The role of cross-modal semantic interactions in real-world visuo-spatial attention

Daria Kvasova

TESI DOCTORAL UPF / 2019

Supervisor:

Dr. Salvador Soto Faraco,

Center for Brain & Cognition

DEPARTAMENT DE CIENCIES EXPERIMENTALS I DE LA SALUD

Universitat Pompeu Fabra Barcelona

to my dear granny

Acknowledgements

This thesis would not have been possible without the generous help of many people.

I am grateful to Salva for giving me the opportunity to be a part of the group and to do my thesis under his supervision. He is the most intelligent person I know, and it was a pleasure to share this scientific journey with him. I also would like to thank him for his patience, for being always there for me and for the priceless sessions of psychoanalysis that a PhD student sometimes needs in a moment of frustration.

I would like to thank all the people from my group, especially Mireia, Luis and Manuela, for all the ideas, suggestions and help that they provided me during these years. Many thanks to master students Laia and Travis with whom I worked together on the projects included in this thesis.

I am grateful to Cristina, Florencia and Bea for the enormous help with paperwork related to science and also to my immigration status.

Many thanks to Xavi and Silvia for their help in the lab, I will never forget how computers start working only because they just enter the room.

Thanks for all the people in the department, I really enjoyed the community and I am very grateful for meeting so many good friends there.

I am grateful to my family for being amazing. Mum, Ilya, Dania, Maya and Paul – thank you for all the support that you gave me, especially for the last months. Without you it would not be possible.

Many thanks for all my dear friends that helped me and took care of me. I know that I was not the most pleasant person while writing the thesis.

At last I would like to thank my dear grandma. She raised me, she believed in me and she passed away just 2 months before this important moment of my life. I am sorry for being so slow and that you do not witness this moment with me… I dedicate my thesis to you.

Abstract

In our everyday life we must effectively orient attention to relevant objects and events in multisensory environments. The impact of cross-modal links for attention orienting to spatial and temporal cues has been widely described. However, real-life scenarios provide a rich web of semantic information through the different sensory modalities. Despite some previous studies have revealed an impact of crossmodal sematic correspondences, the results are mixed with regard to the conditions in which audiovisual semantic congruence can influence attention orienting. Furthermore, the vast majority of the research on crossmodal semantics used simple, stereotyped displays that are far from achieving ecological validity.

The present thesis attempts to close this gap by addressing the role of identity-based crossmodal relationships on attention orienting in scenarios closer to real-world conditions. To this end, the experiments presented here attempt to extrapolate and generalize previous findings in more realistic environments by using naturalistic and dynamic stimuli, and address the theoretical questions of task relevance and perceptual load. The outcome of the three empirical studies in this thesis lead to several conclusions. First, that the effect of audio-visual semantic congruence on attention is not strictly automatic. Instead, they suggest that some top-down processing is necessary for audio-visual semantic congruence to trigger spatial orienting. The second conclusion to emerge is that crossmodal semantic congruence can guide attention under goal-directed conditions in visual search, and also under free

observation in complex and dynamic scenes. Third, that perceptual load is a limiting factor for these interactions. These findings extend previous knowledge on object-based crossmodal interactions with simple stimuli and clarify how audio-visual semantically congruent relationships play out in realistic scenarios.

Resumen

En nuestra vida cotidiana debemos orientar efectivamente la atención a objetos y eventos relevantes en entornos multisensoriales. El impacto que tienen los enlaces intermodales en la orientación de la atención a señales espaciales y temporales ha sido ampliamente descrito. Sin embargo, los escenarios de la vida real proporcionan una rica red de información semántica a través de las diferentes modalidades sensoriales. A pesar de que algunos estudios previos han revelado un impacto de las correspondencias semánticas entre modalidades, los resultados se mezclan con respecto a las condiciones en que la congruencia semántica audiovisual puede influir en la orientación de la atención. Además, la gran mayoría de la investigación sobre semántica intermodal utilizó representaciones simples y estereotipadas que están lejos de alcanzar la validez ecológica.

La presente tesis intenta llenar esta brecha al abordar el papel que las relaciones intermodales basadas en la identidad tienen en la orientación de la atención en escenarios más cercanos a las condiciones del mundo real. Con este fin, los experimentos presentados aquí intentan extrapolar y generalizar hallazgos previos en entornos más realistas mediante el uso de estímulos naturales y dinámicos, y abordar cuestiones teóricas como la relevancia de la tarea y la carga perceptiva. El resultado de los tres estudios empíricos de esta tesis condujo a varias conclusiones. Primero, que el efecto de la congruencia semántica audiovisual en la atención no es estrictamente automático. En cambio, sugieren que es necesario un procesamiento de arriba hacia abajo para que la congruencia

semántica audiovisual desencadene en la orientación espacial. La segunda conclusión que surge es que la congruencia semántica intermodal puede guiar la atención en condiciones de búsqueda visual dirigida a un objetivo, y también bajo observación libre en escenas complejas y dinámicas.

Tercero, la carga perceptiva es un factor limitante para estas interacciones. Estos hallazgos amplían el conocimiento previo sobre las interacciones intermodales basadas en objetos usando estímulos simples y aclaran cómo las relaciones audiovisuales semánticamente congruentes se desarrollan en escenarios realista

Index

1. INTRODUCTION

Imagine yourself in the middle of a busy city street. Many objects that surround you are competing for your attention at the same time – people talking, cars passing by, a barking dog running next to you, billboards, the traffic light sounds, and the music from the street performer. In addition to these external events, you might have internal goals, like trying to find a friend you are meeting up with, or just walking and looking around. Whatever you do, your brain receives (and processes to some extent) a large amount of information from a variety of different sensory modalities.

The question of how we orient attention and, specifically how interactions between sensory modalities can affect this process has been under the spotlight of research for at least the last thirty years. Studies by the end of last century demonstrated that combined audio-visual information affects the deployment of attention (Spence & Driver, 2004; McDonald et al., 2001; Koelewijn et al., 2010; Talsma et al., 2010; Santangelo and Macaluso, 2012). Specifically, a large number of studies provide convincing evidence that an event in one sensory modality (sound) can induce shifts of attention in other modalities based on temporal and spatial proximity (Spence et al, 1998; McDonald et al., 2000; Van der Burg et al., 2008; Van den Brink et al., 2014). At that moment, cross-modal attention studies were an important step in cognitive neuroscience, as they brought attention experiments closer to the multisensory nature of real-life conditions. Despite that, the majority of those studies still used stereotyped stimuli like beeps and noise bursts in audition, or LED flashes and Gabor patches in vision, and they usually investigated attention cueing under

simplified set ups with just the experimentally relevant stimuli, devoid of meaning or context. These features rendered the experimental designs far from the street example that opened this section. If we think about the barking dog in the example, it is evident that our perception of it depends not only on the fact that sound is coming from the dog and it is synchronized with its muzzle movements, but also that barking sound and the sight of the dog share information about identity. The fact that the barking dog attract our attention will also depend on the perceptual context and what is our goal, among other things.

Furthermore, other cross modal studies have also demonstrated the role of meaning shared between two modalities on perception (Chen & Spence, 2011; Molholm, Ritter, Javitt, & Foxe, 2004; Pesquita, Brennan, Enns, & Soto-Faraco, 2013; Iordanescu, Guzman-Martinez, Grabowecky, & Suzuki, 2008; Iordanescu, Grabowecky, Franconeri, Theeuwes, & Suzuki, 2010; List, Iordanescu, Grabowecky, & Suzuki, 2014). Characteristic sounds were proven to increase sensitivity in many like visual detection (Chen & Spence, 2011), object recognition (Molholm et al., 2004) or visual search (Iordanescu et al., 2008; 2010).

Despite studies on cross-modal attention and on cross-modal semantic effects in perception are numerous, still relatively few studies have addressed whether these audio-visual semantic relationships can influence attention orienting. Answering this question is important to understand how these two important multisensory processes (attention and semantics) play out in real world environments. In an attempt to answer this question, previous

studies have demonstrated that cross-modal semantic congruence can affect spatial orienting, but it is still unclear under which conditions this effect appears (Iordanescu, Guzman-Martinez, Grabowecky, & Suzuki, 2008; Iordanescu, Grabowecky, Franconeri, Theeuwes, & Suzuki, 2010; Mastroberardino, Santangelo & Macaluso (2015). What is more important, semantic congruence in spatial orienting so far has been studied using isolated artificial audio-visual stimuli that lacked context and ecological validity.

This thesis attempts to take one step toward understanding the role of audio-visual semantic interactions on attention in realistic conditions. First of all, we study whether crossmodal congruence can summon attention under different task constraints and levels of perceptual load. These are variables relevant to the real-world generalization, and that have not been varied systematically in previous studies (Chapter 2.1). Second, we take a step further to a more realistic and ecologically valid experimental designs and addressed the interaction between attention and audio-visual semantic congruence using complex and dynamic stimuli. In particular, we studied whether characteristic spatially uninformative sounds can guide attention to a corresponding visual event present in the scene (Chapter 2.2). Finally, we also addressed whether this semantic-guided attention orienting in complex dynamic stimuli modulates spontaneous visual exploration under during free observation of complex scenes (Chapter 2.3). The present chapter will introduce the current state of research on the topic, the scope of the thesis and the general hypotheses, whereas the last chapter

(Chapter 3) will discuss the general relevance of the findings and draw the conclusions from this work.

2.1 Multisensory integration and attention

One first relevant issue for discussion is what is the relationship between multisensory integration and attention. The question has been under debate for years. Below, I present a condensed summary.

Traditionally, it was thought that we integrate multisensory events in a bottom-up pre-attentive way, independently from the focus of attention. What is more, as a result of these bottom-up integration processes, attention could be dragged to the multisensory object in an automatic way (Bertelson, Vroomen, De Gelder, & Driver, 2000; Driver, 1996; Van Der Burg, Olivers, Bronkhorst, & Theeuwes, 2008; Vroomen, Bertelson, & de Gelder, 2001). Initially, this effect was demonstrated using low-level cues and spatial cueing between modalities. For instance, McDonald et al 2000 demonstrated that sudden sound increases detection of the flash when it appears in the same location. Along this line, several studies demonstrated that spatially uninformative but temporally correlated coherent sounds can attract attention in an automatic way (Van der Burg et al., 2008; Van den Brink et al., 2014). Remarkably, a variety of results demonstrated that orienting of attention could ocur based on the outcome of multisensory integration, therefore suggestion that integration preceded attention (Driver, 1996; Vroomen et al., 2001a, 2001b).

However, a more recent view on the topic debates pure automaticity of the multisensory integration process, and instead proposes that both bottom-up processing and top-down mediation play a role in the interaction between attention and multisensory processing (Koelewijn et al., 2010; Talsma, 2010; ten Oever et al., 2016; Soto-Faraco et al., 2019). According to this more nuanced view, the answer to this question about attention and multisensory integration will depend on a variety of factors like perceptual load, task relevance and crossmodal correspondences, that could be based not only on spatio-temporal proximities but also on semantic identity-based information.

2.1.1 Cross-modal semantic interactions

The semantic interactions between modalities has been addressed mostly in the context of object recognition or identification literature. In the early 2000s a big body of evidence was produced to support the idea that the representation of objects in the brain is multimodal (Amedi, von Kriegstein, van Atteveldt, Beauchamp, & Naumer, 2005; Beauchamp, Argall, et al., 2004; Beauchamp, Lee, et al., 2004; Molholm et al., 2004; Smith et al., 2007; von Kriegstein et al., 2005). Several of these results showed higher activation in polysensory areas of the temporal cortex (pSTS and MTG) when semantically combined (object-based) audio-visual information was presented, compared to semantically incongruent presentations (or just neutral objects such as scrambled therefore meaningless pictures). Those demonstrations were important to establish the role of object-based information in multisensory integration (for review, Doehrman & Naumer 2008) and led to a

further investigation of the effects that crossmodal sematic correspondences might have. Therefore, researchers started to unfold the impact of object content on crossmodal interactions and its influence on behavior using increasingly more meaningful stimuli and isolating effects of high-level crossmodal correspondences from low-level.

Laurienti et al. (2004) demonstrated how crossmodal semantic congruence improves visual discrimination, although in this case semantic congruence was limited to matching colour patches with colour names. Accordingly, Molholm et al. (2004) found that participants respond faster to the combined semantically congruent auditory and visual cues about objects, in comparison to just single modality ones. This finding in behavior went along with observed modulations of the ERP component associated with early visual object processing (N1), suggesting that visual identification is enhanced when semantically congruent information is provided by visual and auditory modalities (Molholm et al., 2004). Further, naturalistic sounds were proven to increase sensitivity in visual detection (Chen & Spence, 2011), identification (Chen & Spence, 2010), boosting visual events into dominance in binocular rivalry (Chen et al., 2011; Cox & Hong, 2015) and improving performance in picture naming (Mädebach et al., 2017). Together those findings have proven a role that object-based crossmodal information plays in visual perception, above and beyond simple (low-level) stimulus properties such as location, orientation, motion direction, contrast, etc.

2.1.2 Cross-modal semantics and attention orienting

The studies discussed above established that cross-modal combination of semantic cues can have an impact in brain activity related to object recognition, and also improve perception. The present thesis focuses on the question of whether these audio-visual semantic congruence relationships play a role in orienting attention. What does the literature tell us on this subject so far?

Iordanescu et al. (2008) showed that characteristic sounds, even if spatially uninformative, can speed up visual search when consistent with the visual target. In this study, participants were presented a cue word indicating the target of search and then an array of 4 common objects (animals, musical instruments, vehicles, etc.) placed in 4 quadrants of the computer monitor. The presentation of objects was accompanied with a sound that could be either consistent with the search target, consistent with one of the 3 distractors present in the array, or unrelated (a sound that did not correspond to any object in the array). The results showed that participants found the target object significantly faster when the semantically congruent sound had been played. These results, obtained using manual responses, were later supported with eye movement response (Iordanescu et al., 2010) and using rare target objects (Iordanescu et al., 2011). Interestingly, a characteristic sound that was congruent with the search target guided attention to the visual target, however when the sound was congruent with a distractor object, this effect was no longer present. That is, attention was not summoned to the non-target object related to the sound. Iordanescu et al. (2008) suggested that the cue word activates an

attentional template and the semantic network that is related to the object and further the consistent sound cross-modally enhance visual processing of the corresponding target-object. Since target-consistent sound decreased search latencies whereas distractor-consistent sounds did not slow down search, the authors suggested that the effect of crossmodal semantic congruence may appear only in a goal-directed manner. meaning that audio-visual pair that is relevant to the current goal (task) will attract attention whereas the irrelevant will not. This assumption supported a previous finding of Molholm et al. (2004) where it was demonstrated that object identification is enhanced when semantically consistent information is provided by visual and auditory modalities. Therefore, if one wanted to make a parallel to a real word situation, we could say that if you are already looking for a dog on the street (you have set up a search template, and expectation that includes semantic properties), then a characteristic barking sound will facilitate the processes involved.

2.1.3 The importance of task relevance

One interesting question further arises as to whether cross-modal congruence can attract attention despite being irrelevant to a current goal or when there is no specific task at all? That is, in the situation above, if the dog bark will make us orient visually toward the dog even if what we are looking for is our car. Answering this question not only has theoretical importance, but also practical implications for road safety, systems of alerting, or advertising strategies, to provide a few examples.

The first study that addressed the role of semantic crossmodal correspondences on attention was a recent study of Mastroberardino et al. (2015). They demonstrated a small effect whereby semantically congruent events (static image of an object with its corresponding sound) could capture spatial visual attention and thus increase discrimination performance on an upcoming visual target presented at that location. Note that in this study, sounds were presented centrally and could not therefore attract visual attention by means of low-level cues to spatial location. The effect appeared only in some of the condition's tests (high difficulty). The authors concluded that attentional capture was caused by the semantically congruent but task irrelevant audio-visual event. However, it might be the case that visual events in this study were not absolutely task irrelevant, since the study used only two possible objects (cat and dog) and they were always presented in one of the two possible positions where the upcoming visual target would appear. This experimental set up made the competition for attentional processing between stimuli relatively low when the irrelevant objects appeared. It is well known that under these conditions irrelevant stimuli are likely to receive some processing even if not strictly necessary for the task (Lavie et al., 1995). What is more, because visual targets could appear only at one of the possible positions occupied by the irrelevant audiovisual objects, one could argue that the irrelevant objects could have acted as location placeholders for the upcoming targets, hence somehow relevant for the participants.

Another study of Nardo et al. (2014) has offered evidence on the question of crossmodal congruence and spatial orienting under task

irrelevant conditions. Here they demonstrated that semantic congruence (or incongruence) between sounds and visual events did not have any effect on spatial orienting. Interestingly, in this study, no task was used, and participants were observing visual scenes freely. Also, it is important to mention that in this study the main manipulated variable was spatial congruence between visual and auditory stimuli. Semantic congruence was introduced for spatially and temporally correlated events and therefore the possible effects of semantic congruence could not be completely singled out from other low-level correspondences (which was not the goal of the study).

Based on the outcome of these studies, however, no consistency is found so far. Some suggest that semantic congruence, even if irrelevant to the task, has an effect of attracting attention and some suggesting that it does not. Because of the great variability between different studies, an important question remains open: what are the conditions in which cross-modal semantic relationships influence orienting behaviors?

1.1.4 Potentially important differences between studies

If we compare the findings of the studies discussed above, we will see that one of the important differences that varied between studies was the relevance of the audio-visual event to the current goal of the task. The task relevance of the crossmodal stimuli was manipulated in the aforementioned studies from explicitly relevant (REF) to completely irrelevant (Nardo et al., 2014). The studies of Iordanescu (2008; 2010) demonstrated how, in a goal-orienting

paradigm, characteristic sounds drove attention to the visual object that was explicitly relevant to the current task. Instead, when the observer has no particular goal, under free observation and with no task constraints, crossmodal semantic congruence seems to not have an impact on visual spatial orienting (Nardo 2014). On the other hand, the relatively irrelevant audiovisual event summoned attention, if only in some conditions, in the study of Mastroberardino et al. (2015). However, in this study, irrelevant visual objects consistently marked the positions where the upcoming visual target could potentially appear. Also, low uncertainty between stimuli and the sequential presentation of the events leads to the low perceptual load that benefits the processing of irrelevant events (Lavie & Tsal, 1994). Therefore, the assumed irrelevance of crossmodal pair in this study is arguable. Importantly, if completely irrelevant audio-visual semantic pair could attract attention then it would mean that crossmodal semantic congruence can attract attention in an automatic way, similarly to the spatial or temporal crossmodal congruence.

Inferring from these results it might be the case that crossmodal sematic congruence has an effect on attention only when it bears some relevance to the task. The strong automaticity of these effects raises some doubts and needs to be addressed in the new paradigm where irrelevance to the current goal and perceptual load are controlled. This is one of the questions that we want to address in the present dissertation.

1.2 EXTRAPOLATION TO REAL-LIFE SCENARIOS

1.2.1 Cross-modal semantic effects in real-life

So far, the discussion has concentrated on the importance of task relevance for the cross-modal semantic effects of attention. However, there is another dimension in which the few prior studies in the area have differed widely: Ecological validity. Most of the demonstrations of audiovisual semantic effects on attention and behavior in general have used artificial simplified designs without any meaningful context. This detracts from ecological validity. This section briefly addresses the importance of generalizing laboratory findings to realistic contexts.

Traditionally, one hope of scientific research is that the findings from laboratory research are generalized to real life. However, the striking difference between highly controlled experimental setups and reality could sometimes lead to the lack of validity of the findings or simply the inability to apply the results in the real world. It is possible that a phenomenon that is studied in isolation, will change or vanish under the many uncontrolled variables that are brought about in real world conditions. The ongoing trend in cognitive neuroscience now consists of addressing the problem of bringing the research closer to the ecological validity and several researches have already marked the importance if this process (Peelen & Kastner, 2014; Matusz 2018; Soto Faraco et al 2019; Spence & Soto-Faraco, 2019).

Previous studies have already made a point regarding differences in how visual attention operates in naturalistic, real-life scenes

compared to simple and artificial displays that are used traditionally in psychophysical studies (Kingstone et al., 2003; Wolfe, Horowitz, & Kenner, 2005; Nardo et al., 2011; Peelen & Kastner, 2014; Henderson & Hayes, 2017). Here it is important to take into account that when we orient attention in everyday life, we often do so amongst scenes populated with meaningful events embedded in meaningful scenes and not at just basic visual features (like horizontal/vertical bars or flashes) placed on the grey screen. Unlike these simple statistical structures of artificial stimuli, natural environments give us an overwhelming amount of complex information layers. However, despite this, humans are surprisingly efficient at perceiving this information and orient within it, by selecting relevant events. We can extract abundant information from natural scenes at a glance, quickly building up expectations from the spatial layout and functional connections between objects (Biederman, Mezzanotte, & Rabinowitz, 1982; Greene & Oliva, 2009; Peelen, Fei-Fei, & Kastner, 2009; MacEvoy & Epstein, 2011). Also, experience and repetition play a role in visual search within natural environments (Shiffrin & Schneider, 1977; Evans, Georgian-Smith, Tambouret, Birdwell, & Wolfe, 2013; Kuai, Levi, & Kourtzi, 2013). Apart from the simple sensory characteristics of external visual events, internal processes related to current tasks, goals, expectations contribute as well to the distribution of attention (Yantis, 2000).

In the case of multisensory research, the approach to the impact of combined sensory information on behavior and particularly on attention is largely unknown. One of the first attempts to study

spatial orienting in complex and multisensory scenes was performed in the cited study by Nardo and colleagues (2014). In this fMRI study, they presented videos of everyday life scenes containing sounds that could be spatially and/or semantically congruent with particular events in the scene. They found that multisensory brain areas, such as the posterior parietal cortex, displayed an increased BOLD activity when auditory stimuli were spatially congruent to the visual ones (e.g., a character bouncing a ball on the right side of screen, combined with a sound arising from the right side). This crossmodal spatial effect was found to be independent of varying semantic congruence between sounds and visual events. However if we go back to the findings explained in section X, it is clear that these results (regarding lack of semantic effects) seem at odds with previous results in less realistic and more controlled laboratory contexts that clearly state the crossmodal semantic effect on brain responses and behavior (Amedi, von Kriegstein, van Atteveldt, Beauchamp, & Naumer, 2005; Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Beauchamp, Lee, Argall, & Martin, 2004; Iordanescu et al., 2008, 2010; Laurienti et al., 2004). One of the reasons why semantically related audio-visual events were not more salient than unrelated ones might be explained by the presence of spatio-temporal crossmodal correspondences. Possibly congruence based on meaning was less salient than bottom-up salience effects of low-level audio-visual correspondences. Also, it is important to note that the presentation of stimuli in the study of Nardo et al. 2014 was lateralized, and sounds were either spatially congruent or incongruent to the one main visual event on one of the sides. The

imbalance between the saliency of low-level cues such a spatial coincidence, and higher-level cues such as semantic congruence, arises que question of bottom-up versus top-down processes.

1.2.2 The interplay between bottom-up and top-down processes in attention guidance

If we think about a complex audiovisual environment, as we did in the very beginning of the introduction, we will have many events that are correlated to each other temporally and spatially. Some of these coincidences might be spurious, and others signal actual multisensory objects. In such a context, it might be of particular interest to apply cross-modal object-based congruence constraints. These constrains can be beneficial in order to achieve spatial orienting in complex and dynamic scenes where the location of audiovisual events is uncertain and high-level sematic correspondences are isolated from low-level physical ones. However, these two types of information might possibly involve different types of orienting mechanism.

