years old). Data from one additional participant were excluded due to a chance level performance.

Apparatus, stimuli, and procedure

Apparatus, stimuli, and procedure were identical to Experiment 1 with the following exceptions: The retention interval was 1 day. In Experiment 4a, the auditory and the visual tracks of the clips were presented either simultaneously or one immediately after the other in the learning session. In one-third of the trials, the auditory information preceded the visual information whereas the auditory information followed the visual information in another third of the trials. In the remaining third of the trials, auditory and visual track were presented simultaneously. In the testing session, all clips were presented audio-visually (i.e., simultaneously). Prior to the learning session all participants were instructed that subsequent auditory and visual information belong together and that they would be presented simultaneously during testing.

In Experiment 4b, the auditory and visual tracks of a total of 600 clips (in order to maintain 150 targets and 150 distractors per condition) were presented either simultaneously or separately. In contrast to Experiment 4a, matching auditory and visual tracks that were presented separately were randomly intermixed among all clips (i.e., other unimodal as well as audio-visual clips) in the learning session. In the testing session, visual and auditory tracks were again presented simultaneously. As in Experiment 4a, participants were instructed that auditory and visual information of all clips would be presented throughout the learning session and that they would be presented simultaneously in the testing session. Prior to learning, participants were instructed that they would need to retrieve the combined clips during testing.

Results

In Experiment 4a, memory performance was equal for matching auditory and visual tracks that were either presented simultaneously or sequentially (see Fig. S.5 in the Supplementary Material). A repeated measures ANOVA with sensitivity d' as the dependent variable and learning condition (audio first, video first, audio-visual) as well as clip length (3 s, 3.5 s, 4 s) as independent variables revealed no significant effect of the learning condition, $F(2, 46) = 2.91$, $p = .065$, $\eta^2_p = 0.11$, 95 % CI 0–0.27. If anything, performance was (numerically) more accurate in the audio-first condition. In line with the other experiments, however, memory performance increased with an increasing duration of the clip, $F(2, 46) = 28.40$, $p < .001$, $\eta^2_p = 0.55$, 95 % CI 0.33– 0.67. There was no interaction between the learning condition and the length of the clips, $F(4, 92) < 1$.

In Experiment 4b, memory performance was superior for clips that were learned audio-visually to those that were learned separately (see Fig. S.6 in the Supplementary Material). A repeated measures ANOVA with sensitivity d' as the dependent variable and learning condition (audio-visual, separate) as well as clip length (3 s, 3.5 s, 4 s) as independent variables confirmed that memory performance increased with an increasing duration of the clip, $F(2, 44) = 13.00$, $p < .001$, $\eta^2_p = 0.37$, 95 % CI 0.13–0.53. Most importantly, we observed an effect of the learning condition, $F(1, 22) = 15.46$, $p < .001$, $\eta^2_p = 0.41$, 95 % CI 0.10–0.61, with higher sensitivity for clips that were learned audio-visually than those learned separately. The interaction between both variables did not reach significance, $F(2, 44) < 1$.

The results of Experiments 4a and 4b strikingly contrast with each other. Whereas memory for audio-visual clips remained intact when auditory and visual track were presented immediately following each other, memory performance declined when they were presented intermixed separately among the other clips. On the one hand, the results of Experiment 4a therefore confirmed the remarkable robustness of long-term memory performance against violations of audio-visual synchrony. On the other hand, however, the results of Experiment 4b show that memory performance is not immune against large temporal offsets. Memory performance finally declined when other clips of the experiment separated the matching auditory and visual clips. This finding shows that memory performance is sensitive to manipulations of the temporal offset between the two modalities but that the amount of offset that is necessary to corrupt long-term memory is much larger than one would expect based on research on audio-visual integration that explored perceptual tasks and performance. This suggests that the integration of auditory and visual information in memory experiments might arise on a post-perceptual level such as working memory. Here, it is possible that even subsequently presented tracks are integrated into the same memory representation if they match in their semantic structure or their overall gist. We will further elaborate on this possibility in the general discussion.

