Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.
The effect of audition on the deployment of visuo-spatial attention

Daniel Kenneth Rogers
B.A. M.Sc.

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy to the University of Dublin, Trinity College

Supervisor: Professor Fiona Newell

Department of Psychology

November 2010
Declaration

I hereby declare that this thesis, submitted to the University of Dublin, Trinity College, is entirely my own work, unless otherwise stated, and that it has not been previously submitted to this or any other university. I give my permission to the library to lend or photocopy this thesis upon request.
Summary

Vision is commonly accepted as being our most dominant sense. However, our everyday experience informs us that the world is not unisensory. Information from other senses can help us in searching for and identifying objects in our environment. The perceptual processes underpinning visual search have been studied extensively. This thesis investigated specifically the role of audition in attentional deployment on visuo-spatial tasks. Much of the literature in this area is centred around the visual search paradigm and the studies often incorporate basic low-level stimuli. In the present experiments, we used search tasks that incorporated multisensory stimuli in high-level scenes. The primary aim of this thesis was to further our understanding of how sound affects our ability to detect objects in our environment.

The studies reported in this thesis can be divided into two main themes. The first theme concentrated on the role of audition on search for a visual target within static scenes. Here we investigated how sound can be used to guide attention to the location of a static visual target (i.e. Chapter 3). We also investigated the role of characteristic sounds (Chapter 2) and task irrelevant but dynamic visual and audio-visual stimuli in of the deployment of attention in static visual scenes (Chapter 4). The second research theme was an investigation of the role of audition in visual search for a target embedded in complex dynamic scenes. Specifically, we investigated how auditory stimulus which contained either spatial (Chapter 5) and/or temporal (Chapter 6) information affected the detection of visual targets in continually changing environments. Recent studies have found conflicting evidence regarding the role of temporally synchronous sound in visual search tasks. Our research aimed to clarify
whether temporally congruent sound can facilitate target detection in dynamic visual scenes (Chapter 6).

Overall, the findings of this body of research demonstrated that the presence of sound affected visual search performance in static and dynamic scenes. However, the role of sound is less clear for dynamic scenes than it is for static scenes where directionally incongruent sound appeared to benefit task performance. The theoretical implications of the research findings are discussed in Chapter 7.

Multisensory search in high-level dynamic scenes is an emerging area of research and one which could provide many insights into how our senses combine to form a coherent percept of our continually changing environment. There are a number of challenges in studying this in a laboratory environment but clearly there is much scope for future work in this area.
Acknowledgements

I would like to express my gratitude to the people who have helped in the writing of this thesis. Firstly, to my supervisor Professor Fiona Newell: Thank you for the supervision, guidance and opportunities afforded to me during my time in the Multisensory Cognition Lab at TCD. Thank you to my appraisers: Kevin Thomas, Hugh Garavan and Paul Dockree for the insightful feedback, comments and suggestions. Special thanks also to Jason Chan and Annalisa Setti for advice and direction. Special thanks to John Stapleton for programming assistance and to Maeve Barrett for help with data collection. Thanks to Paul McDonald, Simon Dobbyn and Rachel McDonnell for help with stimulus generation. Thanks to all my great lab buddies over the past years, Kate, Joanna, Aisling, Christina, Iwona, Corrina, John and Mark. I am very lucky to have met some lovely people in TCIN over the past few years, Adam, Brendan, Gloria, Erik, Aisling N.S., Jane, Danusia, Ella, Jeni and Sarah to name a few.

Thanks to my parents, Terry and Anto, and my sister Clara Jane, for the love and support. Thanks to friends Paul, Elaine, Richie, Shane, Nick, Sinead, Feek, Simon and Clare for being there. Special thanks to Sophia for encouragement and wisdom when the chips were down, and to Judith for a non-judgemental ear. Thanks to all at DBC for the friendship, wisdom and clarity. Thanks to Melanie for her support in my decision to take on a PhD and for so much more. Thanks to the music and all the people I have the privilege of making it with: Larry, Damo, Paul C., Martin, Shane, Paul S., Jonathan, Sarah, Emma, Aminah and many many more.
This research was funded by Science Foundation Ireland as part of the
Metropolis project: Supercrowds for Multisensory Urban Simulation.

Table of Contents

Declaration ............................................................................................................................. ii
Summary ............................................................................................................................... iii
Acknowledgements ........................................................................................................... v
Table of Contents ............................................................................................................... vi
List of Tables ...................................................................................................................... ix
List of Figures .................................................................................................................... x

1.0  General Introduction .............................................................................................. 2
1.1  The visual search paradigm ....................................................................................... 3
   1.1.1  Principal theories of visual search ................................................................. 3
   1.1.2  Dynamic Visual Search .................................................................................. 7
   1.1.3  Neural basis of visual search ......................................................................... 9
1.2  Multisensory Processes ........................................................................................... 10
   1.2.1.  Neural evidence for multisensory perception ............................................... 11
   1.2.2  Evidence for multisensory interactions in search .......................................... 12
      1.2.2.1  Behavioural evidence for auditory interactions in visual search ............. 12
      1.2.2.2  Neural evidence for multisensory interactions in search ..................... 20
1.3  Evidence for top-down multisensory interactions in search ................................... 21
   1.3.1  The role of crossmodal semantic information in visual search .................... 21
1.4  Scene Perception ..................................................................................................... 23
1.5  Limitations of existing research .............................................................................. 25
1.6  Outline of Thesis ..................................................................................................... 26
   1.6.1  Experimental investigations on search in complex static scenes .................. 26
   1.6.2  Experimental investigations on search in complex dynamic scenes ............. 26

2.0  The effects of characteristic and spatially congruent sounds on visual
search in static visual scenes ............................................................................................ 28
2.1  Introduction ............................................................................................................. 29
2.2  Experiment 1 ........................................................................................................... 31
   2.2.1  Method ............................................................................................................ 32
List of Tables

Table 1.0  A list of scene locations that correspond to centrally and peripherally located targets expressed in terms of visual angle relative to fixation. .................................................................59

Table 2.0 describes the frequencies used for congruent horizontally and vertically moving targets. Horizontally moving targets had a continuous tone that was manipulated using interaural level difference (ILD) whereas vertically moving targets used a frequency shift as described above. 100

Table 3.0 describes the frequencies used in pilot study for congruent horizontally and vertically moving targets. Horizontally moving targets had a continuous tone that was manipulated using interaural level difference (ILD) whereas vertically moving targets used a frequency shift as described above. ..........................................................................................................................181
List of Figures

Figure 1.0 This is an example of two search types: a) a single-feature search where the target is either the red X or the O. Both can be found efficiently based on their single distinguishing features. b) a conjunction search where both features of colour and shape must be used to find the target (a green 'X').

Figure 2.0 This image depicts the six visual targets used in Experiment 1. These targets comprised of (from left to right) a kettle, toaster, clock, duck, cat and dog.

Figure 2.1. An example of visual target a) a clock in an outdoor scene and b) a dog in an indoor scene. Targets were embedded into the visual scenes.

Figure 2.2 This image depicts an example of a spatially congruent (left) and a spatially incongruent (right) scene. Spatial congruency is depicted by sound and visual target occurring in the same hemifield. Spatial incongruency is depicted by the sound and visual target occurring in opposite hemifields.

Figure 2.3 Plot showing the mean response times for the characteristic and uncharacteristic sound conditions for each of the spatial congruency positions of the sound relative to the visual target.

Figure 2.4 Plot showing the mean percentage correct responses across the characteristic and uncharacteristic sound conditions for each of the spatial congruency positions of the sound relative to the visual target.

Figure 2.5 This image illustrates how design differences between Experiments 1 and 2. In Experiment 1, participants searched for one target per block. In experiment 2, the visual targets were presented across trials in a random order. Only one target could appear in each trial (or none in the target absent conditions).

Figure 2.6 Plot showing the mean response times for the characteristic and uncharacteristic sound conditions for each of the spatial congruency positions of the sound relative to the visual target.

Figure 3.0 a) An illustration of four (out of 10) of visual scene backgrounds and b) an image of the visual target used in Experiment 3.

Figure 3.1. An illustration of a scene used in Experiment 3 indicating the possible locations of the visual target. The red markers indicate the central locations of the target whereas the blue markers indicate the peripheral locations of the target relative to fixation. A target could appear in one of these locations in a trial.

Figure 3.2. Plots showing the mean reaction times to the spatially congruent, spatially incongruent and sound absent conditions in Experiment 3. Error bars represent standard error of the mean.

Figure 3.3. Plots showing the two-way interactions on response times between a) AV spatial congruency and target size, b) AV spatial congruency and target size and c) target size and target position in Experiment 3. Error bars represent ± 1 standard error of the mean.

Figure 3.4. Plot showing the interaction between visual target size and AV spatial congruency on the mean correct responses of Experiment 3. Error bars represent ± 1 standard error of the mean.
Figure 4.0  This image depicts examples of male (left) and female (right) ‘point light walkers’ used in this experiment. The white dots represent points on the person’s body for which motion was captured. 

Figure 4.1  This image depicts each of the arrays used. An upright (a) or inverted (b) ‘T’ was presented for 200ms and then an array of crosses only (c) was presented. 

Figure 4.2  This image depicts a typical trial where a fixation cross appears for 500ms, followed by a cue (either ‘<’ or ‘>’), then an inter-stimulus interval of 150ms. Then both the target array and distractor stimulus were presented simultaneously for 200ms. The distracter and the array featuring a ‘+’ instead of the target remained on screen for 4 seconds or until the participant responded. Participants responded to the orientation of the ‘T’ as quickly and as accurately as possible. 

Figure 4.3  Plot showing the interaction in the mean response times for the four distracter types (very masculine, masculine, feminine, very feminine) across the valid and invalid cueing conditions. 

Figure 4.4  Plot showing the interaction on the mean accuracy for the four distracter types (very masculine, masculine, feminine, very feminine) across the valid and invalid cueing conditions. 

Figure 4.5  Plot showing the mean response times for the four audio-visual distracter types (very masculine, masculine, feminine, very feminine) across the valid and invalid cueing conditions. 

Figure 4.6  Plot showing the mean accuracy for the four audio-visual distracter types (very masculine, masculine, feminine, very feminine) across the valid and invalid cueing conditions. 

Figure 5.0  The above images depicts four target present scenes each with a different number of distracters (6, 12, 18, 24). Sample target cars are depicted on the far right: large green car (top) and small red car (bottom). 

Figure 5.1  Plot showing the a) mean response times and b) mean accuracy to locating the target in each sound condition and across each set size for fast moving targets. ‘Congruent’ means the auditory cue was directionally congruent with the visual target, ‘Incongruent’ means the auditory directionally incongruent with the visual target and ‘Sound Absent’ refers to the no sound condition. 

Figure 5.2  Plot showing the a) mean response times and b) mean accuracy to locating the target in each sound condition and across each set size for fast moving targets. ‘Congruent’ means the auditory cue was directionally congruent with the visual target, ‘Incongruent’ means the auditory directionally incongruent with the visual target and ‘Sound Absent’ refers to the no sound condition. 

Figure 5.3  Plot showing the a) mean response times and b) mean accuracy to locating the target in each sound condition and across each set size for fast moving visual targets. ‘Congruent’ means the auditory cue was directionally congruent with the visual target, ‘Incongruent’ means the auditory directionally incongruent with the visual target and ‘Sound Absent’ refers to the no sound condition. 

Figure 6.1  An example of a static image taken from a dynamic scene depicting four roads with cars which were moving left and right. The vertical gray rectangles are “bridges” under which the cars
were occluded. The target object (in this case a green car) in this scene is circled in red for illustration purposes only. The set size is eight in this example, meaning that there were a total of 8 cars in the scene.

Figure 6.2  Plot showing the mean response times to locating the target in each sound condition and across each set size in Experiment 8. The legend refers to each of the auditory conditions: 'con' refers to temporally congruent, and 'incon' to temporally incongruent with the visual target, and 'sa' refers to the sound absent condition.

Figure 6.3  Plot showing the mean percentage accuracy to locating the target in each sound condition and across each set size in Experiment 8.
1.0 General Introduction

In order to search for and select objects in our environment, we must first attend to them. Selective attention has traditionally been studied on a modality specific basis. Our experience of the world is, however, quite different. We receive inputs from several sensory modalities simultaneously. Selection of one input source over another can occur where the information conflicts. A classic example would be the cocktail party phenomenon (Cherry, 1953), where relevant information can be deciphered from multiple overlapping conversations. The Filter Theory proposed by Broadbent (1958) describes an information buffering scenario whereby a participant could encode two kinds of stimuli, both presented at the same time. One of the inputs was allowed through the filter, while the other was waiting in a buffer for later processing. This early work describing the process of selective attention led the way for models of visual search such as the Feature Integration Theory described by Treisman & Gelade (1980) and the Guided Search model described by Wolfe, Cave, & Franzel (1989).

Although the study of audition and vision from a unisensory perspective has amassed a large body of research, it is only in recent years that a literature on sensory integration has emerged. It is now more pertinent than ever to investigate the traditional models of selective attention and visual search from a multisensory perspective. In doing this, we aim to design studies that will provide a greater degree of ecological validity.

This thesis is concerned with how vision and audition interact in our perception of visual scenes. Specifically, the main aim was to investigate whether audition can facilitate visual search for familiar and unfamiliar objects in either static
or dynamic scenes. In this chapter, I will set the theoretical context for the later experimental work by outlining the existing models of visual search and providing a review of the recent literature regarding the role of audition in attentional deployment in the context of visual search.

1.1 The visual search paradigm

The following section describes the main models underpinning visual search theory.

1.1.1 Principal theories of visual search

The ability to locate objects in a cluttered environment is a task which requires focused attention and any theory of visual perception would be of little use if it does not account for the role of attention. Typically we scan the scene searching for an object (target) among a number of other objects (distracters). In a typical experimental context, we are interested in the factors affecting the level of accuracy and time taken to find a target. Traditionally, experimenters have used oriented line segments or coloured letters as target and distracter stimuli. Naturally, we find it easier to locate items that have a different colour to those to which we are searching. Features such as colour, orientation or size enables us to easily discriminate a target from a number of distracters, provided those distracters differ on any one of these dimensions. However, when we find ourselves searching for a target that cannot be defined by a single unique feature, we are forced to use a conjunction of features. For example, searching for a target ‘O’ in an array of X’s (see Fig 1.0a) constitutes a single feature search whereas searching for a green ‘X’ in an array of blue Xs and green diamonds (see Fig 1.0b) constitutes a conjunction search.
Figure 1.0  This is an example of two search types: a) a single-feature search where the target is either the red X or the O. Both can be found efficiently based on their single distinguishing features. b) a conjunction search where both features of colour and shape must be used to find the target (a green 'X').

Feature integration theory (FIT), as proposed by Treisman & Gelade, (1980) suggested that features such as colour and shape are registered pre-attentively and are coded in parallel across the visual field. For example, they described that a red O would 'pop-out' among a number of black X's and O's because it has the discriminative single feature of colour. Treisman suggests that specialised maps exist for the features and these feature maps are thought to correspond to neurons selective for a range of values across the dimensions (i.e. colour, shape and size). A pre-attentive grouping of these feature maps means also that attention could be directed towards groups rather than individual targets as described by Treisman (1982). Pre-attentive processing of features is used to guide the subsequent deployment of attention in a search task. A suggested modification for FIT was that conjunctions with features that were easily discriminated could also be parallel (Treisman & Sato, 1990).

Treisman's FIT has come under some criticism in recent years namely because it fails to explain the nature of the pre-attentive stage sufficiently and also that it does
not account for interactions between serial and parallel processing. For example, search performance is not always simply either parallel (i.e. the target pops out of the scene) or serial but is often somewhere between these extremes. Also, FIT does not account for the integration of features from the same object until a later stage in attentional processing. According to Treisman, the integration of object features occurs at a later stage in processing, as it requires attention in order to integrate information from across the different brain areas.

Assuming the purpose of pre-attentive processes is to guide our attention to relevant object locations in the visual field, we can divide the processes into those which are top-down (knowledge driven) and those which are bottom up (stimulus driven). For example, Bravo & Nakayama (1992) demonstrated how search that employs bottom up processing can maintain a high-level of efficiency in a search task. In a simple experiment, where participants discriminated the presence, the colour or the shape of odd coloured visual target, it was found that reaction times increased with the number of distracters. However, when the colour of the target and distracters was reversed unpredictably, it was found that reaction times decreased as the number of distracters increased. This was an example of where bottom up processing increased task efficiency.

Jeremy Wolfe, in his Guided Search Model (Wolfe et al., 1989; Wolfe, Friedman-Hill, & Bilsky, 1994) described how top down and bottom up processing of a stimulus is used to rank it in terms of attentional priority. This model indicates that attention can be guided towards specific items in a scene that have been ranked with highest priority. If this item is rejected (i.e. a non-target) attention is then directed to the next item.
Although FIT proposed that search performance for targets defined as conjunctions of features would be serial, Nakayama & Silverman (1986) demonstrated that conjunction searches were often more efficient than expected for a serial search. Theeuwes & Kooi (1994) also show evidence that conjunction searches can be as efficient as those incorporating pre-attentive processes. It can however be difficult to ascertain whether a serial or parallel search is taking place. For example, it has been shown in some cases that selective attention is used in order to avoid processing all items in a scene simultaneously (Cave & Bichot, 1999). In this case, a participant focuses on a specific area for periods of time as a strategy to avoid searching the entire display on each trial.

Duncan & Humphreys (1989) proposed that the processing of individual target features did not occur independently. Instead, they suggested that search efficiency could be understood in terms of similarity between targets and distracters and between different types of distracters. The implication here was that larger differences (i.e. less similarity) between targets and distracters in a scene resulted in an easier search while larger differences between different distracters increased search difficulty.

When searching static visual scenes, we would expect observers not to return to an area previously searched or a distracter previously rejected as a non-target. This process, known as Inhibition of return (IOR) was proposed by Posner & Cohen (1984) and is a characteristic of both FIT and Guided Search models and has been demonstrated in a number of search paradigms (Klein, 1988; Muller & von Muhlenen, 2000; Takeda & Yagi, 2000). IOR is generally more particular to small set sizes, and is typically present for the last 4-6 attended items (Snyder & Kingstone, 2000). Dynamic scenes, however, present a scenario where IOR does not seem to
apply (Horowitz & Wolfe, 1998). Observers are unable to track multiple items that have been rejected and recall areas that have previously been searched. Consequently, the same response time patterns are found for searches of varying display sizes for both static and dynamic scenes.

The models of visual search outlined above account for some of the processes that underlie a laboratory based search task. Although such paradigms are useful for revealing the underlying processes, their application in a real world scenario is, however, questionable. Most pertinently, the models are based on unisensory processing and aim to explain search processes within the context of visual attention in particular. A multisensory context is necessary in order to make these models more applicable in real world scenarios.