Current models of attentional deployment focus on either top-down or bottom-up mechanisms, and all need to address the complex interplay between these two processes. One of the traditional and most influential studies on attentional guidance by bottom-up mechanisms was conducted by Koch and Ullman (1985), proposing the idea of a saliency map that is based on luminance, contrast, size, color and orientation. According to this framework, attention guidance could be predicted according to these basic visual features (Itti,Koch & Niebur, 1998; Itti &Koch, 2001; Harel, Koch &

Perona, 2006). Alternatively, cognitive guidance theories state that attention orienting depends highly on the distribution of meaning in the scene, spatial location of meaningful events and previous experience (Potter, 1975; Henderson & Hollingworth, 1999; Wolfe & Horowitz, 2017). The recent framework of Henderson and Hayes (2017) proposed that both meaning and visual salience account for the distribution of attention, but only meaning predicts the unique variance of attention in complex scenes.

However, these models are mostly based on visual studies. To further approximate the reality, an important line for the research is to study how attention operates in situations that include multiple modalities because it is plausible that meaningful audio-visual information can also guide spatial orienting. Therefore, in the second line of research of the present dissertation, we address the question of how the meaningful information shared between auditory and visual modalities influence attention in real-world scenes, when low-level cues provide limited information.

In the following section, we formulate the hypotheses and introduce the studies that address them in the present dissertation.

1.3 SCOPE AND HYPOTHESES

1.3.1 Cross-modal sematic effects on spatial orienting: task relevance and perceptual load

Various studies show that cross-modal semantic relationships play a role in perception, however, it is still unclear if, or under which circumstances, spatial attention orienting can be guided by auditory semantically congruent information. A related question is whether cross-modal semantic congruency automatically attracts attention. At present the outcomes of different studies have been inconsistent (Iordanescu et al., 2008; Nardo et al., 2014; Mastroberardino et al., 2015). The difference in task-relevance of the audiovisual event from explicitly relevant, to completely irrelevant together with variation in perceptual load may account for the previous controversial findings. In the first study of the present thesis, we hypothesized that the effect of crossmodal semantic congruence will occur when at least one of two conditions apply: (a) the multisensory object (or one of its components) is task relevant or, (b) even if the multisensory object is irrelevant, but is presented under low perceptual load. We addressed this hypothesis with experiments using visual search arrays composed of images of everyday life objects (animals, vehicles, musical instruments, etc.) presented with sounds (congruent, incongruent, or neutral). We varied task relevance of the audiovisual object and perceptual load conditions. These experiments are presented in Chapter 2.1.

We found audiovisual semantic congruence influenced attention when it is relevant to the task, or when irrelevant but presented under low perceptual load. This way characteristic sounds guided attention to the corresponding visual image. However, the audiovisual pair was task-irrelevant and perceptual load was high, the sound did not summon orienting to the visual image. These findings lead us to conclude that semantic crossmodal congruence does not attract attention is an automatic way and requires some top-down processing in order to emerge.

1.3.2 Audio-visual semantic effects on visual search in complex scenes

Previous studies demonstrated that characteristic sounds can enhance performance in different visual tasks and that this effect is more likely to emerge in task-relevant and goal-directed paradigms. However, all the previous demonstrations were provided in simple and stereotyped displays that lack ecological validity. Please note that the only study which used close-to-realistic conditions did not address effects on visual search since there was no task (Nardo et al., 2014). Given the importance of generalizing the laboratory research into the real world, in the second study of the present dissertation, we address the identity-based crossmodal congruence effects in naturalistic (close to real world) scenarios. We designed a visual search task using complex, dynamic scenes of everyday life events or footage from movies or video-clips. In this study, participants searched for common objects in these videos, whilst

semantically consistent but spatially uninformative auditory cues were embedded in background noise.

We hypothesized that, if crossmodal semantic congruency guides attention in complex, dynamic scenes, then search times should be faster when the sounds are consistent with the object of search in comparison to when the sounds are consistent with distractor objects, neutral, or when no sound is presented (e.g., target-consistent characteristic sounds will help attract attention to the corresponding visual object). We found that, in these naturalistic scenes, characteristic sounds do improve visual search for task-relevant objects but fail to increase the characteristic sounds salience of irrelevant distracters. Our findings generalize previous results on task-relevant object-based crossmodal interactions with simple stimuli and demonstrate how audio-visual semantically congruent relationships play out in real life contexts.

1.3.3 Audio-visual semantic effects on free observation of real-life scenes

At this point, the first study of this dissertation, presented in Chapter 2.1, demonstrated that crossmodal semantic congruence can influence attention when even if it is irrelevant to the current goal but only when perceptual load is low. The experiment in Chapter 2.2 demonstrated that characteristic sounds can speed up visual search for everyday life objects when embedded in natural and dynamic environments. These results suggest that object-based enhancement occurs in a goal-directed manner. The third study of the current thesis, presented in Chapter 4, investigated crossmodal

semantic congruence drives visual attention also under free-viewing conditions, that is when the observer does not have a specific task. In order to address this question, we designed an eye-tracker study with audio-visual dynamic scenes similar to the second study described in the previous sub-section.

We hypothesized that semantically consistent sounds would increase the salience of the corresponding (irrelevant) visual object, and therefore the probability of directing overt attention toward the visual object would increase. In particular, we computed in how many of the videos the object of interest is looked at under different sound conditions (consistent, neutral or no sound), the total dwell time spent looking at the object, number of fixations made at the object, and time to first fixation inside the area of interest.

We found that characteristic sounds increased the percentage observations, the number of fixations and the total dwell time spent on the object of interest in comparison to the neutral or no sounds. This finding suggests that crossmodal semantic congruence indeed has an effect on gaze and eye movements, and therefore on attention orienting, even under free observation of real-world scenes.

The following section contains 3 experimental studies that are presented in form of submitted articles.

3. EXPERIMENTAL STUDIES

2.1 Not so automatic: Task relevance and perceptual load modulate cross-modal semantic congruence effects on spatial orienting

Kvasova, D., & Soto-Faraco, S. (2019)

Characteristic sounds facilitate object search in real-life scenes

Biorxiv

https://doi.org/10.1101/830679

Not so automatic: Task relevance and perceptual load modulate cross-modal semantic congruence effects on spatial orienting

Daria Kvasova[1] & Salvador Soto-Faraco[1,2]

[1] Center for Brain and Cognition, Universitat Pompeu Fabra, Barcelona

[2] ICREA, Barcelona

Corresponding autor: Daria Kvasova

Pompeu Fabra University

Edifici Merce Rodoreda (Room 24.326)
Carrer de Ramon Trias Fargas, 25-27
08005 Barcelona
Spain

daria.kvasova@upf.edu

Abstract

Recent studies show that cross-modal semantic congruence plays a role in spatial attention orienting and visual search. However, the extent to which these cross-modal semantic relationships attract attention automatically is still unclear. At present the outcomes of different studies have been inconsistent. Variations in task-relevance of the cross-modal stimuli (from explicitly needed, to completely irrelevant) and the amount of perceptual load may account for the mixed results of previous experiments. In the present study, we addressed the effects of audio-visual semantic congruence on visuo-spatial attention across variations in task relevance and perceptual load. We used visual search amongst images of common objects paired with characteristic object sounds (e.g., guitar image and chord sound). We found that audio-visual semantic congruence speeded visual search times when the cross-modal objects are task relevant, or when they are irrelevant but presented under low perceptual load. Instead, when perceptual load is high, sounds fail to attract attention towards the congruent visual images. These results lead us to conclude that object-based crossmodal congruence does not attract attention automatically and requires some top-down processing.

Introduction

Interactions between sensory modalities and their influence on perception and behavior have been convincingly demonstrated over the past decades. For instance, in multisensory contexts, information from different senses influences the deployment of spatial attention (McDonald et al., 2001; Koelewijn et al., 2010; Talsma et al., 2010; Santangelo and Macaluso, 2012). This way, lateralized sounds can produce a shift of attention that facilitates the processing of a visual target presented at that (congruent) location (Spence et al, 1998; McDonald et al., 2000). Even if spatially uninformative, auditory stimuli can enhance the processing of visual events that are temporally congruent (Van der Burg et al., 2008; Van den Brink et al., 2014).

These attention effects by congruent audio-visual stimuli has previously been observed using simple stereotyped objects i.e. Gabor patches, beeps, flashes, by manipulating congruence between low-level attributes such as spatial location or time. Yet, in the real world, multisensory events do not only provide temporally and spatially correlated information but also convey higher-level information about the identity of the object. Like lower level spatio-temporal features, these higher-level attributes can bear congruence relationships, arising from their semantic associations. It is therefore possible that in the natural environment object-based (semantic) relations between sounds and visual events might have an influence on attention orienting. Several recent studies have addressed the role of crossmodal semantic congruence on spatial orienting by investigating how characteristic sounds of objects (musical

instruments, vehicles, animals etc) or semantically congruent tactile information can enhance performance in different visual tasks (e.g., Laurienti et al., 2004; Chen and Spence, 2011; Molholm et al., 2004; Pesquita et al., 2013; Iordanescu et al., 2008; Iordanescu et al., 2010; List et al., 2014). However, the results of these studies are mixed, some suggesting that semantic congruence effectively attracts attention and some suggesting that it does not. Because of the great methodological variability between different studies, an important question remains as to what are the conditions in which cross-modal semantic relationships influence orienting behaviors. Answering this question can shed some light on the underlying processes supporting cross-modal semantic interactions.

Nardo et al. (2014) reported that crossmodal semantic congruency between visual events and sounds had no effect on spatial orienting or brain activity (measured with fMRI) when observers watched videos of everyday life scenes. In contrast, another study by Mastroberardino et al. (2015), using static images of objects, reported that attention was oriented toward the image semantically congruent (albeit spatially uninformative) sound presented at the same time. Along similar lines, Iordanescu et al. (2008, 2010) showed that characteristic sounds, even if spatially uninformative, speeded up visual search when consistent with the target object. Conversely to the study of Nardo et al. (2014) which found no effect, Iordanescu et al. and Mastroberardino et al. used simple static images presented in decontextualized search arrays (Iordanescu et al., 2008, 2010). One could argue that this might be the reason for the different result. Indeed, both the dynamic nature

of natural scenes and their complexity, have been pointed out as important gaps in the generalization of laboratory research findings to real-world contexts (e.g., Hasson et al., 2010). However, a recent study from our laboratory has addressed these potential explanations by demonstrating that characteristic sounds crossmodally enhance visual search of relevant objects even in complex and dynamic real-life scenes (Kvasova et al., 2019). Therefore, the static stimuli and lack of context in previous studies might not fully account for the difference in the results between previous studies. Here, we investigate whether task relevance might be a factor.

Task-relevance is another possibly important variable that has varied significantly across studies in prior research on cross-modal semantic effects on spatial attention. Iordanescu et al have shown that characteristic sounds, even if spatially uninformative, speed up search times for congruent visual targets (Iordanescu et al., 2008, 2010). In these studies, the visual search array contained four competing stimuli and the visual event was a target itself, the audio-visual object was in this case completely task-relevant. A similar method was applied in Kvasova et al. (2019) expect that the objects were embedded in more realistic video clips, with equivalent results. These studies showed consistent effects of cross-modal semantic congruence when the audio-visually congruent object is relevant for the task at hand (see also Iourdanescu et al., 2008, 2010).

What happens when the audio-visual congruent object is task-irrelevant? In Nardo (2014) participants were asked to freely

observe videos without any particular task requirement. In this case, cross-modal semantic relations had null effects on orienting (measured with eye-tracking). In Mastroberardino et al (2015), participants did perform a task, but the audio-visual semantic congruence was putatively task irrelevant. In the cited study, participants were asked to discriminate the orientation of a visual target (a Gabor grating) presented to one side (left or right) of central fixation. However, right before the relevant visual target appeared, a pair of irrelevant images of animals were presented at the corresponding left/right locations where the upcoming targets could appear. What they found is that when a central sound was semantically congruent with one of the two images, then discrimination performance of the visual target presented later at that location improved. Mastroberardino et al. concluded that despite irrelevance to the task, the semantically congruent audio-visual object produced capture, hence summoning attention to that location. Compared to Nardo et al., however, Mastroberardino et al.'s task did not impose a high perceptual load: the presentation the irrelevant animal images was sequential with the relevant visual targets[1], there was only two of them (always the same two, a cat and a dog), and they were presented at two pre-specified locations. One could even think that the images might have acted as placeholders, and therefore not being completely task irrelevant). In any case, according to the perceptual load theory (Lavie and Tsal, 1994), one would expect processing of irrelevant information under these low

[1] That is to say. At the moment the task-irrelevant sound-image combination was presented, there was no other competing task or stimuli.

perceptual load conditions. Inferring from the results of these very different studies, one might be tempted to conclude that under high perceptual load, cross-modal semantic congruence matters only if it bears some relevance to the task at hand. This is precisely the question that we address in the present study.

Here we aim at investigating how task constraints may modulate the effect of cross-modal semantic congruence on attracting attention. We hypothesized that the effect of audio-visual semantic congruence will emerge when at least one of two conditions apply: the multisensory object (or one of its components) carries some relevance to the current goal or, the multisensory object is irrelevant but presented under low perceptual load. We therefore predict that a semantically congruent sounds speed up the explicit search of a corresponding visual target, but when attention is engaged in another task, and therefore search is not explicit, then this cross-modal semantic effect will wane.

We addressed this question in three experiments using the same set of multisensory stimuli, with the only variations being task relevance and perceptual load. For all the experiments we used audio-visual pairs of common objects (e.g. a picture of a cat and a meowing sound, picture of a phone and a ring tone).

In Experiment 1, we studied the effect of audio-visual semantic congruence on spatial attention when the audio-visual pairs are task relevant. To do so, we aimed at replicating the results of Iordanescu et al. (2008, 2010), where the visual component of the multisensory object was task relevant by explicit instruction. In all cases,

participants performed a visual search task for pre-defined visual objects while hearing sounds that could be semantically consistent with the target, consistent with a distracter or not related to any object in the search array. According to these previous results, we expected to find significant effects of cross-modal semantic congruency in the form of shorter search latencies in target-consistent, compared to distractor-consistent trials or neutral trials. Previous findings by Iordanescu et al. (2008, 2010) and Molholm et al. (2004) also show that semantically consistent sounds do not increase the visual salience of a distractor visual object in the search array. In line with this, no difference in reaction time between distractor-consistent and neutral conditions is expected. This would support (and confirm) that the object-based audio-visual facilitation requires some top-down (goal-directed) processing.

In Experiments 2 and 3, we studied the effect of audio-visual semantic congruence on spatial attention when the audio-visual pairs were not task relevant. In these experiments, an array of visual objects and a sound were presented just like in Experiment 1 (and under the same conditions described). Yet, participants did not have to do any task with this array but were instructed to just wait until this array was replaced with a second visual array composed of "T" letters. Participants searched this second array for an upright "T" amongst inverted "T" s. The variable of interest was whether or not the target in the T search task appeared at a location previously occupied by a congruent audio-visual pair. The difference between Experiments 2 and 3 was perceptual load (low *vs*. high respectively).

According to our hypothesis, the predictions are as follows. If cross-modal congruence triggers automatic orienting, even in task-irrelevant conditions, then we expected to find a search advantage (shorter latency) if targets appeared at the location previous occupied by a congruent multisensory object, compared to when the target appeared away from this location. In the case that these interactions were to occur independently of available processing resources, indicating strong automaticity, then we would expect the effect to survive despite of task irrelevance and high perceptual load (Experiment 3). If orienting toward cross-modal semantic congruence breaks down in Experiments 2 and 3, we will conclude that task relevance is a condition for these interactions. If orienting toward cross-modal semantic congruence breaks down only in Experiment 3, then we will conclude that these interactions may happen even if task irrelevant, as long as perceptual load is low. Either of the two latter outcomes will cast doubts on a strong version of the automaticity hypothesis. Finally, by hypothesis we do not expect the effect of cross-modal semantic congruence to be significant in Experiment 3 but not in Experiment 2. Such a pattern of results should lead to a revision of the initial hypothesis.

Experiment 1: Replication of Iordanescu et al., 2008, 2010

In Experiment 1 we aimed at replicating the results of the study of Iordanescu et al. (2008, 2010) but with a new set of audio-visual stimuli. We created a visual search task where participants had to look for a target visual object while hearing characteristic sounds that were either consistent with the target of search, consistent with

a distractor, or neutral (consistent with neither). We conducted three different versions of the experiment (Experiments 1a, 1b and 1c) with variations in measurement: in Experiment 1a and 1b we measured saccadic search times (Iordanescu et al., 2010) and in Experiment 1c participants gave manual responses (Iordanescu et al., 2008). This was done mainly for the replication purposes and also to ascertain if both types of response are reliable in order to be used in other experiments.

Experiment 1a: Saccadic responses with aligned audio-visual stimuli

Methods

Participants

Sixteen volunteers (7 males; mean age 24.56 years, SD = 3.67) took part in the study. They had normal or corrected-to-normal vision, reported normal hearing and were naïve about the purpose of the experiment. All subjects gave written informed consent to participate in the experiment.

Stimuli

A set of 20 different images were obtained from free picture data-bases. Images represented tools, animals, transport, etc. (See supplementary materials). All pictures were edited with Adobe Photoshop CC 2015. Each picture was converted to greyscale and scaled to fit within 4.5 ° x 4.5 ° degrees area. All visual stimuli were presented on a gray background. Characteristic audio clips for each

of the visual objects were obtained from Freesound.org database (See online supplementary materials). The duration of sounds varied due to differences in their natural durations (M = 660 ms with SD = 130 ms). These differences should not have affected our results since the design of our experiments was counterbalanced (see Procedure). The sounds provided no information about the visual target's location, were clearly audible and presented via two loudspeakers, one on each side of the monitor, in order to render them perceptually central. On each trial, the sound was either consistent with the target object (*target consistent*), consistent with a distractor object (*distractor consistent*), or not consistent with any of the four objects included in the search display (*neutral*). All of the objects were randomly selected for each trial.

Procedure

The experiment was programmed and conducted using Psychopy 1.81 (for Python 2.7) running under Windows 7. An Eyetribe eye tracker (60 Hz sampling rate and 0.5° RMS spatial resolution) with a combined chin and forehead rest was used to control for eye movements.

Participants were sitting in front of a computer monitor 22.5'' (Sony GDM-FW900) at a distance of 77cm. In order to start each block of the experiment, participants pressed the space bar. Each trial started with the fixation cross that lasted for 2000 ms. Then a cue word was printed on the screen indicating the target of the visual search for that trial (Figure 1). After 2000 ms, a cue word disappeared, and a search display appeared. Every trial of the experiment contained a

display with 4 black and white pictures of visual objects that were placed in the four quadrants at 4.7° eccentricity. One of these four objects was a visual search target and the rest three were distractors. Visual display with objects appeared simultaneously with the sound that followed one of the three experimental conditions (*target-consistent, distractor-consistent* or *neutral*).

Participants were instructed to look as fast as possible at the visual target. Visual search performance for each subject and condition was determined by the mean Saccadic search time (SST). Once eye gaze entered the quadrant with visual target the trial automatically finished, and the new trial begins. SST was calculated from the beginning of the appearance of the visual search display until the moment the left eye gaze position reached the region of the target. Target object was presented in every trial. The experiment consisted of 4 blocks in total, each block contained: 20 target-consistent, 20 distractor-consistent and 20 neutral trials.

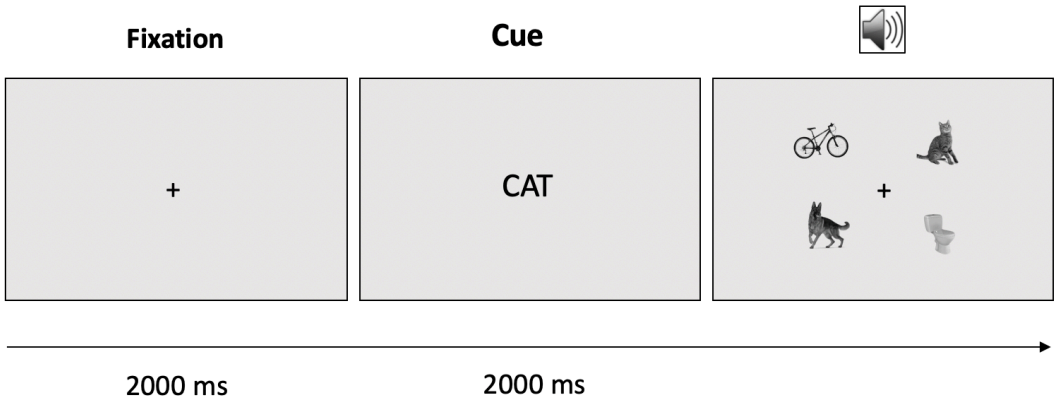

Figure 1. The sequence of event was identical for the Experiments 1a, b and c. First participants were asked to stay fixated in the central cross. The fixation cross was followed by the presentation of a cue word. Then search array of 4 objects appeared together with the sound that in the Experiment 1a was synchronous to the visual onset and in the Experiments 1 b, c preceded the pictures for 100 ms. In the Experiments 1 a, b participants had to look at the target object as fast as possible. Once the gaze was detected inside of the quadrant with target the trial was abrupted and the new trial began. In the Experiment 1c participant had to maintain central fixation throughout the whole trial and press as fast as possible one of the four keys that corresponded to the location of visual target. In contrary to the Experiment 1 a and b, in the Experiment 1c visual objects were presented only for 670 ms. Participant could respond during these 670 ms, otherwise the question mark appeared and stayed until participant presses the response key.

Results

We ran a repeated measures ANOVA on mean search times, with subject as the random effect and condition as the factor of interest. The analysis showed that main effect of condition was not significant ($F(2,30)=0.4$; $p=0.67$). Saccadic search time was not significantly faster in the target-consistent-sound condition (M = 501 ms) compared with both the distractor-consistent- condition (M = 500 ms), $t(15) =0.29$, $p = 0.388$ and neutral condition (M = 497 ms), $t(15) = 1.05$, $p = 0.154$. Neither the difference was found between the distractor-consistent and neutral conditions, $t(15) =0.56$, $p = 0.29$ (Figure 2). Thus, in the Experiment 1a characteristic sounds did not speed up gaze towards the visual target.

The results of this experiment fail to replicate the cross-modal semantic effect, a finding established by several previous studies. One of the reasons for the null result might lie in the nature of processing of complex sounds. Meaningful sounds take a certain (and variable) amount of time to identify given that information needs to be integrated over some hundreds of milliseconds (e.g. Cummings et al., 2006). On the other hand, less time is necessary to access the meaning of visual information in comparison to (Kim et al., 2014; Weatherford et al., 2015). In particular, semantic information can be accessed from visual stimuli within the first 100ms (for a review, see Potter, 2014), whereas processing of the meaning of a complex naturalistic sound can require more time due to the temporal nature of the information (according to some review, approximately 150 ms after onset, Murray and Spierer,

2009). For this reason, the temporal window of audio-visual integration for complex sounds is not the same as for simple artificial sounds (Vatakis and Spence, 2010). Here, because saccades are fast, it might be the case that there was no sufficient integration time for the meaning of the sound to influence visual processing before response. Following the same logic as previous laboratory studies that used complex sounds and visual events, we decided to advance the presentation of sounds by 100ms in Experiment 1b (Vatakis and Spence, 2010, for a review; Knoeferle K. M., Knoeferle P., Velasco and Spence, 2016, Kvasova et al, 2019, for a similar procedure). In Experiment 1c, we used the same procedure but changed the type of response: instead of saccadic search we used manual response.