Please note, that randomly intermixing the clips depicts an extreme variant of increasing the temporal offset between matching auditory and visual information. On average, this results in hundreds of intervening tracks from distinct clips. Consequently, the aim of the experiment was to demonstrate that there are limitations in the integration of asynchronous auditory and visual information, the experiment was neither designed to investigate the mechanism causing this decline (i.e., decay vs. interference) nor to determine its fine grain time course. This issue needs to be addressed by future research evaluating a more systematic (and probably much smaller) increase in the temporal offset. On a related note, our experiment also does not actually show that audio-visual integration is eliminated with intervening other tracks. It remains possible that it just occurs less frequently because we did not include mismatching pairs of tracks in this experiment. Nevertheless, irrespective of these possibilities, Experiment 4b fulfills its major goal by demonstrating that audio-visual integration in memory representations is not immune against manipulations of the temporal synchrony of the perceptual information. Instead, such effects just arise at a larger timescale as would have been expected by the typical temporal profiles of effects of audio-visual integration during perceptual processes.

General discussion

In six experiments, we tested how auditory information contributes to long-term memory performance for dynamic scenes. Our results show that coinciding auditory information increases memory performance and that this increase in performance is larger when the auditory information semantically matches the visual information. Remarkably, memory performance was hardly affected by manipulations of audio-visual synchrony. Neither reversing the visual track of an audio-visual scene nor a sequential presentation of auditory and visual track impaired memory performance for filmic clips tested audio-visually. Only randomly intermixing matching auditory and visual tracks among the other clips finally impaired recognition performance. In principle, our results can be accessed either from a general perspective on human memory or from an audio-visual integration perspective.

From a traditional memory perspective, it is not surprising that memory traces for audio-visual scenes are more accurate than memory for purely visual scenes because the auditory track might provide additional retrieval cues (e.g., Rubin & Wallace, 1989; see also Hyman & Rubin, 1990). Alternatively, the auditory information might provide contextual information for the encoding of the visual track. Because matching contextual information during learning and testing improves retrieval success (study-test congruency; e.g., Godden & Baddeley, 1975; Marian & Neisser, 2000; Smith & Vela, 2001), the presence of auditory information during learning and testing might also explain superior memory performance for audio-visual scenes. In our experiment, we did not systematically manipulate study-test congruency; however, Experiments 4a and 4b did include conditions with and without study-test congruency. In these experiments, auditory and visual information were presented simultaneously during testing but simultaneously (i.e., congruent) or separately (i.e., incongruent) in the learning phase. In Experiment 4a, performance was unaffected by the sequential presentation in the learning phase although this manipulation disrupted study-test congruency. This finding might indicate that auditory information increases the number of retrieval cues rather than providing a context for encoding (note, however, that a

sequential presentation doubles learning time). However, the contrasting results of Experiments 4a (sequential presentation) and 4b (randomly intermixed presentation) also illustrate that matching auditory and visual information interact beyond a pure summation of retrieval cues because the number of retrieval cues does not differ between these conditions.

Because semantically matching auditory enhances long-term memory performance and extending the temporal offset between auditory and visual track by intermixing different tracks impairs memory performance, we argue that both auditory and visual track are integrated into a joint memory representation. However, one might contend that an explanation based on semantic priming (e.g., Collins & Loftus, 1975; Fischler, 1977; Lupker, 1984) provides an alternative approach toward our data. Here, the idea is that preceding information from one modality might facilitate the encoding process of the subsequent track in the remaining modality. On the one hand, semantic priming might well explain why recognition performance is worse for an intermixed presentation of all tracks than for a sequential presentation of matching auditory and visual tracks. On the other hand, other aspects of the data would remain unexplained. For instance, our findings of Experiments 4a and 4b are hard to reconcile with semantic priming only. If the robustness against violations of audio-visual synchrony stems from semantic priming, our findings would imply that auditory and visual primes would be equally efficient in enhancing processing in the remaining modality. This seems unlikely because previous research has identified asymmetries with respect to the order of auditory and visual information (e.g., Baggett, 1984; Huff & Schwan, 2008). Further, it seems unlikely that semantic priming alone is sufficient to boost memory for sequential presentations of auditory and visual information to the level of a simultaneous presentation. This argument is in line with a recent study of Chen and Spence (2010) who showed that semantically matching sound enhanced the encoding of pictures most when auditory and visual information were presented simultaneously. With regard to our experiments, findings such as these would suggest that conditions with audio-visual synchrony outperform conditions with temporal delays. However, because we did not observe such a superiority of the synchronous conditions, we consider it more likely that the matching auditory and visual information are integrated into a multimodal representation during a post-perceptual stage of information processing such as working memory.