1.1.2 Dynamic Visual Search

The manner in which we search for either static or dynamic targets in scenes may differ. For example, McLeod, Driver, & Crisp (1988) reported that search for a target defined by a conjunction of movement and form (e.g. an X moving up among a set of Os moving up and stationary Xs) was parallel. Consequently, the RT for detecting the target stimulus was independent of the number of non-target stimuli or distracters. This is in conflict with Treisman & Gelade’s FIT which clearly stated that search for targets defined by a conjunction of features should be serial. However, some of McLeod’s scenes contained dynamic and static objects and in that case perceptual grouping may explain the segregation of these two sets of items causing the search to be parallel. McLeod, Driver, Dienes, & Crisp (1991) proposed a movement filter in the visual system that allowed visual attention to be directed only to moving objects in a display. This would allow objects to be separated from their
backgrounds or from other moving objects. Treisman, (1988) was not opposed to the concept of grouping, and did note that in displays that allowed parallel conjunction searches, non-targets with one feature in common could be grouped together. Whether a perceptual grouping or a movement filter can offer a satisfactory account of search processes in dynamic scenes is uncertain and further research is necessary in this area.

In assessing the processes involved in searching for a target within a continually changing scene, we must consider the role of memory. Although Klein (1988) provided strong evidence for IOR in serial search of static scenes, it is unlikely to be useful in a dynamic display as the scanned and non-scanned items are continually changing their location. Consequently, we would also expect search times to be longer when a scene is constantly changing. However, Horowitz & Wolfe (1998) demonstrated that there was little difference in search performance in a task when the target was either static or relocated over a time interval of 111ms: performance was serial in both cases. The results of this study suggested that the same search processes were operating in both static and dynamic scenes and that memory was not involved.

The claim that visual search has no memory was later challenged by Shore & Klein (2000) after reanalysing the data set from Horowitz and Wolfe. They highlighted that different ratios for 'target present' to 'target absent' trials were used for the static (approx. 1:2) and dynamic (approx. 1:1) scenes and the false alarm rate was greatly increased in the dynamic condition. Shore & Klein (2000) argued that the differences in these conditions represented different search processes. Subsequently, other researchers have argued for (Kristjansson, 2000) and against (Horowitz & Wolfe, 2001) the role of memory in visual search. However, another explanation may exist for the relatively better search performance reported in dynamic scenes by Horowitz & Wolfe (1998). For example, the absence of memory in visual search is a notion
criticised by von Muhlenen, Muller, & Muller (2003). They outline a “sit and wait” strategy whereby observers attend to one or more locations and wait for the target to appear there. Within this strategy, von Mühlenen et al. suggest that observers may be learning that performance can be improved by shifting attention from time to time to other locations in the scene and memory could be used to retain the path that has been scanned. In a later study, that involved the monitoring of eye movements, Geyer, Von Muhlenen, & Muller (2007) compared the time to search for a target in static and dynamic scenes. They found that for target present trials, search times were almost identical across these different scenes. This replicated the findings of Horowitz & Wolfe (1998). However, the results showed differences in the number of fixations and saccade amplitudes between the static and dynamic conditions. A greater number of fixations in the dynamic condition were found suggesting support for the argument that a memory-guided search strategy is adopted when searching for a target in a static scene, but a passive "sit-and-wait" strategy is adopted when searching for a target in the dynamic scene.

1.1.3 Neural basis of visual search

Early physiological evidence (Hubel & Weisel, 1959) suggests that different visual neurons respond to specific visual features such as colour, orientation, spatial frequency, and movement. It is suggested that these features are analysed in the early stages of vision and are involved in mapping a representations of these features to different brain areas. Later findings show that feature search activates specific visual extrastriate regions of the brain as can be seen in imaging studies by Corbetta, Miez, Dobmeyer, Shulman, & Petersen (1991) and by Heinze et al. (Heinze et al., 1994). Also, fMRI studies have shown activation of the posterior parietal cortex during
It is reasonable to expect that brain areas involved in visual attention are also important for search and this is supported by evidence on search by Donner et al. (2000) where conjunction searches that are attention demanding have shown activation in the right parietal cortex, an area known to be involved in selective attention. A study by de Fockert, Rees, Frith, & Lavie (2004) used fMRI to locate brain areas associated with attentional capture using a visual search task. Participants searched for a unique target shape among different shaped distracters. A search paradigm known to produce attentional capture (Theeuwes, 1992) was used. The results provided evidence that the frontal cortex is involved in mediating control of attention-capturing distracters. Frontal regions were also found to be important when searching for salient targets (Wardak, Vanduffel, & Orban, 2010). For example, Wardak et al. identified ventral prefrontal area 45 as being involved in top down control during efficient search.

Further evidence suggests that coding of spatial location and object identification occur in separate brain regions. With regard to the "where" and "what" stages of visual processing, Monosov, Sheinberg, & Thompson (2010) reported that spatial selection is necessary for object identification in a search task. Specifically, Monosoz et al. recorded single neurons from the prefrontal cortex (involved in visual spatial attention) and inferotemporal cortex (involved in object identification) in monkeys and found a temporal correlation between the two processes.

1.2 Multisensory Processes

While studying the senses in isolation has been the norm for many years, considering the senses in this manner yields an incomplete representation of how we perceive our environment. Moreover, when searching for objects, we tend to use any
sensory input that provides useful information. The existing models of visual search do not account for multisensory influences, though it could be argued that they may be adapted to do so. For example, 'visual feature conjunctions' may be replaced by 'sensory feature conjunctions'. Such an adaptation would necessitate revisiting these models from a multisensory perspective. It is important to consider the impact of other sensory information, in this case, audition on how we search for visual objects in our environment.

1.2.1. Neural evidence for multisensory perception

A large body of neurophysiological evidence exists for the integration of multiple sensory inputs in different brain areas. Multisensory neurons have been shown to be significantly enhanced when stimulated by an auditory and visual source originating at approximately the same location. This has been demonstrated for the mid brain (Meredith & Stein, 1986; Wallace & Stein, 1994) and the polysensory cortex (Stein & Wallace, 1996). From their extensive study of multisensory properties of cells in the superior colliculus of cats, Stein and Meredith outlined the principles of multisensory integration at a cellular level. This pioneering research paved the way for future investigation into crossmodal interaction in the brain. To this end, the neural bases of cross-modal spatial attention have been examined extensively using electrophysiological techniques (Kennett, Eimer, Spence, & Driver, 2001; McDonald, Teder-Salejarvi, Heraldez, & Hillyard, 2001; McDonald & Ward, 2000). Generally, this research supports the behavioural evidence that attending to a stimulus in one modality enhances the neural response to a spatially coincident stimuli presented subsequently in a different modality. Similar neural enhancements are continually reported in event-related potential (ERP) studies of voluntary attention (Eimer &
There is an ongoing debate in the area of multisensory research as to whether the orienting of attention to the location of a target incorporates modality-specific brain mechanisms or supramodal ones. Though it is unclear the extent to which integration occurs, it is likely that involuntary shifts of attention across audition and vision do not operate independently (Farah, Wong, Monheit, & Morrow, 1989). In an ERP study conducted by McDonald et al. (2001), the effects of involuntary attention to a brief flash on the processing of subsequent sounds was investigated. Sound occurred on either the same side of fixation (valid trials) or the opposite side of fixation (invalid trials). The results indicated that the location of a spatially non-predictive visual cue modulated neural (and behavioural) responses to a subsequent target in the visual modality. The findings support those of McDonald & Ward (2000) in suggesting that involuntary shifts in attention are controlled by supramodal brain mechanisms rather than modality specific ones. Such findings should be taken into account in the existing models of visual search.

1.2.2 Evidence for multisensory interactions in search

The following section outlines evidence of multisensory integration in visual search.

1.2.2.1 Behavioural evidence for auditory interactions in visual search

Audio-visual interactions for the purpose of recognition and localisation have been well documented in the multisensory literature. Presenting auditory cues during a visual search task has been shown to facilitate search performance. In the following section, I will outline research that has investigated the role of spatially and
temporally informative sound in visual search. The work is divided according to behavioural and neurological evidence.

The role of spatially congruent sound in visual search

The following section will review several studies which have shown that the presence of a sound can improve search performance for a visual target. Specifically, we focus on how the spatial location of the sound can impact on search performance. From a methodological perspective, studies are sometimes criticised when the auditory and visual stimuli which are described as ‘congruent’ are not perfectly spatially aligned. To address this issue, Rudmann & Strybel (1999) compared auditory and visual stimuli that were spatially coincident, or displaced by 6 degrees, to conditions in which the sound was spatially uninformative. While spatially coincident stimuli were the most effective with regard to reducing search latencies, displaced auditory cues also enhanced performance when compared to performance in the uninformative-cue condition.

In a visual search task, the presentation of auditory cues that are co-located with the visual target have been shown to reduce search times. This effect has been demonstrated for peripherally and centrally placed visual targets. For example, Perrott, Saberi, Brown, & Strybel (1990) reported a 175ms to 1,200ms reduction in search latencies when a 10-hz ‘click’ sound was presented at the same location as the visual target. Target position was varied on the horizontal and vertical planes. This study supports the idea that a spatially co-located sound can improve search performance. Perrot et al. found benefits even when the target was within 10 degrees of the observers initial line of gaze (central). In a later study, Perrott, Sadralodabai, Saberi, & Strybel (1991) reported a benefit for finding targets that had a concurrent
co-located sound over those that had no sound. The task involved locating and identifying one of two visual targets. The benefit for the presence of sound was found to be consistent even when a large number of distracters were present.

Aurally aided visual search has also been investigated in three-dimensional space (Bolia, D'Angelo, & McKinley, 1999; McLntire, Havig, Watamaniuk, & Gilkey, 2010) and similar benefits have been found. Both free-field and virtual auditory cues decreased search times for the visual target. In the study by McLntire et al. (2010), the benefits for auditory cues are compared for both static and dynamic targets. Search times were reduced by 25% for dynamic targets and 22% for static targets and the sound was found to be most beneficial when the target appeared at large eccentricities and on the horizontal plane.

Interestingly, spatially uninformative cues have also been shown to reduce search latencies for visual targets. For example, Dufour (1999) reported that when searching for a conjunction of features, the presence of a tone preceding the visual target improved accuracy in detecting the target. The tone could be either proximal or distal with regards to the location of the visual target. In one experiment, participants discriminated the orientation of a target ‘T’ which was flanked with ‘T’ distracters of different orientations and in the other experiment participants discriminated the orientation of a line segment among distracters. The proximal sound benefitted target discrimination in the first experiment but no difference in sound locations was found for the subsequent orientation discrimination experiment. Dufour argued that the reason for the crossmodal benefit was that the “conjunction of features” task required more focused attention. Consistent with this argument, in the subsequent experiment which was a simple feature discrimination task, no difference was found between the proximal and distal auditory cues. These findings may be evidence that at the early
stages of sensory processing, the auditory and visual inputs operate independently. It is known that attentional mechanisms play an important role in cross-modal interactions. Stein & Wallace (1996) demonstrated how a brief auditory stimulus could significantly enhance the perceived intensity of an LED but only at the location attended to by the participants. Similar cuing benefits to those found by Dufour (1999) have also been found by Perrott & Saberi (1990) and Spence & Driver (1997). In the latter study, benefits were found even when the cue was not predictive of what side of the visual field the visual target would occur.

The automaticity of exogenous spatial orienting in cross-modal tasks, such as when an auditory peripheral cue guides our attention to a visual target, has been widely debated in the literature. Originally, it was thought that the presentation of an auditory cue could automatically orient a participant’s attention to a given visual stimulus. For examples, see Mazza, Turatto, Rossi, & Umilta (2007) and Spence & Driver (1997). This cross-modal facilitation is usually seen in reaction times taken to locate a visual target though it can also impact accuracy as described in the Dufour (1999) study earlier.

There have been some studies which have contradicted the notion that auditory cues will always attract visual attention. In one such study by Ward (1994), participants were required to judge the location of auditory and visual targets that were pre-cued by either auditory or visual cues. The findings were that visual precues reduced response times for localising visual and auditory targets whereas auditory pre-cues affected the time to localise auditory targets only. In their study, which found opposing results to those reported by Spence and Driver (1997), the results were described as being due to response priming effects. Later evidence from
Ward, McDonald, & Lin (2000) supported the findings of response asymmetry but only for complex cue environments.

Inconsistent findings in exogenous crossmodal spatial attention have led to an ongoing debate. However, the use of different experimental paradigms may account for the contrasting results. While both implicit spatial discrimination (ISD) and orthogonal cuing (OC) are useful in studying crossmodal cuing effects, they differ fundamentally. In an OC paradigm, participants are typically required to discriminate the elevation of targets above or below fixation on either side of the display. In an ISD paradigm, participants are typically required to respond to targets at some locations but refrain from responding to targets in other locations (McDonald, Ward, & Kiehl, 1999). Studies which have used the ISD paradigm may be prone to response priming effects in that a participant may be primed to respond to targets located in the required locations.

Until now, we have discussed auditory and visual interactions within the context of static environments. It is important to investigate whether these effects occur in dynamic scenes where visual targets are continuously moving. The use of virtual auditory stimuli has allowed researchers to study the effect of directional congruency (termed Crossmodal Dynamic Capture or CDC) on target search. Congruency effects similar to those found in static scenes are also reported for the dynamic scenes. However, this effect is eliminated when the auditory and visual signals are temporally desynchronized, even by as little as 500ms (Soto-Faraco, Spence, & Kingstone, 2004). Audio-visual integration of motion signals were investigated in an adaption paradigm developed by Kitagawa & Ichihara (2002). In this study, participants observed a visual target moving either toward or away from them. After-effects occurred which affected the perceived direction of the visual and auditory stimuli.
According to Kitagawa and Ichihara, these after-effects were dependent on the directional congruency of the dynamic audio-visual stimuli. These findings are supported by a study conducted by Valjamae & Soto-Faraco (2008), in which they incorporated time-sampled object motion. In this study, directional congruency between modalities resulted in auditory after-effects which caused illusory visual flashes similar to those found in the well documented ‘auditory-flash’ illusion (Shams, Kamitani, & Shimojo, 2002). Historically, not all of the evidence has suggested a benefit for audio-visual congruency. For example, early perceptual studies by Zapparoli & Reatto, (1969) examined the effect of presenting observers with combinations of directionally incongruent visual and auditory apparent motion streams. Despite this manipulation, some observers reported experiencing the auditory and visual modalities as a unified trajectory.

In summary, these findings point towards the presence of an auditory stimulus being beneficial to performance in a static or dynamic visual search task. However, while the majority of research indicates that a spatially co-located or directionally congruent auditory cue improves task performance, this is not always the case. Methodological differences may account for the contrasting evidence.

The role of temporally congruent sounds in visual search

Here we review research where the presence of sound has been shown to improve visual search performance. Specifically, we focus on non-spatial auditory signals that have been shown to affect task performance. For the purposes of this review, and for subsequent experimental chapters, a temporally congruent sound is one which shares the same temporal profile (e.g. onset and offset) as the visual target but does not contain spatial information.
Although many studies have focussed on multisensory interactions for spatially coincident stimuli, there are a number of studies where spatial information was less important for target localisation. Olivers & Van der Burg (2008) demonstrated how the presence of a synchronous sound could cause a visual stimulus to be attended to where it might otherwise have been missed due to the attentional blink. Another well-known example of crossmodal interaction of audition and vision is the auditory-flash illusion (Shams et al., 2002). This illusion occurs when a single visual flash is accompanied by two auditory beeps presented in rapid succession which results in the illusory perception of two visual flashes. For a review of crossmodal influences on visual perception, see Shams & Kim (2010). In the auditory-flash illusion, the sound carried no spatial information. Instead, the perception of a visual stimulus was mediated by the temporal congruency of the auditory cue. In other words, as long as the stimuli were synchronised in time, this was enough for an interaction between vision and audition to occur. Other studies, where either the auditory or visual stimulus were presented shortly before or after each other in time have found similar interaction effects even when the auditory and visual stimuli are presented from separate locations (Arnell & Duncan, 2002; Arnell & Jenkins, 2004).

The concept of temporal ventriloquism was first documented by Morein-Zamir, Soto-Faraco, & Kingstone (2003). The study investigated whether non-meaningful sounds could influence visual perception in a temporal order judgement task. The task for the participant was to judge which of two lights appeared first. The results showed that sound improved performance when presented before the first light and after the second as compared to when sound and light was synchronised. Sound presented between the lights resulted in a deterioration in performance. The authors argued that the perceived order of the visual events shifted towards that of the sound. As the
temporal resolution for audition is thought to be greater than that for vision, this seemed a plausible explanation. Furthermore, Vroomen & Keetels (2006) reported that temporal ventriloquism was not affected by manipulating the spatial location of the auditory signal.

Uninformative auditory cuing has been shown to increase accuracy on a visual target detection task. One such study, reported by Vroomen & de Gelder (2000), found that a high tone embedded in a sequence of low tones could improve the detection of a synchronously presented visual target. In recent years, Van der Burg, Olivers, Bronkhorst, & Theeuews (2008) demonstrated how non-meaningful, non-spatial auditory signals could facilitate visual search performance when the sounds were temporally synchronised with a change in the target stimulus. This phenomenon has become known as the “Pip and Pop” effect. Specifically, Van der Burg et al. asked participants to search for either a visual target which was a vertical or horizontal line segment presented among a number of distracter line segments oriented at ±22.5 degrees. The target and each of the distracter stimuli changed colour regularly from green to red. The auditory cue onset was synchronised with the colour change of the target or there was no sound presented. As there was no spatial information in the auditory cue, any observed benefit on performance could not be attributed to the cross-modal facilitatory effects found in previous spatial cuing studies such as that by Spence and Driver (1997). Van der Burg reported a reduction in search latencies of more than 1000ms when the sound was present relative to when it was not. They argued that the auditory and visual target were being integrated which lead to an increase in the saliency of the visual target causing it to ‘pop out’ of the visual display. Ngo & Spence (2010) replicated the Van der Burg et al. (2008) study but also introduced a spatial component. Their study investigated the benefit of having
auditory cues that are spatially informative and temporally synchronous within a dynamic visual scene. It was found that for temporally synchronous cues, those which were spatially informative resulted in greater performance improvements than those which were temporally synchronous but spatially uninformative.

The literature on auditory alerting effects (e.g. Bertelson, 1967; Posner and Boies, 1971) suggests that an auditory cue can enhance the response to a visual target when a cue is presented between 100 and 300ms prior to the target. Performance benefits due to increased arousal appear to be greatest when the auditory stimulus precedes the visual event. However, as the processing of auditory stimuli is generally faster than the visual counterpart (Wallace, Wilkinson, & Stein, 1996), an asymmetry in performance sometimes occurs when the auditory stimulus is presented after the visual event.

