Do sounds affect how we see?

Oscar Solis[1], with supervision from Cameron Kyle-Davidson[2] and Dr Karla K. Evans[2]

[1]School of Natural Sciences, University of York

[2]Department of Psychology, University of York

ABSTRACT

This project aims to investigate whether cross-modal semantic congruence is advantageous for visual search and clarify and model the audio-visual interactions that make it possible. Participants localised a target object in a naturalistic scene image. The image was presented with a sound that was congruent or incongruent to the target or with no sound, and targets were cued either visually or auditorily. A computational model was also designed to use sounds to localise target objects in scene images. Results showed that response times were faster when images were presented with a congruent sound compared to an incongruent sound, suggesting that cross-modal semantic congruence enhance the salience of the visual object and effectively guides attention. Cross-modal advantages were observed for visual but not auditory cues, suggesting that sounds may activate a visual representation indirectly through an intermediate semantic representation. The computational model failed to generalise from training to validation, possibly due to a limited dataset, highlighting how audio-visual interactions in humans may be due to learned associations.

INTRODUCTION

Each modality we experience the world with provides us with different information. By effectively combining corresponding information from different modalities, we can understand the world in a meaningful way. Spatially congruent sounds improve visual detection [1] and temporally congruent sounds improve visual search [2]. Information can also be congruent through their identity – hearing ‘meow’ is often accompanied by seeing a cat. In a busy environment filled with spatiotemporal congruences, semantic congruence could help make sense of that environment. This project aims to investigate whether semantically congruent sounds improve visual search performance in naturalistic scenes.

In the study by Iordanescu et al. [3], participants searched for a target in an object array. Responses were faster when a semantically congruent sound was played during search compared to incongruent or no sound. This suggests that cross-modal semantic congruence guides attention by enhancing the salience of the visual object. However, the use of arrays in this and prior studies [4,5] lack ecological validity. Objects tend to appear in certain places within a scene [6], therefore cross-modal semantic congruence may be redundant in naturalistic visual search.

Kvasova et al. [7] attempted to address this issue by using videoclips of naturalistic scenes. They showed that participants were faster at correctly detecting targets when a semantically congruent sound was played with the videoclip compared to an incongruent or no sound. This suggests that cross-modal semantic congruence can guide attention in complex, naturalistic scenes. However, videoclips also provide motion information, which may cue object recognition [8]. Some naturalistic visual searches tend to be stationary, such as baggage

screening, so cross-modal semantic congruence may not improve visual search performance in naturalistic scene images.

In the wider literature, semantically congruent sounds have been shown improve performance in various visual tasks such as identification [9], but the nature of this interaction from audition to vision has yet to be understood.

Sounds may directly activate a visual representation, i.e., hearing ‘meow’ could prime the visual features of a cat. Vallet et al. [10] asked participants to categorise a target image. Targets were primed with a semantically congruent sound presented before the image. Responses were slower when the prime was accompanied by a visual mask, suggesting that the sounds directly activated a visual representation which the visual mask would have interfered with.

Hearing a ‘meow’ could also activate a semantic representation of a cat, which is then transformed into a visual representation. Brandman et al. [11] presented participants with blurry images of objects and asked them to press a key whenever the same image was presented twice in a row. Semantically congruent sounds, spoken words naming the object or uninformative noise were played before each image. Brain activity recorded using magnetoencephalography was used to identify how quickly an object category was decoded by the brain. Results showed faster decoding with semantically congruent sounds or spoken words compared to uninformative noise, reflecting an advantage of semantic congruence. The decrease in decoding time was similar for semantically congruent sounds and spoken words, suggesting that, in the same way that words activate semantic representations, audio-visual interactions may also operate indirectly through semantic representations.

Although multiple findings have indicated that semantically congruent sounds advantage visual search, how audio-visual interactions may occur is still debated.

Here we aim to extend findings of cross-modal semantic congruence to visual search in naturalistic scene images. We hypothesise that response times are shorter for congruent sounds than incongruent or no sound, replicating results from previous studies [3,7]. We also hypothesise higher accuracy for congruent sounds than incongruent or no sound. An incongruent sound condition ensures that the audio-visual objects are integrated rather than sounds increasing a participant’s readiness for a task compared to no sound (general alerting).