Multimodal memory representations also have been discussed in the framework of episodic memory (e.g., Rubin, 2006). Because neuroimaging studies have highlighted the importance of the hippocampal structures for the integration and consolidation of episodic memory (Davachi, 2006; see also Eichenbaum, Yonelinas, & Ranganath, 2007), these structures might provide the physiological basis for the enhancing effect of auditory information on memory for dynamic scenes. With regard to this theoretical account, the novel findings of our experiments are that semantically matching information elicits more accurate memory performance than mismatching information and that the integration process is remarkably robust against violations of temporal synchrony. This robustness also contrasts with results from studies exploring audio-visual integration during perceptual and attentional processes.

Indeed, within the literature of audio-visual integration during perceptual tasks, there is remarkable agreement that audio-visual synchrony is the key factor for audio-visual integration (e.g., Chen & Spence, 2010; Meredith et al., 1987; Sekuler et al., 1997, van der Burg et al., 2008). Indeed, there is compelling evidence that audio-visual integration can arise on a perceptual level. For instance, in a study exploring event-related potentials, van der Burg, Talsma, Olivers, Hickey, and Theeuwes (2011) observed an auditory-induced enhancement of the neural response to visual transients as early as 50 ms after stimulus onset. Such early interactions between audition and vision are in line with neuropsychological findings illustrating that the primary visual cortex (V1) is also sensitive to auditory information (Falchier et al. 2002; Giard & Peronnet, 1999) and that audio-visual stimulation increases activity in multimodal areas but decreases activity in the corresponding unimodal areas of the brain (Bushara et al., 2003). Importantly, these early effects of audio-visual integration are highly sensitive to asynchronies between the sensory signals (Sekuler et al, 1997; van der Burg, Olivers, et al., 2008). In order to compensate for offsets in stimulus processing as well as physical transfer times (Lewald & Guski, 2004), audio-visual integration during perceptual tasks tolerates small deviations from perfect synchrony (e.g., Vroomen & Keetels, 2010). Although this temporal window for audio-visual integration varies across different types of stimuli (e.g., Fujisaki, Shimojo, Kashino, & Nishida, 2004; van Wassenhove, et al., 2007), it typically does not exceed a few hundred milliseconds. Therefore, the remarkable temporal tolerance in our data makes it unlikely that our integration effect stems from such perceptual processes. Instead, we suggest that the integration of auditory and visual information in our memory experiments occurred at a later stage of information processing (see also Koelewijn, Bronkhorst, & Theeuwes, 2010).

Note that because we used German-dubbed Hollywood movies, our experiments do not conclusively rule out the possibility that audio-visual integration during perception might also enhance subsequent memory processes. Although dubbing typically remains unnoticed (they are highly synchronous) by the vast majority of the observers, it is still possible that a natural match between lip movements and mouths would have further improved memory performance. Because lip movements were hardly the central aspect of our scenes, there are not enough clips to analyze such effects with our current data, but this might provide an incentive for future

research to assess whether perceptual integration of audio-visual information indeed has no effect on memory performance. At least for our set of experiments, we only observed integration during at the post-perceptual stages of information processing.

A straightforward candidate for audio-visual integration at a later stage than perception is working memory, which has been conceptualized very differently across different theorists. For instance, in Cowan's (1999) embedded-process model, a capacity-limited attentional process allows accessing and manipulating information of a larger subset of currently activated set of representations (McElree, 2001) within a unitary memory system (see also Chein & Fiez, 2010, for evidence in favor of the embedded-process model). With regard to our findings, this model provides a parsimonious explanation for the integration of temporally asynchronous auditory and visual information because it does not assume distinct modules for information from different modalities/codalities. Therefore, as long as the preceding unimodal trace remains within the time-limited state of activation, it might be integrated with the matching information from the remaining modality. Of course, our results can also be resolved within the modular approach of Baddeley (2000). In this conceptualization of working memory, an episodic buffer acts as a capacity-limited component of the working memory that binds originally distinct information into a multimodal representation. In this conceptualization, conscious awareness and focused attention play a major role for binding information in the buffer and for retrieving information from long-term memory (see also Baars, 2002). Assuming that the benefit in recognition performance for audio-visual scenes follows from the integration of matching auditory and visual tracks in the episodic buffer also might explain the temporal tolerance in audio-visual integration and why mainly matching information is integrated into a joint memory representation. Within this account, the first unimodal track might be maintained and integrated with the incoming second track into a common representation.

