Journal of Experimental Psychology:
Human Perception and Performance
2015, Vol. 41, No. 5, 1325–1335

0096-1523/15/$12.00 http://dx.doi.org/10.1037/xhp0000074

Audiovisual Integration Facilitates Unconscious Visual Scene Processing

Jye-Sheng Tan and Su-Ling Yeh
National Taiwan University

Meanings of masked complex scenes can be extracted without awareness; however, it remains unknown whether audiovisual integration occurs with an invisible complex visual scene. The authors examine whether a scenery soundtrack can facilitate unconscious processing of a subliminal visual scene. The continuous flash suppression paradigm was used to render a complex scene picture invisible, and the picture was paired with a semantically congruent or incongruent scenery soundtrack. Participants were asked to respond as quickly as possible if they detected any part of the scene. Release-from-suppression time was used as an index of unconscious processing of the complex scene, which was shorter in the audiovisual congruent condition than in the incongruent condition (Experiment 1). The possibility that participants adopted different detection criteria for the 2 conditions was excluded (Experiment 2). The audiovisual congruency effect did not occur for objects-only (Experiment 3) and background-only (Experiment 4) pictures, and it did not result from consciously mediated conceptual priming (Experiment 5). The congruency effect was replicated when catch trials without scene pictures were added to exclude participants with high false-alarm rates (Experiment 6). This is the first study demonstrating unconscious audiovisual integration with subliminal scene pictures, and it suggests expansions of scene-perception theories to include unconscious audiovisual integration.

Keywords: scene perception, unconscious processing, audiovisual integration, interocular suppression, continuous flash suppression

Supplemental materials: http://dx.doi.org/10.1037/xhp0000074.supp

Our perception of a scene consists of information from different sensory modalities. For example, when we walk along the street, we not only see the moving vehicle but also hear the engine and horn. Similarly, when we dine in a restaurant, we not only see the beautiful decorations but also hear the chatting sounds of people in conversation. Indeed, previous studies have demonstrated audiovisual integration. For example, a beep sound enhances the perceived intensity of simple visual stimuli (Chen, Huang, Yeh, & Spence, 2011; B. E. Stein, London, Wilkinson, & Price, 1996). Also, semantically congruent soundtrack facilitates fearful/disgusting facial discrimination (Collignon et al., 2008) and object localization in visual search (Iordanescu, Guzman-Martinez, Grabowecky, & Suzuki, 2008).

In addition to audiovisual integration with suprathreshold visual stimuli, the unconscious visual processing can also be facilitated by sound. For example, it has been shown that a synchronous sound enables letters in rapid serial visual presentation (RSVP) to escape the attentional blink (Olivers & Van der Burg, 2008) and repetition blindness (Chen & Yeh, 2008, 2009) and be detected more easily. A same-position concurrent sound can improve the perceptual sensitivity of a masked LED light display (Bolognini, Frassinetti, Serino, & Ladavas, 2005; Frassinetti, Bolognini, & Ladavas, 2002) or facilitate detection of interocularly suppressed visual stimuli (Y. H. Yang & Yeh, 2014). A corresponding auditory stimulus can also enhance the dominance period of visual stimuli (car or bird) in binocular rivalry (Chen, Yeh, & Spence, 2011) and in a young and old woman bistable figure (Hsiao, Chen, Spence, & Yeh, 2012), and facilitate a face stimulus to release from interocular suppression (Alsius & Munhall, 2013).

Previous studies about unconscious audiovisual integration have been limited in using single items (such as line drawings, dots, or other objects) as visual stimuli, and little is known about whether unconscious audiovisual integration also occurs for complex scene pictures. It is reasonable to suspect that single items and scene pictures are processed differently. A scene picture contains not only objects but also background, which is visually more complex than objects per se (Epstein, 2005; Henderson, Weeks, & Hollingworth, 1999). If objects and scenes share the same mechanism, the processing of scene should have taken much longer time than that of objects. Previous studies on scene perception suggest that the gist encoded, rather than longer duration of details encoded in object processing, plays an important role in helping us to extract scene information at a very brief glimpse (Sampanes, Tseng, & Bridgeman, 2008). For example, categorization of a visual scene can be accomplished when it is briefly presented within 100 ms (Fei-Fei, Iyer, Koch, & Perona, 2007; Potter, 1976), or sustained in

This article was published Online First June 15, 2015.

Jye-Sheng Tan, Department of Psychology, National Taiwan University; Su-Ling Yeh, Department of Psychology, National Taiwan University, Graduate Institute of Brain and Mind Sciences, National Taiwan University, and Neurobiology and Cognitive Neuroscience Center, National Taiwan University.

This study is supported by grants from Taiwan's Ministry of Science and Technology, MOST98-2410-H-002–023-MY3 and MOST101-2410-H-002–083-MY3. We thank Chih-Chien Cheng and Po-Chun Lien for help conducting part of the experiment, and Ji-Fan Zhou for comments on earlier versions of this article.

Correspondence concerning this article should be addressed to Su-Ling Yeh, Department of Psychology, National Taiwan University, Taipei 10617, Taiwan. E-mail: suling@ntu.edu.tw

an unconscious condition by backward masking (VanRullen & Koch, 2003). More remarkably, the meaning of scene picture can be extracted when presented briefly for 13 ms in RSVP (Potter, Wyble, Hagmann, & McCourt, 2014).

In this study, we used the continuous flash suppression (CFS) paradigm (Fang & He, 2005; Tsuchiya & Koch, 2005) to examine whether a semantically congruent soundtrack enhances the unconscious processing of complex visual scene pictures. Scenery soundtracks need at least several seconds for their meanings to be extracted. Among all the paradigms that render visual stimuli invisible, the CFS paradigm has the advantage that the scene pictures can be presented for a relatively long time without participants' awareness (Faivre, Berthet, & Kouider, 2014). In each trial, the scene picture and the dynamic masks were presented dichoptically in different eyes and the scene was suppressed by the masks at first. A semantically congruent or incongruent scenery soundtrack was presented simultaneously with the scene picture. The time for the scene to release from interocular suppression was measured by requiring participants to press a key as soon as they detected any part of the scene, as this release-from-suppression time (or reaction time; RT) reflects the time for unconscious processing of the scene (Chen & Yeh, 2012; Jiang, Costello, & He, 2007; T. Stein, Senju, Peelen, & Sterzer, 2011; Y. H. Yang & Yeh, 2011). If audiovisual integration occurs, the RT of detecting the scenery picture should be faster when pairing with congruent sound than incongruent sound, because the sound can facilitate the processing of scene by adding to the meaning.

Six experiments were conducted in the current study. In Experiment 1, we used full-scene pictures as visual stimuli and paired them with a congruent or incongruent soundtrack. We demonstrated that the unconscious scene processing was facilitated in the audiovisual congruent condition (Experiment 1) and excluded the possibility that different detection criteria contributed to the results by using a control binocular viewing condition (Experiment 2). Such audiovisual facilitation disappeared when objects-only (Experiment 3) or background-only (Experiment 4) scene pictures were used as visual stimuli. We excluded the possibility that this audiovisual facilitation was generated by consciously mediated conceptual priming (Experiment 5). Finally, participants who had high false alarm rates in catch trials were excluded and the audiovisual congruency effect was also replicated (Experiment 6). The overall pattern of results suggests that unconscious audiovisual integration indeed occurs for complex visual scenes.

