Attention, Perception, & Psychophysics (2022) 84:427–441
https://doi.org/10.3758/s13414-021-02433-8

Auditory enhancement of visual searches for event scenes

Tomoki Maezawa[1] · **Miho Kiyosawa**[1] · **Jun I. Kawahara**[1]

Accepted: 21 December 2021 / Published online: 10 January 2022

Abstract
Increasing research has revealed that uninformative spatial sounds facilitate the early processing of visual stimuli. This study examined the crossmodal interactions of semantically congruent stimuli by assessing whether the presentation of event-related characteristic sounds facilitated or interfered with the visual search for corresponding event scenes in pictures. The search array consisted of four images: one target and three non-target pictures. Auditory stimuli were presented to participants in synchronization with picture onset using three types of sounds: a sound congruent with a target, a sound congruent with a distractor, or a control sound. The control sound varied across six experiments, alternating between a sound unrelated to the search stimuli, white noise, and no sound. Participants were required to swiftly localize a target position while ignoring the sound presentation. Visual localization resulted in rapid responses when a sound that was semantically related to the target was played. Furthermore, when a sound was semantically related to a distractor picture, the response times were longer. When the distractor-congruent sound was used, participants incorrectly localized the distractor position more often than at the chance level. These findings were replicated when the experiments ruled out the possibility that participants would learn picture-sound pairs during the visual tasks (i.e., the possibility of brief training during the experiments). Overall, event-related crossmodal interactions occur based on semantic representations, and audiovisual associations may develop as a result of long-term experiences rather than brief training in a laboratory.

Keywords Crossmodal · Attention · Audiovisual · Auditory enhancement · Visual search

Introduction

When we see a ball bouncing, we also hear a sound generated by the contact between the object and the floor surface. We are surrounded by multisensory information, as this daily-life example illustrates, and such information can be integrated into a supramodal form of representation (e.g., Raij et al., 2000), leading to a coherent perception of objects and event scenes. During this integration, a stimulus that is experienced via one modality (e.g., audition) modulates the early processing of information specific to another modality, such as vision (Chen & Spence, 2011). For example, the processing of visual stimuli is facilitated so that a salient but task-irrelevant sound attracts attention to its source location and improves visual processing at that location (Eimer & Driver, 2001; Feng et al., 2014; Spence & Driver, 1997; Störmer, 2019; Störmer et al., 2009). Earlier findings regarding auditory influences on visual attention have indicated critical determinants of multisensory enhancement in the context of the temporal and spatial proximity of stimuli. Specifically, strong enhancement occurs when the two stimuli are derived from the congruent location simultaneously (e.g., Stein & Stanford, 2008).

Notably, multisensory enhancement not only depends on spatial or temporal congruence but is also associated with uninformative cues regarding the timing and location of the target onset. Auditory-visual interactions are based on higher-level semantic associations (Kvasova et al., 2019), rather than spatiotemporal overlap in perceptions, such that synchronous presentation of auditory cues (e.g., bark) supports identification of the visual features (e.g., picture of a dog) of related objects (Chen & Spence, 2010; Molholm et al., 2004; von Kriegstein et al., 2005). This effect is realized because the characteristic sound is coded into a representation of an individual object (an animal or dog) that is semantically congruent with that obtained from the visual stimulus (Iordanescu et al., 2008). Likewise, sounds improve

✉ Tomoki Maezawa
shortfinned@gmail.com

[1] Department of Psychology, Hokkaido University, N10W7, Kita, Sapporo 060-0810, Japan

detection and localization performance during search tasks when the vision and audition convey the same semantic representation with respect to the target objects (Iordanescu et al., 2008, 2010, 2011). The effects of semantic congruence on audiovisual cues develop through multisensory experiences of co-occurrence of these object features repeated in laboratories and real life (Iordanescu et al., 2011; Smith et al., 2007; Zweig et al., 2015). Strong multisensory enhancement would occur in an object-specific manner so that two stimuli share a similar semantic representation of a particular object derived from the same multisensory experiences (Iordanescu et al., 2008; Laurienti et al., 2004).

The aforementioned findings regarding audiovisual interaction suggest that characteristic sounds aid visual detection of individual target objects amid cluttered scenes in the environment (Iordanescu et al., 2010, 2011; Kvasova et al., 2019). This behavioral benefit may be observed in the laboratory. Some researchers (Iordanescu et al., 2008) have used pictures of archetypal objects as visual targets that were clearly defined, readily identifiable, and comprised of a single exemplar per concept (i.e., nouns such as dogs, coins, or keys). Therefore, a one-to-one correspondence with characteristic prototypical sounds could be formed. It is unclear whether multisensory enhancement occurs only when viewing audiovisual stimuli associated with an object that has salient features. For example, object- and identity-based (Kvasova et al., 2019) crossmodal effects (e.g., auditory facilitation of visual search) may be weakened if the visual targets are required empirical knowledge to form coherent auditory-visual associations. In the present study, we examined the crossmodal effects of search facilitation by simulating the visual search for pictures representing scenes of motion events, the meanings of which were derived from contextual information rather than an object specified by a discrete image. Specifically, we used temporally trimmed pictures of a series of motions, such as running, burning, or cutting, in which the event was conceptualized as verbs (detailed below). The loss of information about temporal changes (i.e., compared to video clips) that is important for analyzing motion scenes weaken the auditory-visual associations. However, based on our daily life experience, we expected that, when such event-related pictures are viewed, the sounds indicating congruent motion can help identify concepts such as "jingling (i.e., verbs)," regardless of what the jingling objects are (i.e., nouns). This implies that comprehensive representations are retrieved when viewing event scenes comprising multiple discrete objects and background elements (Henderson & Hollingworth, 1999), and that auditory information would facilitate this identification process.

Given that semantic congruence is a critical determinant of multisensory enhancement, the audition would interact with the visual processing of an event scene when the two stimuli share the same experience of an event, such as "a key is jingling" or "a person is running." Neuroimaging studies may yield evidence in support of this idea (e.g., Barraclough et al., 2015; Beauchamp et al., 2004), demonstrating that neural activities reflecting the visual integration of object features exhibit a similar pattern to those elicited by the multisensory integration of actions related to events. Specifically, the integration of object features recruits the posterior superior temporal sulcus and middle temporal gyrus, in which significant neural activity is detected when participants receive the presentation of simultaneous audiovisual items (Beauchamp et al., 2004; Hein et al., 2007). However, the polysensory area of the superior temporal sulcus is also activated during recognition of visual events accompanied by sounds such as tearing paper (e.g., Barraclough et al., 2015). This overlap of neural networks relevant to the processing of objects and events (i.e., actions) implies that audiovisual cues can be integrated into a single representation referring to events containing multisensory features (see also Maniglia et al., 2017), leading to multisensory interaction such that one stimulus from a modality would facilitate processing of another stimulus from a modality (Iordanescu et al., 2008, 2010, 2011).

With respect to object-based crossmodal effects, it is important to account for crossmodal interference by distractors when searching for visual targets. This is because involuntary attention shifting to distractors reflects auditory-visual interactions originating from automatic processes. Iordanescu et al. (2008) failed to detect interference by characteristic sounds during visual searches, when reporting that object-based crossmodal facilitation occurred in a top-down goal-directed manner (see also Kvasova et al., 2019). However, this goal-directed control may be situation-dependent (Chen & Spence, 2010), such that the goal-directed mechanism would be insufficiently activated if searching for visual targets that did not predict the features to be retrieved. Accordingly, potential automatic interference by sounds, hindered by the goal-directed mechanism, would occur during the retrieval of conceptual images represented by verbs.

In summary, the present study examined whether visual search performance was improved or degraded by the presentation of a sound associated with an event with respect to specific actions. We used a visual search task similar to that used by Iordanescu et al. (2008), wherein participants were required to indicate the location of a target as swiftly as possible. Participants localized pictures depicting an event scene of actions during the search task. The search array consisted of four pictures comprising one target and three distractors. The display was accompanied by one of three types of sounds: a sound congruent with a target, a sound congruent with a distractor, and a sound unrelated to the pictures in the search display (as the control condition). We expected that presentation of auditory cues would enhance schemes of visual representation, resulting in target-congruent sound

enhancing visual target localization, whereas target-incongruent sound would impair target localization, consistent with the idea of automatic crossmodal interactions. We replicated crossmodal enhancement (or interference) using sound-absent conditions and white-noise presentation as controls. The first set of experiments (1–3) aimed to examine whether crossmodal facilitation would occur when using the same procedure as Iordanescu et al. (2008). In the subsequent experiments (4–6), we modified the visual search task to exclude the potential confounding effect of learning about picture-sound pairs during the task trials. Moreover, these experiments aimed to reveal the consistency in interference effects. Auditory enhancement or interference with the visual search, if any, should be reflected in improved or degraded search performance when a sound corresponding to a target or distractor is presented, relative to the control condition, in which an irrelevant sound accompanies the visual stimulus.