1.2.2.2 Neural evidence for multisensory interactions in search

Non-spatial search is sometimes used to investigate brain areas responsible for temporal processing during visual search. Non-spatial search occurs when stimuli are presented sequentially at the same location. For conjunction search tasks, Shafritz, Gore, & Marois (2002) found selective right parietal activation when objects were presented simultaneously at different spatial locations but this effect was not present when stimuli were presented simultaneously at the same location. Other studies that have used non-spatial displays indicate a greater role for the parietal cortex. For example, bilateral BA7 is activated for non-spatial shifts in attention between colour and form (Le, Pardo, & Hu, 1998) and in non-spatial conjunction search (Wojciulik & Kanwisher, 1999). To our knowledge, neuropsychological research that localises brain areas involved in multisensory non-spatial search has yet to be carried out.
Event-related potential (ERP) studies have found the superior parietal lobule (SPL) to be an area where multisensory spatial processing occurs. Moran, Molholm, Reilley & Foxe (2008) have demonstrated the effect if one brain region over another during unisensory auditory, unisensory visual and multisensory audiovisual stimulation. Specifically, this study found differences in the summed unisensory and multisensory ERPs that were mediated by forward and backward connections to SPL. A negative gain in all connections to SPL during the period of multisensory integration. Gamma-band responses (GBRs), which are hypothesised to reflect neuronal activity relating to object representation, have been used to demonstrate an effect of semantic congruency between auditory and visual information. On a recognition task, it was found that information in the unattended modality could affect behaviour. It was found that irrelevant visual information affected auditory recognition more so than irrelevant auditory information affected visual recognition. This study showed the integration of semantically congruent information across modalities is associated with enhanced synchronized oscillations at the gamma band. These oscillations appear to relate to low level unimodal processing and also higher level multisensory object representation.

1.3 Evidence for top-down multisensory interactions in search

The following section outlines evidence for top-down multisensory interactions in visual search.

1.3.1 The role of crossmodal semantic information in visual search

Studies have reported strong activation in the superior temporal sulcus (STS) when images of objects and their semantically related sounds have been presented
simultaneously as opposed to separately (Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Beauchamp, Lee, Argall, & Martin, 2004). This evidence suggests that although object features differ across modalities, multisensory neurons exist that respond to audio-visual representations of objects, by integrating them into a coherent percept. For example, in a study by Molholm, Ritter, Javitt, & Foxe (2004), participants responded to an image of an animal, a sound of an animal and an audiovisual representation. Participants responded fastest when the item was presented bi-modally. ERP data from the same study showed that the identification of objects was facilitated by scene consistency. Items were more likely to be identified quickly when the auditory and visual information were semantically related. In a recent study by Iordanescu, Guzman-Martinez, Grabowecky, & Suzuki (2008), participants had to indicate the location of a target in one of four quadrants of a visual display. Sounds were either characteristic of the visual target (e.g. a barking dog) or one of the three distracters. The results showed a crossmodal enhancement in searching for the visual target when it was accompanied by a characteristic sound. Interestingly, Iordanescu et al., (2008) replaced the images of objects with text versions of their names and found that the sound then had no effect. The results of these studies indicate that semantically relevant auditory signals that carry no spatial information can reduce response times in an object identification (Molholm et al., 2004) and visual search (Iordanescu et al., 2008) task. A recent follow up study, Iordanescu, Grabowecky, Franconeri, Theeuwes, & Suzuki (Iordanescu, Grabowecky, Franconeri, Theeuwes, & Suzuki, 2010) demonstrated the effect of target-consistent sounds on saccades within a 215-220 ms latency. Using the same search paradigm as their 2008 study, Iordanescu et al (2010) found that not only were saccade latencies reduced, but also that the initial saccade was guided toward the visual target when a semantically
relevant sound was presented relative to a distracter sound or no sound. This evidence further supports the notion that non-spatial but semantically relevant sounds can facilitate visual search.

1.4 Scene Perception

Visual perception of an environment as viewed by an observer incorporates information about individual objects as well their location in space. Marr (1982) described how low-level scene perception allows us to generate surface representations and edges. At an intermediate level, we can extract shape information and spatial relations (Bar & Ullman, 1996) and at a higher level we can map object identity and meaning to the visual representations. Studies of high-level scene perception focus mainly on eye movements, scene context and object identification.

Early studies of eye movements, such as those by Yarbus (1967), suggested that the observer’s eye tended to fixate on information in a scene that was deemed useful to the perceptual task. Mackworth & Morandi (1967) also found that observers fixated on regions that were deemed to be task informative. Loftus & Mackworth (1978) also demonstrated that semantic information in a scene was associated with greater fixation density. Interestingly, it was found that semantically inconsistent objects were fixated on before semantically informative objects though observers were more likely to fixate on semantically informative objects after the first saccade in the scene. This idea has been challenged by De Graef, Christiaens, & d'Ydewalle (1990) who reported that viewers were no more likely to fixate on a semantically informative target than an object which was uninformative. As Hollingworth & Henderson (1998) point out, factors such as image size, viewing time per scene and image type vary in these kinds of studies making it difficult to determine the nature of eye movements in
perception of high-level scenes. Scene context can also influence an observer's ability
to detect and identify an object in a scene. Theories of object recognition, such as
those by Biederman (1987) and Bulthoff, Edelman, & Tarr (1995) suggested that
viewers identified objects based on bottom up visual information alone whereas other
researchers have proposed that object identification can be influenced by the meaning
of a scene (Friedman, 1979; Kosslyn, 1994).

In a review of high-level scene perception, Henderson & Hollingworth (1999)
divide the models of object recognition into three groups. The first proposes that
expectations derived from prior knowledge of a scene interacts with our perceptual
analysis of objects. The second group proposes that this interaction occurs when
perceptual descriptions are matched with memory representations and the final group
 proposes that object identification and scene knowledge are isolated. The majority of
studies have found benefits for object identification when scene information is
consistent (Boyce & Pollatsek, 1992; Rayner & Pollatsek, 1992) rather than that
which is contextually incongruous.

Functional magnetic resonance imaging (fMRI) studies have found that scene
categorisation occurs in the parahippocampal place area (PPA), the retrosplenial
cortex (RSC), the lateral occipital complex (LOC), and primary visual cortex (V1)
(Epstein & Kanwisher, 1998; Walther, Caddigan, Fei-Fei, & Beck, 2009). In an effort
to better understand the neural mechanisms of scene categorisation, Peelen, Fei-Fei,
& Kastner, (2009) used an object categorisation task where observers had to judge the
presence of people or cars in a series of rapidly presented scenes. The results suggest
that a category-specific biasing mechanism in object-selective areas of the cortex
biases information processing in favour of objects belonging to that target object
category.
1.5 Limitations of existing research

This section outlines briefly the limitations of current research regarding the role of audition in visuo-spatial tasks.

There is conflicting evidence regarding the automaticity of directed attention in visual search tasks where an auditory cue is present. It is unclear at what stage the integration of the auditory and visual cue occur, whether it is at the pre-attentive or later attentional stages. Also, the importance of informative auditory information is uncertain. As we have seen, a number of studies demonstrate how non-spatial auditory signals can guide attention to a visual target. Auditory spatial information has been shown to enhance this facilitation further in both static and dynamic environments. However, the scenes that have been used to test these multisensory influences in visual search have either incorporated low level beeps and flashes or oriented line segments. Use of such simple stimuli sometimes neglects the existing body of literature on visual object and surface recognition. Presently, there is little evidence demonstrating auditory facilitation of visual search in natural visual scenes or complex dynamic environments.

The existing main models of visual search, such as Feature Integration Theory by Anne Treisman or Guided Search by Jeremy Wolfe, have offered detailed descriptions of the processes involved in visual search tasks. However, while these models have helped in our understanding of visual search, they have failed to account for multisensory influences that occur when we search for items in real world scenarios. For this reason, the ecological validity of these models is questionable and there is perhaps an opening for a model that incorporates inputs from other senses apart from vision when trying to localise and identify a target in a complex scene.
1.6 Outline of Thesis

The following section outlines the principal aims of the experiments described in the following five chapters in this thesis.

1.6.1 Experimental investigations on search in complex static scenes

Chapters 2 and 3 describe an investigation into the role of spatially co-located sound on searching for a visual target in static scenes. Specifically, Chapter 2 describes a study that investigated the role of characteristic sounds in static scenes. Here we provide an insight into the role of characteristic and spatially congruent sounds in visual search. Iordanescu et al. (2008) previously reported a facilitation effect for characteristic sounds in the form of reduced search latencies. However, in this study we looked at whether this facilitation effect occurs in complex visual scenes and when the sounds were spatially informative. Chapter 3 describes how audition can facilitate the detection of target stimuli of multiple sizes and locations in static scenes. The aim of the experiments in Chapters 2 and 3 is to understand how audition and vision interact in the context of static visual scenes.

1.6.2 Experimental investigations on search in complex dynamic scenes

Chapters 3, 4 and 5 describe experiments that were designed to investigate the role of audition in dynamic visual scenes. Previous studies have argued that non-spatial auditory signals can reduce search latencies in dynamic scenes (e.g. Van der Burg et al., 2008) and others have shown a benefit for spatially informative sound using the same scene (e.g. Ngo and Spence, 2010). The visual stimuli used in reported experiments by Van der Burg et al. and Ngo & Spence were line segments which alternated in colour. In the studies, described in Chapters 3 to 5 we used more high-level dynamic scenes which incorporated both target and distracter motion. We
investigated the role of a temporally synchronous sound and dynamically congruent sound on visual search in these scenes and how audition impacts on performance in visuo-spatial tasks.
2.0 The effects of characteristic and spatially congruent sounds on visual search in static visual scenes

Abstract

Little is known about how a semantically relevant sound can influence attentional deployment in a visual search task. Here we investigated whether characteristic and/or spatially congruent sounds affected visual search performance in a complex visual scene. The task for the participant was to detect the presence or absence of a visual target in a static visual scene. The targets consisted of images of familiar objects and animals and were presented accompanied by a sound. The sound could be either spatially congruent or incongruent to the hemifield in which the target was located. Furthermore, for each sound location the sound could be semantically characteristic or semantically uncharacteristic of the visual target. A characteristic sound was one which was naturally associated with the visual target. In Experiment 1, targets were presented across trials in separate blocks. A significant benefit on search performance was found for a characteristic sound but spatial congruency did not affect performance. In the second experiment, targets were randomly presented across trials within a block, thus increasing task difficulty. Although the benefit for characteristic sounds was not found, when the sound was spatial incongruent this resulted in a further cost in performance when the sound was uncharacteristic. Our findings suggest that while characteristic and spatially congruent sound can benefit visual search performance, this benefit is dependent on task difficulty (one target per block or random targets). These results have important implications for our understanding of multisensory influences on target detection in static visual scenes.
2.1 Introduction

Searching for objects in natural visual scenes can be difficult, particularly if the scene is complex and there is little information available to guide our attention. In the absence of clear bottom up signals, top down information can play a greater role in these kinds of tasks. For example, if a person sees a duck and hears a concurrent sound, they may or may not think the two related. However if they see a duck and hear a corresponding “quack” sound, prior knowledge will inform that these inputs are likely to come from the same source. What if a duck is seen but a dog barking is heard at the same time? This semantic inconsistency (or violation of scene context) may be detrimental to object identification and/or detection. Previous research by Biederman, Mezzanotte, & Rabinowitz (1982) provided evidence that consistent or inconsistent scene context can affect object detection. The study also proposed that the semantic relations among the entities in a scene is not deferred until the completion of spatial and depth processing and object identification. The implication here is that semantic inconsistencies are processed at an early stage of scene perception. As a consequence, we would expect both auditory and visual spatial and semantic consistency to play a significant role in early scene processing.

In the following experiments we investigated the semantic and spatial relationship between auditory information and a visual target by measuring speeded responses to scenes which were exposed for a short period (250ms). In a replication of the earlier Biederman et al. (1982) study, Hollingworth & Henderson, (A. Hollingworth & J. M. Henderson, 1998) found conflicting evidence indicating that object perception was in fact not facilitated by consistent scene context. The principle difference in the later study was that a response bias, which may have been present in
the Biederman et al. study, had been adequately controlled for. However, this object search paradigms adopted by both of these studies was based on the use of line drawings and did not account for multisensory influences on scene context. The present study manipulated scene consistency using cross-modal semantic and spatial congruency and aimed to further investigate the impact of these factors on early processing in the detection of visual targets.

Often, the features that we use to identify objects in scenes differ across the senses but nonetheless, our brains integrate them into a coherent percept. Many studies have been conducted in an attempt to understand the neural basis of this integration for some time. It is now thought that the posterior section of the superior temporal sulcus and the middle temporal gyrus (pSTS/MTG) are areas where the integration of auditory and visual stimuli occurs. Moreover, performance benefits have been found when semantically relevant sounds are simultaneously presented with images of objects.

For example, on the basis of an fMRI study, Beauchamp, Lee, et al. (2004) reported strong activation in the superior temporal sulcus related for multisensory stimuli relative to unisensory auditory and visual stimuli. However, relatively few behavioural studies have been conducted to investigate of combined spatial location and characteristic sounds on visual object identification.

In a behavioural study that also measured the visual evoked potential associated with object processing (specifically the N1 component), Molholm et al. (2004) presented images and sounds of animals in an object recognition task. The stimuli were presented in the auditory modality and the visual modality separately as well as bi-modally. The results showed significantly faster and more accurate responses to
identifying objects when the image and sound were matched than when the target was represented in one modality only. Furthermore, the study found that auditory inputs modulated the processing of visual objects in cortical regions of the lateral occipital cortices affirming that there are multisensory effects occurring in the early stages of visual object-recognition.

A recent study by Iordanescu et al. (2008) provided evidence that characteristic sounds can facilitate the localization of a visual target even when the sound carried no spatial information. In their study, participants were presented with an array of four visual objects and were required to indicate the location of the target within the array. Target objects, which were presented with a consistent sound, were found more quickly than when the sound was consistent with a distracter object or no sound was presented. The study showed object-specific auditory facilitation of visual target location occurred only for target objects that were associated with goal-directed, top-down feedback. In other words, if the goal was to search for a particular visual target and an auditory stimulus is heard, visual salience will increase if the auditory stimulus corresponds to the target. Increased visual salience would not occur if the sound is associated with a distracter object or is unrelated. If a characteristic sound facilitates visual search, as found by Iordanescu and colleagues (2008), we would expect further benefits when the auditory and visual target occur at the same location in space, at least within the same visual hemifield.

2.2 Experiment 1

The following experiment investigated the role of semantic or characteristic sound in the localisation of a visual target embedded in a complex visual scenes. We looked specifically at whether the semantic relevance and spatial co-location of an
auditory stimulus could facilitate target detection in a series of randomly presented natural scenes. Six different images of targets were used. These included familiar objects or animals. The target in this experiment could occur in one of two locations, either left or right of fixation. The experiment contained of six blocks with one target presented across all trials in a block. Our hypothesis was that when the auditory stimulus was both characteristic of the target object and was also spatially co-located with the visual target then performance in locating a visual target would be improved relative to a spatially incongruent or uncharacteristic sound. Performance was measured in terms of response speed and accuracy in detecting the presence or absence of a visual target in each scene.

2.2.1 Method

Participants

Twenty-eight participants (8 male and 20 female with a mean age of 24 years) from the School of Psychology, Trinity College Dublin participated in this experiment for research credits. Their ages ranged from 18 to 26 years. All reported normal or corrected to normal vision and none reported any auditory impairment. Prior approval was obtained from the School of Psychology Ethics Committee. Following a briefing on the experimental protocol, participants provided written consent to take part in the study.

Stimuli

The stimuli used in this experiment incorporated a set of 8 images of natural visual scenes. The images were randomly selected from an online photography
database (http://www.sxc.hu). The images were a combination of six indoor and two outdoor scenes (see Figure 2.1a and 2.1b). The criteria for image selection were as follows: a) images could not contain an object which was similar to the target b) images needed to be non-symmetrical to avoid a symmetry based search strategy and c) images with large areas of solid colour were avoided as target salience would be increased when located in these areas. A total of six visual targets were used: kettle, toaster, clock, duck, cat and dog and are depicted in Figure 2. These images of the targets were sourced from an online graphics database (http://www.sxc.hu), converted to grayscale and positioned in the scenes. Each scene fitted within the dimensions of a 1024 x 768 pixel frame. The target images were positioned either left of right of fixation by a maximum visual angle subtended of 11°. The criteria for target selection was 3 familiar objects and 3 familiar animals and that each target would have a corresponding sound that was easily recognisable and unique to the class of object. It was also preferable that each sound was not associated with a particularly gross motion of the visual target (such as the sound of a moving train) or a change in form of a visual target (such as a balloon bursting).

Figure 2.0 This image depicts the six visual targets used in Experiment 1. These targets were (from left to right) a kettle, toaster, clock, duck, cat and dog.
Each visual target had the maximum dimensions of $4^\circ$ (W) x $4^\circ$ (H) of visual angle. As well as the visual counterpart, we used corresponding auditory cues sourced from *Al free Sound Effects* (http://www.alfreesoundeffects.com) which were sounds that corresponded to each of the targets (e.g. dog bark, kettle whistle, etc.). Each sound clip was edited using Adobe Audition, to be a maximum of 250ms in duration and presented in standard wav format. The sound pressure level was 53dbA relative to the participant’s head. Each target had one corresponding sound. Distracter sounds were sounds that corresponded to the other target objects (but which were not presented in the scene).

![Figure 2.1](image)

*Figure 2.1. An example of visual target a) a clock in an outdoor scene and b) a dog in an indoor scene. Targets were embedded into the visual scenes.*

Each scene was presented with an image of the target (Target Present trials) or without the visual target (Target Absent trials) embedded in the scene. For the target present trials, the image of the target was positioned at one of two locations (left or right of fixation). No attempt was made to place the target object into a meaningful
position on the scene (i.e. on a table-top or on the wall). Therefore, target images were often incongruously located in the scene, as shown in Figure 2.1 a and b.

**Apparatus**

The experiment was presented on a colour CRT monitor (1,024 x 768 pixels) at 75 Hz, and was programmed using a Presentation™ (Neurobehavioral Systems®) script which ran on 3GHZ PC with 1GB RAM. Stereo speakers (Dell HK206) were used to transmit the audio signal and were positioned on stands at the vertical midpoint of either side of the monitor. The speakers were located 18cm from centre of the monitor. A standard PC keyboard was used to record each participant’s responses.