Experiment 1: Full Scenes

In Experiment 1, we examined whether scenery soundtrack can influence the unconscious processing of a subliminal scene picture. Full-scene pictures were presented subliminally under CFS and paired with semantically congruent or incongruent soundtrack in each trial.

To ensure unconscious audiovisual integration was occurring at the scene level but not the object level, full-scene pictures that included background and objects were used in Experiment 1. In addition, objects were presented dispersedly and did not appear at the center of scene pictures, to prevent the objects from being too obvious to attract attention so that objects rather than scenes were integrated with the soundtrack.

Method

Participants. The usual sample size of previous CFS studies (Alsius & Munhall, 2013; Mudrik, Breska, Lamy, & Deouell, 2011; E. Yang, Zald, & Blake, 2007) was around 10 to 30 participants, thus an average number of 20 healthy participants who had normal or corrected-to-normal vision and hearing were recruited in Experiment 1. All participants received credit for a psychology course at National Taiwan University or monetary rewards for a half-hour experiment. All participants gave informed consent before the experiment, and the ethics committee of the Department of Psychology at National Taiwan University approved the experiments conducted in this study.

Apparatus and materials. Visual stimuli were presented by E-prime on a 17-in. calibrated EIZO color monitor with 75 Hz. refresh rate. Participants sat at approximately 70 cm viewing distance from the monitor in a dimly lit chamber.

Using a mirror stereoscope, participants viewed dichoptically scene pictures (5.72° × 5.72°) in randomly chosen eye while simultaneously viewing colorful Mondrians (13.04° × 13.04°) that changed every 100 ms (10 Hz) as suppressor in the other eye. The suppressors were scattered overlapped Mondrians constituted by varying size and color rectangles not to exceed 1.2° × 1.2°. All scene pictures and colorful Mondrians were presented in a square border (13.04° × 13.04°) that served to promote stable binocular eye alignment (see Figure 1).

Four colorful restaurant pictures and four colorful street pictures were used as target pictures (see Figure 2; and see also all pictures used in this study in the Supplementary Materials). All scene pictures had been verified by a pilot test with 17 participants who did not participate in the formal experiment. All participants in the pilot test viewed eight scene pictures used in this study and 12 other scene pictures as distractors in random sequence, and wrote down the name of the scene picture according to their first impression. All participants identified the eight scene pictures used here correctly as restaurant or street scenes in the naming task.

Two scene soundtracks (restaurant and street) downloaded from www.soundsnap.com were adopted as auditory stimuli. Each soundtrack was segregated into four 6-s segments, and each of the segments was paired with each scene picture. External loudspeakers presented each segment in an average of 60 db SPL. Restaurant scene soundtracks included chatting sounds and clinking sounds of plates and bowls, whereas street scene soundtracks included engine and horn sounds of moving vehicles.

Design and procedure. The experiment started with a calibration stage, at which participants were asked to adjust the mirror stereoscope to produce a fused dichoptic viewing of two images from different eyes. After the calibration stage, participants conducted a practice session to become familiar with the experimental procedure. There were eight trials (randomly choosing from 256 trials in the formal test) in one practice session. Participants who obtained above 50% target detection rate with correct spatial localization took the formal test, whereas the others who could not obtain above 50% detection with correct localization within three practice sessions were excluded for further formal experiment. Four, zero, five, 10, three, and zero participants were excluded

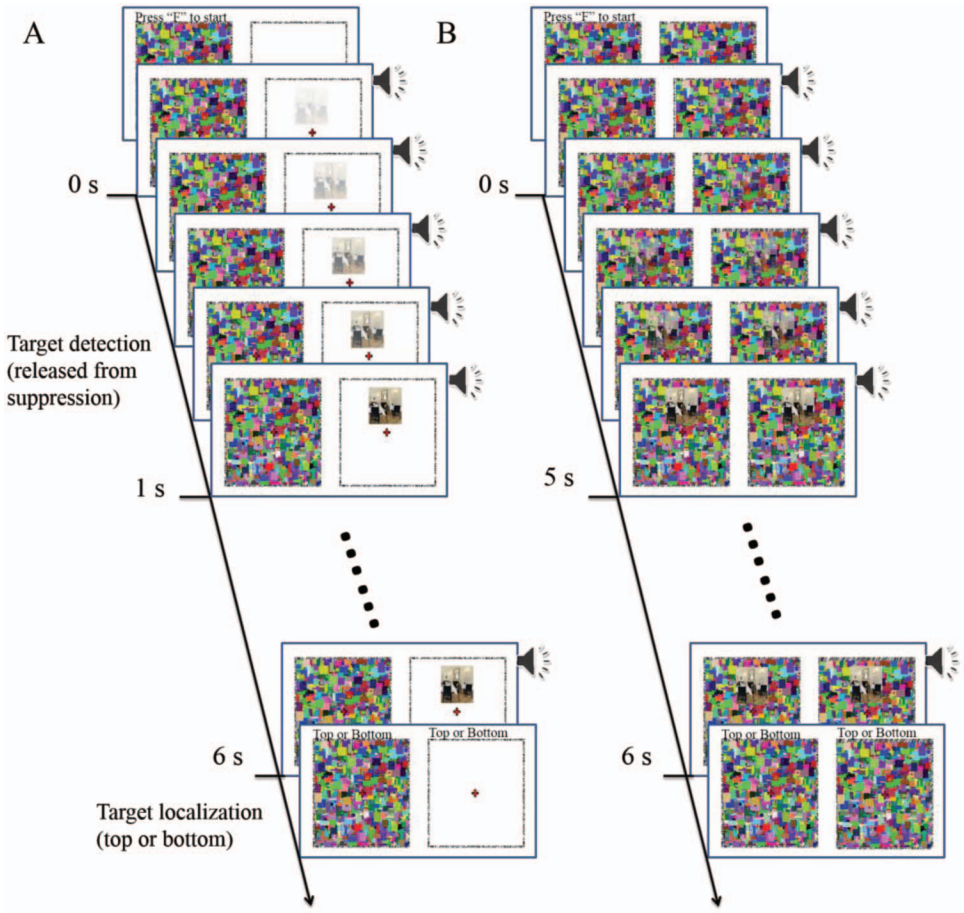

Figure 1. The procedure of CFS paradigm used in Experiment 1 and Experiment 2. A: In Experiment 1, the scene picture was presented in one eye and colorful Mondrains in the other eye; B: in Experiment 2, the scene picture and colorful Mondrains were blended and presented binocularly in both eyes. See the online article for the color version of this figure.

according to this criterion for practice trials in Experiments 1, 2, 3, 4, 5, and 6, respectively.