Experiment 1

Experiment 1 examined whether sound cues would enhance visual search for event scenes semantically or contextually related to the sounds. To achieve this goal, participants searched for a target picture symbolizing concepts of critical actions (e.g., a person running) among distractor pictures. During the search task, participants identified the location in which the target picture was displayed in the search array with synchronous presentation of three types of auditory cues, such as a sound congruent with the target (target-congruent sound), a sound congruent with a distractor (distractor-congruent sound), or a sound unrelated to pictures in the search display (unrelated sound as a control). We predicted that the characteristic sounds would improve visual search performance for the event scenes when the sounds and pictures shared similar semantic representations of the events. By contrast, we expected that performance would decline when the sound cues were congruent with distractor pictures.

Method

Participants Twenty-five undergraduate and graduate students (14 males and 11 females; mean age = 20.0 years, range = 18–23 years) participated in this experiment for monetary compensation or course credit. The sample size was determined based on previous studies of visual searches, which had sample sizes ranging between 16 and 38 participants (e.g., Iordanescu et al., 2008, 2010, 2011; Kvasova et al., 2019). All participants reported having normal color vision and normal or corrected-to-normal visual acuity. None of the participants was hearing impaired, according to self-report. They provided written informed consent prior to each experiment. All experiments in this study were approved by the Human Research Ethics Committee of Hokkaido University.

Apparatus and stimuli Visual stimuli were displayed on an LCD monitor (100-Hz refresh rate, 1,920 × 1,080 pixels; XL2411T, BenQ), and the experiment was controlled using Psychophysics Toolbox 3.0 extensions (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) for MATLAB software (version R2019a, MathWorks). Participants sat in a height-adjustable chair, and a viewing distance of approximately 57 cm was maintained. The auditory stimuli were presented via headphones (5–22,000 Hz; MDR-XB450, SONY) at a comfortable listening level (approximately 65 dB[A]).

This study defined 20 events related to object actions that can be specified by verbs, such as break, write, close, stir-fry, cut, bite, burn, turn (e.g., turn pages), sweep, tear (e.g., tear paper), slap, giggle, boil, fly, run, cry, type (e.g., type letters), sneeze, knock, and pour. Twenty color pictures (Fig. 1) and sound clips related to verb-associated events were collected in advance through web searches. The pictures, subtended 14.06° (W) × 9.44° (H) in visual angle on the monitor, depicted event scenes comprising multiple objects and background elements. The audio clips produced typical sounds corresponding to the event action (e.g., the sound of glass shattering for "break"; the sound of a pen rubbing against paper for "write"; the sound of a door closing and latching for "close," etc.). Owing to the differences in the nature of these events, the sounds varied from 592 to 1,555 ms in duration (mean = 1,367 ms; standard deviation = 274 ms). All sound clips were monophonic and did not provide any spatial information aurally, such as might occur via the interaural time difference of binaural hearing.

Familiarization tasks Prior to the visual search task (see below), two consecutive tasks were administered to familiarize participants with the association between the event scenes and their associated verbs in line with the procedure described in Mädebach et al. (2017) and Zweig et al. (2015). The familiarization aimed to improve participants' compliance with the task requirement for every trial to search for a scene picture that represented the verb that directly proceeded the search display (see Fig. 2). The familiarization was divided into two phases. In the first phase, participants were exposed to the picture-verb associations used in the main search task. The first familiarization phase lasted until all picture-verb pairs had been presented. Figure 2a illustrates a schematic example of the first phase of the familiarization. Every trial began with a fixation cross for 500 ms, followed by presentation of a picture in the center of the screen for 1,000 ms. The picture was accompanied by synchronous presentation of an audio clip that was congruent

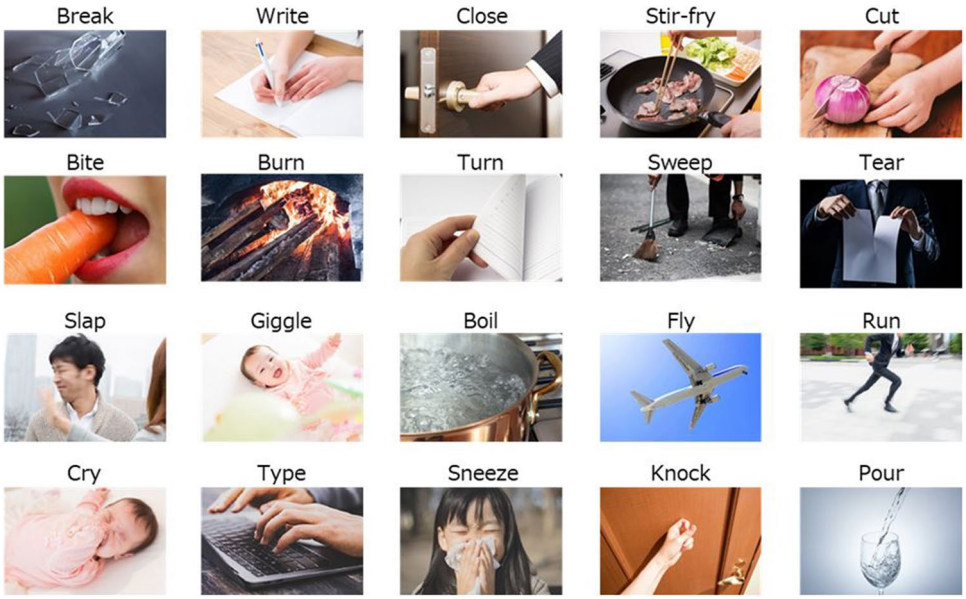

Fig. 1 Pictures of verb-associated event scenes

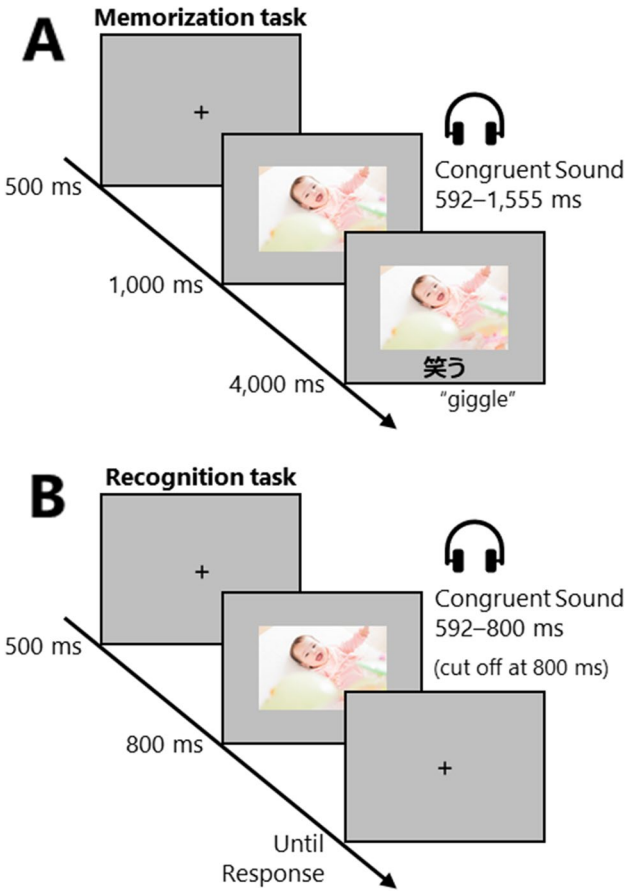

Fig. 2 Schematic examples of a trial sequence in the two familiarization tasks. *Note*. (**A**) The first familiarization phase to associate the sounds with the corresponding images. (**B**) The second phase, recognition of the familiarized pairs

with the scene image. The sounds were not truncated during familiarization, and thus their durations differed depending on the nature of the sound (592–1,555 ms). After the audiovisual stimuli has been presented, a verb associated with the event was visually presented in Japanese below the picture for 4,000 ms. Participants were required to memorize the association between the event and the verb while vocalizing the verb. The order in which the audiovisual stimuli were presented was fixed across participants in the following order: break, write, close, stir-fry, cut, bite, burn, turn, sweep, tear, slap, laugh, boil, fly, run, cry, type, sneeze, knock, and pour. The familiarization's first phase lasted until all pictures had been presented to the participant (i.e., 20 trials).