Finally, the object-oriented episodic record model (Jones, Beaman, & Macken, 1996) also provides some interesting ideas how to explain the integration of desynchronized auditory and visual information. Although originally conceptualized for smaller units of memory such as words, this model makes several assumptions that might be of relevance for the memorization of dynamic scenes in general. According to the model, distinct objects are the basic units of memory. A perceptual change in the state of one sensory stream (e.g., a brief moment of silence) triggers the formation of a new object (Jones, Macken, & Murray, 1993). If a similar organization of memory operates on a scene level, breaks in the semantic structure of the scene might trigger object/event boundaries (see also Zacks, Speer, Swallow, Braver, & Reynolds, 2007). Therefore, as long as subsequent auditory and visual scenes match in their semantic content, they might get bound into a single representation thus exhibiting the memory benefits of audio-visual integration.

Independent of the exact mechanism behind the integration of asymmetric auditory and visual information, such a late integration is reasonable from and ecological point of view because naturalistic events such as thunderstorms might vary in their audio-visual synchrony due to different transfer times of the corresponding physical signals. For events at larger distances, this temporal offset also might exceed the amount of a few hundred milliseconds of the temporal binding window of (perceptual) audio-visual integration. Nevertheless, our data as well as intuition suggests that human observers are able to integrate two events such as lightning and thunder for subsequent memory traces even at larger temporal offsets. Hereby, a central question for future research is to reveal the exact process behind audio-visual integration that is based on semantic congruency. A plausible candidate seems to be a rapid identification of the gist (e.g., Potter, 1976) of the scene as well as a comparison between auditory and visual gist. Only in case of matching gist information would both sources be integrated into a common memory representation. Otherwise, both perceptual streams would be interpreted as distinct events.

Our results provide first evidence that audio-visual integration in a memory task for realistic scenes relies on different processes than audio-visual integration during perceptual tasks. With regard to future research, this observation poses the research questions whether these different levels of audio-visual integration also fulfill distinct psychological functions. While audio-visual integration during perception might primarily contribute to an efficient exploration of the visual scenery by increasing the saliency of a coinciding transient (van der Burg et al., 2008), audio-visual integration during subsequent stages of information processing might foster subsequent memory representations (see also Swallow & Jiang, 2010). Regarding the audio-visually integrated memory representations, future research needs to explore the extent to which unimodal information remains accessible within these supposedly multimodal long-term memory representations as well as the role of attentional processing for the integration of information presented sequentially and potential capacity limitations (see van der Burg, Awh, & Olivers, 2013). Beyond audio-visual integration, our findings also pose questions regarding the perception and memorization of naturalistic scenes in general. For instance, dynamic information does not only enhance memory performance (Matthews et al., 2007), but also synchronizes viewing behavior between different observers (Dorr, Martinetz, Gegenfurtner, & Barth, 2010; Smith & Mital, 2013). Because oculomotor synchrony between encoding and retrieval might also foster memory performance, future research should also address the question how auditory information affects attentional and oculomotor

synchrony in dynamic scenes and how such processes affect memory for objects in a scene (Hollingworth, 2004; Hollingworth & Henderson, 2002).

Beyond the scope of memory research, our results also have implications for the learning sciences. In this field, prominent accounts of multimedia learning argue that integrated auditory and visual information in long-term memory reflects an essential aspect of learning complex concepts (Mayer, 2001; see also Schnotz, 2005). To our knowledge, our study is the first to demonstrate such integrated representations in long-term memory. Another implication for learning sciences might arise from individual differences. Here, the basic memory performance as well as the individual benefit from audiovisual integration might serve as a diagnostic tool to explore the potential of individual adaptions of learning environments (see Pashler, McDaniel, Rohrer, & Bjork, 2008).

Acknowledgments We thank Sandra Hermann for her help in generating stimuli. We also thank Sandra Hermann, Barbara Seitz, and Johanna Donath for their help in conducting the experiments.

Compliance with ethical standards

Author contributions Both authors developed the study concept and contributed to the study design. HSM programmed the experiments. Data collection was performed by student assistants under the supervision of both authors. Data analyses were performed by HSM. HSM drafted the manuscript and MH provided critical revisions. Both authors approved the final version of the manuscript for submission.