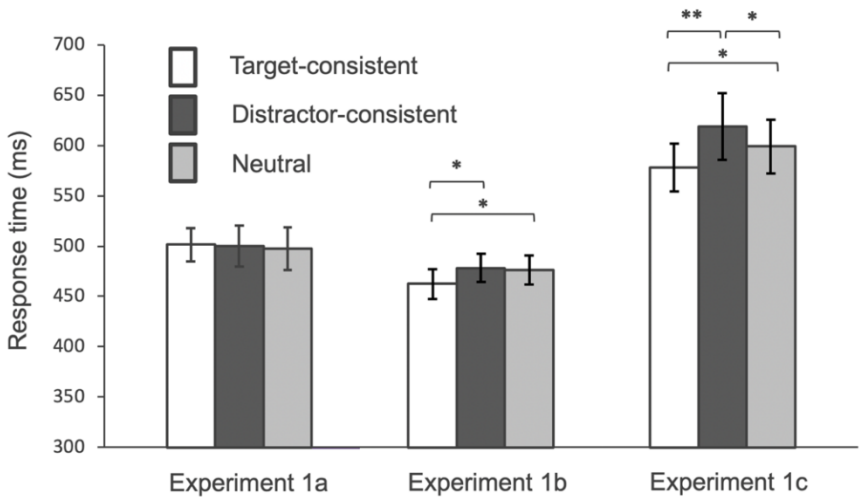

Figure 2. Visual search average reaction times towards a target and error rates were plotted in the *target-consistent* sounds, *distracter-consistent* sounds and *neutral* sounds conditions in Experiment 1a (Saccadic search times, SOA0ms), Experiment 1b (Saccadic search times, SOA100ms) and Experiment 1c (Manual response times, SOA100ms). Error bars indicate the standard error. Asterisks indicate significant difference between conditions (1 asterisk for p-value less than 0.05, 2 asterisks for p-value less than 0.01)

Experiment 1b: Saccadic responses with offset sounds

All apparatuses and stimuli in Experiment 1b (see *Experiment 1a Methods*) were identical to those used in Experiment 1a, except that the sound preceded the onset of visual search array for 100 ms (stimulus onset asynchrony SOA 100ms). For this purpose, we recruited an additional group of 16 participants (7 males; mean age 24.56 years, SD = 3.67).

Results

The analysis returned a significant main effect of condition ($F(2,30)=4.14$; $p=0.025$). Further, we tested the differences between conditions using one tail t-test with Holm-Bonferroni correction for multiple comparisons (Ludbrook, 1998). The analysis showed that saccadic search time was significantly faster in the target-consistent sound condition (M = 462 ms) compared with both the distractor-consistent (M = 478 ms), $t(15) =2.51$, $p = 0.012$, Cohen's $d=0.27$ and neutral condition (M = 476 ms), $t(15) = 2.55$, $p = 0.011$, Cohen's $d=0.23$, see Figure 2). All these results survived the correction for multiple comparisons.

Thus, in Experiment 1b we have shown that during visual search task characteristic sounds, when presented 100 ms in advance, speeded gaze responses towards semantically congruent visual targets. This successfully replicates previous results and establishes the cross-modal semantic effect on visual search. In addition, we compared search times in distractor-consistent and neutral conditions. The additional prediction stated that if audio-visual semantic consistency has an impact only in a goal-directed way then distractor-consistent sounds should not slow down performance compared to other not related sounds. Post-hoc t-tests showed the lack of difference in saccadic search time between the distractor-consistent and neutral conditions, $t(15) =-0.902$, $p=0.381$.

Experiment 1c

An additional group of 16 participants was recruited for this experiment (7 males; mean age 24.56 years, SD = 3.67). The experiment was programmed and conducted using the MATLAB 8.2-R2013b running under Windows 7. The procedure in the experiment 1c was adapted for the use manual response instead of gaze responses. The details of the procedure are as in Experiment 1b, except for the following differences: In experiment 1c participants had to maintain visual fixation throughout the whole trial, and were instructed to press one key out of four possible (1, 7, 9 and 3 keys of the number pad of the keyboard) corresponding to the location of the target (instead of gazing at the target). Visual search performance for each subject and condition was determined by the mean reaction time (RT) of manual responses. RT was calculated from the beginning of the visual search display until the moment the moment subject pressed the response key.

Results

We ran a repeated measures ANOVA on mean RTs (over correct responses), with subject as the random effect and condition as the factor of interest. The analysis showed a significant main effect of condition ($F(2,30)=7.7$; $p=0.002$). We found that reaction time was significantly faster in the target-consistent-sound condition (M = 578 ms) compared with both the distractor-consistent-condition (M = 619 ms), $t(15) =3.05$, $p = 0.004$, Cohen's $d=0.36$ and neutral condition (M = 599 ms), $t(15) =2.51$, $p = 0.012$, Cohen's $d=0.19$

(Figure 2). All comparisons survived the Holm-Bonferroni multiple comparison correction. Furthermore, reaction time was slower in distractor-consistent than in neutral conditions, $t(15) = 2.36$, $p = 0.016$, Cohen's $d = 0.17$. This last result suggests that distractor objects may also attract attention when congruent with the sound, even if these audio-visual events are not relevant to the current goal. This result was not expected, by comparison to previous results in the literature (and with the results of saccadic responses in Experiment 1b). We will come back to this in the General Discussion section. All in all, the results of experiments 1b and 1c allowed us to show that semantically consistent sound attract attention to the visual object when audio-visual event is relevant to the task, i.e. visual search. This replicates the finding from Iordanescu. (2008, 2010) and also Knoeferle et al. (2016) and Kvasova et al. (2019). Because the effect of crossmodal semantic congruence was found only when sound was presented 100ms before the visual stimuli we decided to use SOA100ms in all the following experiments. Also, the effect size was larger when using manual versus eye responses. Therefore, we used manual responses for the subsequent experiments.

Experiment 2

In the previous experiments we found that when audio-visual pair is relevant to the current goal of the task, semantic congruence has an effect on search. As the next step of our study addressed whether

audio-visual semantic congruence attracts attention even when the audio-visual pair is not relevant to the task. Here participants saw the same arrays as in experiments 1A-C, containing images and characteristic sounds of common objects (same conditions), only this time these arrays were completely task-irrelevant. Instead, subjects were asked to wait until the array transitioned into a new display composed of a set of letters T, and then perform a search task for an upright T amongst rotated Ts. In some of the trials the target T appeared at the same spot where the visual object congruent with the sound had been presented before. If audio-visual semantic congruence attracts attention to its location, then we would expect benefits in visual search if the target of the new away falls at that location. Because audio-visual events are task irrelevant, this would mean that crossmodal sematic congruency is able to attract attention in an automatic manner.

Methods

Participants

Fifteen volunteers (6 males; mean age 26.32 years, SD = 4.15) took part in the study. They had normal or corrected-to-normal vision, reported normal hearing and were naïve about the purpose of the experiment. All subjects gave written informed consent to participate in the experiment.

Stimuli and Procedure

For this experiment we used the same set of 20 pictures of common objects and their corresponding sounds as in Experiment 1. We presented sounds 100 ms before visual onset similarly to Experiments 1 b & c. In the beginning of the trial participants were presented with the cross in the middle of the screen for 500 ms and were instructed to maintain visual fixation on it (Figure 3A). Then 4 pictures of common object together with a sound were presented for 670 ms. The sound was either consistent with the object that was located in the quadrant where the following target of search will appear (*cued*); consistent with the object located in one of the other 3 quadrants (*uncued*), or not consistent with any of the four objects (*neutral*). After that, pictures faded out gradually and the search array started to appear on top of it. This transition lasted for 150 ms until pictures of objects completely disappeared and the search array was clearly visible. The transition was used in order to avoid abrupt changes that might induce the reorientation of attention (Remington et al., 1992). The search array contained 16 'T' letters. Participants searched for an upright "T" within inverted "Ts" and were instructed to press a response key as fast as they could when they found the target or withhold response if there was no target (filler trials).

The experiment consisted of 4 blocks of 200 trials each. In total 800 trials: 160 trials with no target (filler trials), 120 cued, 160 neutral and 360 uncued trials. By the low presence of cued trials (15% of the total) we disincentivized the strategy of anticipating targets

where the previous audio-visual congruent event was located, which could artificially generate the result we expected.

This procedure was adapted from the study of Mastroberardino et al. (2015). However, we did several important modifications. In the study of Mastroberardino et al. (2015) authors primed location of the upcoming relevant visual events with 2 object images (always a cat and a dog) that repeated throughout the experiment and were always at the same two locations. Because of this, and the strictly sequential presentation of the images and the visual targets, the perceptual load and the competition for processing resources between stimuli was relatively low. Also, since the initially irrelevant visual objects consistently market the two positions where visual targets could appear could have become relevant to the task. In Experiments 2 and 3 of the current study, displays contained 4 pictures of common objects selected from a set of 20 different objects randomly chosen in every trial. In Experiment 2 we used sequential presentation of events (just like in Mastroberardino et al., 2015), whereas in Experiment 3 (described below) the image array and the search array appeared concurrently. Finally, in the present experiments the target task required visual search with 16 T letters that were equally distributed within the area where previous 4 pictures appeared. This helped us to avoid the role of pictures acting as placeholders and therefore minimize possible relevance to the task which was important to our research question.

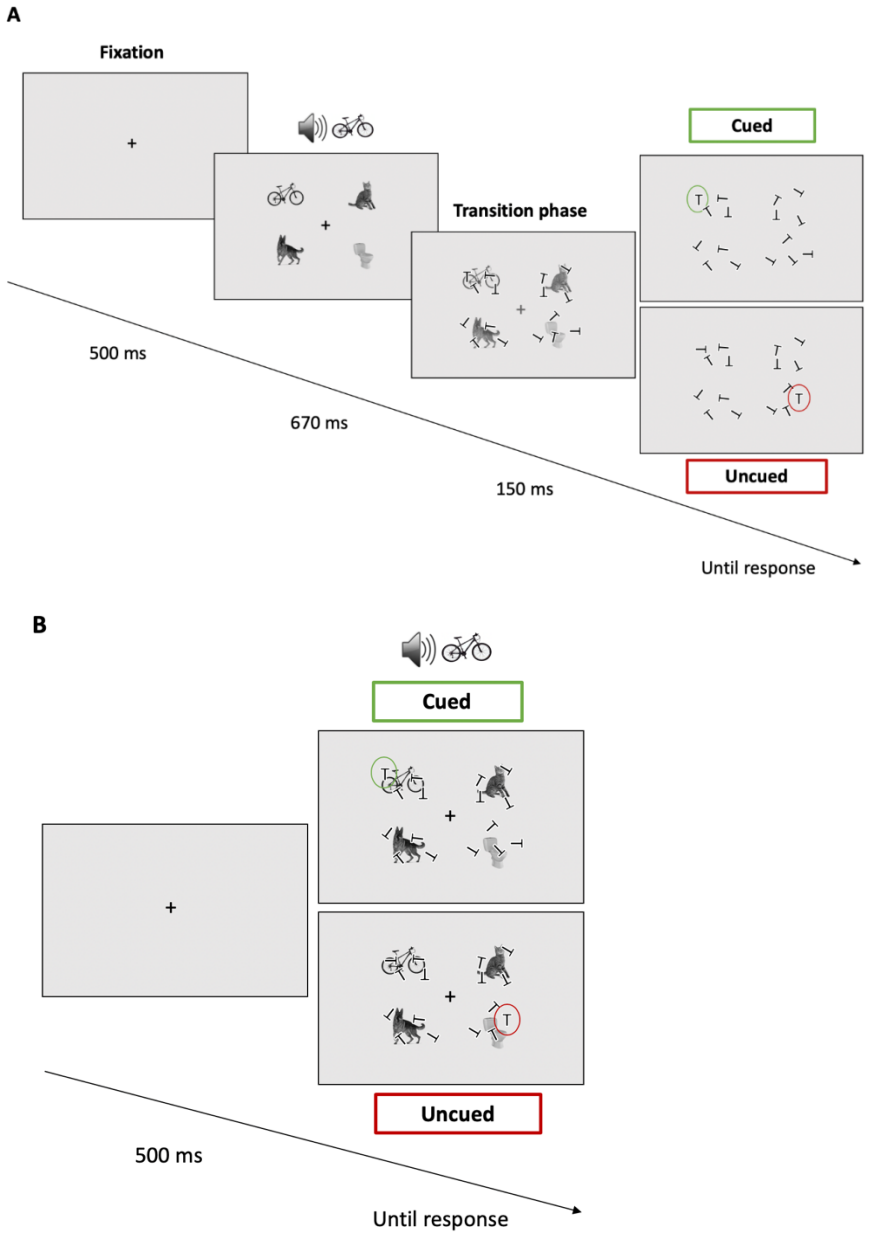

Figure 3. A) Sequence of events in the Experiment 2. At the beginning of the trial fixation cross appeared for 500 ms. 4 pictures of objects were then presented with the centrally presented sound for 670 ms. Sound advanced the presentation of the pictures for 100 ms. Then the trial continued into the transition phase for

150 ms during which the pictures gradually disappeared and search screen with inverted 'T' letters appeared. Search screen stayed until participant responds. B) Experiment 3 included two types of trials: with low (A) and high perceptual load (B). In the high perceptual load trials participants viewed pictures with sounds together with search array of inverted 'T' letters. Search screen stayed until participant responds.

Results

We anticipated that, if crossmodal semantic congruence attracts attention despite being irrelevant to the task then search time in the cued condition should be faster than in uncued or neutral. ANOVA returned a significant main effect of condition, $F(2,28)=5.56$ $p=0.009$. The analysis showed that average RTs in the cued condition (M = 893 ms) were significantly faster than in the uncued (M = 955 ms), $t(14)=3.96$, $p=0.0003$, Cohen's $d=0.34$ or neutral condition (M=937 ms), $t(14) =2.02$, $p=0.031$, Cohen's $d=0.24$ (Figure 4A). All these comparisons are one tail (given the directional hypothesis) and survived the multiple comparison correction using Holm-Bonferroni.

Our results demonstrate that semantically consistent sounds guide attention to its corresponding visual object despite the fact that the audio-visual events are irrelevant to the current task. Although we introduced several important changes in the design to increase uncertainty and irrelevance of audio-visual events to the task, we still have observed the effect of crossmodal semantic congruence on orienting, similarly to the study of Mastroberardino et al. (2015). However, despite we assumed task-irrelevant design, the

presentation of events in the Experiment 2 was still sequential, since sounds and visual objects were always presented before the actual task. Therefore, the perceptual load in this study was relatively low, liberating resources required for the perceptual processing of pictures and sounds that despite being irrelevant to the current goal appear to the participant in the moment when no other events took place. If cross-modal semantic interactions are strongly automatic, then we would expect the effect to survive not only task irrelevance, but also high perceptual load. This was tested in Experiment 3.

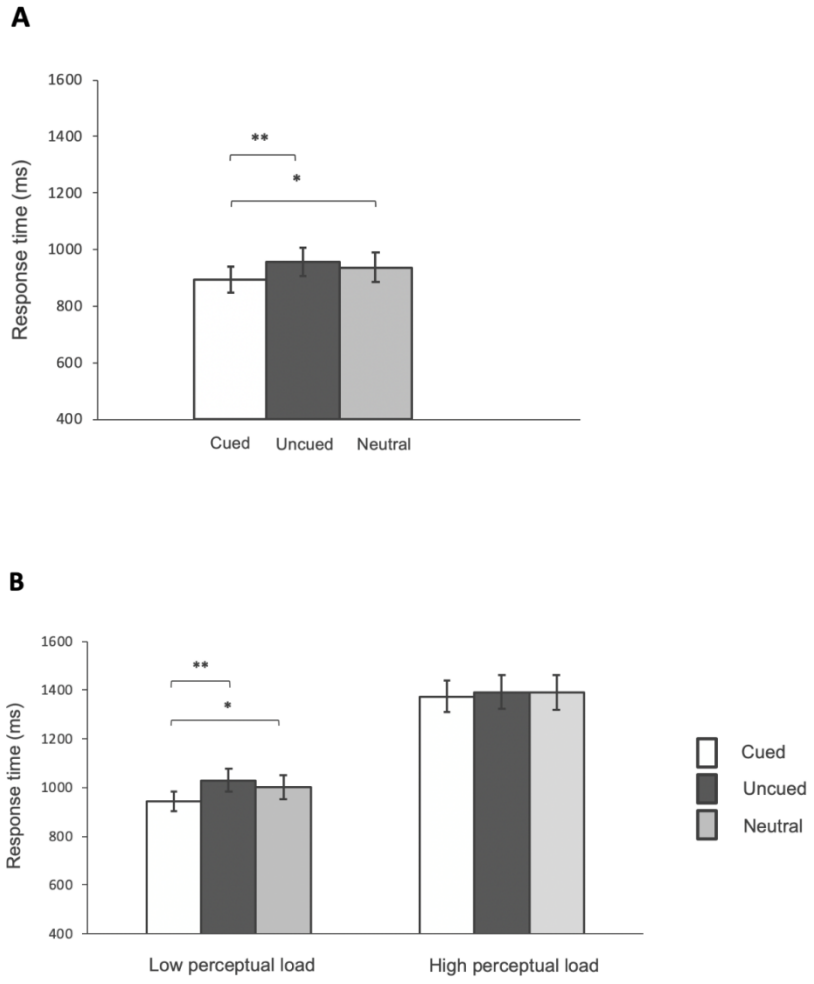

Figure 4. A) Experiment 2: Visual search reaction times towards a target and error rates were plotted in the *cued* (white), *uncued* (black) and *neutral* (grey) conditions. **B) Experiment 3:** Visual search reaction times towards a target and error rates were plotted in the *cued* (white), *uncued* (black) and *neutral* (grey) conditions and separated in two plots. Left plot represents performance in the same three conditions as in the Experiment 2 in the low perceptual load and right plot in the high perceptual load trials. In both experiments error bars indicate the standard error and asterisks indicate significant difference between conditions (1 asterisk for p-value less than 0.05, 2 asterisks for p-value less than 0.01)

Experiment 3

In Experiment 3 we preserved the irrelevance of audio-visual events to the task, but we introduced the additional differentiation between high and low perceptual load. The task was the same as in the Experiment 2 (see *Experiment 2. Methods*) and the presentation of sounds followed the same conditions. However, the perceptual load in this experiment was high since all objects and sounds were presented together with the array of "T" letters upon objects (Figure 3B). To be able to directly compare the effect of perceptual load we included both types of trials *high* vs *low* (intermixed within blocks). An additional group of twenty participants was recruited for Experiment 3 (7 males; mean age 24.45 years, SD = 3.05). The experiment was divided in 8 blocks of 200 trials: 2 types of load (low and high) x (20 no target) + (15 cued) + (20 neutral) + (45 uncued). Hence, this experiment contained a total of 1600 trials (160 no target, 120 cued, 160 neutral, and 360 uncued, per each load condition).

Results

We run a repeated measurements ANOVA separately for high and low perceptual load conditions. The analysis of the low perceptual load data returned a significant main effect of sound condition ($F(2,38)=7.56$; $p=0.002$). Further comparisons showed that in the trials with low perceptual load average RTs in the cued condition (M = 948 ms) were significantly faster than in the uncued (M = 1028 ms), $t(14)=3.57$, $p=0.001$, Cohen's $d=0.40$ or the neutral

condition (M=1005 ms), t(19)=2.41, p=0.013, Cohen's d=0.30 (Figure 4B). This replicates the effects found in Experiment 2. All significant effects survived Holm-Bonferroni correction for multiple comparisons. Instead, in the high perceptual load no effects of sound were found F(2,38)=0.36, p=0.7. Reaction time was not significantly faster in cued (M = 1375 ms) and uncued trials (M = 1393 ms), t(19)=0.72, p=0.24. Neither the difference between cued and neutral (M = 1391 ms) conditions resulted significant t(19)=0.58, p=0.28. These results demonstrate that audiovisual congruent events can summon attention even when task irrelevant, but only in low perceptual load condition. Attention capture by cross-modal semantic congruence does not survive high perceptual load.

Discussion

We addressed whether, and under which conditions, semantic congruence between sounds and visual objects attracts visual spatial attention. We manipulated task relevance of the audio-visual object and perceptual load. The findings to emerge from the experiments presented in this study show that audio-visual semantic congruence can help improve performance when searching for objects that are relevant to the current task goal. When task-irrelevant, the extent to which audio-visual congruence may attract visual attention is limited by perceptual load.

In Experiment 1 (Experiments 1B and 1C) characteristic sounds speeded up search times for the semantically corresponding visual

target in a visual search task. This result is in agreement with the idea that cross-modal semantic congruence can attract spatial attention and confirms prior results (Iordanescu et al, 2008, 2010; Knoeferle et al., 2016; Kvasova et al., 2019). In Experiment 1B, distractor consistent sounds did not slow down responses compared to neural sounds, suggesting that audio-visual congruence benefits goal-directed processes, but not the processing of other potential objects. However, in Experiment 1C we found that distractor-consistent sound slowed down search latencies in comparison to neutral sounds as well. This result is against the hypothesis and rather suggests that despite the irrelevance to the current goal semantically congruent audio-visual distractor attracted attention.

In Experiment 2, we measured search times in a visual array unrelated to the audio-visual objects, presented right after. The results showed that, when perceptual load is low, search times benefit if a visual target appears at a location previously occupied by an audio-visually congruent but task-irrelevant object. This finding suggests that cross-modal semantic congruence can attract spatial attention even if not bearing any particular relevance to the person's task. Therefore, the previous notion that semantic audio-visual enhancements occur only in a goal-directed task-relevant manner is not fully supported (Molholm et al., 2004; von Kriegstein et al., 2005; Iordanescu et al., 2008, 2010). The results of Experiment 2 are in line with the study of Mastroberardino et al. (2015) showing that, despite being irrelevant to the task, crossmodal semantic congruence can still attract attention.

In Experiment 3, we used the same task as in Experiment 2, but used two different perceptual load conditions. In the low load condition, an exact replication of Experiment 2, we found the same results: Task irrelevant audio-visual congruence attracted attention. In the high perceptual load condition, we found that that effect of task-irrelevant semantic congruence just vanished.

What consequences do these results have to interpret prior findings? We believe that the effect of perceptual load might help explain why some studies find task-irrelevant effects of semantic congruence, and why some others do not. For example, this could help explain the unstable effect of distractors in our Experiments 1B and 1C. Contrary to ours, Mastroberardino found crossmodal semantic effects on spatial orienting only for difficult visual targets, in one of the two experiments they reported. The authors suggested that probably, contrary to Iordanescu et al. (2008, 2010) both valid and invalid audio-visual events acted as distractors to the current task. However, we believe that difficulties in finding a stable effect could be rather explained by abrupt change between presentation of audio-visual events and task display. This sudden switch between displays might induce the reorientation of attention (Remington et al., 1992) that further vanished all the cueing effects of semantically congruent audio-visual pair. We believe that the transition phase between displays used in the current study helped to maintain the attention on the location where previous congruent event appeared.