**Design**

This experiment was based on a fully factorial, 2x2 within subjects design with spatial congruency of the sound (congruent or incongruent) and semantic relevance of the sound to the visual target (characteristic or uncharacteristic) as factors. In the semantic sound condition, a characteristic sound was that which was normally associated with the visual target whereas an uncharacteristic sound was any of the sounds associated with the other five targets. The spatial location of the auditory cue was congruent when it was presented in the same visual hemifield as the visual target (see Figure 2.2). Each target was presented 64 times which was a figure derived from the product of 2 (congruency) x 2 (characteristic) x 8 (scenes) x 2 (repetitions). As there were six targets, this led to a total of 384 target present trials in the experiment. There were a total of 96 target absent trials. The proportion of trials was divided across target present (75%) and absent (25%). There were six blocks to
the experiment, and within each block only one target object was presented. Target present and target absent trials were presented in a randomised order within each block to reduce order effects. The dependent variables were response times and accuracy.

**CONGRUENT**

**INCONGRUENT**

*Figure 2.2* This image depicts an example of a spatially congruent (left) and a spatially incongruent (right) scene. Spatial congruency is depicted by sound and visual target occurring in the same hemifield. Spatial incongruency is depicted by the sound and visual target occurring in opposite hemifields.

**Procedure**

Participants began the experiment by completing a series of 12 practice trials to ensure that they were familiar with the visual targets and the task. They were given feedback on their accuracy performance and had the opportunity to pose any questions. Following the practice session, participants were then presented with the test phase which consisted of 6 separate experimental blocks of trials. At the beginning of each block, the participant was shown an image of the visual target for that block.

In the experiment, a trial began with a central fixation cross which was presented for 500ms. An image of a scene was then presented for 250ms and the task
for the participant was to indicate, as fast and as accurately as possible, whether the visual target was present or absent in the scene by pressing the corresponding key on the computer keyboard (‘/’ = present, ‘z’ = absent). At the same time as the scene was presented, an auditory stimulus was also presented for the duration of the scene. The location and type of auditory stimulus was determined by the design of the experiment. The auditory stimulus was present in all trials (target present and absent) throughout the experiment. Subsequent trials occurred only when the participant had input a response. Both accuracy and response times were recorded. Participants could take a self-timed break between blocks. The experiment took approximately 20 minutes for each participant to complete.

2.2.2 Results

The mean accuracy and reaction times were calculated for each of the spatial congruency and semantic sound conditions. Performance in all conditions was greater than chance (50%). For the target present trials, the mean percentage of correct responses was 95.92% (std. dev. was 5.02%) and the mean RT was 481 ms (std. dev. was 64 ms). For the target absent trials the mean percentage of correct responses was 87.08% (std. dev. =10.61%) and the mean RT was 474 ms (std. dev. =95 ms). Only the data from the target present trials were subjected to further analyses as the data from the target absent trials were of no theoretical interest. Separate 2x2 repeated measured ANOVAs were performed on the correct responses and RT data with Spatial Congruency (congruent or incongruent) and Semantic Sound condition (Characteristic or Uncharacteristic) as factors.
Chapter 2

Response Times

The mean response times to each of the characteristic and spatial congruent sounds is depicted in Figure 2.3. For the response time data, we found a main effect of semantic sound \( F(1, 27) = 43.969, p < 0.0001 \) with faster RTs to the visual targets when they were presented with characteristic than an uncharacteristic sound. The effect of spatial congruency was not significant \( F(1, 27) = 2.065, p = 0.1621 \). The interaction between semantic sound and spatial congruency on RTs was not significant.

![Figure 2.3](image)

**Figure 2.3** Plot showing the mean response times for the characteristic and uncharacteristic sound conditions for each of the spatial congruency positions of the sound relative to the visual target. Response to the target absent condition is also shown.

Correct Responses

The mean correct responses made to each of the characteristic and spatially congruent sounds is depicted in Figure 2.4. In terms of accuracy, we found a main
The effect of semantic sound \(F(1, 27) = 16.26, p < 0.001\) with better performance when the sound was characteristic of the visual target than when it was uncharacteristic. The effect of spatial congruency was not significant \(F(1, 27) = 0.08, p = 0.777\). There was no significant interaction on the percentage of correct responses across the spatial congruency and semantic sound conditions.

![Figure 2.4](image)

*Figure 2.4* Plot showing the mean percentage correct responses across the characteristic and uncharacteristic sound conditions for each of the spatial congruency positions of the sound relative to the visual target. Response to the target absent condition is also shown.

### 2.2.3 Discussion

In this experiment it was found that a sound which was semantically characteristic of the visual target object improved performance on a visual search task. However, whether the sound occurred within the same hemifield as the target or
not seemed to have no effect on performance. As participants searched for one target per block in this experiment, it would appear they are ignoring the spatial location of the sound and using only the semantic information. These results suggest that both the auditory and visual stimuli do not need to be co-located in order for task performance to be facilitated.

The results support those of Iordanescu et al. (2008) in that the presence of characteristic sounds have reduced response times in a visual search task significantly. The present study differs from that of Iordanescu et al. as the latter did not have a spatially informative sound condition. These findings indicate that a semantically informative sound can facilitate a visual search task regardless of the hemifield in which the sound occurs (same as target or opposite). The task also differed. In the present study, a target-present/absent task was used whereas Iordanescu et al. study used a target localisation task (target was always present), where participants were required to indicate the screen quadrant in which the target was located. The duration for which the target was presented was significantly longer in the Iordanescu et al. study (670 ms) than the present study (250 ms). It is possible that the exposure time in the present study may not be enough for participants to process the semantic meaning of the sound and attend to its spatial location. It would appear that participants are relying on the former. Recent evidence from Iordanescu et al. (2010) shows that object-specific auditory signals increase the detectability of a seldomly occurring target. This increased visual salience has been shown previously for synchronously presented audiovisual stimuli where the auditory stimulus holds no spatial information about the location of the visual target (Van der Burg et al., 2008). These studies combined, suggest that an audio-visual target is processed more
efficiently when the auditory stimulus is semantically relevant and temporally synchronous. It may be the case that participants adopted a strategy which prioritised semantic information over spatial information. Such a strategy would mean that the spatial location of the auditory stimulus was being successfully ignored while only attending to the semantic content and relying on this to help determine the presence or absence of the target.

This experiment found a beneficial effect of characteristic sound on both Accuracy and RT performance but did not find an effect of spatial congruency of sound. We wished to test the effect of task difficulty on search performance and this prompted a follow up experiment.

2.3 Experiment 2

The purpose of this follow up study was to further investigate the type of search (serial or parallel) that participants may be using in this task. We repeated Experiment 1 with one important design change: any of the six targets could appear across trials in each block. This increased the task difficulty as instead of looking for one target per block, they were now looking for all six. This also reduced the likelihood of the same sound being semantically useful in each block. For example, if the participant was relying on semantic auditory information (rather than spatial congruency as the results of Experiment 1 suggest), it may be relatively easy to disregard non-useful sounds which are unrelated to the target in each block. To ensure that this was not happening, it was necessary to randomise the presentation of the targets within each block. Thus, each sound had the same potential usefulness in any given trial. According to Huang & Pashler (2005), for speeded search tasks, task difficulty can negatively affect search efficiency but not necessarily the limits of
attentional capacity. For this reason, search times were expected to be longer in Experiment 2 but attention to the auditory and visual stimulus not to be impaired.

In a serial search, search latencies usually increase proportionally to the number of distracters (Treisman & Gelade, 1980). While both experiments had the same number of unrelated distracter sounds, it could be the case that disregarding the irrelevant sounds was relatively easy in Experiment 1 because of the experimental protocol (i.e. one target was presented per block). This was not the case in the present experiment, as targets could appear randomly which should result in impoverished performance. Alternatively, the semantically relevant sound may have caused the visual target to “pop out” of the scene in Experiment 1, in which case, we should not see a major difference in task performance for Experiment 2.

2.3.1 Method

Participants

Ten participants (3 male and 7 female with a mean age of 25 years) from the School of Psychology, Trinity College Dublin participated in this experiment for research credits. None of these participants had taken part in the previous study. Their ages ranged from 18 to 26 years. All reported normal or corrected to normal vision and none reported any auditory impairment. Prior approval was obtained from the School of Psychology Ethics Committee. Following a briefing on the experimental protocol, participants provided written consent to take part in the study.

Stimuli and apparatus
The stimuli (i.e. the images of visual scenes and visual targets and the associated sounds) were the same as those used in Experiment 1 in this chapter.

**Design**

The design was the same as that used in Experiment 1 with the exception that instead of presenting one target per block, in Experiment 2, any of the six targets could appear across all trials within a block. An illustration of the differences in the design of these experiments is provided in Figure 2.5. Our hypothesis was that, for target detection where a spatially congruent and semantically relevant sound was synchronously presented, performance would improve. We considered that when the task difficulty increased relative to Experiment 1, performance for the characteristic sound condition would deteriorate.

---

*Figure 2.5* This image illustrates how design differences between Experiments 1 and 2. In Experiment 1, participants searched for one target per block. In experiment 2, the visual targets were presented across trials in a random order. Only one target could appear in each trial (or none in the target absent conditions).
Procedure

As in Experiment 1, participants began the experiment by completing a series of 12 practice trials to ensure that they were familiar with the visual targets and the task. Participants were then shown images of all target objects at the beginning of the experiment. They were not exposed to the individual images of the targets again (apart from in the experimental trials). They were given feedback on their performance and had opportunity to pose any questions. Following the practice session, participants were then presented with the test phase. The structure of a trial was the same as that described in Experiment 1. As before, participants were urged to complete the task as quickly and as accurately as possible. Response times (RTs) and errors were recorded. Trials were divided into six blocks and participants could take a self-timed break between blocks. The experiment took each participant approximately 20 minutes to complete.
2.3.2 Results

The mean accuracy and reaction times were calculated for the spatial congruency and semantic sound conditions. Accuracy performance in all conditions was greater than chance (50%). For the target present trials, the mean correct response was 89.31% (std. dev. =4.83%) and the mean RT was 601 ms (std. dev =115 ms). For the target absent trials the mean correct response was 88.12% (std. dev. =11.6%) and the mean RT was 755 ms (std. dev =110 ms). Only the target present trials were subjected to further analyses. Separate 2x2 repeated measured ANOVAs were performed on the correct responses and RT data with Spatial Congruency (congruent or incongruent) and Semantic Sound (characteristic or uncharacteristic) as factors.

Response times

For the response time data, we found a main effect of spatial congruency \( [F(1, 9) = 7.9125, p < 0.05] \) but not semantic sound \( [F(1, 9) = 1.8516, p = 0.2066] \). We did, however, find a significant interaction between these factors \( [F(1, 9) = 9.3276, p < 0.05] \) which is depicted in Figure 2.6. A Tukey post-hoc analysis revealed that when a visual target was presented with a sound that was uncharacteristic of the target, search times were significantly slower if that sound was spatially incongruent than if it were spatially congruent with the location of the target \( (p<0.05) \). There was no benefit on response times when the sound was both characteristic and spatially congruent, relative to whether it was characteristic but not spatially congruent.
Correct Responses

In terms of accuracy, we found no effect of semantic sound \([F (1, 9) = 1.486, p = 0.2537]\) or spatial congruency \([F (1, 9) = 2.511, p = 0.1474]\). The interaction between spatial congruency and semantic sound on correct responses was not significant.

2.3.3 Discussion

In this experiment it was found that spatially congruent sound affected performance but semantically relevant sound did not. On further analysis of the main effect of congruency, there was a cost on the time it took to search for a visual target when it was presented with a sound which was both uncharacteristic of the target and positioned such that it was spatially incongruent with the hemifield in which the target appeared. This cost in response time performance suggests that an uncharacteristic sound is more distracting when not spatially co-located with the visual target.
Reaction times were slower (120 ms) relative to Experiment 1 and the overall mean accuracy decreased by 6.6% in Experiment 2. The differences in performance may be accounted for by the design differences between the two experiments. In the present experiment, participants searched for one of six targets in each trial as opposed to one specific target per block in Experiment 1. Consequently, in this experiment all sounds had the same probability of being semantically useful compared to Experiment 1 in that searching for one target in a given block, only one of the six sounds is semantically useful. This may be the reason that we are not finding a main effect of characteristic sounds in Experiment 2. It appears that the observer is ignoring the semantic information in the auditory signal and seems to be relying on the spatial location of the sound to guide attention to the target. Perhaps it is because the chances of the sound being semantically useful were now 1 in 36 as opposed to 1 in 6 in Experiment 1. However, in order to accurately determine if the presence of sound had an effect on performance, it would need to be compared to a no-sound condition. The effect of spatial congruency seems to be driven by bottom up processing as no prior knowledge of the scene could allow the observer to predict the hemifield in which the sound would occur.

No effect of accuracy was found in Experiment 2. Neither the location of the target or the semantic relevance of the sound has affected accuracy. If we consider that the semantic information in the sound is being ignored in the present experiment, then we would not expect this to have an affect on accuracy. The longer reaction times suggest that participants are taking longer to attend to the spatial location of the target to achieve the same level of accuracy found in Experiment 1.
2.4 General Discussion

The main aim of the experiments described in this chapter was to investigate the effects of characteristic sounds and spatially congruent sounds on visual search performance in a complex visual scene. In the first experiment, the target object was predictable from one trial to the next (but not whether it was present or absent) and we found that the presence of a sound which is characteristic of the visual target can improve search performance relative to the presence of an uncharacteristic sound regardless of the spatial location of that sound. Indeed, whether the sound occurred within the same hemifield as the visual target or not seemed to have no effect on performance. These results support the findings of Iordanescu et al. (2008) that semantically relevant sound can facilitate search performance. They are also in support of evidence that non-spatial auditory information can guide visual attention (see Van der Burg et al., 2008).

In the second experiment, the task was rendered more difficult as the target object was unpredictable from one trial to the next. As was expected, both accuracy and speed of responses were better when participants were searching for one possible target per block (Experiment 1) as opposed to when they were searching for one of six possible targets per block (Experiment 2). This increased task difficulty was evident in the different response times between both experiments (RTs were, on average, 120 ms longer in the second experiment). In Experiment 2 we found no effect of characteristic sounds. However, we did find a main effect of spatially congruent sound. There was a benefit to search performance when the sound was spatially co-located in the same hemifield as the visual target. These findings support previous studies that have found benefits when auditory and visual information occur at the
same location (Perrott & Saberi, 1990; Perrott et al., 1990; Spence & Driver, 1997). The greatest cost to search time performance in Experiment 2 occurred when the sound was both incongruent and uncharacteristic to the visual target.

The difference in results across the two experiments may be accounted for by the changes in task difficulty. It may be the case that in the first experiment, when searching for one target per block, participants relied only on the semantic information from the sound and not its location. In the first experiment, only one sound could be semantically useful in any given block. This may have caused participants to attend only to the sound relevant to the target they are searching for and ignore the sound in the other trials. For example, if the target for that block was a dog, then the only useful sound in that block would have been a “bark”. In the second experiment however, any of the six sounds could have been associated with a target in any trial within a given block. Therefore, the semantic information in all sounds were potentially useful in that scenario. Participants may be successfully ignoring the semantic content of the sound in the second experiment being unable to predict its semantic usefulness. This finding is interesting as it suggests that the meaning of the sound is not necessarily integrated with the visual component of a target object for the purpose of visual search and that, rather, the semantic content of the sound can be selectively ignored at the input stage.

As target exposure for both experiments was brief (250 ms), the results support the claim that multisensory effects occur at early stages of visual object-recognition (as indicated by Molholm et al., 2004). Audio-visual crossmodal effects were observed in both experiments as auditory and visual attention have combined to facilitate the detection of the target. The processing may however differ across the
experiments. Biederman et al. (1982) also reported that consistent scene context can facilitate object detection. Scene context is also consistent when the auditory and visual information is semantically related. However, this context may no longer be consistent when the sight and sound of a target object are unrelated, such as when presented with an image of a dog but we hear a cat. This inconsistency may account for the effect of semantic sound in our first experiment.

Wolfe at al. (1994) proposed that pre-attentive processes can be used in guiding our attention towards a visual target. It could be that top-down information such as the target’s features and their association with auditory information is being processed at an early stage resulting in an efficient search in Experiment 1 while in Experiment 2, bottom up processing seems to be a more appropriate explanation. Prior knowledge such as target/sound relationship does not seem to have an effect but the effect instead seems to be driven by the location of the auditory stimulus (i.e. stimulus driven).

Finally, the number of spatial locations used in this experiment may have limited the effect of spatial congruency. Though six different targets were used, the target could only appear in one of two locations. Observers may have learned the possible locations of the target. This would explain the high-levels of accuracy in both experiments. An increased number of spatial locations would be necessary to further investigate the effect of spatially congruent sound.

2.5 Summary

In the experiments reported here, we found a benefit for the presence of a sound on reaction times (Exp. 1 & 2) and accuracy (Exp. 1) of search performance for a visual target. Participants were not explicitly aware of any facilitation effect of the sound and were told to ignore it in both experiments. In the first experiment,
participants were faster and more accurate at locating the visual target when the sound was semantically relevant. In the second experiment, an effect of spatial congruency was found where a sound which was both spatially incongruent and uncharacteristic resulted in a cost on search performance. These findings suggest that the effect of characteristic sound on visual search is dependent on task difficulty.
3.0 An investigation of the role of audition in searching for a target in static visual scenes

Abstract

Although most research in the area of scene perception was specific to the visual modality, recently there has been a growth in the number of studies on audio-visual interactions in the perception of scenes. However, in comparison to our knowledge of the effect of vision on sound localization (e.g. the ventriloquist effect), relatively little is known about the role of auditory information on visual perception, particularly on visual search. In particular, although some studies have reported that sound can improve the detection of a synchronously presented visual target, further studies are necessary to determine whether the spatial location of the sound can affect visual performance. Here we investigated whether a spatially congruent sound enhanced the detection of a visual target in a complex scene compared to a spatially incongruent or absent sound. A spatially congruent sound was one which occurred in the same spatial hemifield of the scene as the visual target. The task for the participant was to detect the presence or absence of a visual target in a sequence of randomly presented visual scenes. We found that visual target detection was improved for scenes where the auditory and visual target were spatially congruent (i.e. both were presented in the same hemifield). This finding supports a growing body of studies indicating multisensory interactions, particularly from audition, in visuo-spatial tasks.
3.1 Introduction

Although our knowledge of how scenes are perceived and represented in memory has been well documented in the literature (see Henderson & Hollingworth, 1999 for a review), the vast majority of this work is focussed on the visual domain. Moreover, our knowledge of how visual targets are searched for and located in a scene mainly comes from studies conducted on visual processing alone (Theeuwes, 1992; Treisman & Gelade, 1980; Wolfe et al., 1989). This is in conflict with how we experience the world as humans rarely process one sensory input in isolation for the purpose of every day perception. It is clear from everyday experience that information from other sensory modalities can also affect perception of a scene and, more particularly, can influence how we locate specific target objects (or faces) in a complex visual scene.