References

Amedi, A., von Kriegstein, K., van Atteveldt, N. M., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research, 166,* 559–571. doi:10.1007/s00221-005-2396-5

Baars, B. J. (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences, 6,* 47–52. doi:10.1016/S1364-6613(00)01819-2

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4,* 417–423. doi:10.1016/S1364-6613(00)01538-2

Baggett, P. (1984). Role of temporal overlap of visual and auditory material in forming dual media associations. *Journal of Educational Psychology, 76,* 408–417. doi:10.1037/0022-0663.76.3.408

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

Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science, 24,* 981–990. doi:10.1177/0956797612465439

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. doi:10.1163/156856897X00357

Buratto, L. G., Matthews, W. J., & Lamberts, K. (2009). When are moving images remembered better? Study–test congruence and the dynamic superiority effect. *The Quarterly Journal of Experimental Psychology, 62,* 1896–1903. doi:10.1080/17470210902883263

Bushara, K. O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., & Hallett, M. (2003). Neural correlates of cross-modal binding. *Nature Neuroscience, 6,* 190–195. doi:10.1038/nn993

Chein, J. M., & Fiez, J. A. (2010). Evaluating models of working memory through the effects of concurrent irrelevant information. *Journal of Experimental Psychology: General, 139,* 117–137. doi:10.1037/a0018200

Chen, Y. C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition, 114,* 389–404. doi:10.1016/j.cognition.2009.10.012

Cohen, M. A., Horowitz, T. S., & Wolfe, J. M. (2009). Auditory recognition memory is inferior to visual recognition memory. *Proceedings of the National Academy of Sciences, 106,* 6008–6010. doi:10.1073/pnas.0811884106

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82,* 407–428. doi:10.1037/0033-295X.82.6.407

Cowan, N. (1999). An embedded-process model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). New York: Cambridge University Press.

Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in Neurobiology, 16,* 693–700. doi:10.1016/j.conb.2006.10.012

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, *10*(10), 28, 1–17. doi:10.1167/10.10.28

Eichenbaum, H., Yonelinas, A. R., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience, 30,* 123–152. doi:10.1146/annurev.neuro.30.051606.094328

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *The Journal of Neuroscience, 22,* 5749–5759. doi:10.3758/CABN.4.2.117

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. doi:10.3758/BF03193146

Fischler, I. (1977). Semantic facilitation without association in a lexical decision task. *Memory & Cognition, 5,* 335–339. doi:10.2758/BF03197580

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. Y. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience, 7,* 773–778. doi:10.1038/nn1268

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience, 11,* 473–490. doi:10.1162/089892999563544

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66,* 325–331. doi:10.1111/j.2044-8295.1975.tb01468.x

Grassi, M., & Casco, C. (2010). Audiovisual bounce-inducing effect: When sound congruence affects grouping in vision. *Attention, Perception, & Psychophysics, 72,* 378–386. doi:10.3758/APP.72.2.378

Hollingworth, A. (2004). Constructing visual representations of natural scenes: The roles of short-and long-term visual memory. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 519–537. doi:10.1037/0096-1523.30.3.519

Hollingworth, A. (2005). The relationship between online visual representation of a scene and long-term scene memory. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, 31, 396–411. doi:10.1037/0278-7393.31.3.396

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 113–136. doi:10.1037/0096-1523.28.1.113

Huff, M., & Schwan, S. (2008). Verbalizing events: Overshadowing or facilitation? *Memory & Cognition, 36,* 392–402. doi:10.3758/MC.36.2.392

Hyman, I. E., & Rubin, D. C. (1990). Memorabeatlia: A naturalistic study of long-term memory. *Memory & Cognition, 18,* 205–214. doi:10.3758/BF03197096

Iordanescu, L., Grabowecky, M., & Suzuki, S. (2011). Object-based auditory facilitation of visual search for pictures and words with frequent and rare targets. *Acta Psychologica, 137,* 252–259. doi:10.1016/j.actpsy.2010.07.017

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review, 15,* 548–554. doi:10.3758/PBR.15.3.548

Jones, D. M., Beaman, P., & Macken, W. J. (1996). The object-oriented episodic record model. In S. Gathercole (Ed.), *Models of short-term memory* (pp. 209–238). London: Earlbaum.