Second, we aim to test whether audio-visual interactions occur directly or indirectly by altering how targets are cued – either with a line drawing (providing direct access to some of the visual features used during search) or a sound. If interactions occur directly, we expect the difference in response times between congruent and incongruent (or no sound) conditions to be similar for auditory and visual cues. However, if interactions occur indirectly, we expect these differences to be smaller for the auditory cue because it would require more time to transform the sound into a semantic representation and then into a visual representation.

Cross-modal interactions may result from repeated experiences of naturally co-occurring stimuli from different modalities [12]. This idea influenced Arandjelovic & Zisserman [13] to develop a neural network that learns semantically from audio-visual inputs. They trained a neural network on an audio-visual correspondence task, in which sound-image pairs extracted from videos are inputted and their correspondence (whether they come from the same video) is predicted. Objects heard in a video tend to be seen as well, so the neural network was able to learn different object categories. They used this task to train the Audio-Visual Object Localisation Network (AVOL-Net) [14], a neural network that can use sound to localise objects in an image.

We also aim to model of audio-visual interactions computationally. Will the model, inspired by AVOL-Net, perform similarly to human observers? We expect the model to localise targets more accurately when provided with a congruent sound than an incongruent or no sound. Applications include improving automated machines designed for search, i.e., baggage screening [15].

METHODS

Behavioural Study

Participants

Thirty-one volunteers (16 females), aged 19 to 38, took part. All reported normal or corrected-to-normal vision, no special visual characteristics, and good hearing. Participants were compensated with a £5 Amazon voucher for their time.

Stimuli and Apparatus

The experiment was run on MATLAB R2021a and the Psychophysics Toolbox [16,17].

All images and sounds were obtained from royalty-free websites.

A mixture of composite images, created through Photoshop, and unedited images were used for the experiment, 256 in total. Images were naturalistic scenes containing 1 of 16 possible target objects including ‘cat’, ‘piano’ and ‘scissors’. Images always contained a target that was unambiguously in the left or right of the image. For visual cues, black-and-white line drawings of each target were used. Five-hundred-and-twelve textures, generated using Portilla & Simoncelli’s [18] algorithm, were used for masking. All images were overlaid a grey background displayed centrally on a 21” CRT monitor and scenes subtended approximately 32º visual angle.

Twenty-one sounds were used – 16 corresponded to each target and 5 were unrelated, incongruent sounds. Using Audacity, sounds were normalised by Root-Mean-Square normalisation to a perceived loudness level of -18 dB, trimmed to 1000ms duration, and converted from stereo to mono to remove any spatial information. Sounds were played through headphones at a constant volume of approximately 75db for each participant.

Procedure

Prior to the experiment, it was ensured that all participants were able to identify all sounds. Participants sat alone in a quiet, dimly lit room approximately 57cm away from the screen. Figure 1 shows the structure of each trial.