In a recent study, Van der Burg et al. (2008) reported an investigation which was aimed at improving our understanding of the role of audition in visual search. They demonstrated that searching for a visual target (line segment) in a complex scene containing many different visual elements (i.e. orientated line segments) which randomly appeared and disappeared in the scene, could be more efficient with the presence of an auditory tone. When this tone was presented, such that its onset and offset was synchronised with the appearance and disappearance of the visual target in the scene, then visual search was rendered more efficient with search times resembling the classic ‘pop out’ effect. This effect seemed to be mediated by temporal synchronisation alone since the sound source was not spatially co-incident with the location of the visual target. This effect, now commonly referred to as the “Pip and Pop” effect, demonstrates how a spatially non-specific auditory signal can boost the saliency of a temporally concurrent visual signal. To further test their hypothesis, Van
der Burg and colleagues reduced the likelihood of the auditory event being informative to just 20% of the trials. This meant that for 80% of the trials, the tone was not synchronised with the target but with a distracter event. The result was that the integration of the auditory and visual stimulus still occurred largely automatically yielding quite a robust benefit to visual search. In sum, the experiments reported by Van Der Burg et al. (2008) indicate that when searching for an object in a cluttered scene, a temporally congruent sound will facilitate target detection even when the sound contained no information regarding the specific location or identification of the target.

Attending to objects in a scene is something we do on a daily basis. Whether walking, driving or simply sitting still, we are attending to objects around us using many of our sensory modalities. In some situations, we selectively focus on some objects while ignoring others which may be irrelevant to the goals of the task. Feature Integration theory (Treisman & Gelade, 1980) proposes that visual attention is automatically drawn to visual objects that stand out from the background, particularly when the object differs from other items in the scene on a single basic feature (e.g. orientation or colour). On the other hand, objects which are not uniquely defined by a single feature from others in the scene require, according to the FIT, an effortful, object-by-object analysis of the scene until the target is located. In such cases, it is possible that sound may facilitate visual search. However, it is unknown what effect sound may have on these search processes in high-level complex scenes such as those used in this study.

We know from early work by Biederman (1981) and Schyns & Olivia (1994) that when scenes are presented for 100ms or less (i.e. too little time for the eyes to scan the scene), that this exposure is sufficient to get an overall impression or gist of a
scene although at the cost of being able to identify specific details of the objects. Longer exposure allows us to identify objects in scenes, possibly by the allocation of focused attention on the individual objects. Adding an auditory cue to a complex scene may therefore enhance our ability to attend to and identify objects in the scene (Vroomen & de Gelder, 2000), even at short exposure times of the scene. Visual object identification and target localisation in (relatively impoverished) scenes has been shown to improve with spatially congruent auditory information. For a review of crossmodal spatial attention, see Spence (2010).

Sound localisation is affected by vision, as in the classic ventriloquism effect (Howard & Templeton, 1966) is not a new area of investigation. However, in order to better understand audio-visual interactions in visuo-spatial tasks, further investigation is necessary in the area of visual localization as affected by audition. It is unclear whether a stimulus in one sensory modality (e.g. audition) automatically attracts attention to a spatially coincident stimulus that occurs simultaneously in other modality (e.g. vision). Perrott, Cisneros, McKinley & D'Angelo (1996) found that an auditory stimulus occurring at the same spatial location as a visual target aided visual target detection relative to no sound. However, this facilitation was not found by Doyle & Snowden (2001) in their study. They reported that while the presence of a non-meaningful auditory stimulus reduced errors in a visual search task, the location of the sound (either central to the display or at the same location as the target) did not facilitate performance any further. Therefore, there appears to be conflicting evidence regarding the facilitation of visual target detection by spatially co-located sounds. These studies however did not use meaningful sounds or realistic scenes. Instead they used low-level stimuli such as beeps and flashes. The current study aims to look at these effects in more complex environments.
In this study we investigated whether the covert orienting of spatial attention that is triggered by the presence of a sound can facilitate the detection of a visual target in a complex scene. In a real world scenario, we tend to see and hear objects in our environment simultaneously. We also tend to detect larger objects that are in our central field of vision faster and more easily than those that are small and located in the periphery. This study aims to investigate whether crossmodal facilitation in the detection of visual targets, as demonstrated by previous researchers (McDonald & Ward, 2000; Perrott et al., 1990; Spence & Driver, 1997) holds true for high-level scenes. In particular, on the basis of previous findings, we predicted that the presence of sound would improve visual target detection in complex scenes relative to no sound. Furthermore, we predicted that sound occurring in the same hemifield as the visual target would aid detection of a visual target relative to no sound or to sound occurring in the opposite hemifield.

3.2 Experiment 3

The following study investigated the role of audition on searching for a visual target in complex visual scenes. We looked specifically at how the spatial location of an auditory cue could facilitate the detection of a visual target. The target was constant but could appear in one of two sizes, and in one of 8 locations (4 centrally located and 4 peripherally located) in a scene. The purpose of manipulating these independent variables was to look at what effect visual target size and spatial location had on target detection. The background scenes were of indoor and outdoor images which were randomly presented throughout the experiment. Our hypothesis was that when the auditory stimulus was co-located with the target, search performance in
3.2.1 Method

Participants

Twelve undergraduate students (5 male and 7 female with a mean age of 21 years) from the School of Psychology, Trinity College Dublin participated in this experiment for research credits. Their ages ranged from 18 to 31 years. All reported normal or corrected to normal vision and none reported any auditory impairment. Prior approval was obtained from the School of Psychology Ethics Committee. Following a briefing on the experimental protocol, participants provided written consent to take part in the study.

Stimuli

Ten images of visual scenes were randomly selected from an online photography database (http://www.sxc.hu). The images were the same as those used in Experiment 1 (with two new images added). The images were converted to grayscale and resized to fit a 1024 x 768 pixel frame. Examples of these images are provided in Figure 3a. The visual target used in the following experiment was an image of a mobile phone (see Figure 3b). Two phone sizes were used: Small (2.15° (W) x 4.02° (H) of visual angle; Large: 2.79° (W) x 5.22° (H) of visual angle.

As well as a visual counterpart, we used the sound of a mobile phone as the auditory cue. This auditory cue was a mobile phone ring-tone which was sourced
from *A1 Free Sound Effects* (http://www.a1freesoundeffects.com). The sound used was a standard telephone ringing sound, 200ms in duration and in standard wave file format. It was edited using Adobe Audition and the sound pressure level was 53dbA relative to the participant’s head. The same sound was used throughout the experiment.

![Figure 3.0](image)

*Figure 3.0* a) An illustration of four (out of 10) of visual scene backgrounds and b) an image of the visual target used in Experiment 3.

To create the scene stimuli used in the experiment, each scene was presented with an image of the target embedded into it (i.e. Target Present trials) or without the image of the visual target (i.e. Target Absent trials). For the target present trials, the image of the target was positioned at one of eight locations in each scene. These locations are illustrated in Figure 3.1 and were as follows: four central positions located nearest to the central fixation position and four peripheral locations. Table 1 describes each of the possible locations that the visual target occurred relative to fixation.
Figure 3.1. An illustration of a scene used in Experiment 3 indicating the possible locations of the visual target. The red markers indicate the central locations of the target whereas the blue markers indicate the peripheral locations of the target relative to fixation. A target could appear in one of these locations in a trial.

Table 1.0 A list of scene locations that correspond to centrally and peripherally located targets expressed in terms of visual angle relative to fixation.

<table>
<thead>
<tr>
<th>Location</th>
<th>Up</th>
<th>Down</th>
<th>Left</th>
<th>Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.88°</td>
<td></td>
<td>4.23°</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6.88°</td>
<td></td>
<td>13.05°</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6.88°</td>
<td></td>
<td>13.05°</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6.88°</td>
<td></td>
<td>4.23°</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6.88°</td>
<td>4.23°</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6.88°</td>
<td>13.05°</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>6.88°</td>
<td>13.05°</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6.88°</td>
<td></td>
<td>4.23°</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 3

Apparatus

The experiment was programmed using a Presentation™ (Neurobehavioral Systems®) script which ran on an Intel® Pentium™ 3GHz computer with 512MB RAM and a 20” CRT Monitor. A height adjustable chin rest was used to avoid neck muscle fatigue and also to ensure that each participant was a consistent distance from the screen (i.e. 57cm). Two 10 Watt Boston Stereo Speakers were positioned on stands half way down either side of the screen (as indicated in Figure 3.1) and were used to amplify the sound from the computer at an SPL of 53dbA. The volume was measured using a sound pressure level meter. A standard PC keyboard was used to record participant responses.

Design

The experiment was based on a fully factorial, 3x2x2 within-subjects design, with spatial congruency of the sound (congruent, incongruent or absent), visual target size (small or large) and visual target position (central or peripheral) as factors. Target size was fully counterbalanced across all central and peripheral positions. The spatial location of the auditory cue was congruent when it was presented in the same visual hemifield as the visual target, irrespective of whether the target was presented in a central or peripheral location. This design resulted in a total of 630 trials in the experiment. The proportion of trials was divided across target present (75%) and absent (25%). Trials were presented in a randomised order across participants to reduce order effects. The dependent variables were response times and accuracy.

Procedure

A central fixation cross was presented at the beginning of each trial for 500 ms. The fixation cross was then immediately followed by an image of a scene which was
presented on screen for 200ms. The task for the participant was to maintain focus on the centre of the scene and indicate, as fast and as accurately as possible, whether the target was present in the scene or not by pressing the corresponding key (’/’ = “yes”, z = “no”). Both accuracy and response times were recorded. Overall, the auditory stimulus was present for 66.6% of the target-present trials (i.e. in both the sound congruent and incongruent trials). It was also present for 66.6% of the target absent trials. Prior to testing, each participant completed 10 practice trials to ensure that they understood the task. For each participant, the experiment took about 1 hour to complete.

3.2.2 Results

The mean correct reaction times and correct responses were first calculated across conditions. For target present trials the mean correct response was 87.38% (st. dev. =5.48) and the mean RT was 495 ms (st. dev. =71.34). For target absent trials the mean correct response was 94.01% (st. dev=6.241) and the mean RT was 604 ms (st. dev. =75.85). Performance in all conditions was greater than chance (50%). Only the target present trials were subjected to further analyses as the responses to the target absent trials were of no theoretical value. Separate 3x2x2 repeated measures ANOVAs were performed on the correct responses and RT data with auditory Spatial Congruency (congruent, incongruent or sound absent), Target Size (small or large), and Target Position (central or peripheral) as factors.

Response times

For the response time data, we found a main effect of Spatial Congruency \( [F (2, 22) = 7.254, p < 0.01] \) (see Figure 3.2). There was also a main effect of Target Size \( [F (1, 11) = 486.033, p < 0.001] \) with faster RTs to the large relative to the small target.
There was a main effect of Target Position \([F_{(1, 11)} = 37.266, p < 0.001]\) with faster RTs to targets located in the centre rather than the periphery.

![Figure 3.2](image)

*Figure 3.2.* Plots showing the mean reaction times to the spatially congruent, spatially incongruent and sound absent conditions in Experiment 3. Error bars represent standard error of the mean.

An interaction was found between Spatial Congruency and Target Size \([F_{(2, 22)} = 4.1026, p < 0.01]\) as shown in Figure 3.3a. A Tukey post-hoc analysis revealed that RTs to locate the small target were significantly faster when a spatially congruent sound was presented relative to a spatially incongruent sound \((p<0.0001)\). For small targets, no difference was found between performance for spatially congruent sound and when sound was absent. However, when sound was spatially incongruent with the small visual target, RTs were significantly faster than when sound was absent \((p<0.0001)\). RTs to locate the large target were significantly faster when the sound was spatially congruent \((p<0.0001)\) relative to when a spatially incongruent sound was presented. For the large target, no difference was found between when a spatially congruent sound was presented and when sound was absent. However, when a spatially incongruent sound was presented with a large visual target, RTs were
significantly faster than when the sound was absent (p<0.0001). No other pairwise differences between Spatial Congruency and Target Size were found.

There was also an interaction between Spatial Congruency and Target Position \[F (2, 22) = 5.1229, p < 0.05\] which is shown in Figure 3.3b. RTs were significantly faster for centrally positioned targets, when the sound was spatially congruent relative to spatially incongruent (p<0.0001) and relative to when sound was absent (p<0.001). For peripherally positioned targets, RTs were faster when sound was spatially congruent relative to spatially incongruent sound (p<0.0001). No difference was found between sound present and sound absent for peripherally positioned targets. No other pairwise differences between Spatial Congruency and Target Position were found.

There was also a significant two-way interaction between Target Size and Target Position \[F (1, 11) = 4.8513, p < 0.05\] which is shown in Figure 3.3c. A Tukey post-hoc analysis revealed that RTs to the small targets were faster when that target was presented centrally than when it was presented peripherally (p<0.000). RTs to the large target were significantly faster when that target was presented centrally than when it was presented peripherally (p<0.0001). There were no three-way interactions.

a)
Chapter 3

b)

c)
Figure 3.3. Plots showing the two-way interactions on response times between a) AV spatial congruency and target size, b) AV spatial congruency and target size and c) target size and target position in Experiment 3. Error bars represent ± 1 standard error of the mean.

Correct responses

With regard to accuracy, there was no effect of spatially congruent sound. There was a main effect of target size \([F(1, 11) = 57.234, p < 0.001]\) with better accuracy for the large target over the small target and a main effect of target position \([F(1, 11) = 33.544, p < 0.001]\) with better performance for centrally than peripherally located targets.

A significant interaction was found between target size and target position \([F(1, 11) = 33.295, p < 0.001]\) as can be seen in Figure 3.4. A Tukey Post-hoc analysis revealed significantly better performance for small centrally positioned targets than small targets located in the periphery (\(p<0.001\)). Also, performance was significantly worse for small targets located in the periphery relative to large targets in the same location (\(p<0.001\)). There was no effect of target position when the target was large. No other significant interactions were found.
3.2.3 Discussion

The main aim of this study was to investigate the role of audition in the search for a visual target embedded in a complex scenes. The results indicate that participants were faster at detecting the presence of a visual target when it was presented with a spatially congruent auditory stimulus compared to a spatially incongruent sound or no sound. However, accuracy performance was not significantly affected by the presence or absence of the auditory stimulus, although there was nevertheless a trend in that direction. The scenes were presented for 200ms which was long enough for the participants to detect the target and to perform above chance on the detection task.

The RT data suggests a greater benefit for the spatially congruent sound on search times when the target was small compared to the large target. This suggests that the sound was more likely to facilitate performance when the visual task was relatively difficult, i.e. when the search was for a small target rather than a large target. Furthermore, as we would expect, search times were slower when the sound

Figure 3.4. Plot showing the interaction between visual target size and AV spatial congruency on the mean correct responses of Experiment 3. Error bars represent ± 1 standard error of the mean.
was spatially incongruent to the location of the visual target relative to the congruent sound. This occurred for both large and small target though effect was more pronounced for the smaller targets (see Figure 3.3a). Regarding accuracy, participants made significantly more errors when searching for the small target. This performance deteriorated further for small targets which were located in the periphery (see Figure 3.3b). The presence or absence of sound did not make a difference to accuracy performance for each individual target size. This would suggest that the effect of the sound on detecting the presence of the target is more based on the relative relationship between the location of the sound and the target location rather than the size of the visual target.

Since both the visual and auditory modalities can encode information pertaining to spatial location, it might be assumed that a common crossmodal representation of the target might facilitate its detection. These results suggest that the audiovisual stimuli are acting together to create a unified percept. According to the research of Beauchamp et. al (2004), the integration of such information is likely to be occurring in the superior temporal sulcus (STS).

A pop out effect was possible as the mobile phone had the discriminative feature of a corresponding sound. However, although the number of distracters was not manipulated in this task, the nature of the search performance can be indirectly determined from the comparison between the response times between the target present and target absent trials. If the participants are adopting a serial search strategy, they will need to search each item in the scene, identifying it as a non-target until they find the target. Typically when serial search occurs, responses to 'target present' trials are generally twice as fast as those to the target absent, as more items will be scanned in the scene until the trial is terminated when the target is not present.
We found that the time it took to terminate a target absent trial was 604ms compared to 495ms to locate a target when it is present, which does not fit the typical time ratio between target absent and present trials in a typical serial search strategy. However, the fact that search times were longer when the target was absent suggests such a strategy was employed and that a pop-out effect did not occur.

As Treisman’s Feature Integration Theory (FIT) did not account for multisensory input, we can only suggest what may be happening in the context of this theory. Perhaps FIT would explain the benefit of sound on visual search in that the features of the target (both auditory and visual) were bound together pre-attentively causing the target to become more detectable in the visual scene.

However, Jeremy Wolfe’s guided search model (Wolfe, 1994) may offer an alternative explanation for the results found. Wolfe’s theory proposes that pre-attentive processes can be used to direct attention when searching for a target. As such, a combination of bottom up processing (i.e. the auditory and visual stimuli) and top down processing (i.e. prior knowledge of target’s features and real world scenes) may have directed attention to the visual target. For example, prior knowledge of a bathroom scene that would usually not contain a mobile phone may make such a target easier to detect in such a scene due to its semantic incongruency. Hollingworth and Henderson (1999) have shown that object discrimination performance improves for targets positioned in familiar scenes, particularly when those targets were inconsistent with the schema of the scene. Although we are not looking specifically at the relationship between the target and scene context in this study, it may have been the case that the chosen target in this experiment would be relatively easy to detect because of its inconsistency with the scene background. Moreover, we used the same target throughout the experiment, therefore participants were highly familiar with the
features of this target which could also have facilitated search performance. For example, Castelhano & Heaven (2010) also found that prior knowledge of scene context and the target’s features improved the speed of target detection, with target features particularly affecting the later verification processes involved in object identification.

Based on our finding that the congruency of the sound is affected by both target size and target location, in particular that congruent sounds particularly benefit the detection of small targets located in the periphery, this suggests that sound appears to be more useful when the task difficulty increases. In contrast, sound congruency appeared to have little or no effect on target detection when the target was relatively large and when it was located in central vision. This finding suggests that sound may be used to guide search only when the target is relatively difficult to locate.