Figure 1A shows the procedure. Each trial began with a phrase "press F key to start." After the F key was pressed, colorful Mondrains were presented to one of the participant's randomly chosen eye at full contrast, and continued to flash at 10 Hz from the beginning until the end of each trial. The scene picture was presented to the other eye, at either above or below of the central cross, with contrast increasing gradually from 0% to 100% within 1 s. Target pictures remained at full contrast after 1 s until participants pressed F key again to stop when any part of the scene picture was detected. After this, they were required to press "P" or "L" key to report whether the target was presented above or below the cross after the detection task. A semantically congruent or incongruent soundtrack started and ended synchronously with the scene picture in every trial. Soundtracks were never mentioned throughout the experiment and participants were instructed to focus on the visual detection task only.

Each participant conducted 256 CFS formal trials, which included one self-paced rest interval between two blocks with 128 trials in each block. The 256 trials included 8 (four scene pictures: restaurant or street) × 8 (four soundtracks in each scene: restaurant or street) × 2 (target location of dichoptic viewing: left or right) × 2 (target location: above or below). The total experiment included 128 congruent and 128 incongruent audiovisual trials and all the trials were presented in random sequence.

Results and Discussion

We excluded participants who had higher than 30% miss rate or localization error. Accordingly, four participants were excluded: two of them missed the response within 6 s in more than 30% trials, and the other two had localization errors higher than 30%. Trials that had no response within 6 s, error responses of localization task, and RTs exceeding ±2.5 *SD* from mean of individual participant were excluded from the remaining 16 participants' data (11.31% removed).

A two-tailed paired *t* test showed a significant congruency effect, $t(15) = 2.701$, $p = .016$, Cohen's $d = 0.675$, 95% CI: 20.89 ms, 177.35 ms (see the results of RTs in Figure 4); mean RT was

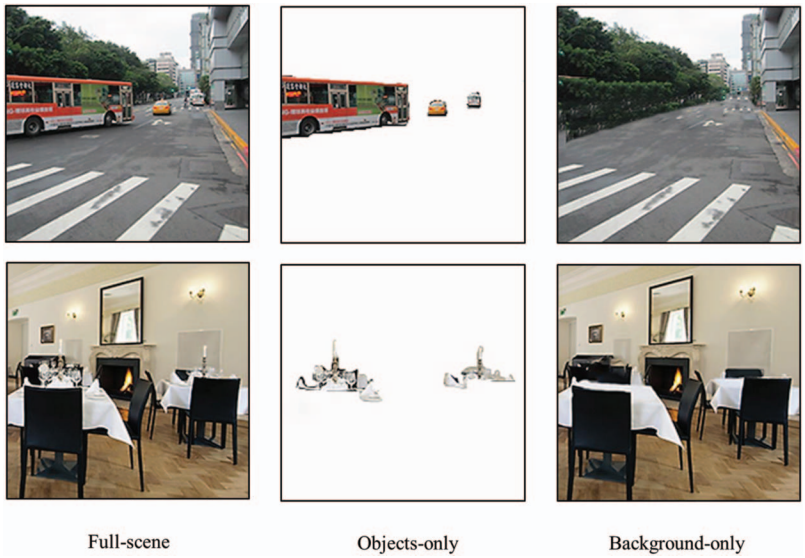

Figure 2. Two types of full-scene pictures (street and restaurant, top and bottom pictures on the left panel) were used as visual stimuli in Experiment 1, 2, 5, and 6. Objects-only pictures (the middle panel) were used in Experiment 3 and background-only pictures (the right panel) were used in Experiment 4. See the online article for the color version of this figure.

significantly shorter in the congruent condition (M = 2,186 ms [restaurant scene = 2,203 ms, street scene = 2,168 ms]) than the incongruent condition (M = 2,285 ms [restaurant scene = 2,354 ms, street scene = 2,215 ms]). Average accuracies in the congruent and incongruent conditions were 93.22% and 93.65%, and there was no difference between the two conditions, $t(15) = 1.187$, $p = .254$, 95% CI: 0.45%, 1.57%, indicating no speed–accuracy trade-off. RT and accuracy data for each individual participant was shown in the Supplementary Materials.

This audiovisual congruency effect obtained in Experiment 1 indicates that the semantically congruent soundtrack facilitates the unconscious processing of subliminal scene pictures. Before we can conclude that the audiovisual integration indeed occurred with scenes that were not consciously processed, however, it is necessary to rule out an alternative possibility that the result obtained was due to different detection criteria corresponding to the audiovisual congruent and incongruent conditions. Experiment 2 was designed to test this alternative hypothesis.

Experiment 2: Binocular Viewing

We adopted a standard control experiment in the CFS paradigm (Alsius & Munhall, 2013; Jiang et al., 2007; Y. H. Yang & Yeh, 2011) by presenting the scene and the Mondrians in both eyes to see if the congruency effect remained. In this binocular viewing condition, the increment of scene contrast mimicked that in the dichoptic viewing condition in Experiment 1, although with a slower increasing speed to obtain similar RTs as in the dichoptic viewing condition (Jiang et al., 2007). If the results we obtained in Experiment 1 could be explained by differences in detection criteria between the congruency and incongruent conditions, we should expect to find the same congruency effect as in Experiment 1. If, however, processing time was reduced for congruency because congruency improved processing efficiency and not differences in detection criteria, no RT difference should be found in this binocular viewing control experiment.

Method

Participants. Twenty new participants were recruited to participate the experiment.

Apparatus, design and procedure. Colorful Mondrains and scene pictures were presented binocularly in two eyes. The Mondrains were presented binocularly at full contrast and continued to flash at 10 Hz from the beginning until the end of each trial; see procedure in Figure 1B. To obtain RTs close to those in the dichoptic viewing condition used in Experiment 1, the contrast of binocularly presented scene pictures was ramped up gradually from 0% to 100% within 5 s. The scene picture was presented binocularly at either above or below the central cross and stopped by participants once they detected any part of the scene picture. Other details were the same as in Experiment 1.

Results and Discussion

All participants' data were included for further analysis. Trials that had no response within 6 s, error responses of localization task, and RTs exceeding ± 2.5 SD from mean of individual participant were excluded (1.97%).