Figure 2b illustrates a schematic example of the second familiarization task. Each trial began with a fixation cross, presented for 500 ms, followed by the presentation of a picture in the center of the screen (for 800 ms) with synchronous onset of an audio clip corresponding to the image. The audio clip was truncated at 800 ms to align the offset of the picture with that of the audio clip. At the end of each test trial, participants verbally reported the corresponding verb. The audiovisual stimuli were presented in the same order as in the first familiarization task. The experimenter orally confirmed that the participants were able to correctly retrieve the associated verbs; the experimenter corrected the participant once if the reported verb differed from the familiarized one (e.g., "closed" rather than "shut" upon viewing the image a of door). The practice was not repeated regardless of whether the participant could or could not correctly retrieve the associated verbs.

Visual search task Each search task trial began with a fixation point (0.33° in diameter) at the center of the screen for 500 ms (Fig. 3), followed by a verb in the center of the screen for 500 ms to inform participants to search for a target during the trial. The verb representing the target event was presented in Japanese (e.g., "割れる" for "break"). A blank screen followed the verb display for 500 ms. After the blank screen, a search display consisting of one target picture and three distractor pictures was presented for 670 ms in the four quadrants at 10.7° eccentricity from the center of the screen to the center of each picture. The centers of the four pictures were separated by 3.33° vertically and 3.33° horizontally. In the search display, a target picture was presented in one of the quadrant positions with equal probability across trials. The search display was accompanied by synchronous onset of an audio clip. Audio clips longer than 670 ms were terminated at that length to align with the offset of the search display used in a previous study (Iordanescu et al., 2008). The sound was randomly assigned to one of three sound conditions: a sound congruent with the target (target-congruent sound), a sound congruent with a distractor that was presented in the quadrant diagonally opposite the target across the fixation point (distractor-congruent sound), and another of the 20 audio clips not included in the search display (unrelated sound). Participants were required to locate the target picture (lower left, upper left, lower right, or upper right) as quickly as possible by pressing a numeric key: 1, 4, 2, or 5, respectively. Participants used the index and middle fingers of both hands to respond (the left forefinger for 1, the left middle finger for 4, the right forefinger for 2, and the right middle finger of the right hand for 5).

The visual search task consisted of four repetitions of an experimental block of 60 trials (240 trials in total). Within the block, each of the 20 sounds was presented once per the sound condition of the target-congruent, distractor-congruent, or unrelated sound. The order of the trials was randomly determined across participants. In every trial, four pictures of event scenes were selected randomly from the pool of pictures without replacement in a given search display. Each of the 20 pictures was presented as a target once for each sound condition within a block. Three remaining images were equiprobably selected from the 20 pictures as distractor stimuli across the trials. Participants completed ten practice trials before undertaking the 240 experimental search trials.

Results

Our primary dependent measure was response time. Error trials (3.68% of all trials) were excluded from the reaction-time analysis. To eliminate outliers, trials with response times that deviated by more than 1.5 times the interquartile range beyond Tukey hinges (i.e., the 25th and 75th

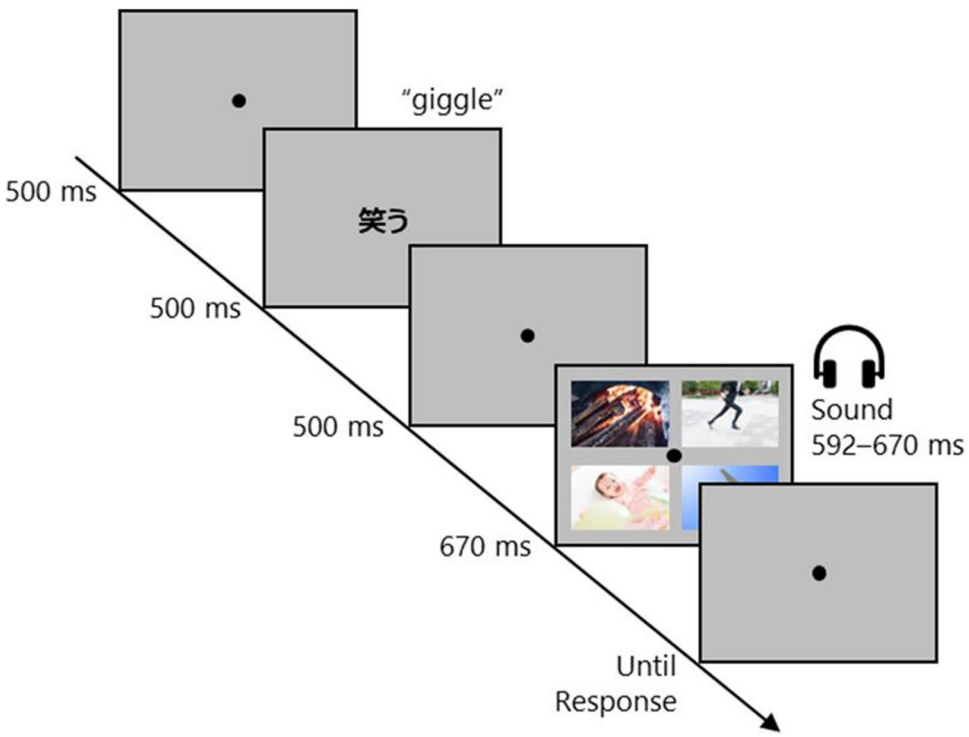

Fig. 3 An example of a trial sequence in a visual search task

percentiles) were excluded from analysis (4.27% of all trials). A mean response time for the correct trials was calculated for each sound condition (target-congruent, distractor-congruent, and unrelated), as shown in Fig. 4.

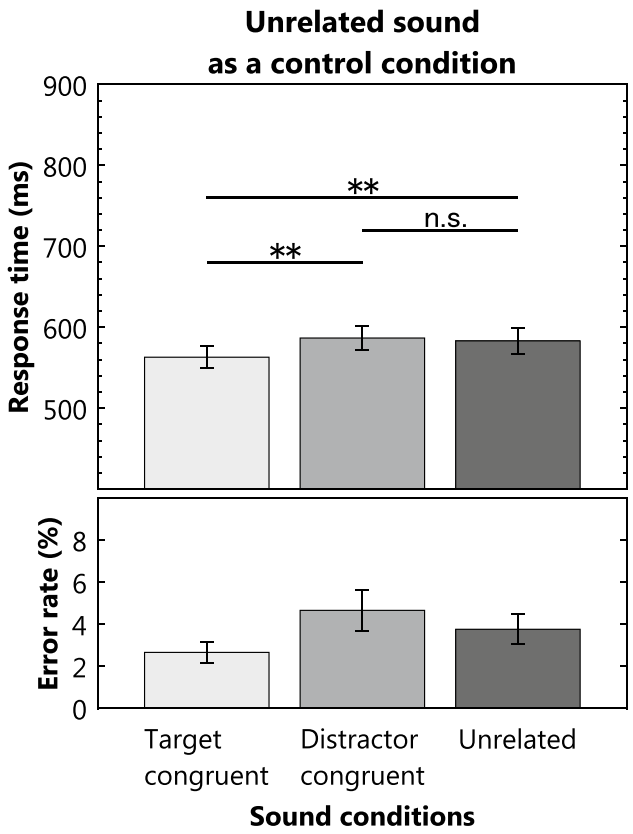

Fig. 4 Mean response times under the target-congruent and distractor-congruent conditions compared to those under the unrelated condition in Experiment 1. *Note*. The error bars represent the 95% confidence interval (**$p < .01$)

A one-way analysis of variance (ANOVA) of mean response times for the correct trials, including the sound condition as a within-subject factor with Chi-Muller's sphericity correction, revealed a significant main effect of sound condition [$F(2, 48) = 9.21$, $p < .001$, $\eta_G^2 = .020$, $\varepsilon = 1.00$]. Holm's multiple comparison tests indicated that the mean response time was shorter under the target-congruent condition (562.93 ms) than under the unrelated (583.12 ms) [$t(24) = 3.42$, $p = .005$] and distractor-congruent (586.63 ms) conditions [$t(24) = 3.68$, $p = .004$]. However, the mean response time under the distractor-congruent condition was comparable to that under the unrelated condition [$t(24) = 0.64$, $p = .528$]. An ANOVA of error rates was performed with sound condition as a within-subject variable, and all results of this and subsequent experiments are summarized in Table 1.