3.3 Summary

In the study reported here, we found a benefit for the presence of a spatially congruent sound on search performance for a visual target, measured in both response times and accuracy, when both were presented in the same spatial hemifield. Participants were not explicitly aware that the study was examining the facilitation effect of the sound but were instructed to ignore the sound and to attend only to the visual stimuli. Overall, participants were faster at locating the target when the sound was spatially congruent with the target than when it was spatially incongruent. This main effect of sound congruency was also qualified by interactions between it and target size and target location. In particular, searching for a small target, located in peripheral vision benefitted more from the presence of a congruent sound than when
the target was large (irrespective of location) or presented in a central position in the scene. These findings indicate that target size and location affect search performance within a complex visual scene.
4.0 The effects of task irrelevant stimuli on the deployment of attention in a spatial cuing task

Abstract

Perceptually salient objects or images in the periphery can distract our attention from the task at hand. In this study, we investigated the role of dynamic, task irrelevant information on the deployment of visual attention in a spatial cuing task. Specifically, we looked at whether dynamic visual images (Experiment 4) or audio-visual stimuli (Experiment 5) of point-light walkers, which varied in gender but were located outside of foveal vision could capture spatial attention. The participant’s task was to judge the orientation of a pre-cued target within a visual array while ignoring other stimuli located in the periphery. In Experiment 4, it was found that performance was facilitated at pre-cued locations of the relevant stimulus relative to the invalid cued locations. However, distracter information affected performance particularly during trials in which the pre-cue was invalid and the point-light walker was “extremely female”. In Experiment 5 performance to the valid cued locations of the targets was not facilitated relative to the invalid locations. Furthermore, the presence of the multisensory distracter stimulus was generally associated with better accuracy performance at the target locations, which was likely due to a response strategy. However, the results suggest that the qualitative nature of a dynamic, task irrelevant stimulus can affect performance on a cognitive task suggesting important multisensory influences can occur even outside the focus of attention.
4.1 Introduction

Both natural and complex scenes can be cluttered, containing many different objects and our ability to attend to multiple objects simultaneously is thought to be limited (see Broadbent, 1958). Attentional mechanisms are necessary to filter out irrelevant information. Guiding attention to a spatial location had been shown to improve response time and accuracy at that location (Posner, 1980). Selective attention has been shown to increase our ability to discriminate target stimuli (Lu & Dosher, 2000) and reduce interference effects from distracters (Shiu & Pashler, 1995).

Posner’s cueing paradigm has often been adopted in studies investigating the role of directed attention on perception and it has been reported that directed attention can act to either facilitate processing in the sensory pathways that receive input from cued locations, or inhibit processing in pathways that receive input from unexpected or irrelevant locations (Hawkins, Shafto, & Richardson, 1988). However, Shaw (1984) rejected Posner’s (1980) findings, arguing that attention does not have an effect on sensory processing relating to visual detection. Shaw explained that the reduced response time associated with spatial cuing could be accounted for by differences in decision making strategies occurring at both the cued and uncued locations. In support of this, Hawkins et al. (1990) reported that as targets are more likely to appear at a cued location, observers may simply require less sensory evidence to make a decision. To test this, Hawkins et al. (1990) ran a number of spatial cuing experiments incorporating central and peripheral cues. Participants had to judge whether a target had appeared at one of four locations and indicate with what confidence this judgement had been made. Their findings suggested that visual spatial attention facilitated the processing of sensory input during target detection either by
increasing sensory gain for inputs at cued locations or instead by prioritising the processing of the cued inputs.

Some have argued that the perception of images and objects in our environment can occur without the use of focussed attention. For example, Li, VanRullen, Koch & Perona (2002) demonstrated how information from a rapidly presented (27ms) image located outside and to one side of fixation can be processed while simultaneously performing an attentional task at another spatial location. Participants focussed their attention on the center of the display where a group of letters were flashed. The task was to indicate if the letters were the same or if there was one different. It was found that even though the observers’ attention was focussed on the letter task, participants could still score above chance in correctly indicating the presence or absence of an animal in the photograph. Using a similar setup, Reddy, Wilken, & Koch (2004) showed that it was possible to identify faces as being male or female even when presented for only 150ms.

Sui and Liu (2009) used a spatial cuing paradigm to test the effect of facial attractiveness on spatial attention. Participants judged the orientation of a cued target which occurred either to the left or right of fixation while a task-irrelevant image of an attractive face was presented at the opposite side of visual space. The results showed that the presence of an attractive face had a detrimental effect on the speed of judging the orientation of the target but not the presence of an unattractive face. The effect of facial attractiveness occurred only when the cue was valid. Their findings demonstrate how a salient visual stimulus can compete with other visual input for spatial attention.
Attractiveness is an important stimulus from a biological and social perspective and has been shown to affect human behaviour (see Langlois et al., 2000). It is evident from Sui and Liu (2009) study that attentional systems prioritize attractive stimuli. Also, studies have shown that neural responses to attractive stimuli occur automatically and can be observed even when participants are engaging in an unrelated task (Aharon et al., 2001; O'Doherty et al., 2003). We therefore used dynamic walker stimuli which, on the basis of previous research, were known to vary on the attractiveness dimension (Johnson & Tassinary, 2007).

Evidence from functional neuroimaging suggests that attentional processes occur at various levels within the visual system. For example, some have found evidence for attentional modulation of visual information already in the lateral geniculate nucleus (LGN) of the thalamus which is one of the earlier stages where visual processing occurs (O'Connor, Fukui, Pinsk, & Kastner, 2002). Later cortical areas, such as V4 and TEO in the macaque brain, have been found to be important for filtering unwanted information (De Weerd, Peralta, Desimone, & Ungerleider, 1999). Such mechanisms of attention appear to be controlled by higher order areas in the frontal and parietal cortex which feedback information to the visual system (see Everling, Tinsley, Gaffan, & Duncan, 2002; Schall & Thompson, 1999). A thorough understanding of the neural basis of attentional processes, particularly those involving cross-modal or multisensory input, has yet to be uncovered. For this reason further imaging and behavioural studies are necessary.

The experiments described below were based on an adaptation of the task used by Sui and Liu (2009), but with one important modification: instead of using static images of faces, we used dynamic images of male and female characters walking. The
aim of Experiment 4 was to investigate the effects of dynamic distracters located in the periphery on the deployment of attention in a spatial cuing task. As crossmodal stimuli are also known to have an effect on spatial attention (for a review, see Spence, 2010), in Experiment 5 we added an auditory stimulus, i.e. the sound of footsteps, which was presented at the same side of visual space as the visual distracter. We expect that both dynamic unisensory and multisensory stimuli would compete with the target stimulus, causing a detrimental effect on response times in excess of those found by Sui and Liu (2009).

4.2 Experiment 4

The following experiment was based on Posner’s cueing paradigm and it was designed to investigate the role of task irrelevant information on attentional deployment in a spatial cuing task. Specifically, we looked at whether task irrelevant dynamic stimuli located in the periphery could capture attention and influence performance on judging the orientation of a target. The visual distracter was a point-light image of human figures walking (male and female). The distracter stimulus was located in one of two locations, either left or right of fixation and always in the opposite field to the visual target. Our main hypothesis was that when the visual target location was validly cued, response times would be faster relative to invalid cued trials. However, we were particularly interested in how the presence of a distracter stimulus would affect overall performance. We varied the information contained in the point-light walker: these were dynamic images of male or female walkers in which the gender was typical (feminine or masculine) or exaggerated (very feminine
or very masculine). Based on the previous study reported by Sui and Liu (2009) we expected that participants would be slower and less accurate at the orientation task when an exaggerated walker was simultaneously presented relative to a typical walker. We expected the attractive movement of the very feminine and very masculine walkers to capture the observer’s attention, resulting in impaired performance on the task. Performance was measured in terms of response speed and accuracy in judging the orientation of the letter ‘T’ in an array of ‘+’s.

4.2.1 Method

Participants

An opportune sample of 42 participants (12 male and 30 female with a mean age of 46 years) who were attendees at a Science Gallery Exhibition in Trinity College took part in this study. All reported normal or corrected to normal vision. Prior approval was obtained from the School of Psychology Ethics Committee. Following a briefing on the experimental protocol, participants provided informed written consent to take part in the study.

Stimuli

The stimuli used in this experiment were a set of dynamic visual images, in particular ‘point light walker’ animations which were based on motion captures of real human (male and female) movement. These dynamic images were developed in conjunction with colleagues from the Graphics and Visualisation Group (GV2), School of Computer Science and Statistics at Trinity College Dublin and were taken from a large database of motion-captured images. The criteria for selecting the ‘walker’ stimuli was as follows: all ‘walkers’ should be consistent in height and
walking trajectory. Each image depicted a human figure, (either male or female) which was viewed from the front and walking in the direction of the camera. These ‘walkers’ did not change in size or direction over the course of their movement.

As is illustrated in Figure 4, the ‘walker’ stimuli were chosen in order that they were classified according to masculinity (masculine, very masculine) and femininity (feminine, very feminine). The exaggerated gender versions of the stimuli were ‘caricatures’ of the original point-light displays which were created by the Graphics and Visualisation (GV2) group. We created further versions of these walker stimuli by varying the weight of the humans (very thin, thin, large and very large build) and speed (very fast, fast, slow and very slow) of the walking motion was varied equally across male and female stimuli. For every ‘walker’ categorised according to sex, there were four variations of weight and speed. This resulted in a total of 64 unique distracter stimuli, with 8 different versions for each of the gender types. Each ‘walker’ was normalised to fit into an area that measured 6.7° (width) x 15° (height) visual angle and was matched according to luminance and contrast to ensure that image properties could not contribute to behavioural differences in performance.

Male walker

Female Walker
Chapter 4

Figure 4.0 This image depicts examples of male (left) and female (right) 'point light walkers' used in this experiment. The white dots represent points on the person's body for which motion was captured.

Each trial consisted of a central fixation point, followed by a central cue which was then, in turn, followed by two position markers comprising two white boxes measuring 3.8° (W) x 3.8° (H). The visual angle subtended between the centre of the display and the outer edge of the white box was 5°. The visual angle subtended by the central cue was a 1° and consisted of either a ' < ' or ' > ' symbol. The target display consisted of a target letter 'T' embedded in the centre of an array of eight crosses. This configuration was consistent across all trials. Each of the target and distracters subtended a visual angle of 1.2° x 1.2° in the display. The target stimulus, i.e. a letter 'T' was briefly presented either upright or inverted (as can be seen in Figure 4.1 a and b) and was replaced by a '+' symbol (as depicted in Figure 4.1c). During the experiment, the onset of the 'walker' stimulus was coincident with the onset of the target array.

Figure 4.1 This image depicts each of the arrays used. An upright (a) or inverted (b) 'T' was presented for 200ms and then an array of crosses only (c) was presented.
Apparatus

The experiment was run on a computer monitor with a resolution of 1024 x 768 pixel frame at 75 Hz, and contained no colour information other than the white target array and distracter stimuli which were all presented on a black background. The experiment was programmed using a Presentation™ (Neurobehavioral Systems®) script which ran on an Apple Macbook with 1GB RAM. A standard PC keyboard was used to record each participant’s responses. Testing took place in a public setting at the Science Gallery, Trinity College Dublin. Participants were seated at a circular testing station and members of the public could watch the experiment from a distance. The headphones used were sound-attenuating to minimise any external distractions.

Design

This experiment was based on a 2x4 within subjects design with cue validity (valid or invalid) and distracter salience (very masculine, masculine, feminine, very feminine) as factors. A valid trial was one in which the spatial location of the target was correctly indicated by the central visual pre-cue. An invalid trial was one in which the visual cue directed attention to the opposite side of space to that of the target. The total number of trials was divided across the valid and invalid pre-cue conditions as follows: 85% and 15% respectively. The central cue pointed to the left or right side of space an equal number of times. The experiment had a total of 192 (164 valid) trials within which each of the four distracters were displayed an equal number of times. An upright ‘T’ was presented in the target array for half of the trials.
and an inverted ‘T’ was presented for the other half. Trial order was randomised across participants. The dependent variables were response times and accuracy.

Procedure

Participants sat approximately 60cm from the computer monitor and began the experiment by completing a series of 12 practice trials to ensure that they were familiar with the visual targets and the task. They were given verbal feedback on their accuracy performance and had the opportunity to pose any questions. Following the practice session, participants were then presented with the test phase which consisted of 3 separate experimental blocks of trials. Participants could take a self-timed break between blocks.

The procedure for each trial can be seen in Figure 4.2. A trial began with a central fixation cross which was presented for 500ms. A central cue was then presented for 200ms, followed by an inter-stimulus interval for 150ms. The array with the oriented target ‘T’ was then presented for 200ms before changing back to a ‘+’. The array (without the target) remained in the display, as did the visual distracter, for a maximum of 4 seconds or until the participant responded. The task for the participant was to indicate, as fast and as accurately as possible, the orientation of the letter ‘T’ in the array of crosses by pressing the corresponding key on the computer keyboard (‘/’ = upright, z = inverted). Participants were told to ignore the ‘walker’ at the opposite side of fixation while judging the orientation of the target and were reminded to maintain central fixation throughout the trial. Participants could take a self-timed break between blocks. The experiment took approximately 20 minutes for each participant to complete.
Figure 4.2  This image depicts a typical trial where a fixation cross appears for 500ms, followed by a cue (either '<' or ' >'), then an inter-stimulus interval of 150ms. Then both the target array and distracter stimulus were presented simultaneously for 200ms. The distracter and the array featuring a '+' instead of the target remained on screen for 4 seconds or until the participant responded. Participants responded to the orientation of the ‘T’ as quickly and as accurately as possible.

4.2.1 Results

The mean accuracy and reaction times were calculated for each of the cueing conditions and distracter salience conditions. Two participants were excluded from
the analysis as their overall performance was below chance (50%), indicating that
they had not performed the task appropriately. This amounted to a removal of a total
of 4.45% of the data. For the valid cue condition, the mean percentage of correct
responses was 70.14% (std. dev.=6.83%) and the mean RT was 796ms (std. dev.=180
ms). For the invalid cue condition the mean percentage of correct responses was
65.88% (std. dev. =8.92%) and the mean RT was 870ms (std. dev. =190ms). Separate
2x4 repeated measures ANOVAs were performed on the correct responses and RT
data with Cue Validity (valid or invalid) and Distracter Salience (very masculine,
masculine, feminine, very feminine) as factors.

*Response Times*

The mean response times to each of the valid and invalid trials across the
distracter conditions are depicted in Figure 4.3. For the response time data, we found
a main effect of cue validity \( [F(1, 41) = 14.2317, p<0.001] \) with faster RTs when the
cue was valid than invalid. We found no main effect of distracter salience \( [F(3, 123) =
2.0308, p= 0.113] \). We found an interaction between cue validity and distracter
salience \( [F(3, 123) = 3.6691, p< 0.05] \) which is suggestive of a different effect of
distracter salience, dependent on the validity of the cue. A Tukey post-hoc analysis
revealed that for the invalid cue condition, RTs to the target were significantly longer
when 'very masculine' distracters were presented relative to 'feminine' distracters
(p<0.005). Also, for the invalid cue condition, RTs were significantly longer when
'very feminine' distracters were presented relative to' feminine' distracters (p<0.05).
Chapter 4

Correct Responses

The mean correct responses to each of the valid and invalid cueing condition is depicted in Figure 4.4. We found a main effect of cueing validity \( F(1, 41) = 7.148, p < 0.05 \) with greater accuracy for valid than invalid cues. We found no main effect of distracter salience \( F(3, 123) = 1.582, p = 0.197 \). We also found an interaction between cue validity and distracter salience \( F(3, 123) = 5.536, p < 0.005 \). A Tukey post hoc analysis revealed no significant differences between the distracter conditions. A post hoc analysis of the invalid cue condition revealed that accuracy was significantly worse when the ‘very feminine’ distracter was presented than when the ‘feminine’ (\( p < 0.005 \)) or the ‘very masculine’ distracter (\( p < 0.05 \)) was presented.

Figure 4.3 Plot showing the interaction in the mean response times for the four distracter types (very masculine, masculine, feminine, very feminine) across the valid and invalid cueing conditions.
3.2.2 Discussion

In Experiment 4, it was found that both accuracy and timing performance was best on a spatial-cuing task improved when the cue pointed to the same hemifield as the target was located. However, we did not find a main effect between the distracters types used. However, distracter type interacted with performance on the task. The post-hoc analysis showed that differences in RT performance for valid and invalid trials could be attributed to the 'very masculine' and 'very feminine' stimuli. Interestingly, the effect of validity appeared to be most prominent for 'very feminine' distracters. The implication is that for an invalid trial where the distracter is 'very feminine', participants took longer and are less accurate at the task at hand, particularly when the central cue was invalid than valid with the same distracter. Considering that in an invalid trial, the cue points to the distracter stimulus, the results suggest that participants take longer and are less accurate at disengaging their attention from the location of the distracter, particularly the 'very feminine' walker in order to attend to the target stimulus. A similar pattern of behaviour can be seen for
the ‘very masculine’ walkers (see Figure 4.3 & 4.4). Though we do not see a main effect of distracter salience in these results, it should be noted that the efficient allocation of spatial attention to the cued target appears to be more attenuated by the more salient distracters (very masculine, very feminine). This is most apparent in the response time data.

4.3 Experiment 5

As the effects of crossmodal stimuli on spatial attention have been demonstrated in static environments (see Ngo & Spence, 2010 for a recent example), we decided to investigate the effects of task-irrelevant crossmodal dynamic stimuli using the same spatial cuing task, to see how performance compared to unisensory stimuli. As both visual and auditory information can contribute to the perception of the sex of a walker, in the following experiment we decided to add the sound of footsteps to the visual distracter. We repeated the design and procedure from Experiment 4 with one important difference. We presented sound simultaneously with the visual distracter. Based on previous findings that multisensory interactions influence the allocation of spatial attention (McDonald & Ward, 2000; Perrott et al., 1990; Spence & Driver, 1997), our hypothesis was that a multisensory stimulus would capture attention leading to longer RTs and reduced accuracy particularly in the invalid cue conditions relative to Experiment 4.
4.3.1 Method

Participants

Eighteen participants (6 male and 12 female with a mean age of 20.2 years) from the School of Psychology, Trinity College Dublin participated in this experiment for research credits. None of these participants had taken part in the previous study. Their ages ranged from 18 to 30 years. All reported normal or corrected to normal vision and none reported any auditory impairment. Prior approval was obtained from the School of Psychology Ethics Committee. Following a briefing on the experimental protocol, participants provided written consent to take part in the study.