Results are shown in Figure 4. Two-tailed paired t test showed no significant difference in RT between the congruent condition (M = 2,175 ms [restaurant scene = 2,197 ms, street scene = 2,153 ms]) and the incongruent condition (M = 2,189 ms [restaurant scene = 2,200 ms, street scene = 2,178 ms]), $t(19) = 1.443$, $p = .165$, 95% CI: −6.44 ms, 35.04 ms. There was also no difference in accuracy between the congruent

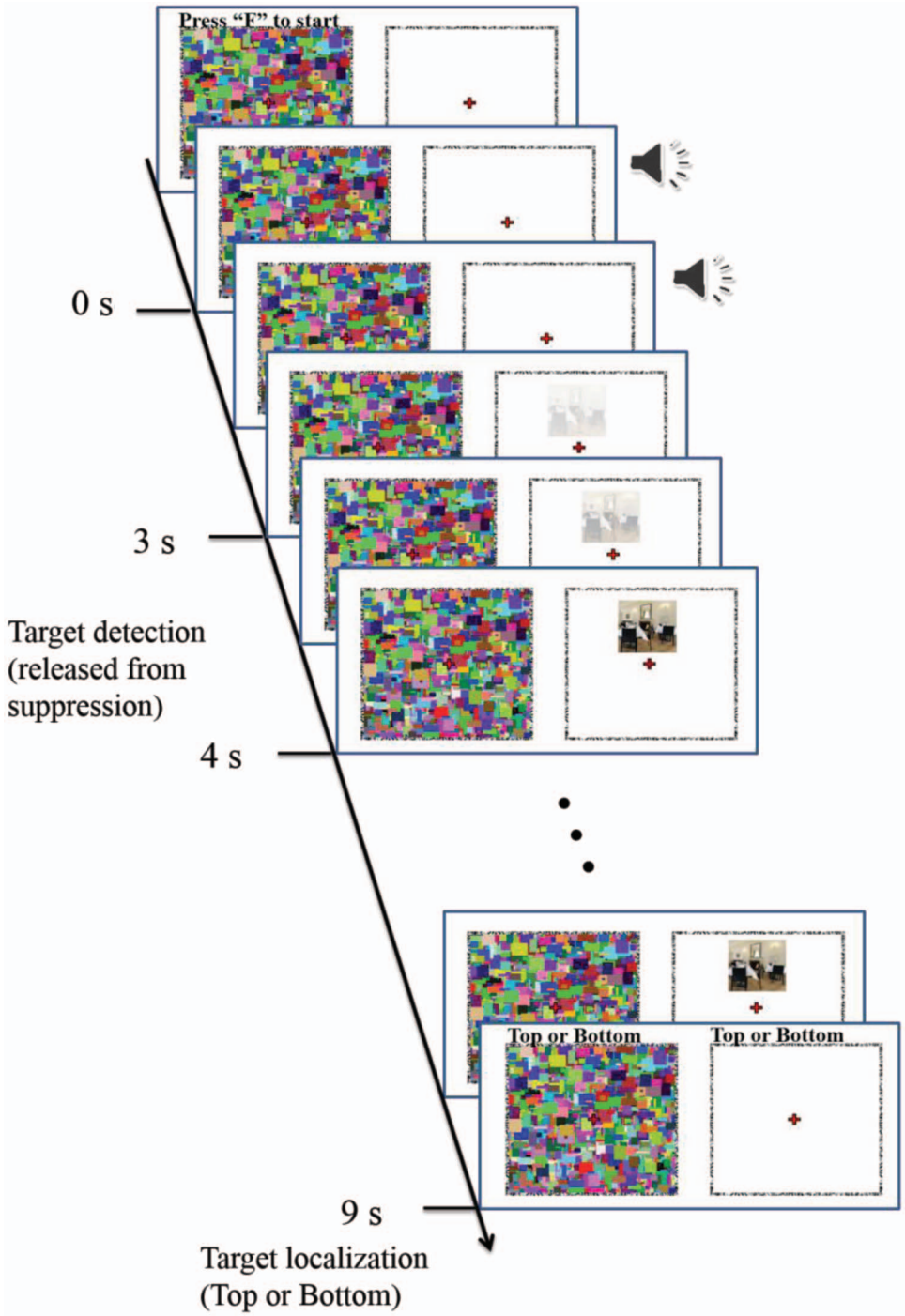

Figure 3. The procedure in Experiment 5. A semantically congruent or incongruent soundtrack was first presented for 3 s, followed by dichoptic viewing of colorful Mondrains and a scene picture until the end of the trial. See the online article for the color version of this figure.

condition (99.02%) and the incongruent condition (99.45%), $t(19) = 1.690$, $p = .107$, 95% CI: −0.11%, 1.01%, indicating no speed–accuracy trade-off.

Results of Experiment 2 clarified that the audiovisual congruency effect found in Experiment 1 was due to the difference in the processing speed between the congruent and incongruent conditions before the scene pictures emerged into consciousness, rather than different detection criteria corresponding to the different conditions (congruent and incongruent) after overcoming the suppression.

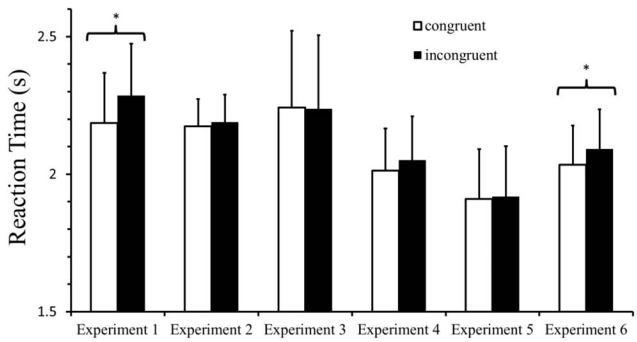

Figure 4. Results of six experiments in this study. White bars represent the semantically congruent condition and black bars represent the semantically incongruent condition between audiovisual stimuli. Asterisk indicates significant difference between conditions ($p < .05$). Error bars represent 1 *SEM* of the data. Experiment 1: full scenes; Experiment 2: binocular viewing; Experiment 3: objects only; Experiment 4: background only; Experiment 5: priming; Experiment 6: catch trials added.

Experiment 3: Objects Only

Unconscious audiovisual integration was found in Experiment 1. In Experiment 3, we aimed at clarifying that this unconscious audiovisual integration found in Experiment 1 occurred at the scene level rather than the object level.

In Experiment 3, for each of the two types of scene pictures, background was removed and only objects (plates and vehicles) were presented as target stimuli. If unconscious audiovisual integration also occurred for objects-only pictures, we should find a significant difference in RT between the congruent and incongruent conditions. If, however, unconscious audiovisual integration occurred only for full-scene but not objects-only pictures, there would be no difference in RT between the congruent and incongruent conditions in the objects-only condition, and this would strengthen the inference that unconscious audiovisual integration in Experiment 1 was due to full-scene pictures but not dispersedly presented objects within the scenes.

Method

Participants. A new group of 21 participants were recruited for Experiment 3.

Apparatus, design and procedure. Experiment 3 shared the same procedure as Experiment 1 except that objects-only pictures were used as visual stimuli. Background was removed from the original full-scene pictures to produce target stimuli using Photoshop software. Same two types of scene soundtracks (restaurant and street soundtracks in Experiment 1) were used as auditory stimuli in Experiment 3.

Results and Discussion

Six participants who exceeded 30% trials of no response within 6 s were excluded in Experiment 3. For the data from the remaining 15 participants, trials included no response within 6 s, error responses of localization task, and RTs exceeding $\pm$ 2.5 *SD* from the mean of individual participants were excluded (13.26%).

A two-tailed paired *t* test showed that average RT was not significantly different between the congruent condition (M = 2,243 ms [restaurant scene = 2,348 ms, street scene = 2,138 ms]) and the incongruent condition (M = 2,237 ms [restaurant scene = 2,307 ms, street scene = 2,167 ms]), $t(14) = 0.181$, $p = .859$, 95% CI: −79.07 ms, 66.78 ms. No difference in accuracy was found between the congruent condition (90.68%) and the incongruent condition (88.87%), $t(14) = 1.552$, $p = .143$, 95% CI: −4.45%, 0.71%, indicating no speed–accuracy trade-off.