In sum, the results of Experiment 2 and 3 have shown cross-modal semantic congruence can attracts spatial attention even if not

bearing any particular relevance to the current task. This pattern of results would suggest that audio-visual congruent objects do have a tendency to attract attention in an automatic manner. However, we also found that perceptual load might act as a limiting condition to this automatic tendency for congruence effects. This speaks against a strong automaticity account of cross-modal semantic interactions.

In order to conclude that audio-visual semantic interactions are fully automatic it would be necessary to demonstrate that the effect appears in task-irrelevant conditions and survives when attention is compromised by high perceptual load. The results of Experiment 3 suggest otherwise. Under high perceptual load when the number of items for processing is high and therefore the amount of resources is exceeded, the effect of audio-visual semantic consistence disappears when task irrelevant. This means that audio-visual semantic congruence necessitates from some top-down regulation in order to guide attention, above and beyond any fast, bottom up cross-modal integration process. This cross-modal interaction can be triggered even in the absence of a particular goal, as long as sufficient processing resources are left available. However, if so, the crossmodal sematic effect should be observed in the high perceptual load condition in Experiment 3. Instead, we found that when attention is fully engaged in different task semantically congruent audio-visual event does not attract attention. Therefore, we believe that semantic-based audio-visual integration requires some attention.

Even if one cannot conclude on automatic cross-modal effects, one interesting question is still open about what is the mechanism of audio-visual semantic interactions. One might think that this pattern of results could be based on the well-known effect of semantic priming, without the need to invoke a different process of fast semantic integration across modalities. Cross-modal facilitation by semantic priming could be explained by the fact that semantic associations across modalities established via prior experience tend to be reinforced. When information in one sensory modality recalls semantic representations, it creates expectation in other modalities which enhances recognition of the upcoming information that is congruent (e.g., Parise and Spence, 2009). Previous studies have demonstrated priming effects across modalities, however, the effect was observed using asynchronous presentation of auditory and visual events and null effect for synchronous or nearly synchronous presentation (Chen and Spence, 2011; 2013). Previous studies generally suggest that cross-modal semantic priming appears when cues (e.g., sounds) are presented prior to targets (e.g., the visual stimulus) (Dehaene et al., 1998; Costello et al., 2009). In our current study, however, consistent sounds were presented only 100 ms before the visual onset, which lead us to suggest that this effect might be caused by a different mechanism than traditional crossmodal semantic priming. Such mechanism would have to be based on interactions between quickly accessed auditory and visually identity information. Possibly, crossmodal semantic effects happen as well due to automatic audio-visual processing based on semantic information. This notion is supported by the findings in

the recent study of Cox et al. (2015). Authors showed that synchronously presented semantically congruent sound boosted visual below threshold image into the awareness during continuous flash suppression (CFS). No effect was found when sounds preceded the image. The authors suggested that these cross-modal semantic effects are due to automatic audio-visual processes, rather than traditional semantic priming.

Despite the interaction mechanisms alluded to in the discussion above provide potential accounts for our effects, the actual impact of traditional priming mechanisms is still difficult to assess. For example, in the previous studies where presentation of visual stimuli was very brief (e.g. 27 ms in the study of Chen & Spence, 2011), synchrony manipulations may have been effective to attribute the effect of priming, which are supposed to unfold in time. In our case, the duration of stimulus presentation was relatively long (approximately 660 ms for sounds and 670 ms for pictures). Given that the temporal overlap between auditory and visual stimuli was large, the effects of semantic priming may still occur over the time-course of synchronized events. However, the methods of the current study do not allow us to conclude on the mechanisms of audio-visual semantic interactions. More studies should be conducted in order to address this question.

Given the results observed in this (and prior) experiments, one important question are the implications for real-life scenarios. Object-based enhancements occur consistently for task-relevant objects and might occur even when task irrelevant but only under

favourable, low load, conditions. However, like in previous demonstrations of the same principle, experiments have typically used rather artificial settings: stimuli are presented under relatively low perceptual load, and without any meaningful context and ecological validity (Iordanescu et al., 2008; 2010; Mastroberardino et al., 2015). This is unlike real world conditions, where functional relationships and statistical regularities between objects (forks are often seen next to dishes), or between an object and its context (cars are rarely part of a submarine scene), are of great importance. Previous visual-only studies have already made a point about the differences in how attention is distributed in naturalistic, real life scenes compared to simple artificial search displays typically used in psychophysical studies (e.g., Peelen and Kastner, 2014, for a review; Henderson and Hayes, 2017).

Recently, the importance of studying multisensory interactions in realistic environments has been highlighted (e.g. Soto-Faraco et al., 2019; Matusz et al., 2019). One particularly relevant point refers to the interaction between these multisensory processes and attention, given that in realistic contexts, perceptual load tends to be high, compared to the idealised conditions of the current (and previous experiments). According to our results, high perceptual load leads to a decrement in the effectivity of crossmodal congruent events to attract attention (see also, Lunn et al. 2019). This could mean that the incidence of these effects in real life contexts could be limited.

As discussed above, another important aspect of real-world scenes, compared to the artificial displays used here, is contextual

information. Whereas in artificial search displays the different elements and their location do not provide any particular constrain on each other, naturalistic scenes are precisely defined by learned relationships that have an impact on object identification (see Peelen and Kastner, 2011). Under this light, in real-world settings that contain information from multiple sensory modalities, semantic relationships might be especially important for orienting. Further studies to understand the limits of crossmodal semantic effects and how they apply to real-life dynamic scenarios are necessary should to clarify this point. In a recent study, we have demonstrated that semantically consistent sounds can speed up search latencies for an object in dynamic and naturalistic visual scenes (Kvasova et al., 2019). This finding proves that audio-visual congruency facilitation effects for task-relevant objects demonstrated with simple and artificial AV stimuli (Experiments 1B and 1C; Iordanescu et al., 2008, 2010) could be generalized to the real-world contexts. However, it is perhaps fair to say that in real-life conditions, most of the sensory information available (including audio-visual congruent objects if present) occur outside the focus of attention and are potentially task irrelevant. Therefore, experiments with realistic scenes should address the effects of cross-modal semantic congruence in task-irrelevant or no-task conditions. These validations in more ecologically valid materials will help understand the relevance of semantic congruence in real life.

Conclusions

In the current study, we examined the constrains under which audio-visual semantic congruence triggers spatial orienting. We found that audio-visual semantic congruence speeded visual search times when the cross-modal objects are task relevant, a phenomenon that had been already described in other studies. Here, we show that even when these audio-visually congruent objects are task irrelevant they can summon attention, but only when presented under low perceptual load conditions. When these audio-visual events are irrelevant to the task and perceptual load is high, then the attention-grabbing effects of audio-visually congruent events vanish. This pattern of results does not support a strict automaticity hypothesis of semantic integration across modalities. Instead, we believe that some top-down processing is necessary for audio-visual semantic congruence to trigger spatial orienting. Further, in order to understand the relevance of semantic congruence in real life more experiments with realistic scenes should address the cross-modal semantic effects in task-irrelevant or no-task conditions.

Acknowledgements

This research was supported by the Ministerio de Economia y Competitividad (PSI2016-75558-P AEI/FEDER), AGAUR Generalitat de Catalunya (2017 SGR 1545). Daria Kvasova was supported by an FI scholarship, from the AGAUR Generalitat de Catalunya. This manuscript has been released as a pre-print at bioRxiv (Kvasova and Soto-Faraco, 2019).

References

Chen, Y.-C. and Spence, C. (2011). Cross-modal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *J. Exp. Psychol. Human* 37, 1554–1568.

Chen, Y.-C. and Spence, C. (2013). The Time-Course of the Cross-Modal Semantic Modulation of Visual Picture Processing by Naturalistic Sounds and Spoken Words. *Multisensory Research* 26 (2013) 371–386

Costello, P., Jiang, Y., Baartman, B., McGlennen, K., and He, S. (2009). Semantic and subword priming during binocular suppression. *Conscious. Cogn*. 18, 375–382.

Cox D. and Hong SW. (2015). Semantic-based crossmodal processing during visual suppression. *Front Psychol.* 6(June):722.

Cummings, A., Čeponienė, R., Koyama, A., Saygin, A. P., Townsend, J., and Dick, F. (2006). Auditory semantic networks for words and natural sounds. *Brain Research*, 1115, 92-107.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Lambertz, G., van de Moortele, P.F. and Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature,* 395, 597–600.

Grill-Spector, K., Henson, R. and Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn Sci*. 2006;10(1):14–23.

Henson, R.N. (2003). Neuroimaging studies of priming. *Prog Neurobiol.* 70:53--81.

Hasson, U., Malach, R., and Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, *14*(1), 40–48.

Henderson, J.M.; Hayes, T.R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nat. Hum. Behav.* 2017, 1, 743–747.

Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, *14*(1), 40–48.

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., and Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review*, *15*(3), 548–54.

Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., and Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception, & Psychophysics*, *72*(7), 1736–1741.

Kim, Y., Porter, A. M., and Goolkasian, P. (2014). Conceptual priming with pictures and environmental sounds. *Acta Psychologica*, 146, 73-83.

Knoeferle, K. M., Knoeferle P., Velasco C., and Spence C. (2016). Multisensory brand search: how the meaning of sounds guides consumers' visual attention. *Journal of Experimental Psychology: Applied,* 22(2):196-210.

Koelewijn, T., Bronkhorst, A., and Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: a review of audiovisual studies. *Acta Psychol.* 134, 372–384.

Kvasova, D., Garcia-Vernet, L., & Soto-Faraco, S. (2019). Characteristic sounds facilitate object search in real-life scenes. *Frontiers in Psychology.* 10:2511.

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., and Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Exp. Brain Res*. 158, 405–414.

Lavie N., and Tsal Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. Percept. *Psychophys*. 56, 183–197.

List, A., Iordanescu, L., Grabowecky, M., & Suzuki, S. (2014). Haptic guidance of overt visual attention. *Attention, Perception, & Psychophysics*, *76*(8), 2221–2228.

Ludbrook, John. Multiple comparison procedures updated. *Clinical and Experimental Pharmacology and Physiology* 25.12 (1998):1032-1037.

Lunn, J., Sjoblom, A., Soto-Faraco, S., and Forster, S. (2019) "Multisensory enhancement of attention depends on whether you are already paying attention." *Cognition* 187: 38-49.

Macaluso, E. (2010). Orienting of spatial attention and the interplay between the senses. Cortex 46:282–297.

Mastroberardino, S., Santangelo, V., & Macaluso, E. (2015). Crossmodal semantic congruence can affect visuo-spatial processing and activity of the fronto-parietal attention networks. *Frontiers in Integrative Neuroscience*, *9* (July), 45.

Matusz, P.J., Dikker S., Huth A.G & Perrodin C. (2019). Are we ready for real-world neuroscience? *Journal of Cognitive Neuroscience*, 31(3), 327-338.

McDonald J.J., Teder-Salejarvi, W.A., and Hillyard, S.A. (2000). Involuntary orienting to sound improves visual perception. *Nature* 407:906–908.

McDonald, J. J., Teder-Sälejärvi, W. A., and Ward, L. M. (2001). Multisensory integration and crossmodal attention effects in the human brain. Science 292, 1791–1791.

Molholm, S., Ritter, W., Javitt, D. C., and Foxe, J. J. (2004). Multisensory Visual-Auditory Object Recognition in Humans: A High-density Electrical Mapping Study. *Cerebral Cortex*, *14*(4), 452–465.

Murray, M. M., Camen, C., Andino, S. L. G., Bovet, P., and Clarke, S. (2006). Rapid brain discrimination of sounds of objects. Journal of Neuroscience, 26, 1293-1302.

Murray, M. M., & Spierer, L. (2009). Auditory spatio-temporal brain dynamics and their consequences for multisensory interactions in humans. Hearing Research, 258, 121-133.

Nardo, D., Santangelo, V., and Macaluso, E. (2014). Spatial orienting in complex audiovisual environments. *Human Brain Mapping*, *35*(4), 1597–614.

Parise, C.V., and Spence, C. (2009) 'When birds of a feather flock together': synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS One* 4:e5664.

Peelen, Marius V., and Sabine Kastner. "A neural basis for real-world visual search in human occipitotemporal cortex." *Proceedings of the National Academy of Sciences* 108.29 (2011): 12125-12130.

Peelen, M., and Kastner, S. (2014) Attention in the real world: toward understanding its neural basis. Trends in Cognitive Sciences 18(5).

Pesquita, A., Brennan, A. A., Enns, J. T., and Soto-Faraco, S. (2013). Isolating shape from semantics in haptic-visual priming. *Experimental Brain Research*, *227*(3), 311–322.

Potter, M. C. (2014). Detecting and remembering briefly presented pictures. In K. Kveraga & M. Bar (Eds.), Scene vision: Making sense of what we see (pp. 177-197). *Cambridge, MA: MIT Press*.

Remington, R. W., Johnston, J. C., & Yantis, S. (1992). Involuntary attentional capture by abrupt onsets. *Perception & Psychophysics*, *51*(3), 279-290.

Santangelo, V., and Macaluso, E. (2012). "Spatial attention and audiovisual processing," in The New Handbook of Multisensory Processes, ed. B. E. Stein *Cambridge,MA: TheMIT Press*, 359–370.

Schneider, T. R., Engel, A. K., and Debener, S. (2008). Multisensory identification of natural objects in a two-way crossmodal priming paradigm. *Exp. Psychol.* 55, 121–132.

Soto-Faraco, S., Kvasova, D., Biau, E., Ikumi, N., Ruzzoli, M., Moris-Fernandez, L., and Torralba, M. (2019). Multisensory interactions in the real world. *Cambridge Elements of Perception,* ed. M. Chun, (Cambridge: Cambridge University Press).

Spence C, Nicholls ME, Gillespie N, and Driver J (1998): Cross-modal links in exogenous covert spatial orienting between touch, audition and vision. *Percept Psychophys* 60:544–557.

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14, 400–410.

Van den Brink, R. L., Cohen, M.X., van der Burg, E., Talsma, D., Vissers, M.E., and Slagter, H. A. (2014). Subcortical, modality-specific pathways contribute to multisensory processing in humans. *Cereb. Cortex* 24, 2169–2177.

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., and Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1053–1065.

Vatakis, A., and Spence, C. (2010), "Audiovisual Temporal Integration for Complex Speech, Object-Action, Animal Call, and Musical Stimuli," in *Multisensory Object Perception in the Primate Brain,* ed. M. J. Naumer and J. Kaiser: Springer, 95-121.

von Kriegstein, K., Kleinschmidt A, Sterzer, P., and Giraud, A.L. Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience,* 2005;17:367–376.

Weatherford, K., Mills, M., Porter, A. M., and Goolkasian, P. (2015). Target categorization with primes that vary in both congruency and sense modality. *Frontiers in Psychology*, 6:20.

2.2 Characteristic sounds facilitate object search in real-life scenes

Kvasova, D., Garcia-Vernet, L., & Soto-Faraco, S. (2019)

Characteristic sounds facilitate object search in real-life scenes

Frontiers in Psychology

https://doi.org/10.3389/fpsyg.2019.02511

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 05 November 2019
doi: 10.3389/fpsyg.2019.02511

Characteristic Sounds Facilitate Object Search in Real-Life Scenes

Daria Kvasova[1], Laia Garcia-Vernet[1] and Salvador Soto-Faraco[1,2]*

[1] Center for Brain and Cognition, Universitat Pompeu Fabra, Barcelona, Spain, [2] ICREA – Catalan Institution for Research and Advanced Studies, Barcelona, Spain

Real-world events do not only provide temporally and spatially correlated information across the senses, but also semantic correspondences about object identity. Prior research has shown that object sounds can enhance detection, identification, and search performance of semantically consistent visual targets. However, these effects are always demonstrated in simple and stereotyped displays that lack ecological validity. In order to address identity-based cross-modal relationships in real-world scenarios, we designed a visual search task using complex, dynamic scenes. Participants searched for objects in video clips recorded from real-life scenes. Auditory cues, embedded in the background sounds, could be target-consistent, distracter-consistent, neutral, or just absent. We found that, in these naturalistic scenes, characteristic sounds improve visual search for task-relevant objects but fail to increase the salience of irrelevant distracters. Our findings generalize previous results on object-based cross-modal interactions with simple stimuli and shed light upon how audio–visual semantically congruent relationships play out in real-life contexts.

Keywords: visual search, attention, semantics, natural scenes, multisensory, real life

OPEN ACCESS

Edited by:
Kielan Yarrow,
City University of London,
United Kingdom

Reviewed by:
Emiliano Macaluso,
Université Claude Bernard Lyon 1,
France
Tiziana Pedale,
Umeå University, Sweden

****Correspondence:***
Daria Kvasova
daria.kvasova@upf.edu;
daria.kvasova@gmail.com

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Psychology

Received: *05 June 2019*
Accepted: *23 October 2019*
Published: *05 November 2019*

Citation:
Kvasova D, Garcia-Vernet L and Soto-Faraco S (2019) Characteristic Sounds Facilitate Object Search in Real-Life Scenes. Front. Psychol. 10:2511. doi: 10.3389/fpsyg.2019.02511

INTRODUCTION

Interactions between sensory modalities are at the core of human perception and behavior. For instance, the distribution of attention in space is guided by information from different sensory modalities as shown by cross-modal and multisensory cueing studies (e.g., Spence and Driver, 2004). Most research on cross-modal interactions in attention orienting has typically employed the manipulation of spatial (Spence and Driver, 1994; Driver and Spence, 1998; McDonald et al., 2000) and temporal (Busse et al., 2005; Van der Burg et al., 2008; van den Brink et al., 2014; Maddox et al., 2015) congruence between stimuli across modalities. However, recent studies have highlighted that in real-world scenarios, multisensory inputs do not only convey temporal and spatial congruence but also bear semantic relationships. The findings of these studies have shown that cross-modal correspondences at the semantic level can affect detection and recognition performance in a variety of tasks, including the distribution of spatial attention (e.g., Molholm et al., 2004; Iordanescu et al., 2008, 2010; Chen and Spence, 2011; Pesquita et al., 2013; List et al., 2014). For instance, in visual search among images of everyday life objects, sounds that are semantically consistent (albeit spatially uninformative) with the target speed up search times, in comparison to inconsistent or neutral sounds (Iordanescu et al., 2008, 2010). However, one paramount question which remains to be answered in this field is, to which extent such multisensory interactions discovered under simplified, laboratory conditions, have an impact under the complexity of realistic, multisensory scenarios (Matusz et al., 2019; Soto-Faraco et al., 2019). We set out to address this question.

Previous findings on cross-modal semantic effects on search behavior so far have used static, stereotyped artificial scenarios that lack meaningful context (Iordanescu et al., 2008, 2010; List et al., 2014). However, searching targets in these simplified displays used in laboratory tasks is very different from the act of looking for an object in complex, naturalistic scenes. As many authors have pointed out before, the generalization of laboratory findings using idealized materials and tasks is often far from trivial (Matusz et al., 2019, for a recent review). Outcomes that are solid and replicable under these simplified conditions may turn out differently in contexts that are more representative of real life (Wolfe et al., 2005; Maguire, 2012; Peelen and Kastner, 2014, for examples in visual research; see Soto-Faraco et al., 2019, for a review concerning multisensory research). First, realistic scenes are usually far more cluttered than stereotyped search arrays. Second, natural scenarios provide organization based on relevant prior experience: When searching for your cat in the living room, you would not expect the cat hovering midway to the ceiling, next to a floating grand piano. Yet, many laboratory tasks require just that: A picture of a (target) cat can be presented within a set of randomly chosen objects that have no relations between them, arranged in a circle, against a solid white background (**Figure 1**).

Previous visual-only studies have already made a point about the differences in how spatial attention is distributed in naturalistic, real-life scenes compared to simple artificial search displays typically used in psychophysical studies (e.g., Peelen and Kastner, 2014, for a review; Henderson and Hayes, 2017). Given that experience and repetition tends to facilitate visual search (Shiffrin and Schneider, 1977; Evans et al., 2013; Kuai et al., 2013), another important difference could lie in our familiarity (and hence, predictability) with natural scenes, compared to laboratory displays. In addition, humans can extract abundant information from natural scenes (gist) at a glance, quickly building up expectations about the spatial layout and relationships between objects (Biederman et al., 1982; Greene and Oliva, 2009; Peelen et al., 2009; MacEvoy and Epstein, 2011).

For example, Nardo et al. (2014) reported that cross-modal semantic congruency between visual events and sounds had no effect on spatial orienting or brain activity during free viewing of videos from everyday life scenes. In contrast, another study by Mastroberardino et al. (2015) with static images reported that visual images could capture spatial attention when a semantically congruent, albeit spatially uninformative sound was presented concurrently. Along with a similar line, Iordanescu et al. (2008, 2010) showed that spatially uninformative characteristic sounds speeded up the visual search when consistent with the visual target. Conversely to the study of Nardo et al. (2014), which found no effect, Iordanescu et al. (2008, 2010) and Mastroberardino et al. (2015) used simple static images presented in decontextualized search arrays (Iordanescu et al., 2008, 2010). Both, these differential features (dynamic nature of natural scenes and their complexity) have been pointed out as important components for the generalization of cognitive psychology and neuroimaging findings to real-world contexts (e.g., Hasson et al., 2010). Another possible important variable in prior research on cross-modal semantic influence on attention is task-relevance. Unlike Nardo et al. (2014) and Mastroberardino et al. (2015) studies, in the study of Iordanescu et al. (2008, 2010) the critical (target) objects were task-relevant, potentially making audio–visual congruence relations also relevant to the task.

FIGURE 1 | Left picture is an example of stimuli used as a typical search array in a search experiment. Figures are randomly chosen and randomly distributed in space without any meaningful connection between them. On the right naturalistic picture some objects are the same as on the left but now they are put into a context with spatial envelope, proportionality, and variety of meaningful and functional connections between objects.

Based on the results of these prior studies, one first outstanding question is whether cross-modal semantic relationships can play a role at all in complex dynamic scenarios. Until now, the only study using such scenarios (Nardo et al., 2014) has returned negative results, in contrast with other studies using more stereotypical displays (Iordanescu et al., 2008, 2010; Mastroberardino et al., 2015). Given that a major difference between these studies was task relevance of the cross-modal events, a second interrelated question is whether the impact of cross-modal semantic relationships, if any, is limited to behaviorally relevant events. Here we present a study using a novel search task on realistic scenes, in order to shed light on these two questions.

In our visual search protocol, targets were everyday life objects appearing in video clips of naturalistic scenes. Spatially uninformative characteristic sounds of objects mixed with ambient noise were presented during search. The relationship between the object sounds and the visual target defined four different conditions: *target-consistent sound*, *distracter-consistent sound*, *neutral sound*, and *no sound*, which was a baseline condition that contained only background ambient noises. Visual search performance was measured with reaction times.

We hypothesized that, if cross-modal semantic congruency guides attention in complex, dynamic scenes, then reaction times should be faster in the target-consistent condition than in the distracter-consistent, neutral, or no sound conditions (e.g., target-consistent characteristic sounds will help attract attention to the corresponding visual object). Regarding the possible task-relevance modulation of cross-modal semantic effects, we hypothesized that if audio–visual semantic congruence attracts attention in natural scenes automatically even when the objects are irrelevant to the current behavioral goal, then one should expect a slowdown in responses to targets in distracter-consistent trials, with respect to neutral sound trials. Else, if audio–visual semantic congruence has an impact only when task-relevant (as we expected), then distractor-congruent sounds should not slow down performance compared to other unrelated sounds.