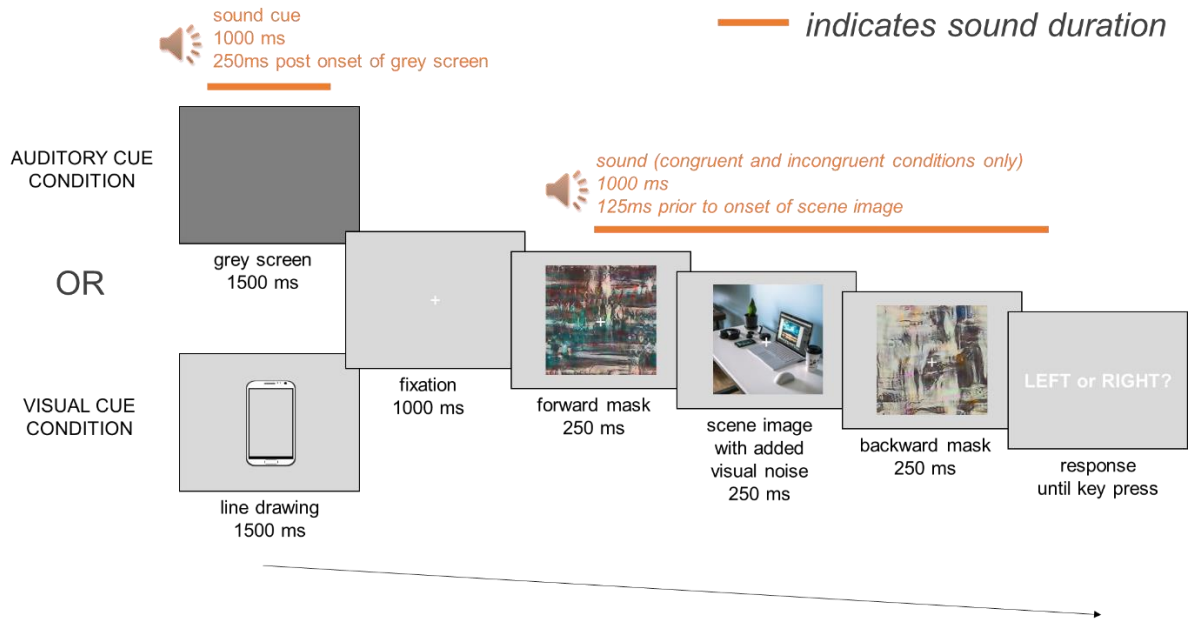


Figure 1. The general structure of a trial. During the experiment, Gaussian noise (mean=0, variance=0.10) was added to these images to impair visibility, preventing ceiling performance.

Each trial started with either a line drawing or a sound cue. A fixation cross was then shown in the centre of the screen, after which a series of images were presented. The task was to indicate whether the cued target object was found on the left or right of the scene image by pressing the left or right arrow keys respectively. Participants were instructed to respond as accurately and quickly as possible.

Sounds were played during search, and three sound conditions were used: congruent (identity of sound and target matched), incongruent (identity of sound did not match any object in the scene) and absent (image presented without sound).

The experiment consisted of a total of 256 trials – 16 practice trials with remaining trials split equally between the three sound conditions. Images and sound conditions were presented in a random order, and an equal number of images had targets in the left and right.

Data Analysis

The two dependent measures were response times (RT) for correct responses and accuracy for the task.

RTs less than 250ms or greater than three standard deviations above the mean were considered outliers and excluded. One participant was excluded due to near-chance performance.

The same 2x3 mixed model ANOVA was used to model the accuracy and RT data to test for statistical significance. Sound condition (3 levels: congruent, incongruent, absent) was a within-subjects factor and cue type (2 levels: auditory, visual) was a between-subjects factor.

Computational Model

Network Architecture

The model consists of three subnetworks, as shown in Figure 2.

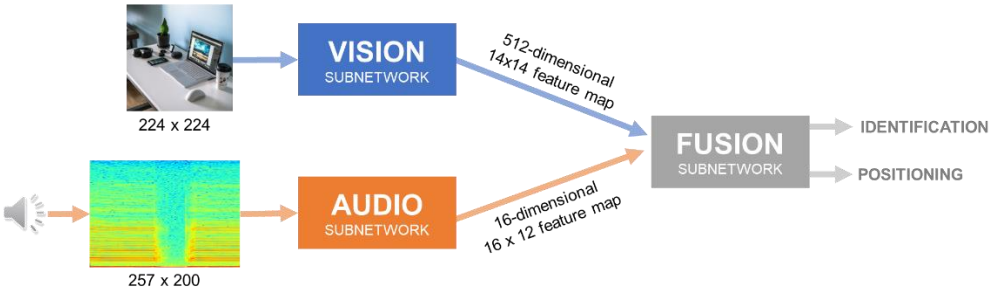


Figure 2. Overall network architecture. A vision subnetwork takes a scene image as its input. An audio subnetwork takes log-spectrograms, a logarithmic plot of frequency over time, generated from sounds as its input. A fusion network combines the outputs of the vision and audio subnetworks to identify the target and predict its position.

For the vision subnetwork, the feature extractor layers of VGG16 [19] pretrained on ImageNet [20] is used with frozen weights.

Sounds are first transformed into log-spectrograms using a method by Frank Zalkow (2012-2013; https://www.frank-zalkow.de/en/create-audio-spectrograms-with-python.html). These are inputted into an audio subnetwork inspired by the one from AVOL-Net [14]. The audio subnetwork that we use has reduced depth and complexity in comparison to accommodate our small dataset.

The fusion subnetwork, also inspired by AVOL- Net [14], was made from scratch. The fusion subnetwork that we use works on fewer features than the one used for AVOL-Net, again to accommodate our small dataset. It reduces the number of dimensions for each subnetwork output to 8 dimensions, performs a pairwise scalar product of the two to generate a 14 x 14 similarity score map used to identify the target object in the image, and predict whether this target is in the left or right of the image.