An independent two-tailed *t* test was also conducted to compare the congruency effect between Experiment 1 and Experiment 3. The magnitude of the congruency effect was calculated by subtracting RTs in the congruent condition from RTs in the incongruent condition. A significant difference in the magnitude of congruency effect was found, $t(29) = 2.097$, $p = .045$, Cohen's $d = 0.754$, 95% CI: −207.93 ms, −2.59 ms; mean congruency effect was significantly greater in Experiment 1 (M = 99.12 ms) than in Experiment 3 (M = −6.15 ms).

The audiovisual integration from semantically congruent audiovisual pairs disappeared when objects-only pictures were used, suggesting that the congruency effect observed in Experiment 1 was not due to audiovisual integration at the object level. This result seems contradictory to previous studies that found unconscious audiovisual integration for objects (Alsius & Munhall, 2013). We suspect that the absence of audiovisual integration at the object level could be explained by the obscure appearance of the objects: the objects were presented dispersedly and not at the center of the pictures; they could fail to attract attention that might be necessary for audiovisual integration of objects. Also, the objects (or collection of objects as used here) embedded in complex scenes were not as discrete, large, and clear as isolated objects. In any case, these did not affect our conclusion: Results from Experiment 3 supports the hypothesis that unconscious audiovisual integration found in Experiment 1 occurred at the scene level rather than at the object level.

Experiment 4: Background Only

In Experiment 4, objects were removed and only background left in the pictures as visual stimuli. Background-only pictures used in Experiment 4 had the same scenery categorization as the full-scene pictures used in Experiment 1, because background-only restaurant and background-only street pictures were still categorized as restaurant and street scenes. However, the background-only pictures represent different details from full-scene pictures; for example the street without vehicles (background-only picture) represents a quiet street but a street full of vehicles (full-scene picture) represents a busy street. This way, we could also test whether unconscious audiovisual integration can be processed up to the level of identification—extracting the detailed semantic information of the scene pictures, and not only constrained by categorization. If identification was done, a street that had no vehicles would be identified as a quiet street and not integrated with the incongruent engine and horn street sounds, leading to no RT difference between the two conditions. If, however, unconscious audiovisual integration required a categorization process but not identification, no RT difference would be seen between the congruent and incongruent conditions, as street without vehicles

categorized as street was semantically congruent with the street sound.

Method

Participants. A new group of 25 participants was recruited for Experiment 4.

Apparatus, design, and procedure. Experiment 4 shared the same procedure as Experiment 1 except it used background-only pictures as visual stimuli. Objects were removed from the original full scene as target stimuli by using Photoshop software. Same two types of scene soundtracks (restaurant or street soundtracks) in Experiment 1 were used as auditory stimuli in Experiment 4.

Results and Discussion

Six participants who exceeded 30% trials of no response within 6 s were excluded. For the data from the remaining 19 participants, trials that had no response within 6 s, error responses of localization task, and RTs exceeding ± 2.5 *SD* from the mean of individual participants were excluded (11.20%).

Two-tailed paired *t* test showed no significant difference in RT, $t(18) = 1.020$, $p = .321$, 95% CI: −39.71 ms, 114.61 ms: average RT had no significant difference between the congruent condition ($M = 2{,}014$ ms [restaurant scene = 1,995 ms, street scene = 2,032 ms]) and the incongruent condition ($M = 2{,}051$ ms [restaurant scene = 1,993 ms, street scene = 2,109 ms]). No significant difference was found in accuracy between the congruent condition (94.42%) and the incongruent condition (93.89%), $t(18) = 0.909$, $p = .376$, 95% CI: −1.74%, 0.69%, indicating no speed–accuracy trade-off.

An independent two-tailed *t* test was also conducted to compare the magnitude of the congruency effect between Experiment 1 and Experiment 4. The magnitude of congruency effect was calculated by subtracting RTs in the congruent condition from RTs in the incongruent condition. No significant difference in the magnitude of congruency effect was found between Experiment 1 ($M = 99.12$ ms) and Experiment 4 ($M = 37.45$ ms), $t(33) = 1.179$, $p = .247$, 95% CI: −168.11 ms, 4.77 ms.

Results of Experiment 4 showed no significant RT difference between the congruent and the incongruent conditions, supporting the idea that identification was accomplished in unconscious audiovisual integration, as the unconscious audiovisual integration was eliminated when only the background was used in Experiment 4. However, when we compared the magnitude of the congruency effect between Experiment 1 and Experiment 4, we found no significant difference in the congruency effects between two experiments (recall that the magnitude of the congruency effect in Experiment 1 and Experiment 3 was significantly different). Comparing the results in Experiments 1, 3, and 4, it is reasonable to infer that gradient of information richness rather than an all-or-none process determines the possibility of audiovisual integration in scene perception. Full-scene pictures contain the richest information to integrate with the semantically congruent scenery soundtrack, followed by background-only pictures, and then objects-only pictures. This is indicated by the results that demonstrated that although background-only pictures generated scenery representations that close to full-scene pictures, richer representations from the full-scene pictures could lead to unconscious audiovisual integration easier. Background-only pictures (such as an empty street) were less likely to be associated with the auditory representation when the objects (vehicles) that produced the sound were absent. More importantly, the audiovisual congruency effect found here is not due to the objects that produce the sound, as inferred from the null results of Experiment 3, as object-only pictures were not effective enough to generate scene representation and be integrated with the congruent scenery soundtrack. Whether this is unique for subliminal audiovisual integration of scene or it is also applied to suprathreshold integration awaits further studies.

Experiment 5: Priming

It is possible that suprathreshold scenery soundtrack acted as a prime to trigger the concept of restaurant or street, and this conceptual priming caused the congruency effect in Experiment 1. It is well known that priming can improve target detection. For example, Evans, Horowitz, and Wolfe (2011) asked participants to judge whether they detected a specific target (e.g., animal, human or others) from the succession of scene pictures that were presented in RSVP. Higher accuracy was found when specific target names were cued before the successive scene pictures, rather than cued after. In our Experiment 1, participants may generate priming for scene representation in the long duration of the presentation of subliminal scene pictures when they were paired with suprathreshold soundtracks simultaneously. The possibility of a consciously mediated conceptual priming effect needs to be excluded to confirm that what obtained in Experiment 1 was indeed unconscious audiovisual integration in scene perception.

In Experiment 5, the scenery soundtrack was presented before the onset of the scene picture to clarify whether conceptual priming from the scenery soundtrack can facilitate the scene to release from suppression more quickly. If congruent facilitation was also found in this experiment, it is reasonable to infer that the audiovisual congruency effect found in Experiment 1 could be generated by priming rather than unconscious audiovisual integration.

Method

Participants. A new group of 20 healthy participants was recruited for Experiment 5.