Stimuli and apparatus

The stimuli (i.e. the ‘walkers’ and target array) were the same as those used in Experiment 4 in this Chapter. In the following experiment we also included a spatially concurrent auditory stimulus which was always presented in the same hemifield as the visual distracter. The auditory stimulus was the sound of footsteps. Male and female sounding footsteps were selected from the online database: *Alf free Sound Effects* (http://www.alfreesoundeffects.com). Each sound clip was edited using Adobe Audition, to be 4 seconds in duration and was presented in a standard ‘wav’ format. While each trial lasted a total of 4 seconds (or until the participant responded), the target ‘T’ was only presented in the array for 200ms before turning to a ‘+’ for the remainder of the trial. The trial structure was the same as that used in Experiment 4 with the exception of the added sound. The sound pressure level was approximately 53dbA. Each distracter category (very masculine, masculine, feminine, very feminine) had one corresponding sound (i.e. four footsteps sounds in total – 2 x male and 2 x
female). Two randomly chosen gender-specific footstep sounds were assigned to the male and female walkers. Footstep that sounded like high-heels were classified as being female and two of these were used (one for feminine, and another for very feminine). Footsteps that sounded like men’s dress shoes were classified as being male and two examples were used (one for masculine, and another for very masculine). The timing of the sound of the footsteps was not exactly synchronous with the walking movement of the visual distracters but it was close to the temporal pattern.

*Design & Procedure*

The design was the same as that used in Experiment 4 with the exception that a sound was also presented which was spatially co-located (in the same hemifield) as the visual distracter for all trials. Again participants were asked to ignore the visual and auditory walking stimuli and to conduct the orientation discrimination task only. As in Experiment 4, a trial began with a fixation cross for 500ms, followed by an inter-stimulus interval of 150ms, after which both the target and distracter stimuli were presented simultaneously. The onset of the auditory stimulus (footsteps) was the same as that of the visual distracter and both were presented for 4 seconds or until the participant responded. However, as mentioned previously, the target ‘T’ was presented for only 200ms before being masked by a ‘+’ symbol.

*4.3.2 Results*

The mean accuracy and reaction times were calculated for each of the cue validity and distracter salience conditions. For the valid trials, the mean percentage of
correct responses was 75.55% (std. dev.=3.98%) and the mean RT was 1091ms (std. dev.=256 ms). For the invalid trials the mean percentage of correct responses was 78.2% (std. dev. =3.54%) and the mean RT was 1068ms (std. dev. =271ms). Separate 2x2 repeated measured ANOVAs were performed on the correct responses and RT data with Cue Validity (valid or invalid) and Distracter Salience (very masculine, masculine, feminine, very feminine) as factors.

Response times

The mean response times to each of the valid and invalid trials are depicted in Figure 4.5. For the response time data, we found no main effect of cue validity \([F(1, 17) = 1.2207, p=0.2846]\) and no effect of distracter salience \([F(3, 51) = 1.8314, p=0.153]\). Also, there was no interaction between Cue Validity and Distracter Salience.

![Figure 4.5](image)

Figure 4.5 Plot showing the mean response times for the four audio-visual distracter types (very masculine, masculine, feminine, very feminine) across the valid and invalid cueing conditions.

Correct Responses

The mean correct responses to each of the valid and invalid cueing conditions are depicted in Figure 4.6. We found no main effect of validity \([F(1, 17) = 3.77, p=\)
0.069]. We did however find a main effect of distracter salience \( F(3, 51) = 2.89, p < 0.05 \). A Tukey post-hoc analysis showed significantly greater accuracy for the 'very feminine' stimuli than the 'very masculine stimuli' \( p < 0.05 \). The interaction between cue validity and distracter salience was not significant \( F(3, 51) = 2.71, p = 0.054 \). It was however close and this appeared to be driven by the valid very masculine trials being less accurate than the invalid very feminine trials.

![Figure 4.6](image)

*Figure 4.6*  Plot showing the mean accuracy for the four audio-visual distracter types (very masculine, masculine, feminine, very feminine) across the valid and invalid cueing conditions.

### 4.3.3 Discussion

In Experiment 5 it was found that a concurrently presented audio-visual distracter competed for visual attention in a spatial cuing task. In contrast to Experiment 4, we found no main effect of cue validity on either response times or accuracy. Interestingly, we found that accuracy for the orientation task increased for the invalid cue particularly when a 'very feminine' distracter was presented at the opposite side of space.

The accuracy scores (valid and invalid) for the 'very feminine' distracter present a striking contrast to those found in Experiment 4. This difference in
performance could perhaps be attributed to the presence of sound as this was the principle design difference between the two experiments. It should be noted that the association between the 'footsteps' sounds and each of the four categories of visual distracter stimuli remained constant throughout the experiment. It is possible that the 'footsteps' sound assigned to the 'very feminine' distracter were qualitatively different. The auditory stimulus for the very feminine distracter category may have been less salient than the other sounds, therefore had less of an effect of attentional capture than the other sound conditions leading to relatively better performance. On the other hand, as we did not find an effect of the validity of the central cue in Experiment 5, it is possible that participants began to ignore this visual cue but that their attention was driven by the location of the peripheral sound source only. The sound consistently occurred at the opposite side of space to the visual target. Consequently, the sound may have acted as an auditory cue informing the participant to look in the hemifield opposite to that where the sound was occurring in order to locate the target. The longer reaction times found in Experiment 5 suggest that participants may be processing the auditory stimulus independently of the visual cue and basing their decision regarding the target location on this instead of the visual spatial cue that preceded each trial.

### 4.4 General Discussion

In this study we found an effect of cue validity on task performance when the task was conducted within one modality, i.e. vision, but no effect of distracter salience. When a multisensory distracter was presented, we found no effect of cue validity but an effect of distracter salience. The results of Experiment 4 are consistent with a body of evidence that suggests that presentation of a cue in a given modality
will lead to the automatic orienting of the observer’s attention in the direction of the cue (see Mazza et al., 2007; Ward, 1994).

The effects of the distracter visual stimuli on attentional allocation may be explained by the results of another recent study: the stimuli used in the experiments described in the present chapter were also used in a study investigating the role of dynamic walking patterns on judgements of attractiveness (Maguinness, unpublished). Maguinness found that the ‘very masculine’ and ‘very feminine’ walkers were rated as more attractive than the typical masculine or feminine walkers. It was for this reason that we expected these stimuli to be more likely to capture spatial attention, since Sui & Liu, (2009) found that more attractive stimuli were effectively more distracting then less attractive stimuli when presented in the periphery. The performance differences between the distracter items were mostly evident for invalid trials in Experiment 4 only. When the cue was invalid (assuming that the cue was effective), participants had to reorient their attention to the target. The time taken for this to occur was 198ms longer when a bimodal stimulus was presented as a distracter relative to the visual stimulus alone. Though we did find an effect of distracter salience in Experiment 5, it was not in the direction of the attentional bias that was expected. Instead, when the ‘very feminine’ distracter was presented with a concurrent footsteps sound, accuracy improved when the central cue was invalid. It is unclear why a highly salient distracter would improve target performance unless, as already mentioned, the specific sound associated with the distracter was acting as a peripheral cue which directed the allocation of attention to the opposite side of space than was cued more so than the other sounds used. Participants were likely to have quickly learned that the target consistently occurred at
the opposite side to the sound, and this effect may have been enhanced by a highly salient distracter sound.

The design of this study may also offer some insight as to why such an unexpected result has occurred. Participants had 200ms in which to view the orientation of the letter ‘T’. This exposure time was justified by the duration of a single eye fixation usually exceeding 200ms (Rayner, 1983) and it was preferable that saccadic orienting would not occur. After 200ms, the upright or inverted ‘T’ turned to a ‘+’. However, for the motion information in the visual distracter stimulus and the semantic content of the auditory stimulus to be perceived, it was necessary to present both for longer periods during both experiments. These distracter stimuli were presented for up to 4 seconds or until the participant responded. The duration of the target stimulus was the same across both experiments (200ms) but the array (without the target) remained present for the duration of the trial. Even though they were requested to respond as quickly and as accurately as possible, it would appear that attention is being reoriented to dynamic distracter stimuli after an orientation judgement has been made but before a response is input. This may help explain the relatively slower RTs in Experiment 5 as the multisensory stimulus may be delaying the response even further after the orientation judgement has been made.

While the results of Experiment 4 fit well within the predictions of Posner’s cuing paradigm, with improved performance for the location where the target it cued, the results of Experiment 5 are in conflict with this prediction. The latter findings can perhaps better be explained by Shaw (1984)’s argument that spatial cuing results in different decision making strategies at the cued and un-cued locations. The strategy
here may be to allow the auditory signal to guide attention which in this case was a reliable cue as the location of the target (which was always in the opposite hemifield).

Potential improvements could be made to a future investigation of the effect of a multisensory distracter on spatial attention (Experiment 5). Firstly, manipulating the spatial location of the sound, so that on some trials the sound was congruent with the target instead of the distracter, would help exclude the possibility that sound was cuing the location of the target. Also, it would be useful to implement a version of the study where sound was the only distracter (i.e. without the walkers). From this, we could determine if the effects in Experiment 5 are as a result the distracter being multisensory and whether this is capturing attention more effectively than either the auditory or visual distracter in isolation.
5.0 An investigation of the role of directionally congruent sound on visual search in dynamic scenes

Abstract

Although performance benefits have been found in search tasks where auditory and visual stimuli occur at the same location (e.g. Perrott et al., 1990; Spence & Driver, 1997), these benefits almost exclusively incorporate static events and scenes. The present study looked at the effects of a moving sound on performance in a dynamic visual search task. The task was to indicate the presence or absence of a dynamic visual target across various set sizes (6, 12, 18) while an accompanying sound was either directionally congruent, directionally incongruent or absent. Items in a scene could move either horizontally or vertically. For targets that moved horizontally, a directionally congruent sound was one in which the envelope of the sound panned from the side of space that the target was moving from to the side that the target was moving towards. The temporal characteristics of this sound were manipulated to match that of the moving object. For objects moving on the vertical axis, an auditory frequency shift was used: an increase in frequency was used for objects moving up and a decrease was used for objects moving down. Incongruent sounds could be any of the alternate sound directions not associated with the target object. In Experiment 6, we found that an incongruent sound benefitted search performance for a target object over other sound conditions. In a second experiment, we manipulated the increased the proportion of sound congruent relative to sound incongruent trials. Again, we failed to find evidence that a directionally congruent sound benefitted search performance. Nevertheless, overall we found that the presence of sound benefitted target detection but in an unpredictable manner.
5.1 Introduction

Sound emitted from a moving object can help us determine the direction that the object is coming from. A common example is traffic sounds that help us judge the direction and distance of oncoming traffic before we cross the road. In doing this, we are combining dynamic auditory and visual information to form a coherent percept of our continually changing environment. When investigating crossmodal enhancements resulting from the integration of auditory and visual stimuli, intersensory conflict is often used. For example, as was shown in the experiments reported in Chapter 3, auditory and visual signals occurring at the same side of space are more likely to guide attention to that side of space relative to those that are presented on opposing sides of space. However, intersensory conflict is also used to demonstrate how one sensory modality can dominate over the other. Perhaps the most common examples of this would be the ventriloquist effect (Howard & Templeton, 1966) and McGurk illusion (McGurk & MacDonald, 1976). Both effects are usually investigated in the context of static auditory events that imply no apparent auditory motion even if their visual counterpart does. It is as yet unclear whether multisensory interactions occur in our perception of motion and if these interactions are comparable to those occurring in static audio-visual event.

In a previous study into intersensory conflict, Zaparoli & Reatto (1969) reported that when presented with incongruent visual and auditory apparent motion streams, some observers experienced the bimodal stimuli moving in a unified trajectory. Other studies such as that by Anstis (1973), also found that motion in one modality could cause illusory motion in another modality. In that study, participants wore a microphone on each hand for an extended period, which was connected to
stereo headphones. The purpose was to temporarily shift the location of the participant’s ears. It was found that hand movements (when the subjects’ eyes were closed), resulted in apparent movements of sound sources. While this was a subjective perceptual experience, it does provide an interesting insight into audio-visual motion conflict and suggests that multisensory interactions are likely to occur in relatively complex experiments.

Allen and Kolers (1981) reported that a correctly timed visual flash could facilitate the perception of auditory apparent motion. These crossmodal effects of vision on the direction of apparent auditory motion were found to be quite strong whereas the opposite effects of audition on vision were found to be weak. The Allen & Kolers study demonstrated a crossmodal interaction between audition and vision in motion perception independent of directional congruency between the two modalities. The present study looked for evidence of a multisensory benefit when presenting directionally congruent bi-modal stimuli.

Soto-Faraco, Lyons, Gazzaniga, Spence & Kingstone (2002) designed a study which investigated multisensory integration of dynamic information. The factors examined were synchrony (sound and light presented synchronously or asynchronously) and congruency (sound and light moving in the same or opposite directions). The stimuli were delivered via loudspeakers onto which LEDs were placed in front. Participants were presented with two apparent motion streams (auditory and visual) and asked to indicate the direction of the auditory stream. Soto-Faraco et al. (2002) found that temporally coincident lights moving in the opposite direction to the sound reversed the direction in which the sound appeared to move. This effect was not present in asynchronous trials indicating that temporal coincidence is important for the dynamic capture of motion information. If vision can
influence the perceived direction of an auditory stimulus, it is logical to wonder whether an auditory stimulus can influence the perceived direction or indeed increase the salience of a visual stimulus. The present study differs from that of Soto-Faraco et al. (2002) in that it used virtual stimuli rather than LEDs. Also, the task in the Soto-Faraco et al. study was to indicate the direction of a target whereas we investigated the location of a moving visual target in a dynamic scene while directionally congruent or incongruent auditory stimuli were presented.

In the following experiment, we used a visual search paradigm to investigate how audition might influence performance in a dynamic search task. Our hypothesis was that a directionally congruent auditory stimulus would aid the detection of a dynamic visual target. A limitation of many studies in audio-visual spatial perception is the spatial coincidence of the stimuli. However, if sound is delivered via headphones, efforts can then be made to simulate auditory motion using interaural time and level differences. The present study used virtual visual stimuli and interaural level differences to simulate auditory motion.

5.2 Experiment 6

The following experiment investigated the role of directionally congruent sounds in the search for a visual target in complex dynamic scenes. We looked specifically at how a directionally informative sound can facilitate visual target detection in a series of randomly presented dynamic scenes. Four display sizes were used. The task for the participant was to indicate the presence or absence of a predefined target object in the scene. Our hypothesis was that when the auditory cue was directionally congruent with the dynamic visual target that performance in locating the target would be improved relative to a directionally incongruent sound or the
absence of sound. Performance was measured in terms of response speed and accuracy in detecting the presence or absence of a visual target in each scene.

5.2.1 Method

Participants

An opportune sample of 130 participants (with a mean age of 37 years and including 74 males) who were attendees at an exhibition in the Science Gallery, Trinity College Dublin took part in this experiment. All reported normal or corrected to normal vision. Prior approval was obtained from the School of Psychology Ethics Committee. Following a briefing on the experimental protocol, participants provided written consent to take part in the study.

Stimuli

The stimuli used in this experiment were a set of dynamic visual scenes. In the scenes were virtual images of moving cars. These scenes were developed using Adobe After Effects (motion graphics application). Each scene depicted a set of horizontally and vertically moving virtual ‘cars’ with the constraint that none of the cars collided or overlapped each other. Four target stimuli were used: a ‘large green car’, a ‘small red car’, a ‘large red car’ and a ‘small green car’. The large cars differed from each other only in colour. The same was the case for the small cars. Targets were resized to subtend a visual angle of 0.23° (width) x 0.64° (length) for the large car and 0.34° (width) x 0.42° (length) for the small car. Each scene fitted into a 1024 x 768 pixel frame which subtended visual angles of 36.12 degrees horizontally and 27.09 degrees vertically. Examples of scenes and target objects are provided in Figure 5. Each scene was presented for 8 seconds (or until the participant responded). The
graphical representations of virtual ‘cars’ were sourced from an online graphics database SXC (http://www.sxc.hu).

The target ‘car’ and the distracters moved either left or right on the horizontal axis, and either up or down on the vertical axis and always crossed the central fixation point. Two speeds were used for the target cars: fast (4°/sec) and slow (1.25°/sec). The speed of distracter cars was assigned randomly.

The auditory stimulus was a sine wave presented at varying frequencies and was generated using NCH Tone Generator software. For directionally congruent targets which were moving horizontally, a continuous tone was used. Auditory motion
for these targets was simulated by dynamically manipulating the interaural level difference between the visual target’s start and end point. For directionally congruent vertically moving targets, a frequency shift was used (e.g. 300-450 for a slow target moving up). A detailed description of the tones used can be found in Table 2.0. The sound intensity level of the auditory stimuli was kept constant across all stimuli and all participants, at an average of 12dB SPL. A pilot study was conducted to confirm that that participants could accurately determine directional information from the sounds used (see Appendix A).

<table>
<thead>
<tr>
<th></th>
<th>Horizontal (Hz)</th>
<th>Vertical (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Left</td>
<td>Right</td>
</tr>
<tr>
<td>Fast</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>Slow</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

Table 2.0 describes the frequencies used for congruent horizontally and vertically moving targets. Horizontally moving targets had a continuous tone that was manipulated using interaural level difference (ILD) whereas vertically moving targets used a frequency shift as described above.

**Apparatus**

The experiment was programmed using a Presentation™ (Neurobehavioral Systems®) script which ran on an Intel® Apple Macbook with 1GB RAM and an external 20” LCD Monitor. The scenes were presented with a 1024 x 768 pixel resolution. Sennheiser HD headphones were used to transmit the auditory information. Correct responses and reaction time data were recorded using a standard PC keyboard. The participant was seated so that their head was positioned at about 60cm from the monitor and they were asked to maintain that position for the duration of the experiment. At this distance, each visual scene subtended a visual angle of 100°.
36.12° in the horizontal axis and 27.09° in the vertical axis. Testing took place in a public setting: participants were seated at a three-sided testing station/sound attenuated booth and members of the public could watch the experiment from a distance. The headphones used were sound-attenuating to minimise any distractions.

Design

This experiment was based on a 3x2x4 within subjects design with directional congruency of the sound (congruent, incongruent or absent), speed of target (fast or slow), and number of distracters (6, 12, 18 and 24) as factors. In the directionally congruent sound condition, a congruent sound was one in which the interaural level difference was manipulated to match the direction of the moving visual target. A directionally incongruent sound was one in which the interaural level difference did not correspond with the direction of the target but instead corresponded with any one of the other directions (i.e. a direction in which a distracter object was moving). The proportion of trials was divided evenly across target present (50%) and absent (50%). Within the target present trials, the proportion of congruent, incongruent and sound absent trials were divided evenly (32 trials per condition). This led to a total of 96 target present and 96 target absent trials in the experiment. For each participant, trials were presented in a randomised order. However, each participant was presented with one of the four target types and targets were counterbalanced across participants. Target direction was nested within the congruency condition. The dependent variables were response time and accuracy.