Implementation and Testing Protocol of the Network

To train and test this model, the set of 256 images (32 for testing), labelled with target category and position, and 21 sounds from the behavioural experiment were used. Congruent sound-image pairs were used for training. Stratified cross-validation was used: the training set was split into 7 equal groups, each containing 2 exemplars for each target. Each group took turns acting as a validation set. Adam optimiser was used with a learning rate of 0.0005. The sum of cross entropy losses between predicted and veridical target categories and positions was used to train the network. The network was implemented in PyTorch and trained on CPU with batch size of 32 for 20 epochs.

Data Analysis

The network's performance was evaluated through accuracy in target identification and predicting position. Two further metrics were used: loss and accuracy predicting target position given a correctly identified target.

RESULTS

Behavioural Study

Accuracy

Analyses revealed a significant main effect of cue, $F(1,28)=7.150$, $p=.012$, $\eta_p^2=.203$, revealing better accuracy with a visual cue. There was no significant effect of sound, $F(2,56)=1.472$, $p=.238$, $\eta_p^2=.050$, indicating that sounds did not affect target localisation accuracy. No significant interaction effect between cue and sound was observed, $F(2,56)=.130$, $p=.878$, $\eta_p^2=.005$, suggesting that accuracy in different sound conditions did not differ according to cue. Figure 3 shows mean accuracy over the different sound and cue conditions.

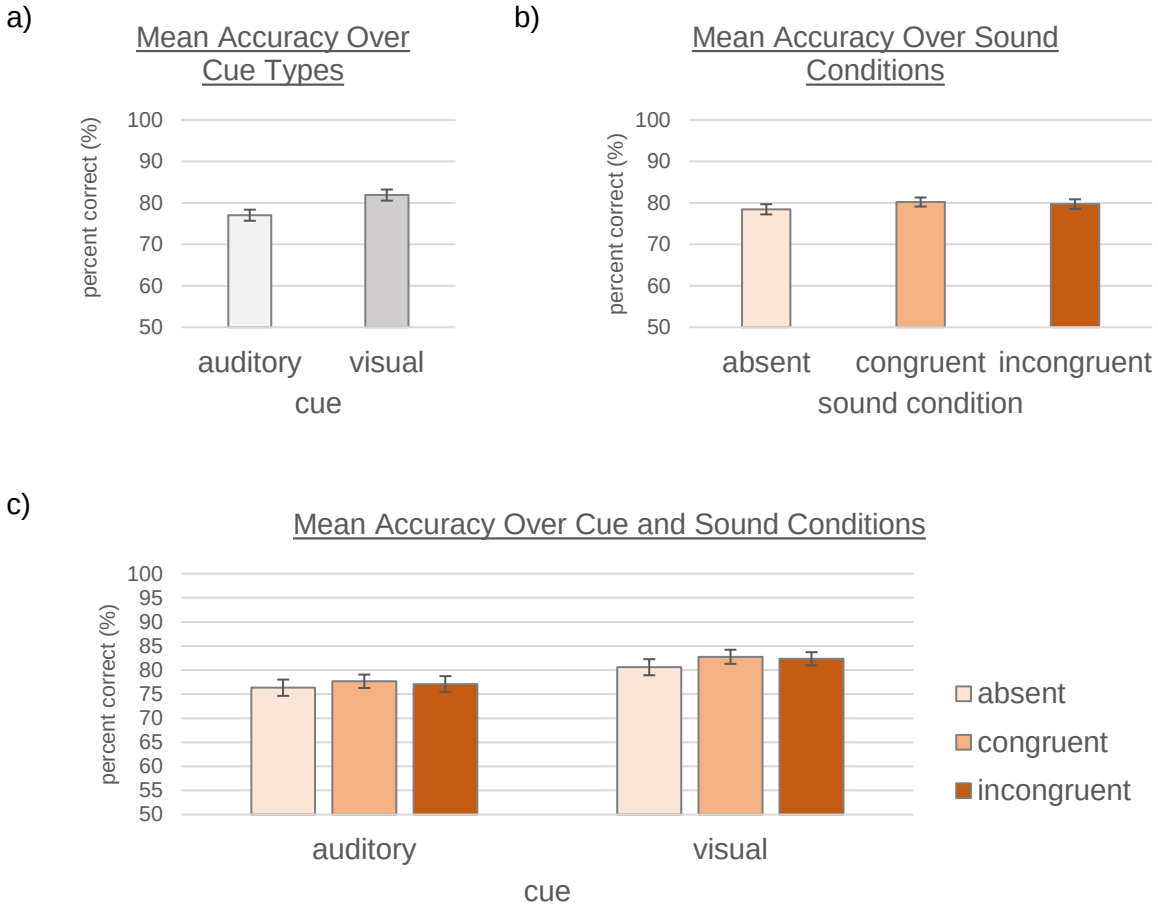
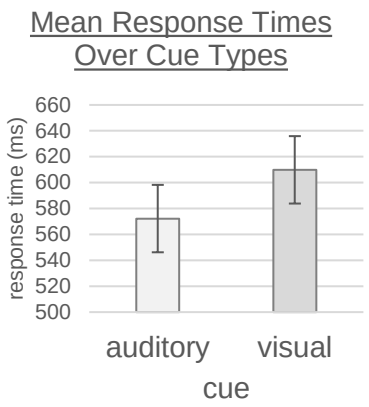