Discussion

In Experiment 1, visual searches for event-action scenes were swifter when the search stimuli were accompanied by target-congruent sounds than when they were accompanied by distractor-congruent sounds and sounds unrelated to the search stimuli. Error rates were not modulated by the sound types, and no speed-accuracy trade-off was found during the task. Thus, auditory enhancement of visual search performance occurred when the target and its semantically related sound represented the same event concept. Although sound-image-verb associations may be developed by practicing familiarization tasks, in Experiments 4–6 (detailed below), we demonstrated that

Table 1 The results of ANOVAs on error rates

Study	ANOVA				Post-hoc test			*d*
	F-ratio	p-value	η_G^2	ε	Pairs	t-value	p-value	
Experiment 1	$F(1.79, 42.94) = 2.68$	.085	.046	.895				
Experiment 2	$F(2, 48) = 3.13$	.053	.038	1.00				
Experiment 3	$F(1.79, 42.99) = 0.44$	.623	.006	.896				
Experiment 4	$F(2, 48) = 3.94$	.026*	.054	1.00	Target-congruent vs. Control	$t(24) = 0.85$	.404	.182
					Target-congruent vs. Distractor-congruent	$t(24) = 2.57$	.051	.536
					Distractor-congruent vs. Control	$t(24) = 1.89$	.141	.370
Experiment 5	$F(1.97, 47.18) = 7.57$	.002**	.089	.983	Target-congruent vs. Control	$t(24) = 2.28$	.063	.445
					Target-congruent vs. Distractor-congruent	$t(24) = 4.06$	.001**	.776
					Distractor-congruent vs. Control	$t(24) = 1.63$	.117	.312
Experiment 6	$F(1.68, 40.3) = 3.46$	.049*	.060	.840	Target-congruent vs. Control	$t(24) = 0.14$	.889	.033
					Target-congruent vs. Distractor-congruent	$t(24) = 1.92$	.133	.470
					Distractor-congruent vs. Control	$t(24) = 2.36$	.080	.506

the auditory enhancement may be due to long-term experiences, as well as short-term learning of specific sound-picture pairs (Iordanescu et al., 2011; Smith et al., 2007; Zweig et al., 2015).

Contrary to the effects of semantic congruence, an interference effect of semantically incongruent sounds was not observed in Experiment 1. Specifically, response times under the distractor-congruent condition were comparable to those under the control condition, in which unrelated sounds accompanied the visual stimuli. This pattern of results is similar to that reported by Iordanescu et al. (2008), who suggested that visual searches for distinct visual target objects, such as dogs or cats, designated by preceding words, would be driven in a goal-directed top-down manner (e.g., Tomita et al., 1999). In other words, participants in Iordanescu et al. (2008) might activate the stored knowledge of dogs and evoke prototypical visual representations of the objects while protecting against interference from unrelated information (e.g., non-targets and unrelated sounds) by using the word designating the target object at the beginning of a trial (Chen & Spence, 2010).

However, we argue that goal-directed control would not be activated in the present experiment, despite the provision of a target event by a verb presented before the search display. Because the event scenes in Experiment 1 contained multiple visual elements in terms of backgrounds and objects (Fig. 1), the verb would not elicit clear imagery (i.e., prototypes of the presentation) regarding the scenes before undertaking the search. In fact, the interference effects of multisensory interactions have been demonstrated using visual tasks such as object identification (Laurienti et al., 2004; Suied et al., 2009), indicating that task-irrelevant sounds cannot be ignored. Notably, whether the interference effects are observed (e.g., Suied et al., 2009) or not (e.g., Molholm et al., 2004) depends on baseline performances under control conditions. Chen and Spence (2010) found that semantically incongruent sounds interfered with visual identifications compared to performances under a control condition in which white noise was presented, whereas the interference did not occur under a no-sound control condition. The apparent inconsistency of the results in terms of interference might be attributed to a shortened response-time baseline in the sound-absent condition because participants were more alert in response to the onset of stimulus when sounds were present than when no sound was presented. Specifically, response times under a silent control condition increased to a comparable level with those under a semantically incongruent condition, suggesting no auditory interference. Experiment 1 may have been afflicted by this potential baseline inflation. That is, Experiment 1's result implies that performances under the unrelated sound condition might have been delayed to a level comparable to those under the distractor-congruent condition because the unrelated sounds were selected from the 20 audio clips that were semantically inconsistent with the target images.

To eliminate the possibility of baseline inflation, we modified a control condition in Experiment 2 such that search stimuli appeared with a synchronous white-noise presentation. The white noise under the control condition should exclude semantic or contextual influences on a performance baseline as compared with the unrelated condition in Experiment 1. The procedure of the visual search task in Experiment 2 was identical to that used in Experiment 1, with the exception of the white-noise condition.

Experiment 2

Experiment 2 was designed to replicate the audiovisual interaction observed in Experiment 1 during visual searches for event scenes. We examined whether visual search was enhanced or degraded when semantically congruent or incongruent sounds were synchronously presented in a search display. Participants performed the search task in a procedure identical to that followed in Experiment 1, with the exception that the control condition was accompanied by presentation of white noise. We predicted that auditory enhancement would occur when target-congruent sounds were present, and that auditory interference would occur when distractor-congruent sounds were presented during the search display. Alternatively, if the pattern of the results in Experiment 1 reflected the goal-directed control suggested by multisensory studies (Iordanescu et al., 2008, 2010; Molholm et al., 2004), task-irrelevant sounds should be ignored during the search in the distractor-congruent condition.

Method

A new group of 26 students (16 males and ten females; mean age = 20.34 years, range = 18–25 years) participated in this experiment. One participant was excluded due to failure to remain awake and comply with the study tasks. The apparatus, stimuli, and procedure were identical to those used in Experiment 1, with the exception that a search display under the control condition was accompanied by Gaussian white noise for 670 ms. Prior to undertaking the visual search task, participants completed the two familiarization tasks to associate the event scenes with the corresponding verbs.

Results

Error trials (3.80% of all trials) were excluded from the reaction-time analysis. Moreover, response times outside the 1.5-times interquartile range beyond the 25th and 75th

percentiles were defined as outliers (3.07% of all trials). Figure 5 shows the mean response times for the correct trials averaged separately for the target-congruent, distractor-congruent, and white-noise conditions. An ANOVA of the mean response times revealed a significant main effect of sound condition [$F(2, 48) = 27.46$, $p < .001$, $\eta_G^2 = .033$, $\varepsilon = 1.00$]. Multiple comparisons found that the mean response time was shorter under the target-congruent condition (564.86 ms) than under the white-noise (579.23 ms) [$t(24) = 2.41$, $p = .024$] and distractor-congruent (612.80 ms) conditions [$t(24) = 7.01$, $p < .001$]. Furthermore, the mean response time was longer under the distractor-congruent condition than under the white-noise condition [$t(24) = 4.75$, $p < .001$].

Discussion

Experiment 2 included the white-noise condition as a control, in which participants performed visual searches for event scenes with synchronous presentation of white noise. The results showed that response times decreased in the presence of target-congruent sounds compared to those in the presence of white noise, indicating auditory enhancement of the search (Iordanescu et al., 2008, 2010; Molholm et al., 2004). Contrary to the pattern of results in Experiment 1, in Experiment 2 response times were longer under the distractor-congruent condition than under the noise condition, indicating interference effects of semantically incongruent sounds (Chen & Spence, 2010; Laurienti et al., 2004; Suied et al., 2009). Accordingly, the event-related auditory stimuli interacted with the visual representation based on semantic congruence/incongruence.

Although unlikely, the interference effects in Experiment 2 might not have reflected multisensory interactions. Instead, the interference might have reflected shortened response times in the control condition owing to the sharper auditory onset of white noise than that of the redundant target- and distractor-congruent sounds (Iordanescu et al., 2011). Specifically, the participants may have been more alert under the white-noise condition than under the other conditions (Chen & Spence, 2010). To rule out this alternative explanation for the results of Experiment 2, we conducted Experiment 3, in which no sound was presented as an alternative to the white-noise presentation. In Experiment 3, the search task was identical to that used in Experiments 1 and 2, with the exception of sound being absent.