Apparatus, design, and procedure. In Experiment 5, the soundtrack was presented for 3 s after participants pressed the F key to start, and then the scene picture and colorful Mondrains were presented dichoptically in different eyes after the soundtrack ended (see Figure 3). The colorful Mondrains were presented at full contrast and continued to flash at 10Hz until the end of the trial. The contrast of the scene pictures was ramped up from 0% to 100% from the time point of 3 s to 4 s, and remained at full contrast after 4 s until participants detected any part of the picture and pressed the key. Participants were told to ignore the soundtrack and focus on the visual task only. Based on the results from Experiment 1 to Experiment 4, approximate 3 s was enough for participants to detect the scene pictures.

Results and Discussion

We excluded three participants who exceeded 30% trials with no response within 6 s. For the data of the remaining 17 partici-

pants, the trials that had no response within 6 s, error responses of localization task, and RTs exceeding ± 2.5 *SD* from mean of individual participant were excluded (10.48%).

The result of Experiment 5 is shown in Figure 4. No significant difference in RT was found between the congruent condition (M = 1,910 ms [restaurant scene = 1,885 ms, street scene = 1,935 ms]) and the incongruent condition (M = 1,918 ms [restaurant scene = 1,848 ms, street scene = 1,988 ms]), $t(16) = 1.413$, $p = .685$, 95% CI: −34.35 ms, 50.95 ms. There was no significant effect in accuracy between the congruent condition (97.16%) and the incongruent condition (97.12%), $t(16) < 0.000$, $p = 1$, 95% CI: −1.47%, 1.47%, indicating no speed–accuracy trade-off.

An independent two-tailed t test was also conducted to compare the magnitude of the congruency effect between Experiment 1 and Experiment 5. The magnitude of congruency effect was calculated by subtracting RTs in the congruent condition from RTs in the incongruent condition. A significant difference in the magnitude of congruency effect was found between Experiment 1 (M = 99.12 ms) and Experiment 5 (M = 8.30 ms), $t(31) = 2.205$, $p = .035$, 95% CI: −174.82 ms, −6.83 ms.

To determine if the effect might be mediated entirely by a consciously mediated conceptual priming effect, we did Experiment 5 by presenting the scenery soundtrack before the onset of the scene picture. Based on the results that the congruent effect was absent in Experiment 5, and that the results of Experiment 5 differed significantly from those of Experiment 1, it seems unlikely that conceptual priming can account for the audiovisual congruency effect we observed in Experiment 1.

Experiment 6: Catch Trials Added

In Experiment 6, catch trials that contained no scene pictures were included to exclude participants who showed a high false-alarm rate and see whether the audiovisual congruency effect could still be replicated. In Experiment 1, the participants were required to press a button whenever they saw "any part of the scene" as soon as possible. Under this circumstance and because of the characteristics of CFS paradigm that the dynamic changing Mondrians were always present, it is sometimes ambiguous to differentiate the Mondrian patterns and the target scene pictures, especially when the scene pictures were at low contrast. Therefore, participants might take a generally lenient criterion or even an unreliable one to respond quickly for the detection task even without seeing (any part of) the scene. By adding catch trials and informing participants not to respond when there were no scene pictures, it is more likely that participants were indeed "subjectively seeing" the scene pictures when they pressed the key.

Method

Participants. A new group of 25 healthy participants was recruited.

Apparatus, design, and procedure. Experiment 6 shared the same procedures as Experiment 1 except for adding catch trials that did not contain scene pictures. Participants were required to respond as soon as possible when they saw any part of the scene and not to respond when they saw nothing except the Mondrians.

Experiment 6 included 320 trials interleaved with one self-paced rest interval and 160 trials in each of two blocks. The 64 catch trials (20% of all trials), 128 congruent and 128 incongruent audiovisual trials were presented in random sequence and randomly paired with eight soundtracks as in Experiment 1.

Results and Discussion

We excluded eight participants in Experiment 6. One exceeded 30% trials with no response within 6 s. The other seven participants had false alarm rates higher than 50% (chance level) and their data were excluded from further analysis. From the seven participants who had high false alarm rates in catch trials, two of them also had low average accuracy (65%), lower than 70%, one of the criteria that we used for excluding participants. However, the remaining five participants were found to have high average accuracy (87%) that would not be excluded in Experiment 1.

For the data from the 17 participants included, the trials that had no response within 6 s, error responses of localization task, and RTs exceeding ± 2.5 *SD* were excluded (11.60%).

The result of Experiment 6 is shown in Figure 4. The t test revealed significant difference in RT, $t(16) = 2.299$, $p = .035$, Cohen's $d = 0.557$, 95% CI: 4.39 ms, 108.48 ms. Average RT was significantly shorter in the congruent condition (M = 2,035 ms [restaurant scene = 1,953 ms, street scene = 2,116 ms]) than in the incongruent condition (M = 2,091 ms [restaurant scene = 1,983 ms, street scene = 2,199 ms]). There was no significant effect in accuracy between the congruent condition (97.03%) and the incongruent condition (96.54%), $t(16) = 0.636$, $p = .534$, 95% CI: −1.78%, 0.96%, indicating no speed–accuracy trade-off.

Five participants were found to have accuracy higher than 70% and high false alarm rates. It is reasonable to infer that these five participants were those who took a lenient criterion and misjudged the Mondrains as target. These participants were excluded and unconscious audiovisual facilitation was also replicated, indicating a real audiovisual congruency effect found in this study.

General Discussion

To examine whether semantically congruent soundtrack could influence the unconscious processing of complex visual scenes, we conducted six experiments in this study and the answer was positive: RTs were faster in the audiovisual congruent condition than the incongruent condition (Experiment 1), even when catch trials were added to include only participants who followed the instructions and responded when they indeed detected the scene (Experiment 6). This audiovisual congruency effect was found only with full-scene pictures (Experiment 1), but not with objects-only (Experiment 3) or background-only (Experiment 4) pictures. We also made efforts to exclude the possibilities that the audiovisual congruency effect found here was caused by participants' adopting different detection criteria in the congruent and incongruent conditions after the suppression was overcome (Experiment 2) or from a consciously mediated conceptual priming effect after the pure meaning of the scenery soundtrack was extracted (Experiment 5).

This is the first study demonstrating unconscious audiovisual integration with subliminal scene pictures. Previous studies have demonstrated that semantic information can be extracted from subliminal scene pictures presented briefly in RSVP (Potter et al., 2014) or interocularly suppressed under CFS as used here (Mudrik

et al., 2011). Our results expand the finding and suggest that the semantic information extracted from invisible scene pictures can further be integrated with congruent scenery soundtracks. In the past decades, researchers have explored the limitations of unconscious processing using varieties of methods (Kim & Blake, 2005) and fruitful results have been obtained regarding how far unconscious processing can go (Faivre et al., 2014; E. Yang & Blake, 2012). Our result moves the limits at which visual and auditory information of a complex scene can be extracted and integrated to an earlier stage without consciousness, which is more close to real-life visual worlds. Accordingly, our finding argues against the prominent theory that consciousness is indispensable for constructing information from diverse source into a unified concept (Marcel, 1983; Tononi & Edelman, 1998).