Figure 3. Mean accuracy plotted over a) cue, b) sound and c) cue and sound conditions. Error bars denote mean standard errors.

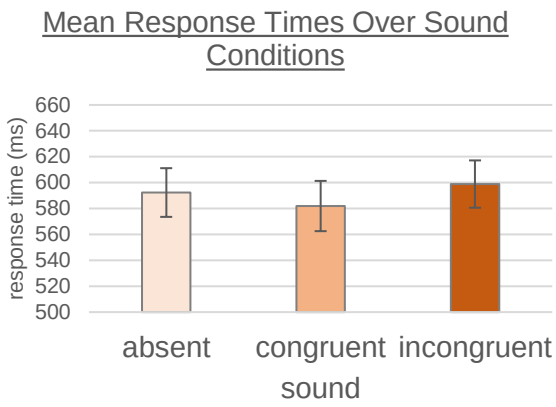
Response Times

Analyses reveal no significant main effect of cue, $F(1,28)=1.046$, $p=.315$, $\eta_p^2=.036$, suggesting that cues did not affect how quickly participants localised targets. The effect of sound was significant, $F(2,56)=3.346$, $p=.042$, $\eta_p^2=.107$. Planned contrasts revealed that RTs for the congruent condition were not significantly faster than for the absent condition, $F(1,28)=2.965$, $p=.096$, $\eta_p^2=0.96$, but were significantly faster than for the incongruent condition, $F(1,28)=7.069$, $p=.013$, $\eta_p^2=.202$. The interaction effect between cue and sound on RTs for correct responses was close to statistical significance, $F(2,56)=2.537$, $p=.088$, $\eta_p^2=.083$. Figure 4 shows mean RTs for correct responses over the different sound and cue conditions.

a)



b)



c)

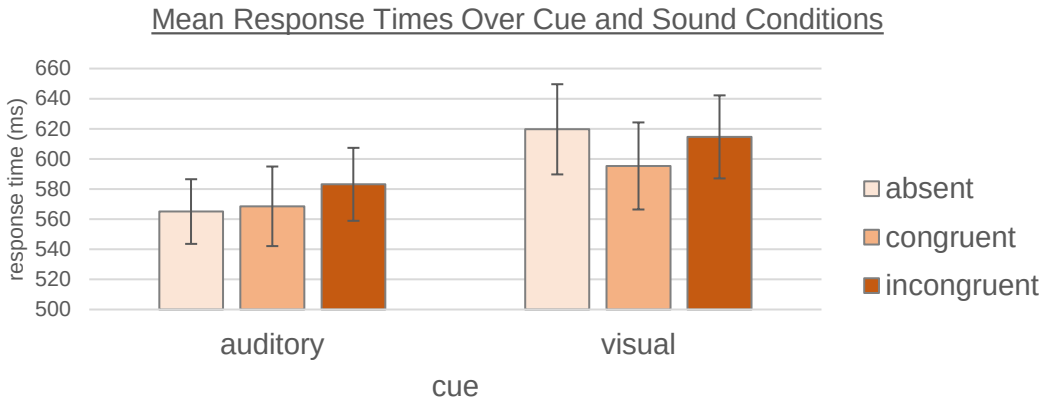
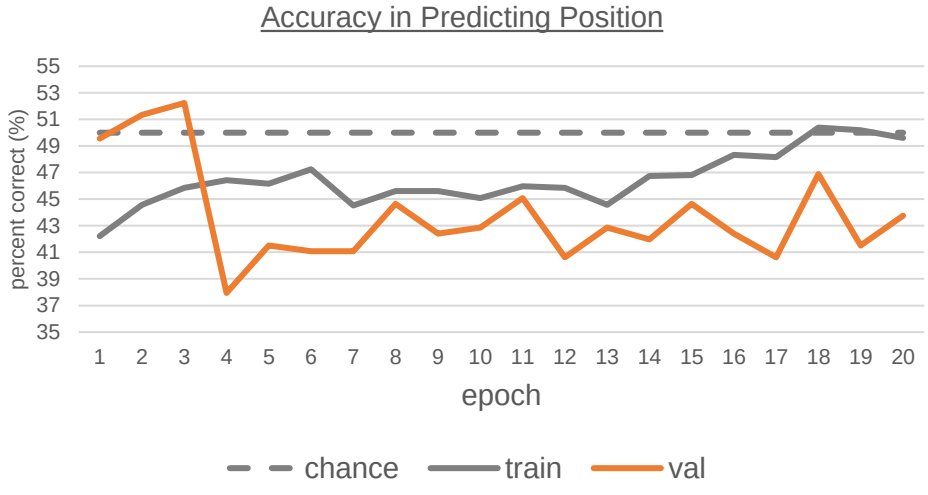