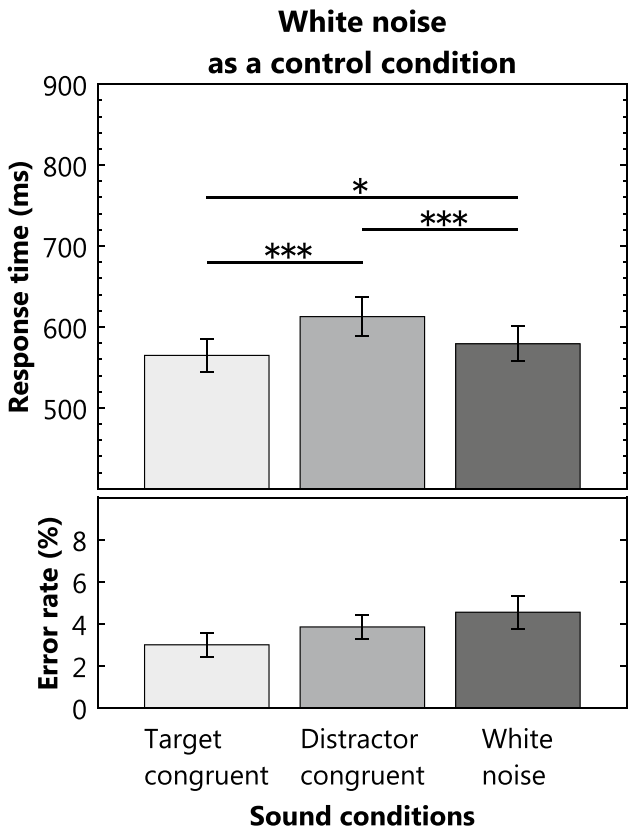

Fig. 5 Mean response times under the target-congruent and distractor-congruent conditions compared to those under the white-noise condition in Experiment 2. *Note*. *$p < .05$, ***$p < .001$

Experiment 3

Experiment 3 was designed to replicate the audiovisual interaction shown in Experiment 2 during visual searches for event scenes. We examined whether visual search was enhanced and/or degraded when semantically congruent and incongruent sounds were synchronously presented with a search display. This experiment included the sound-absent condition in a visual search task, during which participants searched for a target scene with no sound onset, to measure participants' baseline performance. We predicted that we would observe both enhanced and degraded performances.

Method

A new group of 25 students (13 males and 12 females; mean age = 20.20 years, range = 18–30 years) participated in this experiment. The apparatus, stimuli, and procedure were identical to those used in Experiments 1 and 2, with the exception that no sound was presented under the control (sound-absent) condition. The two familiarization tasks were conducted for each participant before they undertook the visual search task.

Results

Error trials (3.57% of all trials) were excluded from the reaction-time analysis. Data from trials with response times that deviated by more than 1.5 times the interquartile range beyond the 25th and 75th percentiles were excluded (4.58% of the trials). The mean response times for the correct trials were calculated per the sound condition (target-congruent, distractor-congruent, and sound-absent; in Fig. 6). An ANOVA on the mean response times revealed a significant main effect of the sound condition [F (1.44, 34.53) = 23.73, $p < .001$, $\eta_G^2 = .024$, $\varepsilon = .72$]. Multiple comparisons revealed that the response time was shorter under the target-congruent condition (526.05 ms) than under the sound-absent (542.98 ms) [$t(24) = 4.26$, $p < .001$] and distractor-congruent (565.17 ms) [$t(24) = 5.43$, $p < .001$] conditions. The mean response time was longer under the distractor-congruent condition than under the sound-absent condition [$t(24) = 4.08$, $p < .001$].

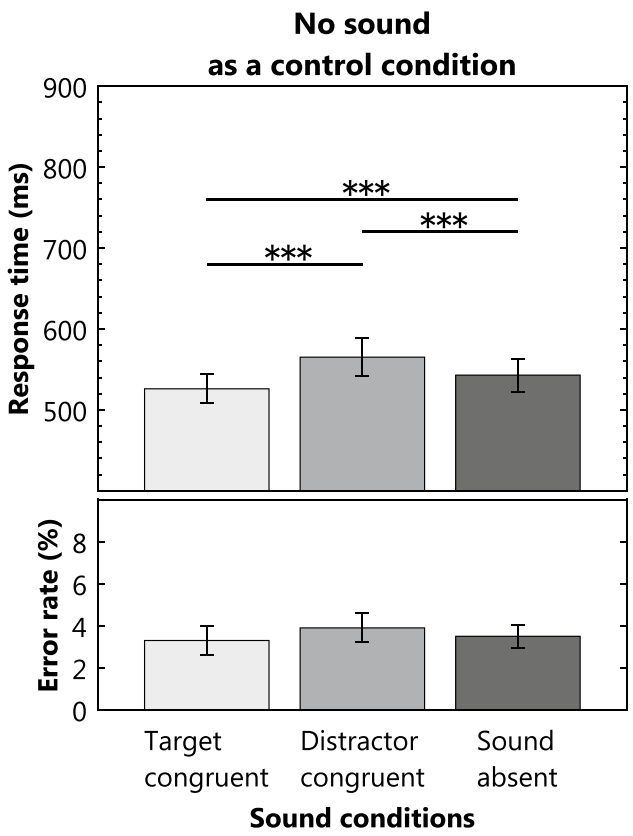

Fig. 6 Mean response times under the target-congruent and distractor-congruent conditions compared to those under the sound-absent condition in Experiment 3. *Note*. ***$p < .001$

Discussion

Experiment 3 replicated both auditory enhancement and interference of visual searches, implying that task-irrelevant auditory stimuli interacted with visual representation to enhance semantic representation of action events. Notably, response times were delayed when the auditory cues were congruent with a distractor picture compared with when the cues were absent. The interference effect reflects performance deficits by semantic incongruence during multisensory interactions (Chen & Spence, 2010), rather than baseline inflation.

Although the results demonstrate enhanced search for event scenes, the question remains as to whether audiovisual associations (e.g., a picture of a person running is associated with the sound of their footsteps) develop through long-term multisensory experiences encountered in real life (e.g., Smith et al., 2007) or short-term training (e.g., Zweig et al., 2015) during the familiarization and subsequent search tasks (see the Method section in Experiment 1). For example, integration of repeatedly co-occurring audiovisual signals can increase their intensity and coherence (Iordanescu et al., 2008; Smith et al., 2007; Zweig et al., 2015). Associating auditory signals with visual events (or objects) is an ecologically important mechanism to enhance visual identification and search for targets of interest (Chen & Spence, 2010; Iordanescu et al., 2011). However, Zweig et al. (2015) reported that audiovisual associations, such as novel face-voice pairs, could be established in the course of only two brief training sessions in which participants memorized and identified pairs in approximately 20 trials per task. The two familiarization tasks in the present study were similar to the training tasks used by Zweig et al. (2015) in that participants were exposed to a combination of event-related pictures accompanied by sounds and corresponding verbs prior to undertaking the visual search task. Furthermore, the same auditory and visual stimuli as those used in the familiarization were presented multiple times during the search (at least 12 trials per sound clip as a target).

Thus, we conducted Experiments 4–6 to rule out the possibility that participants learned picture-sound pairs during the familiarization and visual search tasks, and would replicate the audiovisual interactions in terms of enhancement and interference demonstrated in Experiments 1–3. We used multiple exemplars of a single scene for each of the 20 events in these experiments. That is, we used different multiple pictures representing the same event scene of "break," and we also did for the other events (i.e., "write," "close," etc.). This manipulation was intended to avoid repetitive presentation of the same 20 pictures between and within the familiarization and search

tasks. With the exception of the diverse pictures of event scenes, the search task procedure was the same as that followed in Experiments 1–3, wherein the search was accompanied by target-congruent, distractor-congruent, and control sounds (i.e., unrelated sounds in Experiment 4, white noise in Experiment 5, and no sounds in Experiment 6).

Experiments 4, 5, and 6

Experiments 4–6 were designed to replicate Experiments 1–3 with three types of synchronous sounds: the target-congruent, distractor-congruent, and control sounds. The control sounds varied across the experiments so that the sound unrelated to the target was presented in Experiment 4, white noise was presented in Experiment 5, and no sound (sound absent) was presented in Experiment 6. These experiments never presented identical pictures during visual search tasks. If the auditory enhancement and interference effects of sounds on visual searches demonstrated in Experiments 1–3 were attributed to participants' learning of specific picture-sound pairs through the familiarized and experimental trials, no such effects should have been obtained in Experiments 4–6, wherein each picture differed while semantic congruency was maintained (e.g., photographs of individuals against different backgrounds although all were running). By contrast, if the effects resulted from the automatic interaction of task-irrelevant sounds with visual representation about events of actions, similar enhancement and interference would be expected in Experiments 4–6.

Method

Participants A group of 75 students (47 males and 28 females; mean age = 20.21 years, range = 18–25 years) participated in the experiments and were randomly assigned to one of the three experiments (25 each). All participants reported having normal color vision and normal or corrected-to-normal visual acuity. None of the participants was hearing impaired according to self-report.

Apparatus and stimuli We used the same audio clips representing 20 events as those used in Experiment 1. To avoid repetitive presentation of an identical picture, the search task included 16 pictures for each of the 20 critical verbs (320 pictures in total, available on the Open Science Framework at https://osf.io/kgzqd/). Twelve of the 16 pictures were randomly assigned to the target-congruent, distractor-congruent, and control conditions (i.e., four each). Four of the 16 pictures were assigned to a distractor image congruent with a sound in the distractor-congruent condition. Moreover, the search task included 640 non-target images (i.e., filler pictures) representing events or objects irrelevant to the critical 20 verbs. The 640 fillers were selected from a pool of 670 images that consisted of 626 non-overlapping images and 22 pairs of identical photos. We previously clarified that 30 other students who did not participate in the present experiments (15 males and 15 females; mean age = 20.60 years, range = 18–25 years) were able to associate the 320 verb-relevant pictures with correct verbs with more than 80% accuracy (mean accuracy = 97.36%, *SD* = 5.65).