In order to check the potential unspecific effects of object sounds on visual search times, such as alerting (Nickerson, 1973), we included neutral sound condition as a control. Neutral sounds were sounds that did not correspond to any object in the video of the current trial. Thus, we expected that differences due to general alerting of sounds, if any, would equally affect target-consistent, distractor-consistent, and neutral sound conditions, but not the no-sound baseline.

MATERIALS AND METHODS

Participants

Thirty-eight volunteers (12 males; mean age 25.22 years, *SD* = 3.97) took part in the study. They had normal or corrected-to-normal vision, reported normal hearing, and were naïve about the purpose of the experiment. All subjects gave written informed consent to participate in the experiment. Two subject-wise exclusion criteria were applied before any data analysis. (1) If the false alarm rate in catch trials (trials in which the search target was not present) was above 15%. (2) If accuracy in one or more conditions was <70%. After applying these criteria, we retained data from 32 participants.

Stimuli

Visual Stimuli

A set of 168 different video-clips were obtained from movies, TV shows, and advertisement, and others were recorded by experimenters from everyday life scenes. The video clips, size 1024 × 768 pixels, and 30 fps were edited with Camtasia 9 software[1] to 2 s duration fragments. No fades were used during the presentation. Ninety-six videos were used for the experimental conditions described below, and 72 videos for catch trials. For all of the videos, the original soundtrack was replaced with background noise created by the superposition of various everyday life sounds (see example video clips and sounds in the **Supplementary Materials**).

Each video clip used for experimental (target-present) conditions contained two possible visual targets, which were always visual objects which have a characteristic sound (such as musical instruments, animals, tools, etc.). The criteria to choose the target objects in the videos was that, although they were visible (no occlusions, good contrast), they were not part of the main action in the scene. For instance, if a person is playing guitar and this is the main action of the scene, the guitar could not be a target object. However, in a scenario with a band playing different instruments, the guitar could be a possible target. Both target and distractor objects are presented from the beginning till the end of the video except for the catch trials where neither target or distractor are presented. We applied this criterion to make the search non-trivial. Nevertheless, in order to compensate for potential biases related to particular objects or videos, we counterbalanced the materials so that each video and object contributed as a target and as a distractor in equal proportions across participants (see the section "Procedure").

[1] https://www.techsmith.com/camtasia/

Auditory Stimuli

We used characteristic sounds that corresponded semantically to the target/distractor objects (e.g., barking dog). However, they gave no information about the location of the object (sounds were always central) or its temporal profile (the sound temporal profile did not correlate with visual object motion or appearance). All the sounds were normalized to 89 dB SPL and had a duration of 600 ms. Sounds were delivered through two loudspeakers placed at each side of the monitor, in order to render them perceptually central.

Procedure

The experiment was programmed and conducted using the Psychopy package 1.84.2 (Python 2.7) running under Windows 7. Participants were sitting in front of a computer monitor 22.5″ (Sony GDM-FW900) at a distance of 77 cm. We calibrated the video and sound onset latencies using The Black Box Toolkit[2] (United Kingdom), within an error of *SD* = 7.34 ms.

In order to start each block of the experiment, participants pressed the space bar. Each trial started with a cue word printed on the screen indicating the target of the visual search for that trial. After 2000 ms, a video clip with the background noise plus, if applicable, a characteristic object sound of the corresponding condition (target-consistent, distracter-consistent, neutral) were presented. Following previous laboratory studies that used complex sounds and visual events we decided to desynchronize presentation of the audio–visual event, by presenting the sound 100 ms before the video onset (Vatakis and Spence, 2010, for review; Knoeferle et al., 2016, for a similar procedure).

The participant's task was to judge whether or not the pre-specified target object was present in the video clip as fast as possible and regardless of its location. If the video ended before participants' response, a question mark showed up on the screen and stayed there until the participant responded. The next trial started 200 ms after the participant had responded (**Figure 2**). Half of the participants had to press A key (QWERTY keyboard) as soon as they found the target object. In case the object was not present on the scene, they pressed L key. For the other half it was the other way around. Visual search performance for each subject and condition was determined by the mean response time (RT) of correct responses.

Four types of sound–target conditions were used: *target-consistent*, *distractor-consistent*, *neutral*, and *no sound*. In the target-consistent condition, the identity of the sound matched with the target object. In the distractor-consistent condition, the sound matched a non-target (distracter) object present in the scene. In the neutral condition, the object sound did not match any of the objects in the scene. Finally, in the baseline condition, no particular object sound (an auditory cue) was present (besides the background noise) (**Figure 3**).

Due to the high heterogeneity of the video-clips, we decided to counterbalance them across conditions and participants. Each participant saw each video-clip once, but overall, each video clip appeared in each of the four experimental conditions the same number of times (across subjects), except for trials which were

[2] www.blackboxtoolkit.com

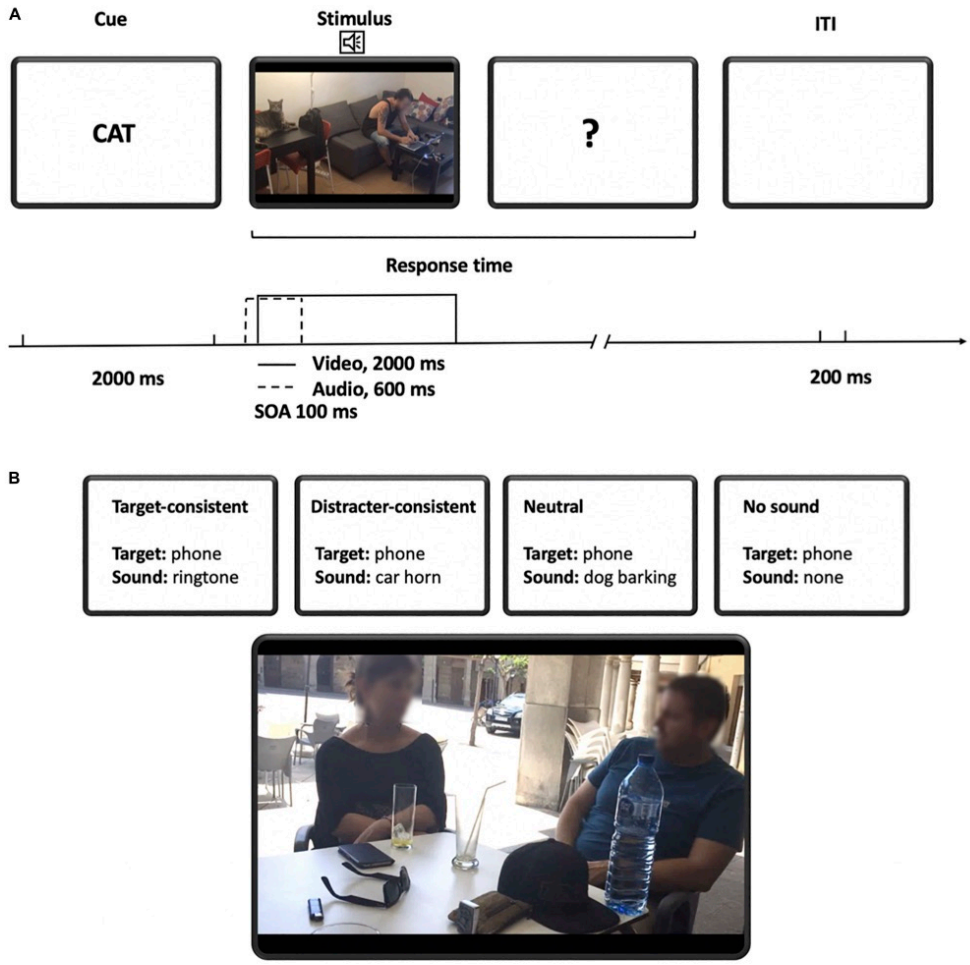

FIGURE 2 | (A) Sequence of events in the experiment. The trial started with the presentation of target word for 2000 ms. The target word was followed by the auditory cue and video. Auditory cue was presented 100 ms before the video was shown (SOA = 100 ms) and lasted for 600 ms while the video lasted for 2000 ms. There was no time limit for the participant response. 200 ms after the participant had responded a new target word was presented. **(B)** Example of conditions. In this example of stimulus, the possible targets are a mobile phone and a car. If the target is a mobile phone, in the target-consistent condition the sound will match the target, in the distractor-consistent condition the sound will match the distractor (a car), in the neutral condition the sound will not match any object of the scene (e.g., dog barking), and in the no sound condition there will be just background noise and no auditory cue. The image is a frame of the video clip filmed by the research group.

the same for all participants. To achieve this, we created a total of eight different versions of the experiment (in order to equate the number of times each of the two objects in each video was the target). In order to make sure that participants understood the task, they ran a 14-trial training block before the beginning of the experiment. The training set used video clips that were equivalent to, but not contained in, the experiment and included examples of the four experimental conditions as well as catch trials.

The experiment contained a total of 168 trials (24 trials per experimental condition plus 72 catch trials; hence, the overall proportion of target-present trials was ∼57%). The experiment was divided into six blocks of 28 videos with a representative

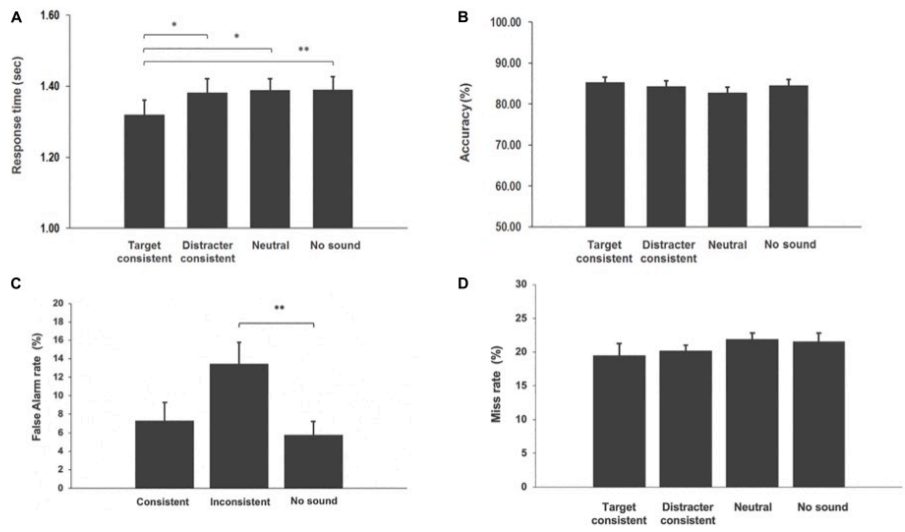

FIGURE 3 | (A) Visual search reaction times toward a target and error rates were plotted in the target consistent sounds, distracter-consistent sounds, neutral sounds, and no sound conditions. Error bars indicate the standard error. Asterisks indicate significant difference between conditions (*p-value < 0.05, **p-value < 0.01). **(B)** Visual search accuracy toward a target and error rates were plotted in the target-consistent sounds, distracter-consistent sounds, neutral sounds, and no sound conditions. Error bars indicate the standard error. **(C)** False alarm rates were plotted in the conditions when sound was consistent with the cue, inconsistent with the cue, and in the no sound condition. **(D)** Miss rates were plotted in the target-consistent sounds, distracter-consistent sounds, neutral sounds, and no sound conditions. Error bars indicate the standard error.

number of trials of each condition and catch. Each participant received a different random order of videos.

RESULTS AND DISCUSSION

We ran a repeated measures ANOVA on mean RTs (for correct responses), with subject as the random effect and condition as the factor of interest. The analysis returned a significant main effect of condition [$F(3,93) = 3.14$; $p = 0.0289$]. Given this significant main effect, we went on to test our specific *a priori* predictions using t-tests. In particular we had hypothesized that target-consistent characteristic sounds will help attract attention to the corresponding visual object. Based on this hypothesis, we predicted that reaction times should be faster in the target-consistent condition than in the distractor-consistent, neutral, and no sound conditions. The analysis demonstrated that responses in the target-consistent condition were faster than in distracter-consistent [$t(31) = 2.36$, $p = 0.012$, Cohen's $d = 0.27$], neutral [$t(31) = 2.33$, $p = 0.013$, Cohen's $d = 0.39$], and no sound [$t(31) = 2.53$, $p = 0.008$, Cohen's $d = 0.32$] conditions. All these comparisons are one tail (given the directional hypothesis) and survived the multiple comparison correction using Holm–Bonferroni (Ludbrook, 1998).

The second prediction stated that if audio–visual semantic congruence attracts attention in natural scenes automatically even when the objects are irrelevant to the current behavioral goal, then one should expect a slowdown in responses to targets in distracter-consistent trials, with respect to neutral sound and no sound condition. *Post hoc* t-test showed the lack of difference between distractor-consistent and neutral conditions $t(31) = 0.28$, $p = 0.39$. For completion, we also performed non-planned t-tests (two-tails) between distractor-consistent and no sound $t(31) = 0.28$, $p = 0.39$, and between neutral and no sound conditions $t(31) = 0.33$, $p = 0.37$. Neither of these comparisons resulted significant. The latter comparison suggests that no cross-modal effect was observed in this experiment due to unspecific general alerting influence of sounds.

To ensure that there was no speed–accuracy trade-off we analyzed error data. The analysis showed that there was no difference in performance between conditions (**Figure 3B**). Since catch trials do not contain target and distractor objects, the false alarm rate was calculated between three conditions: consistent (when sound corresponds to the search cue word), inconsistent (when the sound does not correspond to the search cue word), and no sound (**Figure 3C**). The analysis showed no difference in consistent vs. inconsistent trials [$t(31) = 1.37$, $p = 0.09$] and consistent vs. no sound [$t(31) = 0.44$, $p = 0.33$]. However, in inconsistent trials participants had higher false alarm rate in comparison to the no sound condition [$t(31) = 2.74$, $p = 0.005$]. Analysis of miss rates showed no difference between conditions (**Figure 3D**). The increase in false alarms for catch trials in the

inconsistent condition is surprising, because it would mean that participants tend to respond more when the cue word and the characteristic sound are different, rather than the same. Recall that in these trials, there are no visual objects that correspond to either. If this result was to reflect an actual response bias toward being more liberal in inconsistent trials (hence, make more false detections and/or responding faster), this bias would be against the main result detected in the experimental trials.

Over all, the results to emerge from the present study show that, when searching for objects in real-life scenes, target-consistent sounds speed up search latencies in comparison to neutral sounds or when only background noises are present. Instead, distracter-consistent sounds produced no measurable advantage or disadvantage with respect to these baseline conditions (albeit, responses were slower than for target-consistent conditions). This finding demonstrates, for the first time, that characteristic sounds improve visual search not only in simple artificial displays (Iordanescu et al., 2008, 2010) but also in complex dynamic visual scenes with contextual information. In general, and according to previous studies (Iordanescu et al., 2008, 2010), we can affirm that the results obtained in this study are due to object-based and not due to spatiotemporal correspondences since we avoided any kind of spatiotemporal congruence. Semantic relationships between the objects in a complex visual scene can guide attention effectively (Wu et al., 2014, for review), our results suggest that this semantic information did not make congruent auditory information redundant. Semantically consistent sounds can indeed benefit visual search along with available visual semantic information. This is the novel contribution of this study.

Despite research on attention orienting has been dominated primarily by low-level spatial and temporal factors (salience), recent research has focused on the role of higher-level, semantic aspects (e.g., Henderson and Hayes, 2017). Visual-only studies have highlighted, for example, the importance of functional relationships between objects (Biederman et al., 1982; Oliva and Torralba, 2007; MacEvoy and Epstein, 2011), expectancies regarding frequent spatial relations (Peelen and Kastner, 2014, for review), and cues to interpersonal interactions (Kingstone et al., 2003; Papeo et al., 2017, 2019) as important in determining some aspects of visual scene perception. These factors are to play an especially important role in real-life naturalistic scenarios, where these high-level relationships are often abundant (Peelen and Kastner, 2014). Adding to this evidence from visual-only experiments, in the present study we demonstrated that high-level cross-modal (auditory–visual) semantic relations may as well exert an impact in spatial orienting and guide attention in visual search for objects in real-life, dynamic scenes. In fact, one could speculate that especially in complex and noisy environments where many visual and auditory events are spatially and temporally coincident, semantic information might become a leading predictor of object presence, and hence, guide attention.

The visual and auditory materials we used in our study are highly heterogeneous; therefore, it is very challenging to control for all the possible compounds such as movement, presence of people in videos, size, and position of objects, physical salience, and meaning of the scene. We addressed these differences between videos by counterbalancing them across subjects. However, this does not allow us to completely discard the possible influence of the stimulus properties on orienting behavior and therefore on the results of the study. Another possible issue might be the absence of distinction in our study between sounds that either physically or semantically are close to each other, e.g., sound of a guitar and sound of the piano (the same semantic group of musical instruments) or the sound of the coins or keys (physically similar). This way we cannot be sure that sound from the same semantic category or sound that is physically similar could play a proper role of a distractor or neutral sound.

In the current study, we used a detection task (pressing the button as soon as the target object is found). One may argue that this design does not allow us to assure that participants respond to the target and not for the distractor. Since the videos are very heterogeneous, it was not possible to design discrimination instead of a detection task while preserving control of the relevant variables. Catch trials were introduced in the experiment specifically to avoid (and control) excessively liberal response criteria (high proportion of "yes" guessing responses). However, we did not anticipate any particular hypothesis regarding false alarms in different conditions and because of this catch trials did not contain sound-congruent distractor objects. This way our design does not allow to calculate false alarm rate for the distractor-consistent trials. One possible concern which could be raised is that participants were responding to the sound rather than cue-word, which would still generate correct responses in the target-consistent and target-inconsistent trials. However, if this happened, we should observe a difference in reaction time data between all target-present trials (consistent and inconsistent) and the neutral sound condition. In particular, since in neutral trials the presented sound does not correspond to any object in the scene, it will probably take more time for participants to respond since they will be looking for something that is not there. This effect is not present in the data of the current study.

Another possible limitation of our design is that distractors that are consistent with the characteristic sound could have induced responses. These responses would compete with the actual correct detection in the target-inconsistent sound condition but could be counted as correct in the target-consistent conditions, hence generating the observed difference between these two conditions in our data. How can we address this possible limitation? If this effect of cue-sound competition had a sizable effect on response patterns, then reaction times and accuracy should decay in the target-inconsistent condition in comparison to the neutral condition (in which no visual objects coincided with the distractor sound and competed for response). However, no differences in reaction times or accuracy between distractor-consistent and neutral conditions were found (we elaborate on this point in the next paragraph). False-alarm data could be potentially informative in this case but unfortunately the design of this study does not allow us to calculate false alarm for distractor-consistent condition (see above). One prior study by Knoeferle et al. (2016) measured false alarm rates in a similar visual search task with simpler scenes and the same conditions for characteristic sounds. Knoeferle et al. (2016)

reported no differences in false alarms between conditions in five experiments with an exception of marginal tendency for distractor-congruent sound compared to the no-sound condition in two of the experiments. Therefore, based on that study it seems that incongruent sound does not strongly bias participant to confuse target with the distractor. However, we must be careful in extrapolating these assumptions to the current data.

Another open question is why target-consistent sounds benefit search, but distracter-consistent sounds do not slow down reaction times (in comparison to neutral or no sounds). If cross-modal interactions were strictly automatic and pre-attentive, then distractor sounds should increase the saliency of their corresponding, yet irrelevant objects present in the scene. However, the evidence we found is not consistent with the strong pre-attentive view of cross-modal semantic effects. Despite the interplay between attention and multisensory interactions is far from resolved (Talsma et al., 2010; Ten Oever et al., 2016; Hartcher-O'Brien et al., 2017; Lunn et al., 2019; Soto-Faraco et al., 2019, for some reviews), many studies illustrate that multisensory interactions tend to wane when the implicated inputs are not attended (e.g., Alsius et al., 2005, 2014; Talsma and Woldorff, 2005). For example, Molholm et al. (2004) demonstrated that object-based enhancement occurs in a goal-directed manner, suggesting that while a characteristic sound of a target will facilitate its localization, a characteristic sound of a distracter will not attract attention to the distracter. In line with Molholm et al. (2004) and other previous studies (Iordanescu et al., 2008, 2010; Knoeferle et al., 2016) in our study we demonstrated that in visual search task semantically consistent sound helps to find a visual target faster. This might be due to the fact that auditory encoding of a sound, e.g., a barking sound enhances visual processing of all the features that are related to a dog. This way all the auditory and visual semantic associations are likely to develop simply because of repeated coincidence when experiencing the multisensory object. At first, the cue word activates the semantic web of the target of search and creates an attentional template for the search. Further, the characteristic sound reinforces this activation and therefore the object is found faster. However, it remains unknown if the semantically congruent audio–visual event can attract attention in an automatic way when it is not relevant to the task or when there is no task at all (e.g., free observation).

Consistent with the idea of automaticity [and therefore, contrary Molholm et al. (2004) and to our results], a study by Mastroberardino et al. (2015) showed that audio–visual events can capture attention even when not task-relevant. Here, it is important to note that our design was not necessarily optimized to detect such distractor-consistent effect (e.g., as discussed above, it was not sensitive enough in terms of detecting distractor-induced false alarms). There are other important differences between the present study and Mastroberardino et al. (2015), which could account for the fact that task-irrelevant semantic audio–visual congruency could have had a larger impact. For example, Mastroberardino et al. (2015) used a low perceptual load situation with a very limited range of possible semantic relationships (just two). We believe that object-based cross-modal enhancements might eventually occur even when task-irrelevant, under favorable low load conditions. Further studies to understand the limits of cross-modal semantic effects and how they apply to real-life dynamic scenarios should be run to clarify this point. For example, in line with the present study, a possible next step would be to use eye-tracking with free-viewing of the video-clips to investigate if cross-modal semantic congruency attracts visual behavior and can be, therefore, responsible for the visual search effects seen here.

CONCLUSION

In conclusion, we have demonstrated that semantic consistent sounds can produce an enhancement in visual search in complex and dynamic scenes. We suggest that this enhancement happens through object-based interactions between visual and auditory modalities. This demonstration not only generalizes (and confirms) previous laboratory findings on semantically based cross-modal interactions but also expands it to the field of research in natural scenes.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Clinical Research Ethics Committee (CEIC) of Parc de Salut Mar UPF. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DK, LG-V, and SS-F contributed to the conception and design of the study. DK and LG-V prepared the stimuli and collected the data. DK and LG-V performed the statistical analysis. DK wrote the manuscript. All authors contributed to the manuscript revision, and read and approved the submitted version of the manuscript.