Figure 4. Mean response times plotted over a) cue, b) sound and c) cue and sound conditions. Error bars denote mean standard errors.

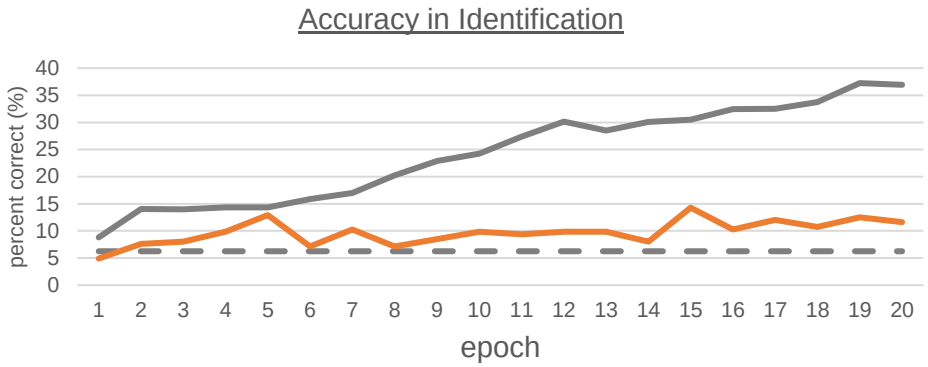
Computational Model

The model's performance during training and validation are shown in Figure 5. Results show that the model was unable to generalise, as loss did not decrease and mean accuracy for identification did not increase in the same ways for validation and training. Mean accuracy for predicting position, even given a correctly identified target, remains near chance.

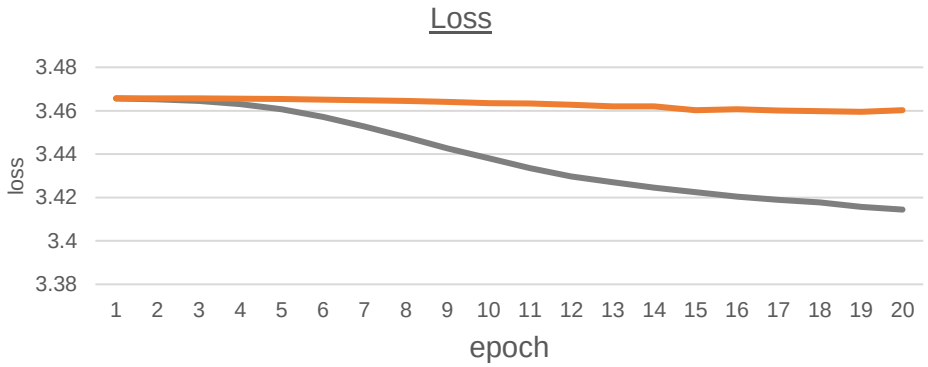
a)



b)



c)



d)