Visual search tasks. Participants completed the two familiarization tasks using the same apparatus, stimuli, and procedure as Experiment 1, before a modified version of the visual search task was initiated (240 trials). The procedure was identical to that used in Experiments 1–3. However, pictures were not repeated in the main visual search task.

Figure 7 illustrates a schematic example of the search task. As in Experiments 1–3, participants were required to locate the target picture (lower left, upper left, lower right, or upper right) as quickly as possible by pressing a numeric key – 1, 4, 2, or 5 – after a target verb was presented in the center of the screen. In every trial, the search display consisted of four pictures containing a single target and three non-target items. Under the target-congruent and control conditions, the three non-target locations of the search display were located by filler pictures. Under the distractor-congruent condition, one of the three non-targets was presented as a distractor that was congruent with a sound that accompanied the search display, and was located in a quadrant diagonally opposite to a target across the fixation point. The other two locations under the distractor-congruent condition were occupied by filler pictures. The targets, distractors, and fillers were randomly selected from a pool of visual stimuli to appear once in the trials without any duplications. Before performing the 240 main trials, participants completed ten practice trials using a different set of pictures (20 targets or distractors and 160 fillers). As in Experiments 1–3, we primarily analyzed response times.

Error-pattern analysis To provide further evidence regarding the interference effects of task-irrelevant sounds, Experiments 4–6 focused on error patterns in visual localization to examine whether participants attended to distractor pictures when they performed under the distractor-congruent condition. If participants randomly responded to one of three non-target locations, the probability of error responses to one distractor and two filler locations should have fallen within a ratio of 1:2 under the distractor-congruent condition. If the distractor-congruent sounds directed participants' attention to the distractor items rather than the two filler locations, the number of error responses to that location should have exceeded chance (33.3% of probability of all incorrect responses).

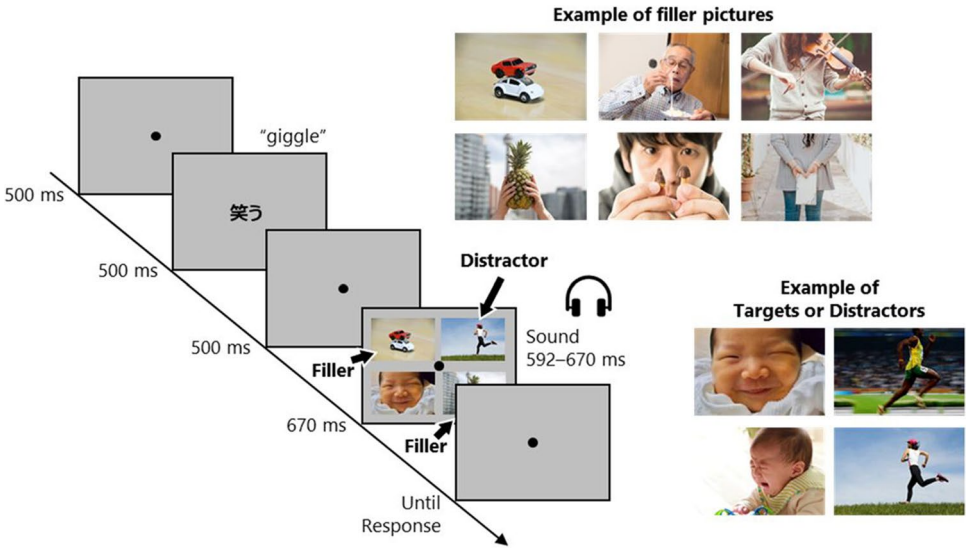

Fig. 7 An example of a trial sequence in a visual search task modified for Experiments 4–6

Results

Experiment 4 Error trials (4.48% of all trials) were excluded from the analysis of reaction times. Correct trials with response times outside the 1.5-times interquartile range beyond the 25th and 75th percentiles were excluded from the analysis (3.55% of all trials) as outliers. Response times for the correct trials were averaged for each sound condition (i.e., the target-congruent, distractor-congruent, and unrelated sound conditions) (Fig. 8). An ANOVA on the mean response times revealed a significant main effect of sound condition [$F(2, 48) = 29.63$, $p < .001$, $\eta_G^2 = .032$, $\varepsilon = 1.00$]. Further comparisons using Holm's method revealed that the mean response time was shorter under the target-congruent condition (642.29 ms) than under the unrelated (680.76 ms) [$t(24) = 5.89$, $p < .001$] and distractor-congruent (698.26 ms) conditions [$t(24) = 7.29$, $p < .001$]. Although weak, the mean response time was longer under the distractor-congruent condition than under the unrelated condition [$t(24) = 2.18$, $p = .039$]. For the error-pattern analysis, we calculated the numbers of incorrect responses to each location of the distractor and fillers across participants (Table 2). A binomial test revealed that the number of error responses at the distractor location was beyond the 33.3% probability of all incorrect responses ($p = .003$).

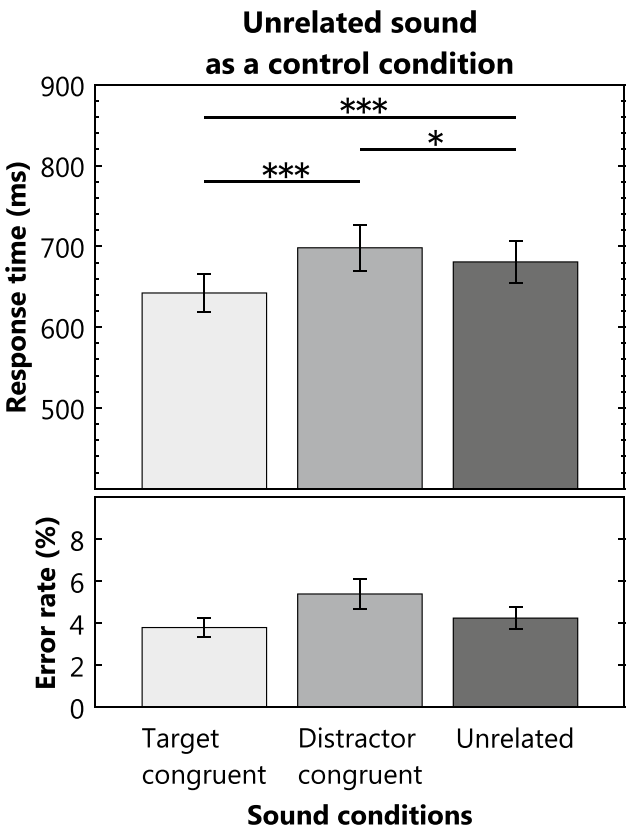

Fig. 8 Mean response times under the target-congruent and distractor-congruent conditions compared to those under the unrelated condition in Experiment 4. *Note*. *$p < .05$, ***$p < .001$

Experiment 5 Error trials (5.03% of all trials) were excluded from the reaction-time analysis. Correct trials with response times outside 1.5 times the interquartile range beyond the 25th and 75th percentiles were excluded from the analysis (2.60% of all trials). Mean response times from the remaining trials were calculated for the target-congruent, distractor-congruent, and white-noise conditions (Fig. 9). An ANOVA on the mean response times found a significant main effect of sound condition [$F(1.94, 46.64) = 41.39$, $p < .001$, $\eta_G^2 = .085$, $\varepsilon = .97$]. Multiple comparisons revealed that the

Table 2 The numbers and percentages of incorrect responses to a distractor location

Study	Number of responses			
	Distractor location	%	Filler locations	%
Experiment 4	58	44.96 %	71	55.04 %
Experiment 5	62	45.93 %	73	54.07 %
Experiment 6	52	49.52 %	53	50.48 %