FUNDING

This research was supported by the Ministerio de Economia y Competitividad (PSI2016-75558-P AEI/FEDER) and the AGAUR Generalitat de Catalunya (2017 SGR 1545). DK was supported by an FI scholarship, from the AGAUR Generalitat de Catalunya. This manuscript has been released as a pre-print at bioRxiv (Kvasova et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02511/full#supplementary-material

REFERENCES

Alsius, A., Mottonen, R., Sams, M. E., Soto-Faraco, S., and Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: evidence from ERPs. *Front. Psychol.* 5:727. doi: 10.3389/fpsyg.2014.00727

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Curr. Biol.* 15, 839–843. doi: 10.1016/j.cub.2005.03.046

Biederman, I., Mezzanotte, R. J., and Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cogn. Psychol.* 14, 143–177. doi: 10.1016/0010-0285(82)90007-x

Busse, L., Roberts, K. C., Crist, R. E., Weissman, D. H., and Woldorff, M. G. (2005). The spread of attention across modalities and space in a multisensory object. *Proc. Natl. Acad. Sci. U.S.A.* 102, 18751–18756. doi: 10.1073/pnas.0507704102

Chen, Y., and Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *J. Exp. Psychol.* 37, 1554–1568. doi: 10.1037/a0024329

Driver, J., and Spence, C. (1998). Cross-modal links in spatial attention of cognitive. *Philos. Trans. R. Soc. B* 353, 1319–1331. doi: 10.1098/rstb.1998.0286

Evans, K. K., Georgian-Smith, D., Tambouret, R., Birdwell, R. L., and Wolfe, J. M. (2013). The gist of the abnormal: above-chance medical decision making in the blink of an eye. *Psychon. Bull. Rev.* 20, 1170–1175. doi: 10.3758/s13423-013-0459-3

Greene, M. R., and Oliva, A. (2009). Recognition of natural scenes from global properties: seeing the forest without representing the trees. *Cogn. Psychol.* 58, 137–176. doi: 10.1016/j.cogpsych.2008.06.001

Hartcher-O'Brien, J., Soto-Faraco, S., and Adam, R. (2017). a matter of bottom-up or top-down processes: the role of attention in multisensory integration. *Front. Integr. Neurosci.* 11:5. doi: 10.3389/fnint.2017.00005

Hasson, U., Malach, R., and Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends Cogn. Sci.* 14, 40–48. doi: 10.1016/j.tics.2009.10.011

Henderson, J. M., and Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nat. Hum. Behav.* 1, 743–747. doi: 10.1038/s41562-017-0208-0

Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., and Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Atten. Percep. Psychophys.* 72, 1736–1741. doi: 10.3758/APP.72.7.1736

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., and Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychon. Bull. Rev.* 15, 548–554. doi: 10.3758/PBR.15.3.548

Kingstone, A., Smilek, D., Ristic, J., Friesen, C. K., and Eastwood, J. D. (2003). Attention, researchers! it is time to take a look at the real world. *Curr. Direct. Psychol. Sci.* 12, 176–180. doi: 10.1111/1467-8721.01255

Knoeferle, K. M., Knoeferle, P., Velasco, C., and Spence, C. (2016). Multisensory brand search: how the meaning of sounds guides consumers' visual attention. *J. Exp. Psychol.* 22, 196–210. doi: 10.1037/xap0000084

Kuai, S. G., Levi, D., and Kourtzi, Z. (2013). Learning optimizes decision templates in the human visual cortex. *Curr. Biol.* 23, 1799–1804. doi: 10.1016/j.cub.2013.07.052

Kvasova, D., Garcia-Vernet, L., and Soto-Faraco, S. (2019). *Characteristic Sounds Facilitate Object Search in Real-Life Scenes. bioRxiv*. [Priprint]. doi: 10.1101/563080

List, A., Iordanescu, L., Grabowecky, M., and Suzuki, S. (2014). Haptic guidance of overt visual attention. *Atten. Percep. Psychophys.* 76, 2221–2228. doi: 10.3758/s13414-014-0696-1

Ludbrook, J. (1998). Multiple comparison procedures updated. *Clin. Exp. Pharmacol. Physiol.* 25, 1032–1037. doi: 10.1111/j.1440-1681.1998.tb02179.x

Lunn, J., Sjoblom, A., Ward, J., Soto-Faraco, S., and Forster, S. (2019). Multisensory enhancement of attention depends on whether you are already paying attention. *Cognition* 187, 38–49. doi: 10.1016/j.cognition.2019.02.008

MacEvoy, S. P., and Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nat. Neurosci.* 14, 1323–1329. doi: 10.1038/nn.2903

Maddox, R. K., Atilgan, H., Bizley, J. K., and Lee, A. K. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. *Elife* 4:e04995. doi: 10.7554/eLife.04995

Maguire, E. A. (2012). Studying the freely-behaving brain with fMRI. *Neuroimage* 62, 1170–1176. doi: 10.1016/j.neuroimage.2012.01.009

Mastroberardino, S., Santangelo, V., and Macaluso, E. (2015). Crossmodal semantic congruence can affect visuo-spatial processing and activity of the fronto-parietal attention networks. *Front. Integr. Neurosci.* 9:45. doi: 10.3389/fnint.2015.00045

Matusz, P. J., Dikker, S., Huth, A. G., and Perrodin, C. (2019). Are we ready for real-world neuroscience? *J. Cogn. Neurosci.* 31, 327–338. doi: 10.1162/10.1162/jocn_e_01276

McDonald, J. J., Teder-Salejarvi, W. A., and Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature* 407:906. doi: 10.1038/35038085

Molholm, S., Ritter, W., Javitt, D. C., and Foxe, J. J. (2004). Multisensory visual-auditory object recognition in humans: a high-density electrical mapping study. *Cereb. Cortex* 14, 452–465. doi: 10.1093/cercor/bhh007

Nardo, D., Santangelo, V., and Macaluso, E. (2014). Spatial orienting in complex audiovisual environments. *Hum. Brain Mapp.* 35, 1597–1614. doi: 10.1002/hbm.22276

Nickerson, R. S. (1973). Intersensory facilitation of reaction time: energy summation or preparation enhancement? *Psychol. Rev.* 80, 489–509. doi: 10.1037/h0035437

Oliva, A., and Torralba, A. (2007). The role of context in object recognition. *Trends Cogn. Sci.* 11, 520–527. doi: 10.1016/j.tics.2007.09.009

Papeo, L., Goupil, N., and Soto-Faraco, S. (2019). Visual search for people among people. *Psychol. Sci.* 30, 1483–1496. doi: 10.1177/0956797619867295

Papeo, L., Stein, T., and Soto-Faraco, S. (2017). The two-body inversion effect. *Psychol. Sci.* 28, 369–379. doi: 10.1177/0956797616685769

Peelen, M. V., Fei-Fei, L., and Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature* 460:94. doi: 10.1038/nature08103

Peelen, M. V., and Kastner, S. (2014). Attention in the real world: toward understanding its neural basis. *Trends Cogn. Sci.* 18, 242–250. doi: 10.1016/j.tics.2014.02.004

Pesquita, A., Brennan, A. A., Enns, J. T., and Soto-Faraco, S. (2013). Isolating shape from semantics in haptic-visual priming. *Exp. Brain Res.* 227, 311–322. doi: 10.1007/s00221-013-3489-1

Shiffrin, R. M., and Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychol. Rev.* 84:127. doi: 10.1037/0033-295X.84.2.12

Soto-Faraco, S., Kvasova, D., Biau, E., Ikumi, N., Ruzzoli, M., Moris-Fernandez, L. et al. (2019). "Multisensory interactions in the real world," in *Cambridge Elements of Perception*, ed. M. Chun (Cambridge: Cambridge University Press).

Spence, C., and Driver, J. (2004). *Crossmodal Space and Crossmodal Attention*. Oxford: Oxford University Press.

Spence, C. J., and Driver, J. (1994). Covert spatial orienting in audition: exogenous and endogenous mechanisms. *J. Exp. Psychol.* 20, 555–574. doi: 10.1037/0096-1523.20.3.555

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14, 400–410. doi: 10.1016/j.tics.2010.06.008

Talsma, D., and Woldorff, M. G. (2005). Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. *J. Cogn. Neurosci.* 17, 1098–1114. doi: 10.1162/0898929054475172

Ten Oever, S., Romei, V., van Atteveldt, N., Soto-Faraco, S., Murray, M. M., and Matusz, P. J. (2016). The COGs (context, object, and goals) in multisensory processing. *Exp. Brain Res.* 234, 1307–1323. doi: 10.1007/s00221-016-4590-z

van den Brink, R. L., Cohen, M. X., van der Burg, E., Talsma, D., Vissers, M. E., and Slagter, H. A. (2014). Subcortical, modality-specific pathways contribute to multisensory processing in humans. *Cereb. Cortex* 24, 2169–2177. doi: 10.1093/cercor/bht069

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., and Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1053–1065.

Vatakis, A., and Spence, C. (2010). "Audiovisual temporal integration for complex speech, object-action, animal call, and musical stimuli," in *Multisensory Object Perception in the Primate Brain*, eds M. J. Naumer, and J. Kaiser (Berlin: Springer), 95–121. doi: 10.1007/978-1-4419-5615-6_7

Wolfe, J. M., Horowitz, T. S., and Kenner, N. M. (2005). Cognitive psychology: rare items often missed in visual searches. *Nature* 435:439. doi: 10.1038//435439a

Wu, C., Wick, F. A., and Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Front. Psychol.* 5:1–13. doi: 10.3389/fpsyg.2014.00054

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, EM, declared a past collaboration, with one of the authors, SS-F, to the handling Editor.

2.3 The Impact of Audio-Visual Semantic Information on Spontaneous Orienting in Real-life Scenes

Daria Kvasova, Travis Stewart & Salvador Soto-Faraco

The Impact of Audio-Visual Semantic Information on Spontaneous Orienting in Real-life Scenes

In prep.

The impact of audio-visual semantic information on spontaneous orienting in real-life scenes

Daria Kvasova[1], Travis Stewart[1] & Salvador Soto-Faraco[1,2]

[1] Center for Brain and Cognition, Universitat Pompeu Fabra, Barcelona

[2] ICREA, Barcelona

Corresponding autor: Daria Kvasova

Pompeu Fabra University

Edifici Merce Rodoreda (Room 24.326)
Carrer de Ramon Trias Fargas, 25-27
08005 Barcelona
Spain

daria.kvasova@upf.edu

Abstract

Real-world events provide a rich web of semantic correspondences about object identity in different sensory modalities. These correspondences help us parse sensory information and make sense of the environment. For example, a sudden car honk just as we are about to cross the street brings the image of an approaching vehicle into sharp focus. Previously, we were able to demonstrate that characteristic sounds speed up visual search for everyday life objects in natural and dynamic environments. Furthermore, our results suggest that object-based enhancement occurs in a goal-directed manner. In the current study, we investigated if this crossmodal semantic congruency effect drives visual attention under free-viewing condition, without any specific task. We addressed this question in an eye-tracker study using a set of 108 video clips from realistic complex scenes (YouTube, TV, movies, news, etc.) presented alongside various sounds of varying semantic congruency with objects within the videos. We found that characteristic sound increased the percentage of observed corresponding visual objects, number of fixation and the total dwell time inside the area of interest in comparison to the neutral sounds or when video are presented with background noise only. The results suggest that crossmodal semantic congruence indeed have an effect on gaze and eye movements and therefore attention in a free viewing paradigm. Our findings extend previous findings on object-based crossmodal interactions with simple stimuli and shed more light upon how audio-visual semantically congruent relationships play out in realistic scenarios.

Introduction

In real world contexts, we are constantly bombarded with all sorts of sensory information. Our brains need to make sense of these signals and deploy resources efficiently by orienting attention on what is relevant. Previous studies have highlighted that sensory cues from different modalities can affect visuo-spatial orienting. For instance, cross-modal spatial cueing demonstrates that abrupt sounds summon not only auditory attention but also visual attention towards their location (Driver, 1996; Spence et al., 1998; McDonald et al., 2000; see Spence & Soto-Faraco, 2019 for an applied perspective). In addition, temporal congruence between multisensory events can also attract attention and enhance responses to subthreshold stimuli (REF). Previous studies have also shown that a sound synchronized with a visual transient can boost the salience of that visual event, improving search performance (Van der Burg et al., 2008; Van den Brink et al., 2014). Looking at the literature, it is clear that spatio-temporal information has a robust influence on perception and attentional orienting (Santangelo and Spence, 2007).

However, going back to real world contexts, in our everyday life environments we do not only focus attention based on spatial and/or temporal attributes. These environments are rich in behaviorally relevant information about the identity and semantic attributes of objects. Hence, they contain a rich web of semantic correspondences about object identity in different sensory modalities. These correspondences also help us parse sensory

information and deploy processing resources in an efficient manner. In line with this idea, prior studies have demonstrated that high-level semantic information (as opposed to low-level spatio-temporal cues) can also influence visuospatial attention (Iordanescu et al., 2008,2010; Mastroberardino et al., 2015; Kvasova and Soto-Faraco, 2019). However, most of these studies have used simplified situations which allow for optimal experimental control, but do not capture the complexity of relevant real-world situations (e.g., Soto-Faraco et al., 2019; Kvasova et al., 2019; Matusz et al., 2019, for a similar argument). The present work is precisely about the potential of cross-modal semantic cues to summon attention under conditions closer to real-life scenarios.

Previous studies on crossmodal semantic congruence have demonstrated cross-modal semantic effects on visual processing and attention. For example, Molholm et al. (2004) found that performance in an audio-visual object-recognition task was enhanced with respect to uni-modal trials, when the auditory and visual semantic cues were congruent. Similar enhanced object-recognition was also seen with haptic-visual cross-modal stimuli by Pesquita et al. (2013). In addition to cross-modal effects on identification, it has been demonstrated that crossmodal semantic congruence can also improve performance in visual detection task (Chen and Spence, 2011), picture naming (Mädebach et al., 2017), and especially relevant for the question at stake in this study, cross-modal semantic congruence influences the spatial distribution of visual attention (Iordanescu et al., 2008;2010; Mastroberardino et al., 2015; Kvasova et al., 2019).

These previous findings suggest that semantic information may play a role in spatial orienting and making sense of multisensory information. However, most of these studies are characterized by a reductionist approach which trades ecological validity for experimental control (Soto-Faraco et al., 2019; Blanton & Jaccard, 2006; Burgess et al., 2006; Kayser, Körding, & König, 2004; Kingstone et al., 2003; Neisser, 1976; 1982). For example, the studies mentioned so far have used simplified scenarios in which isolated visual objects are presented in the absence of any meaningful or structured context.

The meanings of objects and their role on the guidance of attention have recently come under the spotlight of visual attention research (Peelen et al., 2014; Henderson and Hayes, 2017). Previous studies have already addressed the question about the differences in human performance, comparing simple meaningless artificial search displays versus naturalistic scenes that are rich in meaningful context (Peelen et al., 2014). It has been shown that humans can extract complex information from just a brief glance at a scene, and can then predict which types of objects, which likely spatial arrangements, and what semantic connections between objects are likely to be found in this type of scenes (Wu et al., 2014). In addition, real world scenes are typically characterized by high perceptual load. Given these differences, it would be fair to assume that visual search and eye movement within a complex scene is significantly different from that of simple search paradigms using well controlled but idealized search arrays (like those used in most previous studies). Hence, generalization studies are warranted.

The studies discussed above addressing attention orienting in complex, realistic scenarios, have focused on visual-only tasks. Little is known about whether the cross-modal semantic effects on orienting discovered might generalize from the laboratory to complex real-life scenes. In one of the first studies to address this question, Nardo and colleagues (2014) used video-clips of real-life scenes whilst participants passively watched. They measured eye movements and BOLD responses using fMRI. In this study, semantic congruence between sounds and visual events did not produce any particular modulation in gaze distribution or brain activations (despite other aspects of the stimuli, such as spatial congruence between visual and acoustic cues did). The null result for semantic congruence in realistic scenes contrasts with the significant outcomes from the various, more controlled laboratory studies discussed at the beginning. One could think that this could constitute a case in which laboratory findings magnify an effect that is not very significant for real life.

A recent visual search study from our group (Kvasova et al., 2019), also using realistic video clips, found that semantically consistent (but spatially uninformative) sounds speeded up visual search times (in comparison to the distractor-consistent, neutral sounds, or no sound conditions). However, one important aspect of this study is that it used a goal-oriented search paradigm. Given that in Nardo et al.'s experiment, subjects did not have a specific task other than to view the video clips, one could think cross-modal congruence simply is inefficient when task irrelevant. However, work from Mastroberardino and colleagues (2015) demonstrated that audio-

visual semantically congruent events can indeed influence and attract attention despite being unrelated to the task. Another possible explanation for the reduced impact of semantic congruence whilst freely observing real scenes is that, in Nardo et al.'s study, the presence of strong cues to spatial and temporal congruence of audio-visual events overrode potential cross-modal effects based on object identity.

After these results, one is left to wonder whether semantic coincidences between modalities in everyday life scenes, even if not necessarily relevant to our current goals, do exert some attraction effect on our attention. Here, we have used eye-tracking to investigate the impact of cross-modal congruence as people watch real world scenes without a specific task. However, in contrast to Nardo et al., we focused on scenes where spatio-temporal cues are not as salient, and therefore semantic cues may become more salient. In particular, we measured whether semantically consistent sounds can produce faster, longer and/or more frequent spatial orienting towards the corresponding visual object in real-life dynamic scenes. By using eye-tracking, we could test the influence of sounds on attention while participants are freely observing visual scenes and orienting behavior unfolds spontaneously.

We presented subjects with a series of short video-clips of complex everyday life scenes (from movies, video-clips, or self-produced footage) mixed with background noise plus object characteristic sounds that were chosen according to the experimental conditions described below. For each video, an area of interest (AOI)

containing an object of interest (such as a car, a telephone, a guitar, etc.) was defined. The AOI was used to measure eye-gaze parameters such as fixation number, duration (dwell time) and time to first fixation. Each participant was presented with all videos but with the three different sound conditions counterbalanced: *consistent* (the characteristic object sound is consistent with the object in the AOI), *neutral* (the characteristic object sound corresponds to an object that is not presented in the video) and *no sound* (videos are presented only with background noise).

We hypothesized that semantically consistent sounds would increase the salience of the corresponding visual object if present, and therefore the probability of directing overt attention toward the visual object would increase. In order to test this hypothesis, we measured various eye-gaze parameters. First, we computed in how many of the videos the AOI is looked at in each condition. Based on our hypothesis we predicted that percentage of videos where AOI is observed by participant will be higher in consistent than in neutral or no sound conditions. Second, we hypothesized that consistent auditory semantic information will strengthen the effect of spatial orienting and increase exploration of the corresponding visual object. This way we predicted that the total time spent looking (dwell time) and number of fixations inside the AOI in the consistent condition will be higher in comparison to the neutral or no sound conditions. Further, we measured how fast the gaze goes towards the AOI. We predicted that in the consistent condition gaze would be directed in the AOI faster in comparison to control conditions.

Methods

Participants

45 subjects (12 males, mean age 23.3 (between 19-35) participated in the study. All gave consent to take part in this experiment. All participants had normal or corrected to normal vision, normal hearing and were naïve about the purpose of the study. Each participant performed a calibration run with the eye-tracker software before each experiment block to ensure that accuracy of gaze tracking was acceptable (within <0.5º of visual angle). Participants were excluded is calibration failed to reach the criterion after 3 attempts.

Stimuli

We created a set of 108 video clips for the experiment. To avoid cross-modal spatio-temporal correspondences we removed original audios from videos and replaced them with background noises composed of generic sounds typical of the corresponding visual scene. We decided to present the characteristic sounds of each condition embedded in background noise, in order to avoid alerting effect of the sound. Background noise was tailored of the videos. For example, if the video contained scenes from the concert, we added a noise of the crowd to it (see example video clips and sounds in the online supplementary materials). This was done for all videos to maintain the ecological validity.

Videoclips were taken from movies, television, YouTube or were recorded by the authors. All videos, size 1024x768 pixels and 30 fps were edited with iMovie software 10.1.10 to the 2 seconds duration clips. Each video contained a target object (e.g.: a guitar, a bicycle, a dog, etc.). All target objects were chosen based on the subjective criteria. Although objects of interest were visible (no occlusions, good contrast), they were not part of the main action in the scene and never centrally presented. Objects of interest were always presented from the beginning till the end of the video. Nevertheless, in order to compensate for potential biases of highly heterogeneous stimuli, we counterbalanced the materials so that each video contributed in all conditions in equal proportions across participants The area of interest (AOI) in each video was defined around the target object. Due to the high heterogeneity of object's size and location, all the areas of interest were defined subjectively and manually, by creating a rectangular area around the object. This method of AOI definition has proved useful when dealing with complex and heterogenous visual scenes (Hessel et al., 2016 for review of AOIs methods). We chose only videos where the target object had a fixed screen position throughout the duration. This allowed us to define AOI for only one frame of each video and keep it fixed because objects were not changing its location in space. Slight movements (camera, or object) were always inside the defined AOI.

Sounds for each visual object were obtained from Freesound.org database. We used characteristic sounds that corresponded semantically to the target objects (e.g. sound of barking

corresponded to a dog). However, those sounds provided no information about the location of the object (as they were always presented from the same central location) or its temporal profile (the sound onset was fixed for all videos and conditions). All the sounds were normalized to 89 dB SPL and their duration varied due to differences in their profile (M = 600 ms with SD = 145 ms). Sounds were delivered through two loudspeakers placed at each side of the monitor, in order to render them perceptually central. Background noise was normalized to 72 dB SPL.

The three sound conditions were used in this experiment: *consistent*, *neutral*, and *no sound*. For the control no sound condition, the video was presented with no characteristic sounds, but still contained background noise suitable to the content of the scene. For the consistent condition, a characteristic sound of the target object of that video was embedded in the background noise (e.g.: the sound of a barking dog when a dog is the target object). For the neutral condition, a sound characteristic of a different object than the target in the video was presented (e.g.: the sound of a piano when a car is the target object).

We used 23 different sounds, since objects were repeated across different videos (different dogs, different guitars, etc.). All the sounds were divided in 5 semantic groups: animals, vehicles. electronics, musical instruments and other (not related to any of the groups). We created an exclusion criteria in our design so that sounds within the same semantic category could not be presented as neutral (e.g.: a video with the target object of a guitar could not be

presented with a sound of a piano for the neutral condition, as guitar and piano fall within the same semantic category of "musical instruments"). Moreover, after the automatic exclusion procedure, we inspected manually that videos did not contain objects that could be related semantically to any of the sounds from the pool in the neutral condition. If such an object was found, we restricted all the sounds from the related semantic category to be used as neutral. For instance, the dog is our object of interest, but there is also a guitar presented in the scene. In this case none of the musical instrument's sounds were used as neutral for this video. This way we avoided the creation of unwanted distractor events.

Procedure

The experiment was programmed and conducted using the Psychopy package 1.84.2 (Python 2.7) running under Windows 7. Participants were sitting in front of a 22.5" computer monitor (Sony GDM-FW900) at a distance of 70cm. Two loudspeakers, delivering the sounds stereophonically, were placed at each side of the monitor. Eye movements were recorded using the EyeTribe eyetracker (60 Hz sampling rate and 0.5° RMS spatial resolution) and PyGaze 0.5.1 open-source software.

Participants were instructed to watch the videos as if they are watching television. This way we attempted to minimize any task related behavior. The only restriction was to watch within the screen area (e.g., avoid looking aside of the screen or closing the eyes). First, participants underwent an individualized calibration of

the eye-tracking equipment and sensors. In addition to this general calibration, embedded within the experiment there were separate calibration runs before each of the six blocks of the experiment, to ensure proper measurements would be taken.

Due to the high heterogeneity of visual stimuli it was necessary to counterbalance videos across conditions and participants. Each participant saw each video-clip once, but overall, each video clip appeared in each of the three experimental conditions the same number of times (across subjects). To achieve this, we created 3 versions of the experiment (15 participants in each group). This was done to equate the number of times each video appears in each condition. The experiment consisted of 108 trials, divided into six blocks of 18 videos. Six videos per condition were used in each block, 36 videos per condition for each participant. The order of blocks and the order of appearance of videos within blocks were randomized between participants.