Our finding of audiovisual congruency effect in full-scene pictures that included background and objects (Experiment 1) but not in objects-only pictures (Experiment 3) suggests that gist plays an important role when extracting scenery information even without awareness. This is consistent with previous studies that objects can be identified with greater accuracy when accompanied by a coherent background rather than when objects were presented alone without background (Biederman, 1972; Biederman, Mezzanotte, & Rabinowitz, 1982; Boyce & Pollatsek, 1992; Boyce, Pollatsek, & Rayner, 1989; Davenport & Potter, 2004; T. E. Palmer, 1975). However, previous studies all concerned with scene perception in visual aspect only, without taking into consideration the fact that in daily life visual and auditory scenes usually appear together and we perceive audiovisual scenes constantly. Our study here thus helps extend the studies of scene perception from single sensory modality to multiple modalities.

Plass, Guzman-Martinez, Ortega, Grabowecky, and Suzuki (2014) demonstrated that invisible lip movements under CFS facilitated the classification of target spoken word when the lip movements were congruent with the spoken word. Because invisible lip movements did not produce a McGurk effect (T. D. Palmer & Ramsey, 2012), Plass et al. (2014) concluded that the audiovisual congruency effect they found did not occur at the syllable level; rather, it occurred at the word level. We have shown in the current study that objects alone did not produce the congruency effect, but the full scene did. Therefore, it may seem similar to the finding of Plass et al. (2014) that the unconscious audiovisual congruency effect occurs at the "whole" level but not the "component" level. However, the overall pattern of results in Experiments 1, 3, and 4 indicate that unconscious audiovisual integration of scene perception is more likely to be determined by the gradient of information richness provided in the scene pictures. That is, full-scene pictures (e.g., that contained street and vehicles), although no more likely than background-only pictures (e.g., that contained empty street) to successfully integrate with the congruent scenery soundtracks (e.g., engine and horn sound), they are more likely than objects-only pictures (e.g., that contained vehicles) to successfully integrate with the congruent soundtracks. This is important because it indicates that it is indeed the gist extracted from the scene (but not the objects producing the sounds) that contributes to the audiovisual integration we found here. Recent studies have demonstrated unconscious audiovisual integration of semantically congruent objects under the competition of binocular rivalry (Chen, Yeh et al., 2011) or bistable figure (Hsiao et al., 2012). Also, a matched face-voice pair has been shown to release from interocular suppression earlier compared to mismatched pair (Alsius & Munhall, 2013). The absence of unconscious audiovisual integration with objects-only pictures in our Experiment 3 may be due to the fact that the objects we used were a collection of the same type of objects distributed in the scene rather than a single isolated object as used in previous studies. More importantly, these results suggest that it is at the scene level but not the object level that contributes to the audiovisual congruency effect obtained in this study, and that it is the gradient of information richness rather than an all-or-none process that determines whether audiovisual integration would occur in scene perception.

What is the mechanism underlying the unconscious audiovisual integration of scene perception we found in this study? By using the backward masking paradigm assumed to be devoid of high-level feedbacks, VanRullen (2007) demonstrated that categorization of visual scene pictures can be accomplished within 150 ms without participant's awareness. However, when stimuli were presented for several seconds under the CFS paradigm we used here, high-level feedbacks may also be involved. The congruency effect we found here indicates that there is a link between suprathreshold auditory information and subthreshold visual information when a congruent soundtrack is played, and that link allows the visual information to overcome the interocular suppression faster. It takes time for the accumulation and extraction of information from visual and auditory processing to establish the link between them. During this period of time, enough recurrency may work to hold information online for semantic processing, but not enough for the access to awareness.

Based on the finding that a spoken word (e.g., "pumpkin") facilitated the release time of a subsequently presented invisible object (e.g., the picture of a pumpkin presented under CFS), Lupyan and Ward (2013) proposed that feedback signals sending from verbal labels to early visual processing can boost visual features diagnostic for objects to break through visual awareness faster. Similarly, our participants may extract from the meaning of the scenery soundtrack to form an abstract semantic representation, which then sends feedback signals to the ongoing processing to boost the otherwise invisible complex scene into visual awareness. According to Lupyan and Ward (2013), the ongoing process under CFS maintains at the perceptual level. However, based on previous studies showing that semantic processing survives CFS (Costello, Jiang, Baartman, McGlennen, & He, 2009; Y. H. Yang & Yeh, 2011) and that meanings of complex scenes can be extracted within 13 ms (Potter et al., 2014), the unconscious processing of the unseen complex scene may very well be semantic in nature, and the audiovisual link occurs at the semantic level.

Previous studies and theories about scene perception focused on visual aspect only, without taking into consideration the fact that in daily life our scene perception is multimodal. Our study here thus helps extend the studies of scene perception from single sensory modality to multiple modalities. For example, how would the global image features (i.e., the degree of naturalness, openness, roughness, expansion, and color) detected in natural scenes (Oliva & Torralba, 2001) be combined with the auditory soundtrack to help our unconscious processing of scene perception? Or, alternatively, is it the semantic regularity provided by the context of the scene (Oliva & Torralba, 2007) that combines with the semantically congruent soundtrack? Although answers to these questions need further investigations, our finding here suggests an expansion

of theories of scene perception to include unconscious audiovisual integration.

The audiovisual integration without waiting for the time-consuming consciousness feedback provides instantaneous activation to enable faster reactions in real life. For example, on a rainy or foggy day, the engine sound accompanied with the car rushing out from the roadside can facilitate the unconscious processing of subliminal visual stimuli and make us react faster. So, turn off the music or roll down your window to listen to your surroundings if you care about safe driving in foggy days.

References

Alsius, A., & Munhall, K. G. (2013). Detection of audiovisual speech correspondences without visual awareness. *Psychological Science, 24,* 423–431. http://dx.doi.org/10.1177/0956797612457378

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177,* 77–80. http://dx.doi.org/10.1126/science.177.4043.77

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14,* 143–177. http://dx.doi.org/10.1016/0010-0285(82)90007-X

Bolognini, N., Frassinetti, F., Serino, A., & Làdavas, E. (2005). "Acoustical vision" of below threshold stimuli: Interaction among spatially converging audiovisual inputs. *Experimental Brain Research, 160,* 273–282. http://dx.doi.org/10.1007/s00221-004-2005-z

Boyce, S. J., & Pollatsek, A. (1992). Identification of objects in scenes: The role of scene background in object naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 531–543. http://dx.doi.org/10.1037/0278-7393.18.3.531

Boyce, S. J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 556–566. http://dx.doi.org/10.1037/0096-1523.15.3.556

Chen, Y. C., Huang, P. C., Yeh, S. L., & Spence, C. (2011). Synchronous sounds enhance visual sensitivity without reducing target uncertainty. *Seeing Perceiving, 24,* 623–638. http://dx.doi.org/10.1163/187847611X603765

Chen, Y. C., & Yeh, S. L. (2008). Visual events modulated by sound in repetition blindness. *Psychonomic Bulletin & Review, 15,* 404–408. http://dx.doi.org/10.3758/PBR.15.2.404