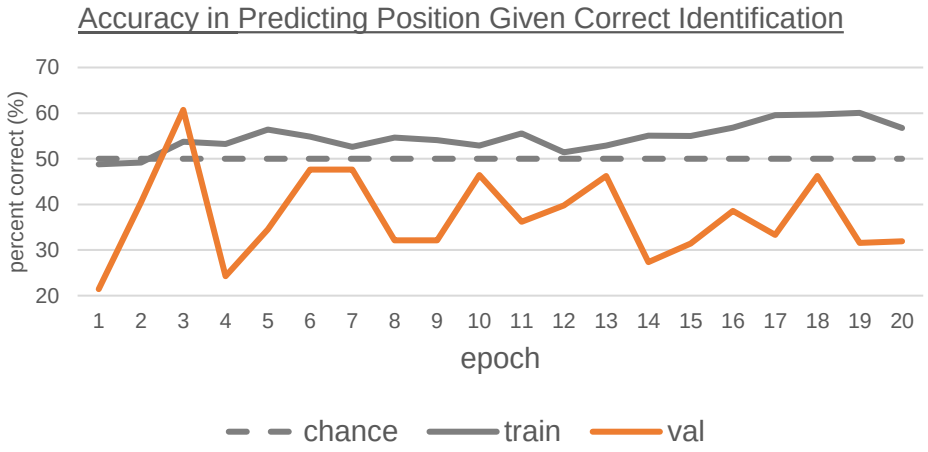


Figure 5. Line graphs showing the performance of the model over epoch, evaluated through four metrics (a-d).

Table 1 shows the model's performance on the test set provided congruent, incongruent or no sound. The model was unable to identify targets in the test set or predict their positions. Since few images were correctly identified, the likelihood that the position of the target in these images was correctly predicted is even lower.

sound	accuracy for identification (%)	accuracy for predicting position (%)	accuracy for predicting position, given correct identification (%)
congruent	3.125	37.5	0
incongruent	3.125	34.375	0
absent	5.625	33.75	0

Table 1. Accuracy for three metrics used to evaluate the performance of the model on the test set.

DISCUSSION

The current project aimed to investigate whether semantically congruent sounds facilitate visual search in real-world scenes. Our results show that accuracy was higher when targets were cued visually than auditorily but the semantic congruence of sounds presented with the scene image did not affect accuracy. Response times were similar for visual and auditory cues, but semantically congruent sounds reduced response times compared to incongruent sounds regardless of cue. This result supports the hypothesis that cross-modal semantic congruence improves visual search performance in real-world scene images. Furthermore, our findings suggest that cross-modal semantic congruence is not redundant despite the availability of other semantic information, such as scene guidance [6], and lack of motion information [8].

Our results suggest that semantically congruent sounds enhance the visual object's salience compared to other objects in the scene, effectively guiding attention in search. This supports the idea that visual attention is guided by semantic information [19] even across modalities. Audio-visual objects may integrate to reduce uncertainty about the visual scene [20]; multiple objects in a scene must be searched so enhanced salience may reduce target misses [21]. Previous studies have shown that sounds congruent to a distractor or task-irrelevant object in the search image does not capture attention and does not lead to a deficit compared to when sound is absent [3,7]. We also find that incongruent sounds did not disadvantage participants' performance compared to no sound. Based on these findings we conclude that cross-modal semantic congruency may only be advantageous for search when it is task relevant.

The lack of response time difference between different modality cues as well as no interaction between cue and sound condition would suggest that the cross-modal interaction has a direct route. However, we see a significant difference in accuracy between visual and auditory cues, which indicated that visual cues prime search for visual targets much better than auditory cues. Although the difference was not statistically significant, response times for the visual cue were slower also indicating a possible speed-accuracy trade-off. In addition, a cross-modal advantage was not observed when cued auditorily, further lending support to the hypothesis that sounds presented with the search image may not directly provide a visual representation that can be used to guide search. Together these findings are more indicative that audio-visual semantic interaction operate indirectly through an intermediate semantic representation.

Although all participants identified all sounds correctly before the experiment, it is also possible that the sounds we used did not map naturally to their intended target. This may explain why we fail to see the effects of a direct link between semantically congruent cross-modal signals. This ambiguity may have also reduced accuracy with auditory cues. Audio-visual stimuli can integrate at the semantic level in multimodal brain regions [22]. Because some sounds in our study may not have been processed semantically, the integration mechanisms may not have been fully engaged, limiting the cross-modal advantages that we could observe.

We also aimed to model audio-visual interactions computationally through a neural network that combined sound-image inputs. Our model was unable to generalise and so could not replicate performance in human observers. However, the model was trained on a very small dataset, with only a few exemplars for each target. This, alongside the fact that the task was complex, resulted in the model's inability to integrate sound and image information effectively. This highlights how the cross-modal advantages that we find in human observers most likely arise from repeated exposure to naturally co-occurring audio-visual stimuli, as opposed to the limited dataset used to train the learning model.