Each trial began with a fixation cross presented in the center of a blank screen. Participants had to fixate on the cross in order for the next trial to start. After it was determined that the participants were indeed fixating on the cross, then the trial began. The sound lasted for approximately 600 ms (the natural duration of sound varied with standard deviation of 145 ms), was presented centrally and gave no information about the location of corresponding visual object. 100 ms after auditory onset the video with embedded background noise started and lasted for 2000 ms and was followed by the fixation cross again. In the *no sound* condition the presentation of fixation

cross was followed by the direct presentation of video with background sounds only.

We recorded eye movements from the onset of the video. After each trial, there was a brief rest period of 1s before the next fixation cross would appear, followed by the next video (***Figure 1A***). Participants were encouraged to rest between blocks, but they could take additional self-paced rest periods by not fixating on the cross between trials.

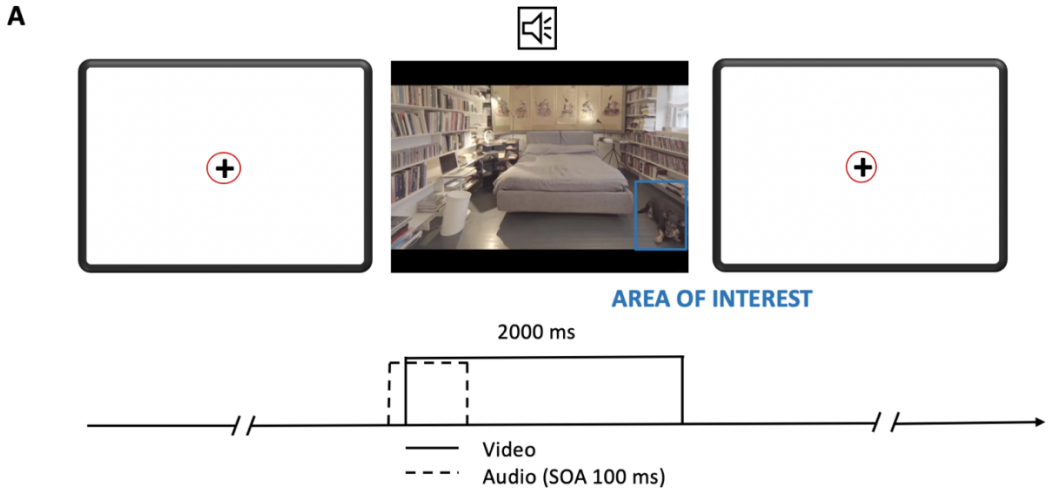

Figure 1. A) Sequence of events in the experiment. The trial started with the presentation of a fixation cross. The gaze allocation of the participant is represented as a red circle on the screen. Participants were asked to fixate on the cross, so the circle will move to the cross. Once participant was fixated on the center of the screen the trial started. In the *consisten*t and *neutral* condition, the trial started with the sound that was either characteristic to the one of the objects (*consistent*) or not characteristic to any object on the video (*neutral*). Participants were instructed to fixate on the cross but once the video started, they could move their eyes and freely watch the video. Blue quadrant represents area of interest. It was not presented to the participant and was applied on the visual frame on the analysis stage. Here it is placed on top if the video for the illustrative purposes. *B)* One example heat map out of 108. Heat plots indicate fixation behavior of 45 participants over one stimulus video in 3 experimental conditions (15 participants per condition). This visualization is used to display the general distribution of gaze points overlaid on presented on the frame of experimental video. Colors from yellow to blue represent in descending order the amount and duration of fixations on the screen with ratio from 0 to 250.

Results

We performed detection of fixations on the raw eye-gaze data, in relation to the AOI using MATLAB_R2017a. Fixation was defined as a period of stable gaze within 1 degree of visual angle that lasted for 50 milliseconds minimum. We calculated four measurements: number of fixations inside AOI, dwell time inside the AOI, time to first fixation inside AOI and percentage of observed AOIs per each condition (at least one fixation should be found inside AOI). All the results were calculated for each participant, and then averaged across participants.

For a preliminary visualization, we generated heat maps to visualize the general distribution of gaze points in all experimental videos in three sound conditions (*Figure 1B*). From the representative heat maps (see an illustrative example in Figure 1B), it appears that there might be a tendency to spend more time fixating within the AOI in congruent sound conditions in comparison to neutral or no sound conditions. As a next step, we examined and analyzed the data quantitatively (*Figure 2*), and ra n repeated measurements ANOVA separately on each of the four measurements listed above, with subject as the random effect and auditory condition as the factor of interest.

The analysis on the percentage of videos with fixation in the AOIs returned a significant main effect $F(2,88)=14.37$; $p=0.000004$. Based on the significant main effect of condition on the percentage of attended videos with a fixation in the AOIs we went to test our specific prediction using t-tests. We had hypothesized that during

free observation of ecologically valid and dynamic videos, semantically congruent audio-visual event will attract attention. Based on this hypothesis, we predicted that percentage of observed objects of interest should be higher in consistent condition in comparison to the neutral or no sound conditions. The analysis demonstrated that percentage in the consistent condition was higher M=41% than in neutral M=33% [t(44)=2.22, p=0.00015, Cohen's d=0.45] and no sound M=32% [t(44)=2.89, p=0.00002, Cohen's d=0.5] conditions (*Figure 2A*). In summary, an object was 8-9% to be looked at if its characteristic sound was presented.

The analyses on the number of fixations inside the AOI F(2,88)=10.91; p=0.00006 and dwell time inside AOI F(2,88)=17.23; p=0.0000005 also returned significant effects. (These two variables are logically correlated). We hypothesized that consistent auditory semantic information will strengthen the effect of spatial orienting towards visual object and increase exploration inside the AOI. Therefore, we predicted that dwell time and number of fixations inside the AOI in consistent condition would be higher in comparison to the neutral or no sound conditions. One-tailed t-tests showed that number of fixations was higher in the consistent condition (M=0.68) in comparison to neutral (M=0.52), [t(44)=3.12, p=0.0016, Cohen's d=0.44] and no sound (M=0.5), [t(44)=4.22, p=0.00006, Cohen's d=0.5] conditions (*Figure 2B*). Accordingly, dwell time was higher in the consistent condition (M=207 ms) in comparison to neutral (M=153 ms), [t(44)=3.94, p=0.0016, Cohen's d=0.55] and no sound (M=143), [t(44)=5.01, p=0.00006, Cohen's

d=0.66] conditions (*Figure 2C*). All the above comparisons are one-tailed (given the directional hypothesis) and survived the multiple comparison correction using Holm-Bonferroni (Ludbrook, 1998).

The analysis on the time to first fixation was not significant $F(2,88)=1.31$; $p=0.27$. Because our last hypothesis stated that semantically consistent sound will guide attention to the visual object faster than neutral or no sound conditions, we had predicted that participants would look at the AOI faster in the consistent condition than on control conditions. This did not happen. Further exploratory t-tests showed only marginal difference (M=61ms) between consistent sounds and no sound [$t(44)=1.49$, $p=0.072$, Cohen's d=0.16] conditions in the expected direction (*Figure 2D*).

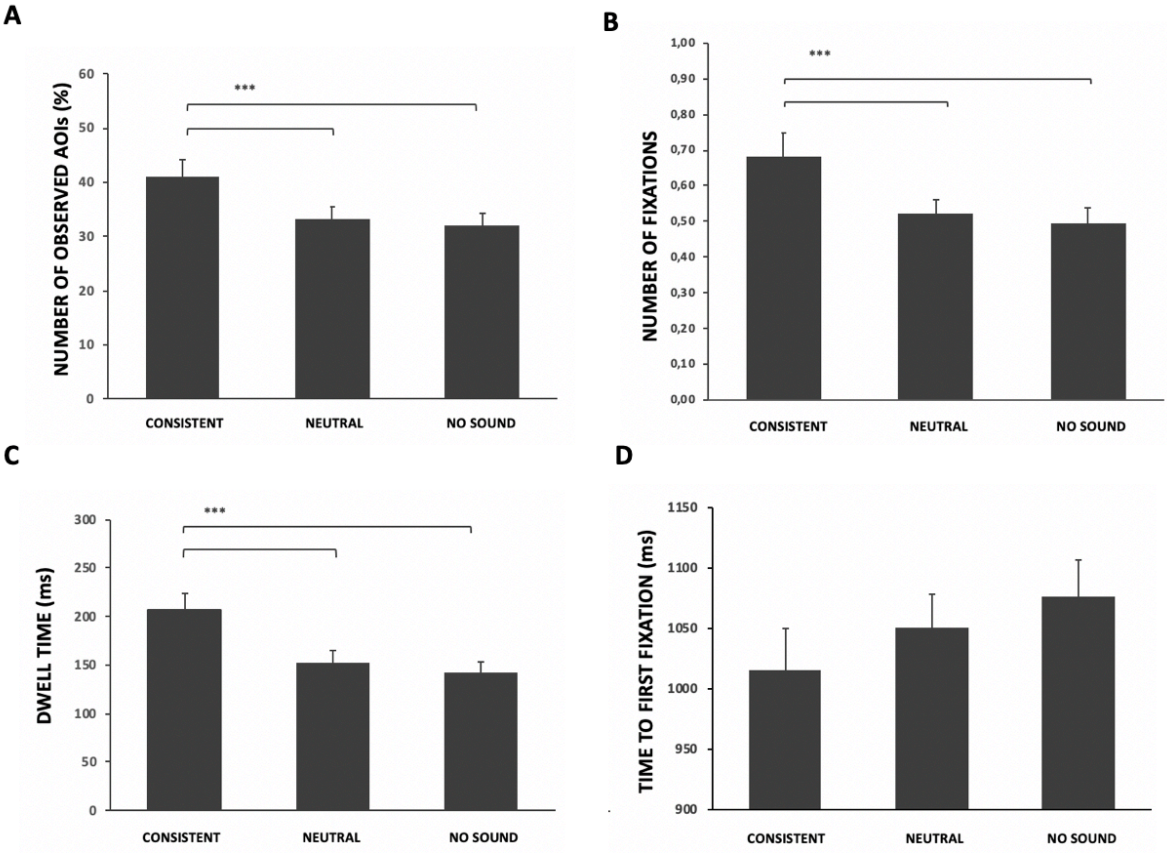

Figure 2. A) The percentage of videos where participants looked at the area of interest. observed areas of interest. B) Number of fixations detected inside of the region of interest C) Total dwell time spent looking inside of the area of interest D) How much time participant spend before the first fixation is detected inside the area of interest

All measurements and error rates were averaged across 45 participants and plotted in the *consistent*, *neutral* and *no sound* conditions Error bars indicate the standard error. Asterisks indicate significant difference between conditions (1 asterisk for p-value less than 0.05, 2 asterisks for p-value less than 0.01, 3 asterisks for p-value less than 0.001)

Discussion

We addressed whether semantic correspondences between sounds and visual objects influence spatial orienting during free viewing of real-life dynamic scenes. Our results suggest that cross-modal semantic congruence indeed has an effect on gaze behavior and on the distribution of spatial attention, in a free viewing paradigm. In particular, hearing the characteristic sound of an object increases the likelihood that the corresponding visual object will be looked at in the scene. Visual objects in control conditions, with neutral sounds or no sound, were looked at only in approximately 32% of times the video was shown. Please note that we intentionally chose videos where the objects of interest had relatively low salience in the image, hence they did not always attract gaze. Despite low salience of stimuli, when presented together with a semantically congruent sound, this percentage increases up to 41%. We also found that cross-modal semantic congruence increased the number of fixations and total dwell time spent looking on the object of interest (inside area of interest) by significant amounts. However, semantically congruent sound did not make participants look at the object of interest faster than in control conditions.

These results indicate that characteristic sounds might increase the salience of corresponding visual object and drive spatial orienting towards them. Many studies have investigated the potential benefits of cross-modal congruence using a variety of attributes (e.g. Bolognini et al., 2005; Koelewijn et al., 2010; McDonald et al., 2000, 2001; Vroomen & de Gelder, 2000), including experiments

addressing semantic correspondences (Chen and Spence, 2011; Iordanescu et al., 2008, 2010; Molholm et al., 2004; Pesquita et al., 2013). As mentioned in the introduction, previous findings have shown a facilitation effect of cross-modal object-based congruence on visual search (Iordanescu et al., 2008;2010; Knoeferle et al., 2016; Kvasova et al., 2019). However, in these studies, participants were given the goal to find objects and therefore, actively used the semantic cues for the task. Here, we have investigated if these cross-modal semantic effects do have an impact on spontaneous viewing behavior (under no task constrains).

The present result has implications about the degree to which voluntary attention guides these cross-modal effects. For example, in the study of Iordanescu (2008) it was demonstrated that semantically congruent sound attracts attention towards visual target, but not to the distractor object. The authors then claimed that cross-modal facilitation can occur only in goal-directed manner, meaning that sounds can only enhance visual representations if an attentional template is activated for a visual target search. This would imply a strong role of voluntary attention. This finding was further supported by our previous study (Kvasova et al., 2019) where we demonstrated that cross-modal semantic effects extrapolate to search tasks into real-life scenes (as opposed ordered to search arrays). Here, the result of Nardo et al. (2014) is also relevant, given that in their experiment using eye-tracking and free observation found no particular effects of cross-modal semantic congruence on orienting. This would support the same idea, namely,

that cross-modal semantics matters only when for goal relevant objects, but wanes when they are task irrelevant.

Our results, however, point to a different conclusion, because the observers oriented more frequently and for longer time to objects whose characteristic sound was playing, even if they were not searching for them or had been primed in any other way. The potential for cross-modal congruence to attract attention even when task irrelevant has been also pointed out by other studies (Mastroberardino et al., 2015; Kvasova and Soto-Faraco, 2019). In both these studies audio-visual objects appeared as task-irrelevant objects before or in parallel with a different primary task. In the two studies, the main finding was that under certain conditions the position of the irrelevant cross-modal congruent objects attracted spatial attention and influenced performance in the primary task.

Previous results then seem to point in opposite directions with regard to the question of whether cross-modal semantic congruence have an effect on spontaneous spatial orienting in real life. How can we reconcile these findings? In the light of the present evidence, it would seem that cross-modal semantic congruence effects on orienting do not abide to a strictly automatic process, yet under some conditions they can percolate behavior even if irrelevant. We believe that the question is not whether or, but under which conditions cross-modal semantic congruence has an influence on orienting. For example, previous studies have varied widely in terms of perceptual load, which seems to be a determining factor for whether irrelevant information is processed or not (Lavie, 2005).

The effect of cross-modal attentional capture with simple stimuli is sensitive to perceptual load (Lunn et al., 2019). In addition, a recent study, Kvasova and Soto-Faraco (2019) shows that this perceptual load modulation also applies to cross-modal semantic congruence effects on in visual search.

However, perceptual load is not the only relevant variable to explain whether cross-modal semantic effects happen or not in free viewing of natural scenes. The high perceptual load typical of the crowded, dynamic scenes did not eliminate the impact of cross-modal congruence in the present study. In natural scenes, the impact of the amount of elements present would not be that high as if we perceive artificial set of events (Peelen & Kastner, 2011). This is because real-world scenes are meaningful, contain context and learned structural relationships between elements (e.g., pendant lamps often appear hanging from ceilings), which are not present in artificial arrays (Kaiser et al., 2014). It was demonstrated previously that not only low-level visual salience but also semantic relationships between the objects in a complex visual scene can guide attention effectively (Wu et al., 2014, for review). In particular, Henderson and Hayes (2017) showed that while calculating salience of visual scenes one must take into account not only low-level features but high-level object and context information, since both low- and high-level information participate in guiding attention.

These structural properties of natural scenes apply to the multisensory case investigated here, and could motivate the fact that, despite the perceptual load, cross-modal congruence still

played a role in task-irrelevant conditions. This constitutes an example of how the outcomes of laboratory experiments with simplified set ups might vary as conditions approach real life (Soto-Faraco et al., 2019; Matusz et al., 2019; Maguire et al., 2012).

In natural scenes, cues regarding low level features (such as the location of a sound and a visual event) as well as the higher-level features regarding structural correlations or semantic relationships are all present at the same time. Nardo et al. (2014) already demonstrated that low-level spatial correspondences between sounds and visual objects affect spontaneous spatial orienting in real-life scenes. Yet, they did not find any effect of semantic congruence. We reasoned that the low level properties in that study overwhelmed the possibly weaker effects of semantic congruence. In our study we minimized the role of low-level cross-modal congruence by making spatial and temporal properties equal across conditions. Admittedly, this is unrepresentative of natural audio-visual events, but a necessary measure in order to isolate the semantic effects from any direct attention-grabbing effect of spatio-temporal correspondences.

Conclusions

In the current study, we examined whether crossmodal sematic congruence guides attention in real-life scenes. We found that characteristic sounds increase the probability to look at the corresponding visual objects and increase total time spent looking and number of fixations at the object of interest. All in all, the results demonstrate that cross-modal semantic congruency can play a role when watching everyday life scenes.

References

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, 61(1), 27–41.

Bolognini, N., Frassinetti, F., Serino, A., and Làdavas, E. (2005). 'Acoustical vision' of below threshold stimuli: interaction among spatially converging audiovisual inputs. *Experimental Brain Research*, 160(3), 273–82.

Burgess, P. W., Alderman, N., Forbes, C., Costello, A., Coates, L. M-A., Dawson, D. R. and Channon, S. (2006). The case for the development and use of 'ecologically valid' measures of executive function in experimental and clinical neuropsychology. *Journal of the International Neuropsychological Society*, 12(02), 194–209.

Chen, Y.-C. and Spence, C. (2011). Cross-modal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *J. Exp. Psychol. Human* 37, 1554–1568.

Driver J. (1996) Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. Nature, 381, 66-68.

Henderson, J.M. and Hayes, T.R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nat. Hum. Behav.* 2017, 1, 743–747.

Hessel, R.S., Kemner, C., van den Boomen, C., and Hoogel, I. (2016). The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli. *Behav. Res.* 48:1694–1712

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., and Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review*, *15*(3), 548–54.

Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., and Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception, & Psychophysics*, *72*(7), 1736–1741.

Kaiser, D., Stein, T. and Peelen, M. V. (2014). Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. *Proceedings of the National Academy of Sciences.* 111:30, 11217-11222.

Kayser, C., Körding, K. P., & König, P. (2004). Processing of complex stimuli and natural scenes in the visual cortex. *Current Opinion in Neurobiology*, 14(4), 468–73.

Kim, Y., Porter, A. M., and Goolkasian, P. (2014). Conceptual priming with pictures and environmental sounds. *Acta Psychologica*, 146, 73-83.

Kingstone, A., Smilek, D., Ristic, J., Kelland Friesen, C. and Eastwood, J. D. (2003). Attention, researchers! It is time to take a look at the real world. *Current Directions in Psychological Science*, 12(5), 176–80.

Knoeferle, K. M., Knoeferle P., Velasco C., and Spence C. (2016). Multisensory brand search: how the meaning of sounds guides consumers' visual attention. *Journal of Experimental Psychology: Applied,* 22(2):196-210.

Koelewijn, T., Bronkhorst, A., and Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: a review of audiovisual studies. *Acta Psychol.* 134, 372–384.

Kvasova, D., Garcia-Vernet, L., & Soto-Faraco, S. (2019). Characteristic sounds facilitate object search in real-life scenes. *Frontiers in Psychology*. 10:2511.

Kvasova, D. and Soto-Faraco, S. (2019). Not so automatic: Task relevance and perceptual load modulate cross-modal semantic congruence effects on spatial orienting. *bioRxiv*

Lavie N (2005) Distracted and confused?: selective attention under load. *Trends Cogn Sci* 9:75–82.

Ludbrook, John. "Multiple comparison procedures updated." *Clinical and Experimental Pharmacology and Physiology* 25.12 (1998):1032-1037.

Lunn, J., Sjoblom, A., Soto-Faraco, S., and Forster, S. (2019). Multisensory enhancement of attention depends on whether you are already paying attention. *Cognition* 187: 38-49.

Mädebach, A., Wöhner, S., Kieseler, M.-L., and 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*. Advance online publication.

Maguire, E. A. (2012). Studying the freely-behaving brain with fMRI. *NeuroImage*, 62(2), 1170–6.

Mastroberardino, S., Santangelo, V., & Macaluso, E. (2015). Crossmodal semantic congruence can affect visuo-spatial processing and activity of the fronto-parietal attention networks. *Frontiers in Integrative Neuroscience*, *9* (July), 45.

Matusz, P.J., Dikker S., Huth A.G & Perrodin C. (2019). Are we ready for real-world neuroscience? *Journal of Cognitive Neuroscience*, 31(3), 327-338.

McDonald J.J., Teder-Salejarvi, W.A., and Hillyard, S.A. (2000). Involuntary orienting to sound improves visual perception. *Nature* 407:906–908.

McDonald, J. J., Teder-Sälejärvi, W. A., and Ward, L. M. (2001). Multisensory integration and crossmodal attention effects in the human brain. *Science,* 292, 1791–1791.

Molholm, S., Ritter, W., Javitt, D. C., and Foxe, J. J. (2004). Multisensory Visual-Auditory Object Recognition in Humans: A High-density Electrical Mapping Study. *Cerebral Cortex*, *14*(4), 452–465.

Nardo, D., Santangelo, V., and Macaluso, E. (2014). Spatial orienting in complex audiovisual environments. *Human Brain Mapping*, *35*(4), 1597–614.

Neisser, U. (1976). Cognition and reality. Principles and implication of cognitive psychology. San Francisco: WH Freeman and Company.

Neisser, U. (1982). Memory: what are the important questions? In J. U. Neisser & I. E. Hyman (eds.), *Memory observed* (pp. 3–18). New York: Worth.

Peelen, Marius V., and Sabine Kastner. "A neural basis for real-world visual search in human occipitotemporal cortex." *Proceedings of the National Academy of Sciences* 108.29 (2011): 12125-12130.

Peelen, M., and Kastner, S. (2014) Attention in the real world: toward understanding its neural basis. *Trends in Cognitive Sciences,* 18(5).

Pesquita, A., Brennan, A. A., Enns, J. T., and Soto-Faraco, S. (2013). Isolating shape from semantics in haptic-visual priming. *Experimental Brain Research*, *227*(3), 311–322.

Santangelo, V., & Spence, C. (2007). Multisensory cues capture spatial attention regardless of perceptual load. *Journal of Experimental Psychology: Human Perception and Performance*, 33(6), 1311–21.

Soto-Faraco, S., Kvasova, D., Biau, E., Ikumi, N., Ruzzoli, M., Moris-Fernandez, L., and Torralba, M. (2019). Multisensory

interactions in the real world. *Cambridge Elements of Perception,* ed. M. Chun, (Cambridge: Cambridge University Press).

Spence C, Nicholls ME, Gillespie N, and Driver J (1998): Cross-modal links in exogenous covert spatial orienting between touch, audition and vision. *Percept Psychophys* 60:544–557.

Spence, C. & Soto-Faraco, S. (2019). Crossmodal attention applied: Lessons for and from driving. To appear in M. Chun (Ed.), *Cambridge Elements of Attention*. Cambridge, UK: Cambridge University Press.

Van den Brink, R. L., Cohen, M.X., van der Burg, E., Talsma, D., Vissers, M.E., and Slagter, H. A. (2014). Subcortical, modality-specific pathways contribute to multisensory processing in humans. *Cereb. Cortex* 24, 2169–2177.

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., and Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1053–1065.