Chen, Y. C., & Yeh, S. L. (2009). Catch the moment: Multisensory enhancement of rapid visual events by sound. *Experimental Brain Research, 198,* 209–219. http://dx.doi.org/10.1007/s00221-009-1831-4

Chen, Y. C., & Yeh, S. L. (2012). Look into my eyes and I will see you: Unconscious processing of human gaze. *Consciousness and Cognition: An International Journal, 21,* 1703–1710. http://dx.doi.org/10.1016/j.concog.2012.10.001

Chen, Y. C., Yeh, S. L., & Spence, C. (2011). Crossmodal constraints on human perceptual awareness: Auditory semantic modulation of binocular rivalry. *Frontiers in Psychology, 2,* 212. http://dx.doi.org/10.3389/fpsyg.2011.00212

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research, 1242,* 126–135. http://dx.doi.org/10.1016/j.brainres.2008.04.023

Costello, P., Jiang, Y., Baartman, B., McGlennen, K., & He, S. (2009). Semantic and subword priming during binocular suppression. *Consciousness and Cognition: An International Journal, 18,* 375–382. http://dx.doi.org/10.1016/j.concog.2009.02.003

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15,* 559–564. http://dx.doi.org/10.1111/j.0956-7976.2004.00719.x

Epstein, R. A. (2005). The cortical basis of visual scene processing. *Visual Cognition, 12,* 954–978. http://dx.doi.org/10.1080/13506280444000607

Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). When categories collide: Accumulation of information about multiple categories in rapid scene perception. *Psychological Science, 22,* 739–746. http://dx.doi.org/10.1177/0956797611407930

Faivre, N., Berthet, V., & Kouider, S. (2014). Sustained invisibility through crowding and continuous flash suppression: A comparative review. *Frontiers in Psychology, 5,* 475. http://dx.doi.org/10.3389/fpsyg.2014.00475

Fang, F., & He, S. (2005). Cortical responses to invisible objects in the human dorsal and ventral pathways. *Nature Neuroscience, 8,* 1380–1385. http://dx.doi.org/10.1038/nn1537

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision, 7,* 10. http://dx.doi.org/10.1167/7.1.10

Frassinetti, F., Bolognini, N., & Làdavas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research, 147,* 332–343. http://dx.doi.org/10.1007/s00221-002-1262-y

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 210–228. http://dx.doi.org/10.1037/0096-1523.25.1.210

Hsiao, J. Y., Chen, Y. C., Spence, C., & Yeh, S. L. (2012). Assessing the effects of audiovisual semantic congruency on the perception of a bistable figure. *Consciousness and Cognition: An International Journal, 21,* 775–787. http://dx.doi.org/10.1016/j.concog.2012.02.001

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

Jiang, Y., Costello, P., & He, S. (2007). Processing of invisible stimuli: Advantage of upright faces and recognizable words in overcoming interocular suppression. *Psychological Science, 18,* 349–355. http://dx.doi.org/10.1111/j.1467-9280.2007.01902.x

Kim, C. Y., & Blake, R. (2005). Psychophysical magic: Rendering the visible 'invisible.' *Trends in Cognitive Sciences, 9,* 381–388. http://dx.doi.org/10.1016/j.tics.2005.06.012

Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 110,* 14196–14201. http://dx.doi.org/10.1073/pnas.1303312110

Marcel, A. J. (1983). Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology, 15,* 238–300. http://dx.doi.org/10.1016/0010-0285(83)90010-5

Mudrik, L., Breska, A., Lamy, D., & Deouell, L. Y. (2011). Integration without awareness: Expanding the limits of unconscious processing. *Psychological Science, 22,* 764–770. http://dx.doi.org/10.1177/0956797611408736

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42,* 145–175. http://dx.doi.org/10.1023/A:1011139631724

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences, 11,* 520–527. http://dx.doi.org/10.1016/j.tics.2007.09.009

Olivers, C. N., & Van der Burg, E. (2008). Bleeping you out of the blink: Sound saves vision from oblivion. *Brain Research, 1242,* 191–199. http://dx.doi.org/10.1016/j.brainres.2008.01.070

Palmer, T. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition, 3,* 519–526. http://dx.doi.org/10.3758/BF03197524

Palmer, T. D., & Ramsey, A. K. (2012). The function of consciousness in multisensory integration. *Cognition, 125,* 353–364. http://dx.doi.org/10.1016/j.cognition.2012.08.003

Plass, J., Guzman-Martinez, E., Ortega, L., Grabowecky, M., & Suzuki, S. (2014). Lip reading without awareness. *Psychological Science, 25,* 1835–1837. http://dx.doi.org/10.1177/0956797614542132

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

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics, 76,* 270–279. http://dx.doi.org/10.3758/s13414-013-0605-z

Sampanes, A. C., Tseng, P., & Bridgeman, B. (2008). The role of gist in scene recognition. *Vision Research, 48,* 2275–2283. http://dx.doi.org/10.1016/j.visres.2008.07.011

Stein, B. E., London, N., Wilkinson, L. K., & Price, D. D. (1996). Enhancement of perceived visual intensity by auditory stimuli: A psychophysical analysis. *Journal of Cognitive Neuroscience, 8,* 497–506.

Stein, T., Senju, A., Peelen, M. V., & Sterzer, P. (2011). Eye contact facilitates awareness of faces during interocular suppression. *Cognition, 119,* 307–311. http://dx.doi.org/10.1016/j.cognition.2011.01.008

Tononi, G., & Edelman, G. M. (1998). Consciousness and complexity. *Science, 282,* 1846–1851. http://dx.doi.org/10.1126/science.282.5395.1846

Tsuchiya, N., & Koch, C. (2005). Continuous flash suppression reduces negative afterimages. *Nature Neuroscience, 8,* 1096–1101. http://dx.doi.org/10.1038/nn1500

VanRullen, R. (2007). The power of the feed-forward sweep. *Advances in Cognitive Psychology, 3,* 167–176. http://dx.doi.org/10.2478/v10053-008-0022-3

VanRullen, R., & Koch, C. (2003). Visual selective behavior can be triggered by a feed-forward process. *Journal of Cognitive Neuroscience, 15,* 209–217.

Yang, E., & Blake, R. (2012). Deconstructing continuous flash suppression. *Journal of Vision, 12,* 8. http://dx.doi.org/10.1167/12.3.8

Yang, E., Zald, D. H., & Blake, R. (2007). Fearful expressions gain preferential access to awareness during continuous flash suppression. *Emotion, 7,* 882–886. http://dx.doi.org/10.1037/1528-3542.7.4.882

Yang, Y. H., & Yeh, S. L. (2011). Accessing the meaning of invisible words. *Consciousness and Cognition, 20,* 223–233. http://dx.doi.org/10.1016/j.concog.2010.07.005

Yang, Y. H., & Yeh, S. L. (2014). Unmasking the dichoptic mask by sound: Spatial congruency matters. *Experimental Brain Research, 232,* 1109–1116. http://dx.doi.org/10.1007/s00221-014-3820-5

Received July 15, 2014
Revision received April 14, 2015
Accepted April 17, 2015 ■