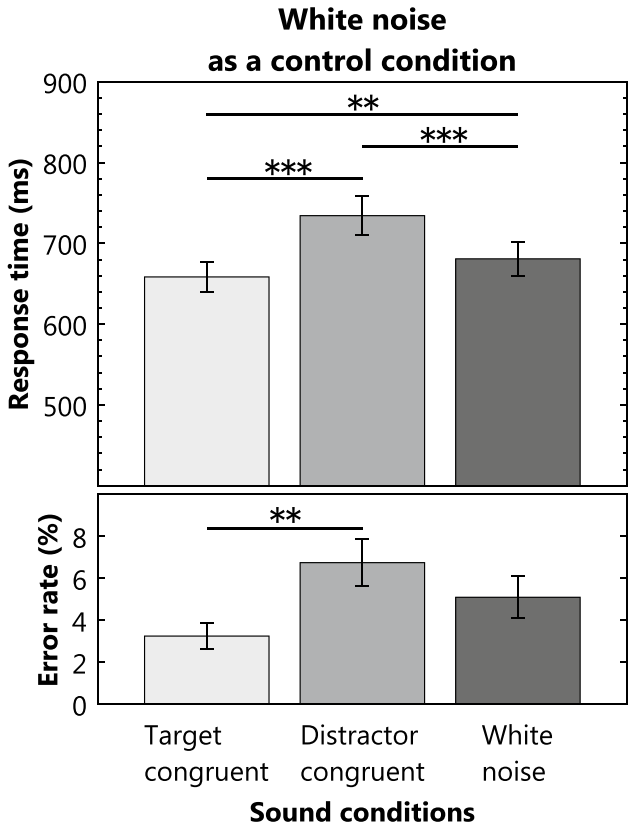

Fig. 9 Mean response times under the target-congruent and distractor-congruent conditions compared to those under the white-noise condition in Experiment 5. *Note*. $^{**}p < .01$, $^{***}p < .001$

response times were shorter under the target-congruent condition (658.44 ms) than under the white-noise (680.86 ms) [$t(24) = 2.83$, $p = .009$] and distractor-congruent (734.29 ms) [$t(24) = 7.78$, $p < .001$] conditions, and the response times were longer under the distractor-congruent condition than under the white-noise control condition [$t(24) = 6.78$, $p < .001$]. Moreover, the error-pattern analysis revealed that the number of error responses at the distractor location was greater than the 33.3% probability of all incorrect responses (a bimodal test, $p = .001$).

Experiment 6 Error trials (3.92% of all trials) were excluded from the reaction-time analysis. Trials with response times that deviated by more than 1.5 times the interquartile range of the 25th and 75th percentiles were excluded from the analysis (3.02% of all trials). Mean response times for the correct trials were calculated for the target-congruent, distractor-congruent, and sound-absent conditions (Fig. 10). An ANOVA of the mean response times found a significant main effect of sound condition [$F(1.71, 41.02) = 46.60$, $p < .001$, $\eta_G^2 = .072$, $\varepsilon = .85$]. Multiple comparisons revealed that the response times were shorter under the target-congruent condition (618.46 ms) than under the sound-absent (641.99 ms) [$t(24) = 4.52$, $p < .001$] and distractor-congruent (682.28 ms) [$t(24) = 9.65$, $p < .001$] conditions, and the response times were longer under the distractor-congruent condition than under the sound-absent control condition [$t(24) = 5.06$, $p < .001$]. Furthermore, the error-pattern analysis revealed that the number of error responses at the distractor location was greater than the 33.3% probability of total incorrect responses (a bimodal test, $p < .001$).

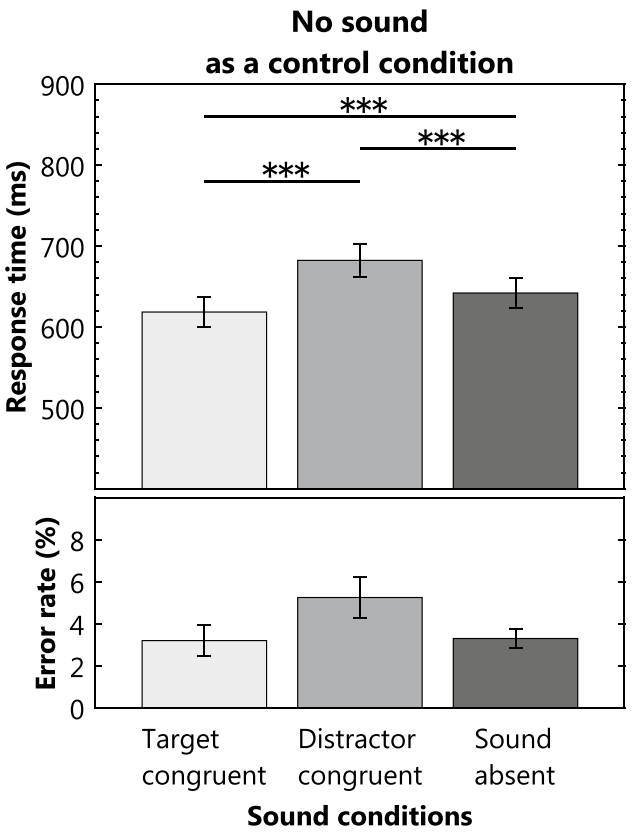

Fig. 10 Mean response times under the target-congruent and distractor-congruent conditions compared to those under the sound-absent condition in Experiment 6. *Note*. $^{***}p < .001$

Discussion

Experiments 4–6 replicated the results pattern of Experiments 1–3, in that the target-congruent sounds resulted in swift responses to the target-event scenes, and semantically incongruent sounds delayed the visual searches. The swift responses indicate that auditory enhancement of visual localization occurred consistently even when the possibility of learning specific picture-sound pairs was excluded. This finding implies that audiovisual interactions based on semantic congruence regarding event-related stimuli may develop as a result of long-term experiences other than as a result of brief training during experiments (Zweig et al., 2015).

Moreover, we examined auditory-interference effects on visual search by measuring whether the number of error responses to the distractor location deviated by more than the chance probability (33.3%) of all incorrect responses under the distractor-congruent condition. Across the three experiments, the error patterns of visual localization consistently showed that the number of responses to the distractor locations exceeded the chance probability, indicating participants' attention to the distractor items during the distractor-congruent condition. Consequently, we suggest that goal-directed control would not have been activated in the present experiments; rather, both enhanced and degraded responses were observed when an event scene was used as a target.

General discussion

This study examined whether event-related sounds facilitated or inhibited visual searches for relevant scenes based on semantic congruence/incongruence. Across six experiments, we compared visual performance under three types of control conditions (i.e., the unrelated-sound, white-noise, and sound-absent conditions) with that under the target-congruent and distractor-congruent conditions. All experiments revealed that when sounds related to target scenes appeared in a search display, visual-search performance was auditorily enhanced compared to conditions in which such congruent sounds were absent (Experiments 1–6). Importantly, search performance was impaired by semantic incongruence of the audiovisual stimuli when sounds were related to distractor event scenes (Experiments 2–6). Error-pattern analysis revealed that participants directed their attention toward to a distractor location in a search display when distractor-congruent sounds appeared synchronously (Experiments 4–6). Furthermore, the automatic interaction between sounds and pictures is based on semantic representations that may develop as a result of long-term experiences other than short-term training periods (Experiments 4–6).

This study's main objective was to semantically extend the multimodal enhancement (Iordanescu et al., 2008, 2010, 2011) used to detect and localize event-scenes representing the interactions of objects, rather than merely detecting readily identifiable Individual objects. The visual targets used in Experiments 1–3 were temporally trimmed pictures of series of motions, requiring extrapolation of contextual information. The contextual meanings were likely to be rapidly extrapolated during the search when sounds were congruent with visual representations. Using multiple exemplars of a single scene in Experiments 4–6, we ruled out visual images forming a one-to-one correspondence with characteristic sounds. Nevertheless, we found that the auditory stimuli facilitated visual searches, consistent with the object-based crossmodal effects. Our findings imply that multimodal enhancement occurs when audiovisual stimuli are viewed in association with more complex and interactive stimuli consisting of multiple objects against the background in a real-life situation. Importantly, audiovisual interaction helped observers to quickly explore what was happening in a cluttered scene in addition to finding isolated objects. The observed improvements in event detection and/or localization performance using semantic cues are consistent with the hypothesized effect of knowledge-based information on scene recognition (for a review, see Võ, 2021), indicating that meaningful contexts retrieved by audio cues help to improve coherent perception of event occurrence.

One of this study's most noteworthy findings is that auditory stimuli were automatically integrated with distractor scenes during the search task. Specifically, we observed impaired performance under the distractor-congruent condition, in contrast to earlier findings on multimodal interaction demonstrating no or weak interference effects on visual performance under the same condition (Iordanescu et al., 2008, 2010, 2011; Molholm et al., 2004; von Kriegstein et al., 2005). The error-pattern analysis demonstrated an increase in the number of incorrect localizations at the distractor location and this result is incompatible with the model proposed by Iordanescu et al. (2008), wherein multimodal interactions rely on goal-directed top-down control. Instead, performance deficits associated with exposure to semantic-incongruent stimuli may be caused by the reduced coherence of visual representation, such that semantic-incongruent sounds impair the identification of visual items (Chen & Spence, 2010; Glaser & Glaser, 1989). For example, Chen and Spence (2010) argued that participants automatically and quickly retrieve meaningful contexts of auditory and visual inputs and integrate them into critical semantic representation, while multisensory incoming information is retained in a short-term buffer for semantic processing. If

the activated semantic representation contains highly coherent information, it is advantageous for the participants to perform the current task (Potter, 1999). According to this idea, the incongruence of audiovisual stimuli should impair the formation of semantic representations and exert a negative impact on task performance.