Vroomen, J., & Gelder, B. de. (2000). Sound enhances visual perception: Cross- modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance.* American Psychological Association.

Wu, C.-C., Wick, F. A., and Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology*, 5, 54.

3. GENERAL DISCUSSION

This dissertation aimed at assessing the role of audiovisual semantic correspondences in visuo-spatial orienting. Despite some previous studies have already studied this question, the approach here was to understand better how this cross-modal congruence effects on attention would play out in complex, close to real life scenarios. The thesis addressed the following three main hypotheses:

1. The effects of crossmodal semantic congruence emerges when at least one of two conditions apply: the audio-visual object (or one of its components) carries some relevance to the current goal or, it is irrelevant but presented under low perceptual load.
2. Crossmodal semantic congruence can guide spatial orienting in real-life scenes in goal directed manner.
3. Semantically consistent sounds would increase the salience of the corresponding visual object under free observation of real-life scenes, and therefore the probability of directing overt attention toward the visual object would increase.

In this final chapter I will discuss the findings of the empirical studies reported in Chapters 2.1-2.3 with respect to these three hypotheses and their implications regarding current literature.

3.1. Cross-modal semantic congruence speeds up search under task relevance

The findings from Chapter 2.1 (the first study of the thesis) show that audio-visual semantic congruence can help improve

performance when searching for visual objects. Visual search represents in this case an example of task-relevant conditions since the audio-visual objects are explicitly relevant for the current goal (find the visual object). The results demonstrated, using a visual search task for object images in an array, that characteristic sounds shorten search latencies of the corresponding visual targets in comparison to sounds semantically congruent with a distractor or just neutral sounds that did not correspond to any object in the array. This result replicates prior findings and suggests that cross-modal semantic congruence can attract attention in these goal-directed tasks (Iordanescu et al, 2008; 2010; Knoeferle et al., 2016). In Experiment 1b distractor-consistent sounds did not slow down responses compared to neutral sounds, suggesting that cross-modal congruence only benefits processing of relevant visual object and not the irrelevant distractor. This crossmodal effect on attention orienting under goal-directed conditions was further supported by the results of visual search but in more naturalistic conditions. In particular, the second study presented in the thesis (Chapter 2.2) used dynamic more ecologically valid visual scenes with similar results (Kvasova et al. 2019a).

3.2 The effects of cross-modal semantic congruence in task-irrelevant objects

One question of relevance, in order to settle the automaticity of these effects was whether sounds congruent with distractors would hinder performance by attracting attention to non-target objects.

Here, similar to prior literature, the results were mixed. Whereas in Experiment 1b of Chapter 2.1 and of Chapter 2.2 distractor-congruent sounds did not hinder performance (hence, suggesting a non-automatic process), the results of Experiment 1c (in Chapter 2.1) revealed an effect of distractor-consistent sounds. These sounds increased search times in comparison to neutral sounds, suggesting attention capture. This finding goes against the hypothesis and suggests that despite the irrelevance to the current goal semantically congruent audio-visual distractor attracted attention. Yet, after the results of other experiments in Chapter 2.1 (Experiments 2 and 3) we believe that any automatic effect of cross-modal semantic congruence might not be strong, since neither of the previous studies nor the studies presented in this thesis replicates the difference in search time between semantically inconsistent and neutral sounds. This may be linked to a limitation in terms of perceptual load, as is discussed further, below.

The results of Experiment 2 (Chapter 2.1) showed that task relevance is indeed not a necessary requirement for crossmodal semantic effects on attention to occur. However, Experiment 3 revealed that this is true only when perceptual load is low. These two experiments, which specifically addressed cross-modal semantic effects under task irrelevant conditions, used a visual search task unrelated to the audio-visual objects. In this case, the subjects task was to search for a T in an array of tilted Ts, just after or at the same time as a completely irrelevant array of visual objects was presented together with a sound. When perceptual load was low, search times were faster if the target of the main visual search

task (the T letter) appeared at a location previously occupied by an audio-visually congruent but task-irrelevant audio-visual object (valid trials). This benefit was found in comparison to the invalid condition (audio-visual pair primed wrong location) and neutral (sound was not congruent to any of the visual objects). These results are in line with the study of Mastroberardino et al. (2015), although the adapted paradigm was significantly different. The effect of irrelevant audio-visual event was found despite the higher uncertainty and greater competition between stimuli then in the study of Mastroberardino et al. (2015).

3.3. Perceptual load and automaticity of cross-modal semantic effects

Previous authors have suggested that, unlike the strongly automatic effects of crossmodal spatial congruence, high-level semantic congruence can only produce enhancements in behavior in a goal-directed task-relevant manner (Molholm et al., 2004; von Kriegstein et al., 2005; Iordanescu et al., 2008; 2010). In contrast, the results of Experiment 1c and Experiment 2 suggest that even task-irrelevant crossmodal semantic congruence can attract attention (incidentally, this conclusion is also supported by the results of the free observation study, in Chapter 2.3, which will be discussed later). The pattern of results would suggest that audio-visual congruent objects do have a tendency to attract attention even if task irrelevant. This tendency, however, wanes as perceptual load increases. This is what we learn from the results of Experiment 3 (Chapter 2.1), which showed that perceptual load might act as a

limiting condition to the automatic tendency for cross-modal congruence effects on orienting. Altogether, we believe that this combination of results speak against a strong automaticity account of cross-modal semantic interactions.

Perceptual load has been demonstrated as an important factor for irrelevant information to be consequential to behavior or not in visual attention (Lavie et al., 2005). In addition, the effect of cross-modal attentional capture with simple stimuli has been shown to be sensitive to perceptual load (Lunn et al., 2019). It seems that perceptual load modulation also applies to cross-modal semantic congruence effects observed in Chapter 2.1. We could conclude that irrelevant audio-visual congruent events do not attract attention when the number of items for processing is high and therefore the amount of resources is exceeded. This means that crossmodal semantic congruence necessitates from top-down regulation in order to guide attention, above and beyond any fast, bottom up cross-modal integration process. This audio-visual interaction can be induced in the absence of a particular relevance to the task, as long as sufficient processing resources are available. As it was found in the experiments of Chapter 2.1, when attention is fully engaged in different task due to high perceptual load semantically congruent audio-visual event does not attract attention. Perceptual load might also account for the unstable effect of distractor consistent sounds in Experiment 1c, as well as perhaps in other experiments in the literature. Task irrelevance of the audio-visual distractor events in Experiment 1 is similar to ones in the high perceptual load condition of Experiment 3. In both cases distractor and target are presented at

the same time and intervene with the attentional template activated for a different object. Taken together these results suggest that semantic-based audio-visual integration is not strictly automatic and requires some attention in order to emerge.

3.4. Cross-modal semantic congruence in real-world scenarios

Another important contribution of the present dissertation is the demonstration of how audio-visual semantic congruence influences visuo-spatial orienting in real-life scenes. Findings in the second study (Chapter 2.2) support previous demonstrations of crossmodal semantic effect on visual search (Iordanescu 2008;2010; Knoeferle et al., 2016; Kvasova et al. 2019b). However, this study demonstrated for the first time that semantically congruent and spatially uninformative sounds speed up visual search times not only in simple and artificial displays but also when searching for an object in complex and dynamic scenes that contained contextual information. Hence, the hypothesis about the potential of crossmodal semantic congruence to guide spatial orientin in a goal directed manner was confirmed in real life scenes.

Reaction times in a search task were faster in the target-consistent condition than in the distracter-consistent, neutral or no sound conditions (target-consistent characteristic sounds help attract attention to the corresponding visual object). Distracter-consistent sounds produced no measurable advantage or disadvantage with respect to the control condition, supporting again the weak effects

of task-irrelevant crossmodal semantic congruence under (putatively) high perceptual load, as explained above. Importantly, effect of the characteristic sound was demonstrated in isolation of possible spatio-temporal audio-visual correspondences, therefore the observed effect is due to object-based correspondences only. The beneficial effect of semantically congruent auditory information in visual search neither could be explained by the general alerting of the sound (Nickerson, 1973) since no difference between distractor-consistent, neutral and visual only trials has been found.

Previous studies have already highlighted the difference in how visual attention operates in naturalistic, real-life scenes compared to simple and artificial displays that are used traditionally in psychophysical studies (Kingstone et al., 2003; Wolfe, Horowitz, & Kenner, 2005; Nardo et al., 2011; Peelen & Kastner, 2014; Henderson & Hayes, 2017). Additionally, semantic relationships between the objects in a complex visual scene were proved to be effective in guiding attention (Wu, Wick, & Pomplun et al., 2014), Therefore, the findings in Chapter 2.2 fill an important gap between crossmodal semantic effects in artificial and realistic environments. First, the results of the second study (Chapter 2.2) showed that crossmodal semantic congruence can guide attention to the target of visual search despite the higher perceptual complexity of realistic scenes. Second, this finding demonstrates that the predictive semantic, functional and structural relationships in complex scenes do not make characteristic acoustic information redundant. These structural and functional relationships have been shown to facilitate

search in visual only studies (e.g. Hershler & Hochstein, 2009; Preston et al., 2013; Peelen & Kastner, 2014; Kaiser et al., 2014). Here, it is shown that characteristic sounds might benefit visual search along with, or in addition to, the available visual semantic structure of the scene.

3.5. The impact of cross-modal semantic congruence during free observation

Together with crossmodal semantic congruence during goal-directed tasks, such as in visual search, this dissertation also addressed the role characteristic sounds in visuo-spatial orienting during free observation of real-life scenes. The results of the third study (Chapter 2.3) suggest that crossmodal semantic congruence has an effect on gaze behavior in free viewing. In particular, it was found that hearing a characteristic sound of a visual object increases the likelihood of that object to be looked at, if present in the scene. Using different measures derived from eye-tracking during free observations of videos, the experiment showed that sound-congruent objects were observed by participants significantly more times, compared to when sounds were inconsistent or when no sound was presented. It was also shown that that cross-modal semantic congruence increased the number of fixations and total dwell time spent looking on the object of interest.

Previous studies already demonstrated benefits of crossmodal congruence in a variety of tasks, by manipulating congruence between different attributes such as space and or time (e.g.

Bolognini et al., 2005; Koelewijn et al., 2010; McDonald et al., 2000, 2001; Vroomen & de Gelder, 2000). The role of semantic congruence in particular has been less investigated. Importantly, the role of semantic information shared between modalities was proved to be beneficial for visual detection, identification and (as shown also in this dissertation Chapters 2.2 and 2.3) search tasks (Chen and Spence, 2011; Iordanescu et al., 2008, 2010; Molholm et al., 2004; Pesquita et al., 2013). However, these studies addressed the role of cross-modal congruence in paradigms with goal-directed tasks. In Chapter 2.3 we addressed the impact of audio-visual semantic congruence on the spontaneous orienting behavior under free observation of visual scenes. Demonstrating the effect of cross-modal semantic congruence on attention without any particular task was important in order to raise a doubt on the claim that this effect only operates in goal directed manner, meaning when audio-visual event is explicitly relevant to the current goal.

The fact that characteristic sounds increase the processing of the visual object only when relevant suggests a strong role of top-down voluntary attention. So far, the result of the study by Nardo et al., 2014 would suggest so. However, the results of the first and third studies (Chapters 2.1 and 2.3), together with the study of Mastroberardino et al. (2015), lead to a different conclusion: even though the strong automaticity of the effect of crossmodal semantic congruence on attention was disproved by the combination of high perceptual load and irrelevance of the audio-visual event to the task, the impact of voluntary attention in this process is much less then it was thought before.

3.6. Real-world scenes and perceptual load

High perceptual load was highlighted as a limiting factor for the effects of crossmodal semantic congruence on orienting in the first study (Experiments 2 and 3). However, it seems not to play a big role in the highly cluttered but realistic and meaningful scenes employed in Chapter 2.2 and, especially, under the free observation conditions of Chapter 2.3. It is arguable that in natural scenes, the impact of the amount of elements present would not be as detrimental as in artificial sets of elements (Peelen & Kastner, 2011). For instance, Li et al. (2002) demonstrated that participants easily detect meaningful peripheral stimuli despite the high perceptual load of a dual task, whereas performance drops at chance level when artificial stimuli are presented instead of meaningful ones. In a similar way, the rich semantic structure of natural scenes provides basis for more efficient parsing, so it is feasible that the effect of perceptual load on crossmodal semantic congruence seen in artificial displays, tends to vanish when perceiving natural scenes. This would provide a valid illustration of how the outcomes of laboratory experiments, with simplified set ups, might change when played out in real life scenarios (Maguire et al., 2004; Soto-Faraco et al., 2019; Matusz et al., 2019).

4. Conclusions

The results of this dissertation extend our knowledge about the role that crossmodal correspondences play in spatial orienting in real life scenes. Previous visual only studies demonstrated how low-level visual salience together with high-level meaningful structure of the visual scene account for distribution of attention (Henderson & Hollingworth, 1999; Wolfe & Horowitz, 2017; Henderson and Hayes, 2017). Further, Nardo et al. (2014) demonstrated how low-level spatial congruence between auditory and visual modalities affect distribution of attention. In the present dissertation a further step toward ecological generalization was made. It was found that meaningful information shared between auditory and visual modalities contributes to visuo-spatial orienting in natural scenes. In addition, the present findings extend previous knowledge on conditions under which semantic crossmodal interaction influence attention, regarding task relevance and perceptual load. Taken together, the results of the empirical studies presented in this thesis demonstrate that cross-modal semantic congruency can play a role when searching through, or simply watching, everyday life scenes. In more concrete terms, this dissertation advances two main conclusions on how crossmodal semantic congruence influence attention.

First, it was demonstrated that crossmodal sematic congruence attracts attention in task-relevant and no task conditions in complex, dynamic scenes. This demonstration not only generalizes and confirms previous laboratory findings on semantically based crossmodal interactions but expands it to the field of research in natural scenes.

Second, it was found that irrelevant audio-visual semantically congruent events can summon attention, but only when presented under low perceptual load conditions. When these audio-visual events are irrelevant to the task and perceptual load is high, then their attention-grabbing effect vanishes. This pattern of results does not support a strict automaticity hypothesis of semantic integration across modalities, and instead suggests that some top-down processing is necessary for audio-visual semantic congruence to trigger spatial orienting.

5. Future directions

To this point, evidence suggests that crossmodal congruence plays a role in visuo-spatial orienting. In the future work, it is of a particular interest to further investigate how characteristic sounds can influence visual salience and meaning maps in natural scenes.

To further approximate the real world, an important step for future research will be to study neural mechanisms of crossmodal semantic interactions in complex scenes. Additionally, it is important to demonstrate if semantically congruent and spatially uninformative visual cues could produce an enhancement through the object-based interactions between auditory and visual modalities while performing an auditory task. Such demonstration will help to test whether audio-visual semantic effects on attention operate in a bi-directional way.

6. References

Amedi, A., von Kriegstein, K., van Atteveldt, N. M., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research*, 166(3–4), 559–71.

Bertelson, P., Vroomen, J., De Gelder, B., & Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics*, 62(2), 321–32.

Biederman, I., Mezzanotte, R. J., and Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cogn. Psychol*. 14, 143–177.

Bolognini, N., Frassinetti, F., Serino, A., and Làdavas, E. (2005). 'Acoustical vision' of below threshold stimuli: interaction among spatially converging audiovisual inputs. *Experimental Brain Research*, 160(3), 273–82.

Chen, Y.-C. and Spence, C. (2011). Cross-modal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *J. Exp. Psychol. Human* 37, 1554–1568.

Doehrmann, O., &Naumer, M. J. (2008). Semantics and the multisensory brain: How meaning modulates processes of audio-visual integration. *Brain Research*, 1242, 136–50.

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, *381*(6577), 66.

Evans, K.K., Georgia-Smith, D., Tambouret, R., Birdwell, R.L., & Wolfe, J.M. (2013) The gist of the abnormal: above-chance medical decision making in the blink of an eye. *Psychon. Bull. Rev.* 20, 1170–1175.

Greene MR, Oliva A (2009) The briefest of glances: The time course of natural scene understanding. *Psychological Science,* 20: 464–472.10.

Harel, J., Koch, C. & Perona, P. (2006). Graph-based visual saliency. in *Advances in Neural Information Processing Systems (NIPS 2006)* Vol. 19, 1–8.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. Annual Review of Psychology, 50, 243–271.

Henderson, J.M. and Hayes, T.R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nat. Hum. Behav.* 2017, 1, 743–747.

Hershler, O. and Hochstein, S. (2009). The importance of being expert: top-down attentional control in visual search with photographs. *Atten. Percept. Psychophys*. 71, 1478–1486

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., and Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review*, *15*(3), 548–54.

Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., and Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception, & Psychophysics*, *72*(7), 1736–1741.

Itti, L., Koch, C., & Niebur, E. (1998) A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell* 20, 1254–1259 (1998).

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews, Neuroscience*, 2, 194–203.

Kingstone, A., Smilek, D., Ristic, J., Kelland Friesen, C., & Eastwood, J. D. (2003). Attention, researchers! It is time to take a look at the real world. *Current Directions in Psychological Science*, 12(5), 176–80.

Koelewijn, T., Bronkhorst, A., and Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: a review of audiovisual studies. *Acta Psychol.* 134, 372–384.

Kuai, S.G., Levi, D., & Kourtzi, Z. (2013) Learning optimizes decision templates in the human visual cortex. *Curr. Biol.* 23, 1799–1804.

Kvasova, D., Garcia-Vernet, L., & Soto-Faraco, S. (2019). Characteristic sounds facilitate object search in real-life scenes. *Frontiers in Psychology.* 10:2511.

Kvasova, D. & Soto-Faraco, S. (2019). Not so automatic: Task relevance and perceptual load modulate cross-modal semantic congruence effects on spatial orienting. *bioRxiv*

Laurienti, P., Kraft, R., Maldjian, J., Burdette, J., & Wallace, M. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. Experimental Brain Research, 158(4), 405–14.

Lavie N., and Tsal Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. Percept. Psychophys. 56, 183–197.

Lavie, N. (1995) Perceptual load as a necessary condition for selective attention. *J. Exp. Psychol. Hum. Percept. Perform*. 21, 451–468

Li, F.F. et al. (2002) Rapid natural scene categorization in the near absence of attention. *Proc. Natl. Acad. Sci.* U.S.A. 99, 9596–9601.

List, A., Iordanescu, L., Grabowecky, M., & Suzuki, S. (2014). Haptic guidance of overt visual attention. *Attention, Perception, & Psychophysics*, *76*(8), 2221–2228.

Lunn, J., Sjoblom, A., Soto-Faraco, S., & Forster, S. (2019). Multisensory enhancement of attention depends on whether you are already paying attention. *Cognition,* 187: 38-49.

MacEvoy, S.P. & Epstein, R.A. (2011) Constructing scenes from objects in human occipitotemporal cortex. *Nat. Neurosci*. 14, 1323–1329.

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*. Advance online publication.

Maguire, E. A. (2012). Studying the freely-behaving brain with fMRI. *NeuroImage*, 62(2), 1170–6.

Mastroberardino, S., Santangelo, V., & Macaluso, E. (2015). Crossmodal semantic congruence can affect visuo-spatial processing and activity of the fronto-parietal attention networks. *Frontiers in Integrative Neuroscience*, *9* (July), 45.

Matusz, P. J., Dikker, S., Huth, A. G., & Perrodin, C. (2018). Are we ready for real-world neuroscience? *Journal of Cognitive Neuroscience*,1–12.

McDonald J.J., Teder-Salejarvi, W.A., and Hillyard, S.A. (2000). Involuntary orienting to sound improves visual perception. *Nature* 407:906–908.

McDonald, J. J., Teder-Sälejärvi, W. A., and Ward, L. M. (2001). Multisensory integration and crossmodal attention effects in the human brain. *Science,* 292, 1791–1791.

Molholm, S., Ritter, W., Javitt, D. C., and Foxe, J. J. (2004). Multisensory Visual-Auditory Object Recognition in Humans: A High-density Electrical Mapping Study. *Cerebral Cortex*, *14*(4), 452–465.

Nardo, D., Santangelo, V., & Macaluso, E. (2011). Stimulus-driven orienting of visuo-spatial attention in complex dynamic environments. *Neuron,* 69(5), 1015–28.

Nardo, D., Santangelo, V., and Macaluso, E. (2014). Spatial orienting in complex audiovisual environments. *Human Brain Mapping*, *35*(4), 1597–614.

Nickerson, R. S. (1973). Intersensory facilitation of reaction time: Energy summation or preparation enhancement? *Psychological Review*, 80 489 ^ 50.

Peelen, M. V., Fei-Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, 460(7251), 94–97.

Peelen, M., & Kastner, S. (2014) Attention in the real world: toward understanding its neural basis. *Trends in Cognitive Sciences,* 18(5).

Pesquita, A., Brennan, A. A., Enns, J. T., & Soto-Faraco, S. (2013). Isolating shape from semantics in haptic-visual priming. *Experimental Brain Research*, *227*(3), 311–322.

Preston, T.J. et al. (2013) Neural representations of contextual guidance in visual search of real-world scenes. *J. Neurosci*. 33, 7846–7855

Potter, M. C. (1975). Meaning in visual search. *Science*, 187, 965–966.

Santangelo, V., and Macaluso, E. (2012). Spatial attention and audiovisual processing. *in The New Handbook of Multisensory Processes*, ed. B. E. Stein *Cambridge, MA: TheMIT Press*, 359–370.

Shiffrin, R., and Schneider, W. (1977). Controlled and automatic human information processing: 2. Perceptual learning, automatic attending and a general theory. *Psychol. Rev*. 84, 127–190

Spence, C., Nicholls, M.E., Gillespie, N., & Driver, J. (1998). Cross-modal links in exogenous covert spatial orienting between touch, audition and vision. *Percept Psychophys* 60:544–557.

Spence, C., & Driver, J. (2004). Crossmodal space and crossmodal attention. *Oxford University Press.*

Spence, C. & Soto-Faraco, S. (2019). Crossmodal attention applied: Lessons for and from driving. To appear in M. Chun (Ed.), *Cambridge Elements of Attention*. Cambridge, UK: Cambridge University Press.

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14, 400–410.

ten Oever, S., Romei, V., van Atteveldt, N., Soto-Faraco, S., Murray, M. M., & Matusz, P. J. (2016). The COGs (context, object, and goals) in multisensory processing. *Experimental Brain Research*, 234(5).

Van den Brink, R. L., Cohen, M.X., van der Burg, E., Talsma, D., Vissers, M.E., & Slagter, H. A. (2014). Subcortical, modality-specific pathways contribute to multisensory processing in humans. *Cereb. Cortex* 24, 2169–2177.

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1053–1065.

von Kriegstein, K., Kleinschmidt A, Sterzer, P., & Giraud, A.L. Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience,* 2005;17:367–376.

Vroomen, J., Bertelson, P., & De Gelder, B. (2001). Directing spatial attention towards the illusory location of a ventriloquized sound. *Acta psychologica*, *108*(1), 21-33.

Vroomen, J., Bertelson, P., & De Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & psychophysics*, 63(4), 651-659.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature,* 435(7041), 439–40.

Wolfe, J. M. & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nat. Hum. Behav.* 1, 0058.

Wu, C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology, 5*(February), 1–13.

Yantis, S. (2000). Goal-directed and stimulus-driven determinants of attentional control. *Attention and Performance* XVIII, eds S. Monsell and J. Driver (Cambridge, MA: MIT Press), 73–103.