Jones, D. M., Macken, W. J., & Murray, A. C. (1993). Disruption of visual short-term memory by changing-state auditory stimuli: The role of segmentation. *Memory & Cognition, 21,* 318–328. doi:10.3758/BF03208264

Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychologica, 134,* 372–384. doi:10.1016/j.actpsy.2010.03.010

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

Lewald, J., & Guski, R. (2004). Auditory-visual temporal integration as a function of distance: No compensation for sound-transmission time in human perception. *Neuroscience Letters, 357,* 119–122. doi:10.1016/j.neulet.2003.12.045

Lupker, S. J. (1984). Semantic priming without association: A second look. *Journal of Verbal Learning and Verbal Behavior, 23,* 709–733. doi:10.1016/S0022-5371(84)90434-1

Marian, V., & Neisser, U. (2000). Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology: General, 129,* 361–368. doi:10.1037/0096-3445.129.3.361

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

Mayer, R. E. (2001). *Multimedia learning*. Cambridge, UK: Cambridge University Press.

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 817–835. doi:10.1037/0278-7393.27.3.817

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *The Journal of Neuroscience, 7,* 3215–3229.

Papenmeier, F., Huff, M., & Schwan, S. (2012). Representation of dynamic spatial configurations in visual short-term memory. *Attention, Perception, & Psychophysics, 74,* 397–415. doi:10.3758/APP.72.3.628

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles concepts and evidence. *Psychological Science in the Public Interest, 9*, 105–119. doi:10.1111/j.1539-6053.2009.01038.x

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442. doi:10.1163/156856897X00366

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

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

Rubin, D. C. (2006). The basic-systems model of episodic memory. *Perspectives on Psychological Science, 1,* 277–311. doi:10.1111/j.1745-6916.2006.00017.x

Rubin, D. C., & Wallace, W. T. (1989). Rhyme and reason: Analyses of dual retrieval cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 698–709. doi:10.1037/0278-7393.15.4.698

Santangelo, V., & Spence, C. (2007). Multisensory cues capture spatial attention regardless of perceptual load. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 1311–1321. doi:10.1037/0096-1523.33.6.1311

Schnotz, W. (2005). An integrated model of text and picture comprehension. In R. E. Mayer (Ed.), *Cambridge hand- book of multimedia learning* (pp. 49–69). Cambridge: Cam- bridge University Press.

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385,* 308. doi:10.1038/385308a0

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature, 408,* 788. doi:10.1038/35048669

Smith, T. J., & Mital, P. K. (2013). Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision*, *13*(8), 16, 1–24.

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review, 8,* 203–220. doi:10.3758/BF03196157

Spence, C. (2010). Crossmodal spatial attention. *Annals of the New York Academy of Sciences, 1191,* 182–200. doi:10.1111/j.1749-6632.2010.05440.x

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73,* 971–995. doi:10.3758/s13414-010-0073-7

Standing, L. (1973). Learning 10000 pictures. *The Quarterly Journal of Experimental Psychology, 25,* 207–222. doi:10.1080/14640747308400340

Stevenson, R. A., Zemtsov, R. K., & Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Journal of Experimental Psychology: Human Perception and Performance, 38,* 1517–1529. doi:10.1037/a0027339

Swallow, K. M., & Jiang, Y. V. (2010). The attentional boost effect: Transient increases in attention to one task enhance performance in a second task. *Cognition, 115,* 118–132. doi:10.1016/j.cognition.2009.12.003

Unsworth, N., Heitz, R. P., & Parks, N. A. (2008). The importance of temporal distinctiveness for forgetting over the short term. *Psychological Science, 19,* 1078–1081. doi:10.1111/j.1467-9280.2008.02203.x

van der Burg, E., Awh, E., & Olivers, C. N. (2013). The capacity of audiovisual integration is limited to one item. *Psychological Science, 24,* 345–351. doi:10.1177/0956797612452865

van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: Non-spatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 1053–1065. doi:10.1037/0096-1523.34.5.1053

van der Burg, E., Talsma, D., Olivers, C. N., Hickey, C., & Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *NeuroImage, 55,* 1208–1218. doi:10.1016/j.neuroimage.2010.12.068

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception.

Neuropsychologia, 45, 598–607. doi:10.1016/j.neuropsychologia.2006.01.001

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception, & Psychophysics, 72,* 871–884. doi:10.3758/APP.72.4.87

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,* 273–293. doi:10.1037/0033-2909.133.2.273

Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005). Audio-visual simultaneity judgments. *Perception & Psychophysics, 67,* 531–544. doi:10.3758/BF03193329