In conclusion, semantically congruent sounds improve visual search performance in naturalistic scene images by effectively guiding attention through enhanced target salience. The interaction between audition and visual may be indirect, whereby sounds activate a semantic representation which is then transformed into a visual representation. A computational model designed to make use of auditory and visual information to localise a target in a scene is unable to show cross-modal semantic congruence advantages, possibly because, for this semantic cross-modal semantic advantage to emerge, there needs to be associative learning across multiple experiences.

ACKNOWLEDGEMENTS

I cannot express enough how grateful I am for the mentorship I have received from Dr Karla K. Evans, Cameron Kyle-Davidson, and the rest of the Complex Cognitive Processing Lab at the University of York. I am also thankful for my friends and fellow Laidlaw Scholars for their support in this project. Finally, I thank my family – I would not be who I am today without them.

BIBLIOGRAPHY

1. Bolognini N, Frassinetti F, Serino A, Làdavas E. 'Acoustical vision' of below threshold stimuli: Interaction among spatially converging audiovisual inputs. Exp Brain Res. 2005;160(3).
2. van der Burg E, Olivers CNL, Bronkhorst AW, Theeuwes J. Pip and Pop: Nonspatial Auditory Signals Improve Spatial Visual Search. J Exp Psychol Hum Percept Perform. 2008;34(5).
3. Iordanescu L, Guzman-Martinez E, Grabowecky M, Suzuki S. Characteristic sounds facilitate visual search. Psychon Bull Rev. 2008;15(3).
4. Kvasova D, Soto-Faraco S. Not so automatic: Task relevance and perceptual load modulate cross-modal semantic congruence effects on spatial orienting. bioRxiv. 2019.
5. Maezawa T, Kiyosawa M, Kawahara JI. Auditory enhancement of visual searches for event scenes. Atten Percept Psychophys. 2022;84(2).
6. Oliva A, Torralba A. The role of context in object recognition. Vol. 11, Trends in Cognitive Sciences. 2007.
7. Kvasova D, Garcia-Vernet L, Soto-Faraco S. Characteristic Sounds Facilitate Object Search in Real-Life Scenes. Front Psychol. 2019;10.
8. Stone J v. Object recognition using spatiotemporal signatures. Vision Res. 1998;38(7).
9. Chen YC, Spence C. When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. Cognition. 2010;114(3).
10. Vallet GT, Riou B, Versace R, Simard M. The Sensory-dependent nature of audio-visual interactions for semantic knowledge. In: Proceedings of the 33rd Annual Conference of the Cognitive Science Society. 2011.
11. Brandman T, Avancini C, Leticevscaia O, Peelen M v. Auditory and Semantic Cues Facilitate Decoding of Visual Object Category in MEG. Cerebral Cortex. 2020;30(2).

12. Iordanescu L, Grabowecky M, Suzuki S. Object-based auditory facilitation of visual search for pictures and words with frequent and rare targets. Acta Psychol (Amst). 2011;137(2).

13. Arandjelovic R, Zisserman A. Look, Listen and Learn. In: Proceedings of the IEEE International Conference on Computer Vision. 2017.

14. Arandjelović R, Zisserman A. Objects that Sound. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2018.

15. Mouton A, Breckon TP. A review of automated image understanding within 3D baggage computed tomography security screening. Vol. 23, Journal of X-Ray Science and Technology. 2015.

16. Brainard DH. The Psychophysics Toolbox. Spat Vis. 1997;10(4).

17. Pelli DG. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spat Vis. 1997;10(4).

18. Portilla J, Simoncelli EP. Parametric texture model based on joint statistics of complex wavelet coefficients. Int J Comput Vis. 2000;40(1).

19. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2015.

20. Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In 2010.

21. Wu CC, Wick FA, Pomplun M. Guidance of visual attention by semantic information in real-world scenes. Vol. 5, Frontiers in Psychology. 2014.

22. Alais D, Burr D. The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. Current Biology. 2004;14(3).

23. Solman GJF, Cheyne JA, Smilek D. Found and missed: Failing to recognize a search target despite moving it. Cognition. 2012;123(1).

24. Calvert GA, Campbell R, Brammer MJ. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Current Biology. 2000;10(11).