A study by Steinweg and Mast (2017) proposed another multimodal interference possibility relating to response biases or decision strategies. Specifically, during a speeded-up, two-alternative, forced-choice decision task, participants may cautiously select a target picture from the stimulus array to avoid false responses when semantically inconsistent sounds appear. This bias toward cautious responses would elicit a negative impact on visual search performance, irrespective of the multimodal interaction process. However, it should be noted that most multisensory studies methodologically ruled out influences of the response bias by adopting accuracy measures rather than response times (Chen & Spence, 2010). Moreover, the error analysis in Experiments 4–6 demonstrated that participants increased erroneous localization to distractor pictures to above chance level when sounds corresponding with those scenes were presented simultaneously. Thus, we argue that response bias is a possible explanation for the impaired performance in the distractor-congruent condition, rather than a single critical determinant.

We consider that the inconsistency of the incongruence effects between the present study (Experiments 2–6) and earlier studies (Iordanescu et al., 2008, 2010, 2011; Molholm et al., 2004; von Kriegstein et al., 2005) can be attributed to the complexity of the visual stimuli. Because the event scenes consisted of multiple layers of information, such as individual objects and background elements, participants in the present study might not have been able to readily access prior knowledge, such as prototypical visual representations of the targets, in a goal-directed manner (see Chen & Spence, 2010). Thus, mechanisms underlying the filtering out of irrelevant stimuli may have been weakened or deactivated. Rather, the auditory stimuli interacted automatically, consistent with the idea of semantic incongruency effects (e.g., Chen & Spence, 2010; Potter, 1999). This finding highlighted the involuntary guidance provided by spatial orientation based on auditory-visual semantic aspects. This automaticity was impaired when the auditory cues were embedded in the video clips during searches for realistic objects (Kvasova et al., 2019). However, we speculate that the crossmodal interference effect was only reduced, rather than eliminated, in the video-using-study, due to activation of the goal-directed control mechanism.

Auditory facilitation/interference showed a pattern similar to that of the characteristic sounds that improved or degraded object identification performance (e.g., Chen & Spence, 2010; Molholm et al., 2004; von Kriegstein et al., 2005). This similarity suggests that the same mechanism may underlie the two tasks; we proposed that while the auditory-visual interactions of high-level semantic aspects facilitated identification of relevant contextual representations, low-level processes receive this feedback in the form of visual enhancement (salience) during concurrent searching or discrimination (Iordanescu et al., 2008). If this mechanism was involved in the present findings, auditory facilitation would be expected to decrease the search slope due to visual salience.

In conclusion, the present study revealed that auditory enhancement and interference affect visual search for event scenes. We demonstrated that visual searches for a target are robustly enhanced when accompanying sounds are related to the target. However, performance was measurably degraded when the sound corresponding to a distractor scene appeared synchronously with a target item. The audiovisual interactions involving semantic processes support coherent and swift interpretation of events that occur in surrounding environments.

Acknowledgements This work was supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (20H01779) to JK and Graduate Grant Program of Graduate School of Letters, Hokkaido University, Japan, and a Grant-in-Aid from the Japan Society for the Promotion of Science Fellows (20J20490) to TM.

References

Barraclough, N. E., Xiao, D., Baker, C. I., Oram, M. W., & Perrett, D. I. (2015). Integration of Visual and Auditory Information by Superior Temporal Sulcus Neurons Responsive to the Sight of Actions. *Journal of Cognitive Neuroscience*, *17*, 377-391.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, *41*, 809-823.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433-436.

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.

Chen, Y. C., & Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 1554.

Eimer, M., & Driver, J. (2001). Crossmodal links in endogenous and exogenous spatial attention: Evidence from event-related brain potential studies. *Neuroscience and Biobehavioral Reviews*, *25*(6), 497–511. https://doi.org/10.1016/S0149-7634(01)00029-X

Feng, W., Störmer, V. S., Martinez, A., McDonald, J. J., & Hillyard, S. A. (2014). Sounds activate visual cortex and improve visual discrimination. *Journal of Neuroscience*, *34*, 9817-9824.

Glaser, W. R., & Glaser, M. O. (1989). Context effects in Stroop-like word and picture processing. *Journal of Experimental Psychology: General*, *118*(1), 13–42. https://doi.org/10.1037/0096-3445.118.1.13

Hein, G., Doehrmann, O., Müller, N. G., Kaiser, J., Muckli, L., & Naumer, M. J. (2007). Object familiarity and semantic congruency

modulate responses in cortical audiovisual integration areas. *Journal of Neuroscience*, *27*, 7881-7887.
Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, *50*, 243–271. https://doi.org/10.1146/annurev.psych.50.1.243
Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review*, *15*, 548-554.
Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., & Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception, & Psychophysics*, *72*, 1736-1741.
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.
Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception, 36*(14), 1–16, ECVP Abstract Supplement.
Kvasova, D., Garcia-Vernet, L., & Soto-Faraco, S. (2019). Characteristic sounds facilitate object search in real-life scenes. *Frontiers in Psychology*, *10*, Article 2511. https://doi.org/10.3389/fpsyg.2019.02511
Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research*, *158*(4), 405–414. https://doi.org/10.1007/s00221-004-1913-2
Mädebach, A., Wöhner, S., Kieseler, M. L., & Jescheniak, J. D. (2017). Neighing, barking, and drumming horses—object related sounds help and hinder picture naming. *Journal of Experimental Psychology: Human Perception and Performance*, *43*, 1629.
Maniglia, M., Grassi, M., & Ward, J. (2017). Sounds are perceived as louder when accompanied by visual movement. *Multisensory Research*, *30*(2), 159–177. https://doi.org/10.1163/22134808-00002569
Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory Visual-Auditory Object Recognition in Humans: A High-density Electrical Mapping Study. *Cerebral Cortex*, *14*(4), 452–465. https://doi.org/10.1093/cercor/bhh007
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437-442.
Potter, M. C. (1999). Understanding sentences and scenes: The role of conceptual short-term memory. In V. Coltheart (Ed.), MIT Press/Bradford Books series in cognitive psychology. Fleeting memories: Cognition of brief visual stimuli (p. 13–46). The MIT Press.
Raij, T., Uutela, K., & Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron*, *28*(2), 617–625. https://doi.org/10.1016/S0896-6273(00)00138-0
Smith, E. L., Grabowecky, M., & Suzuki, S. (2007). Auditory-visual crossmodal integration in perception of face gender. *Current Biology*, *17*(19), 1680–1685.
Spence, C., & Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. *Perception & Psychophysics*, *59*, 1-22.
Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, *9*(4), 255–266. https://doi.org/10.1038/nrn2331
Steinweg, B., & Mast, F. W. (2017). Semantic incongruity influences response caution in audio-visual integration. *Experimental brain research*, *235*, 349-363.
Störmer, V. S. (2019). Orienting spatial attention to sounds enhances visual processing. *Current Opinion in Psychology*, *29*, 193–198. https://doi.org/10.1016/j.copsyc.2019.03.010
Störmer, V. S., McDonald, J. J., & Hillyard, S. A. (2009). Cross-modal cueing of attention alters appearance and early cortical processing of visual stimuli. PNAS Proceedings of the National Academy of Sciences of the United States of America, 106(52), 22456–22461. https://doi.org/10.1073/pnas.0907573106
Suied, C., Bonneel, N., & Viaud-Delmon, I. (2009). Integration of auditory and visual information in the recognition of realistic objects. *Experimental Brain Research*, *194*(1), 91–102. https://doi.org/10.1007/s00221-008-1672-6
Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I., & Miyashita, Y. (1999). Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature*, *401*(6754), 699–701. https://doi.org/10.1038/44372
Võ, M. L.-H. (2021). The meaning and structure of scenes. *Vision Research*, *181*, 10–20. https://doi.org/10.1016/j.visres.2020.11.003
von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of Face and Voice Areas during Speaker Recognition. Journal of Cognitive Neuroscience, 17(3), 367–376. https://doi.org/10.1162/0898929053279577
Zweig, L. J., Suzuki, S., & Grabowecky, M. (2015). Learned face–voice pairings facilitate visual search. *Psychonomic Bulletin & Review*, *22*(2), 429–436. https://doi.org/10.3758/s13423-014-0685-3

Open practices statement The data and significant program code will be made available after acceptance via the Open Science Framework (https://osf.io/kgzqd/), and none of the experiments were preregistered.