Procedure

In the experiment, a trial began with a central fixation cross which was presented for 500ms. A scene was then presented for 8 seconds or until the participant
responded. The task for the participant was to search for the moving target in a scene and to indicate, as fast and as accurately as possible, the presence or absence of the visual target by pressing the corresponding key (‘/’ = present, z = absent). The direction of the auditory stimulus was determined for each trial according to the design of the experiment. A subsequent trial occurred only when the participant had responded to the previous trial. Both accuracy and response times were recorded. Participants could take a self-timed break between blocks. The experiment took approximately 15-20 minutes for each participant to complete.

5.2.2 Results

The results of a pilot study in which sound direction was discriminated showed that the mean accuracy performance across all participants was 71.2% which was above chance level (25%). Thus the ability to accurately identify the direction of the auditory motion stimuli was ensured (see Appendix A). As frequencies below 400Hz are more difficult to localise than those above 400Hz, and the frequency of the auditory stimuli used for the slow targets was below 325hz, the decision was made to analyse the fast and slow targets separately. Furthermore, due to a programming error which caused an incorrect number of trials to occur for the largest display size with 24 distracters, this display size was excluded from further analysis.

Fast Targets

The mean accuracy and reaction times were calculated for each of the directional congruency and display size conditions. Performance in all conditions was greater than chance level of 50%. For the target present trials, the mean correct
response was 94.04% (std. dev. = 5.52%) and the mean RT was 989ms (std. dev. = 173 ms). For target absent trials the mean correct response was 97.81% (std. dev. = 4.03%) and the mean RT was 1472ms (std. dev. = 289ms). Only the target present trials were subjected to further analyses. Separate 3x3 repeated measures ANOVAs were performed on the correct responses and RT data with Directional Congruency (congruent, incongruent or sound absent) and Number of Distracters (8, 12, or 18) as factors.

Response Times for Fast Targets

For the response time data, we found a main effect of directional congruency \([F(2, 258) = 101.740, \ p < 0.001]\). A Tukey post-hoc analysis revealed that search times were fastest when sound was absent and slowest when the sound was directionally congruent, as can be seen in Figure 5.1a. There was a main effect of display size \([F(2, 258) = 19.524, \ p < 0.001]\). A Tukey post-hoc showed slowest RTs when 18 distracters were presented and fastest RTs when 6 distracters were presented.

A significant interaction was found between congruency and number of distracters \([F(4, 516) = 16.566, \ p < 0.001]\). A Tukey Post-hoc analysis revealed that RTs to congruent trials when 6 distracters were presented were significantly faster than congruent (p<0.0001) and incongruent (p<0.0001) trials when 12 distracters were presented, and than congruent (p<0.0001), incongruent (p<0.0001) and sound absent (p<0.0001) trials when 18 distracters were presented. RTs to incongruent trials when 6 distracters were presented were significantly slower than sound absent (p<0.0001) when the same number of distracters were presented, and than congruent (p<0.0001) and sound absent (p<0.0001) trials when 12 distracters were presented, and than congruent (p<0.0001), incongruent (p<0.0001) and sound absent (p<0.0001) trials when 18 distracters were presented. RTs to sound absent trials when 6
distracters were presented were significantly faster than congruent (p<0.0001) and incongruent (p<0.0001) trials when 12 distracters were presented, and than congruent (p<0.0001), incongruent (p<0.0001) and sound absent (p<0.0001) trials when 18 distracters were presented. When 12 distracters were presented, RTs to sound absent trials were significantly faster than those to congruent (p<0.0001) and incongruent sound trials (p<0.0001). When 18 distracters were presented, RTs to incongruent trials were significantly slower than congruent (p<0.001) and sound absent trials (p<0.05) with the same number of distracters, but faster than sound absent trials when twelve distracters were presented (p<0.001).

a)
In terms of accuracy, we found no effect of directional congruency \( F(2, 258) = 0.91, \ p = 0.4 \). We did find a main effect of display size \( F(2, 258) = 7.04, \ p < 0.01 \). A Tukey post-hoc analysis of this effect revealed greater accuracy for displays with 6 distracters than displays with 18 distracters \( (p<0.005) \). Greater accuracy also occurred for display sizes with 12 distracters relative to those with 18 distracters \( (p<0.005) \). A significant interaction was found between congruency and number of distracters \( F(4, 516) = 6.93, \ p < 0.0001 \) as can be seen in Figure 5.1b.

A Tukey post-hoc analysis of this interaction revealed that responses to incongruent trials when 6 distracters were presented were significantly less accurate than sound absent trials when the same number of distracters were presented \( (p<0.0001) \) and sound absent trials when 18 distracters were presented \( (p<0.0001) \). Responses to sound absent trials where 6 distracters were presented were also found to be more accurate than congruent \( (p<0.001) \) and incongruent \( (p<0.05) \) trials where
12 distracters were presented, and more accurate than congruent (p<0.01) and sound absent (p<0.05) trials when 18 distracters were presented. For displays where 12 distracters were presented, accuracy was greater for incongruent than congruent trials (p<0.01). Responses for sound absent trials where 12 distracters were presented were more accurate than those for congruent trials where 18 distracters were presented (p<0.05).

**Slow Targets**

The mean accuracy and reaction times were calculated for the directional congruency and display size conditions. Performance in all conditions was greater than chance (50%). For the target present trials, the mean correct response was 92.44% (std. dev. =6.23%) and the mean RT was 877ms (std. dev. =150 ms). For target absent trials the mean correct response was 97.81% (std. dev. =4.03%) and the mean RT was 1423ms (std. dev. =276ms). Only the target present trials were subjected to further analyses. Separate 3x3 repeated measured ANOVAs were performed on the correct responses and RT data with Directional Congruency (congruent, incongruent or sound absent) and Number of Distracters (8, 12, or 18) as factors.

**Response Times for Slow Targets**

For the response time data, we found a main effect of directional congruency \([F (2, 258) = 150.740, \ p < 0.0001]\). Overall, RTs were fastest when the sound was directionally incongruent and slowest when the sound was directionally congruent. We also found a main effect of display size \([F (2, 258) = 17.932, \ p < 0.0001]\). A Tukey post-hoc analysis showed that RTs to displays with 6 distracters were significantly
faster than those with either 12 (p<0.005) or 18 (p<0.0001) distracters. RTs to displays with 12 distracters were also found to be faster than those with 18 distracters (p<0.05). A significant interaction was found between congruency and number of distracters $[F(4, 516)= 27.159, \ p < 0.0001]$ as can be seen in Figure 5.2a.

A Tukey post-hoc analysis on the interaction revealed that RTs to the sound congruent trials when 6 distracters were presented were significantly slower than those to the incongruent (p<0.0001) or sound absent (p<0.0001) trials when the same number of distracters were presented. RTs to the congruent trials when 6 distracters were presented were significantly faster than congruent trials when 12 distracters were presented (p<0.0001), and than congruent (p<0.0001), incongruent (p<0.0001) and sound absent (p<0.05) trials when 18 distracters were presented.

RTs to the sound incongruent trials when 6 distracters were presented were significantly faster than those to the sound congruent trials with the same number of distracters (p<0.0001). RTs to the sound incongruent trials when 6 distracters were presented were also significantly faster than responses to the congruent (p<0.0001) and sound absent (p<0.0001) trials when 12 distracters were presented and congruent (p<0.0001), incongruent (p<0.0001) and sound absent (p<0.0001) trials when 18 distracters are presented. RTs to the sound absent trials when 6 distracters were presented were also significantly faster than those to the sound congruent trials with the same number of distracters (p<0.0001), and also to the sound congruent (p<0.0001) and sound absent (p<0.0001) trials when 12 distracters were presented. RTs to the sound absent trials were faster to the congruent (p<0.0001), incongruent (p<0.0001) and sound absent (p<0.0001) trials when 18 distracters were presented.
For trials with 12 distracters, RTs to incongruent trials were faster than those to congruent trials ($p<0.0001$). For trials with 18 distracters, RTs to the sound absent trials were faster than those to the sound congruent ($p<0.0001$) or incongruent ($p<0.0001$) trials with the same number of distracters, and slower than to the incongruent trials when the number of distracters was 12.
Figure 5.2 Plot showing the a) mean response times and b) mean accuracy to locating the target in each sound condition and across each set size for fast moving targets. ‘Congruent’ means the auditory cue was directionally congruent with the visual target, ‘Incongruent’ means the auditory directionally incongruent with the visual target and ‘Sound Absent’ refers to the no sound condition.
Correct Responses for Slow Targets

In terms of accuracy, we found an effect of directional congruency \( F_{2, 258} = 10.04, \ p < 0.0001 \): responses to the sound absent condition were significantly more accurate than those to the sound congruent \( p < 0.001 \) condition and responses to the sound incongruent condition were found to be more accurate than those to sound absent condition \( p < 0.001 \). We found a main effect of display size \( F_{2, 258} = 6.80, \ p < 0.01 \) with greater accuracy for displays with 6 distracters relative to those with 12 distracters \( p < 0.01 \) and greater accuracy for displays with 12 distracters than displays with 18 distracters \( p < 0.01 \). No interaction was found between sound congruency and display size.

5.2.3 Discussion

In Experiment 6, it was found that for fast moving visual targets, participants responded more quickly and were most accurate when the sound was absent, except in scenes with a large number of distracters (i.e. 18) where an incongruent sound yielded best performance. For slow moving targets, a directionally incongruent sound improved search performance for the visual target relative to congruent and sound absent trials.

For fast moving targets, it would appear that the presence of sound in general (i.e. whether congruent or incongruent) was resulting in a cost to search performance as the absence of sound benefitted RTs and accuracy dramatically for the display sizes with six and twelve distracters. In contrast the presence of sound improved performance for the largest display size. However, interestingly, it was the
directionally incongruent sound that yielded best performance in this condition rather than the expected congruent sound. It may be the case that participants were relying on visual cues alone for the smaller display sizes, but as the task increases in difficulty (i.e. no of distracters increase), auditory cues were also being used. The results show no difference between performance for congruent and sound absent trials for the largest display size indicating that there was something uniquely useful about the directionally incongruent sound in this condition.

One possible explanation for this result would be that for the largest display size target motion in the visual modality could have caused an illusory motion in the auditory modality similar to that found by Anstis (1973). If this were the case, opposing or incongruent audio-visual signals would be processed as a unified percept with vision dominating the overall directional judgement of both events. However, this idea would predict equal performance across the sound congruent and incongruent conditions, which was not found. Therefore, although visual capture of sound direction may have occurred, this does not explain why there was a relative benefit on performance when sound was in the opposite direction to the visual target than when it was in the same direction.

For the slow targets, the incongruent sound benefited performance relative to congruent sound for the displays with 6 and 12 targets but was least beneficial for the largest display size. If we consider that participants were attending to the sound for the largest display size, then the differences across performance for the fast and slow targets within this condition may be explained by the information in the corresponding fast and slow sounds. Mean response times were below 1 second for both the fast and slow targets. Within this time period, a fast auditory motion stream is more informative than a slow auditory stream as more (virtual) distance has been
travelled by fast than the slow sound making it easier to resolve. For example, in a vertically moving sound a greater frequency range is covered (see Table 2.0) within this time frame. This may account for the shift in performance between target speeds.

As an effect favouring the incongruent auditory condition was an unexpected result in this experiment, a follow up study was run where the number of congruent trials was increased relative to the number of incongruent trials. In the present experiment, a directionally congruent sound was equally likely to occur as a directionally incongruent sound and this may have affected the results. The purpose of the following experiment was to increase the likelihood of a directionally congruent sound being useful in searching for the visual target in the scenes.

5.4 Experiment 7

5.4.1 Method

Participants

Eighteen participants (8 male and 10 female with a mean age of 25 years) from the School of Psychology, Trinity College Dublin participated in this experiment for research credits. Their ages ranged from 18 to 59 years. All reported normal or corrected to normal vision and none reported any auditory impairment. Prior approval was obtained from the School of Psychology Ethics Committee. Following a briefing on the experimental protocol, participants provided written consent to take part in the study.

Stimuli and apparatus

The stimuli (i.e. the images of 'virtual cars' and the associated sounds) and apparatus were the same as those used in Experiment 6 in this chapter. Testing took
place in a behavioural testing laboratory in the Institute of Neuroscience, Trinity College.

Design and Procedure

The design was the same as that used in Experiment 6 with the exception that instead of presenting an equal number of directionally congruent, incongruent and sound absent trials, we increased the relative number of congruent trials presented. The target present trials were now presented according to the following ratio: congruent (2): incongruent (1): sound absent (3). This resulted in twice as many congruent trials as incongruent trials thus increasing the likelihood of the congruent sound being useful in the search task relative to the incongruent sound. Overall, there were as many target present trials which were accompanied by sound as those which were not. The experimental procedure was the same as that used in Experiment 6.

5.4.2 Results

For target absent trials the mean correct response was 96.88% (std. dev. = 2%) and the mean RT was 1660 ms (std. dev. =297 ms). For the target present trials, the mean correct response was 94.04% (std. dev. = 5.5%) and the mean RT was 989 ms (std. dev. =150 ms). Only the target present trials were subjected to further analyses.

Fast Targets

The mean accuracy and reaction times were calculated for the directional congruency of the sound and display size conditions. Performance in all conditions was greater than chance (50%). For target present trials, the mean correct response to locating the fast target was 92.28% (std. dev. =5.31%) and the mean RT was 1157 ms
(std. dev. = 163 ms). Separate 3x3 repeated measured ANOVAs were performed on the correct responses and RT data with Directional Congruency of the sound to the visual target (congruent, incongruent or sound absent) and Number of Distracters (6, 12, or 18) as factors.

Response Times for Fast Targets

For the response time data, we found a main effect of directional congruency \( F(2, 34) = 41.8115, \ p < 0.0001 \). A Tukey post-hoc analysis revealed that RTs to the congruent sound trials were faster than those to the incongruent sound trials (\( p < 0.0001 \)). However, response times to the target in the sound absent trials were found to be significantly faster than to either the congruent (\( p < 0.0001 \)) or incongruent (\( p < 0.0001 \)) sound conditions as can be seen in Figure 5.3a. We also found a main effect of display size \( F(2, 34) = 8.6723, \ p < 0.0001 \). A Tukey post-hoc analysis revealed that RTs to displays with 6 distracters were significantly faster than to those with 18 distracters (\( p < 0.01 \)) and RTs to displays with 12 distracters were significantly faster than RTs to displays with 18 distracters (\( p < 0.01 \)). The interaction between sound congruency and number of distracters was not significant \( F(4, 68) = 1.7571, \ p = 0.47675 \).

Correct Responses for Fast Targets

In terms of accuracy, we found an effect of directional congruency \( F(2, 34) = 5.834, \ p < 0.01 \) of the sound relative to the visual target. Search times to the visual target when sound was absent were significantly more accurate than those when the sound was congruent (\( p < 0.01 \)) or incongruent (\( p < 0.05 \)) to the direction of the moving
visual target. No difference was found between responses to the congruent and incongruent sound trials. As can be seen in Figure 5.3b, we did find a main effect of display size \( F(2, 34) = 6.994, p < 0.01 \) with greater accuracy for displays with 12 distracters relative to those with 18 distracters \( (p<0.01) \). The interaction between sound congruency and number of distracters was not significant \( F(4, 68) = 1.243, p = 0.3009 \).

Figure 5.3 Plot showing the a) mean response times and b) mean accuracy to locating the target in each sound condition and across each set size for fast moving visual targets. ‘Congruent’ means the auditory cue was directionally congruent with the visual target, ‘Incongruent’ means the auditory directionally incongruent with the visual target and ‘Sound Absent’ refers to the no sound condition.
Slow Targets

The mean accuracy and reaction times were calculated for the directional congruency and display size conditions. Performance in all conditions was greater than chance (50%). For target present trials, the mean correct response was 94.75% (std. dev. = 4.03%) and the mean RT was 1005 ms (std. dev. = 173 ms). Separate 3x3 repeated measured ANOVAs were performed on the correct responses and RT data with the directional congruency of the sound relative to the visual target (congruent, incongruent or sound absent) and number of distracters (8, 12, or 18) as factors.

Response Times for Slow Targets

For the response time data, as depicted in Figure 5.4a, we found a main effect of directional congruency \[ F(2, 34) = 14.3635, \ p < 0.0001 \]. A Tukey post-hoc analysis revealed that RTs to locating the visual target accompanied by the congruent sound were slower than those to the incongruent (p<0.001) and sound absent conditions (p<0.001). We also found a main effect of display size \[ F(2, 34) = 23.6370, \ p < 0.0001 \]. A Tukey post-hoc analysis revealed that RTs to displays with 6 distracters were faster than those with 12 (p<0.05) and 18 (p<0.001) distracters. RTs to displays with 12 distracters were found to be significantly faster than those with eighteen distracters (p<0.05). The interaction between sound congruency and number of distracters was not significant \[ F(4, 68) = 1.4499, \ p = 0.2271 \].

Correct Responses for Slow Targets

In terms accuracy, there was no main effect of directional congruency \[ F(2, 34) = 0.988, \ p = 0.3828 \] or display size \[ F(2, 34) = 1.551, \ p = 0.2267 \] when searching for
a slow moving target (see Figure 5.4b). The interaction between sound congruency and number of distracters was not significant \( F(4, 68) = 0.773, p = 0.5462 \).

\[ \text{Figure 5.4 Plot showing the a) mean response times and b) mean accuracy to locating the target in each sound condition and across each set size for slow moving targets.} \]
5.4.3 Discussion

Similar to the results found in Experiment 6, RTs and performance accuracy in searching for a moving target when sound was absent was more efficient relative to sound present trials. Unlike the previous experiment, this effect was consistent across the fast and slow moving targets. However, while the time to locate a fast moving visual target was faster when the scene was accompanied by a sound that was congruent relative to one which was incongruent to the direction of the visual target this effect was reversed for slow moving targets (see Figure 5.4 a and b). Therefore, as was also found in Experiment 6, the presence of sound appeared to be associated with a deterioration in search performance relative to the absence of the sound. This discrepancy may have been due to the relative numbers of trials in each of the sound conditions: although there were equal numbers of sound present to sound absent trials, the sound present trials were more variable in that the sound could be either congruent or incongruent with the direction of the moving visual target. Although there were more congruent than incongruent sound trials in Experiment 7 (relative to Experiment 6), this qualitative difference between the sound present and absent conditions may have had the effect of making the sound a less reliable source of information when present, thus affecting less efficient performance in that condition.

Surprisingly, for fast targets, RTs to the largest display size do not differ whether the sound is present or absent. However, for the same display size, a 10% decrease in accuracy can be seen when searching for a visual target accompanied by an incongruent sound. The differences in accuracy for the fast targets accompanied by congruent or incongruent sounds are consistent for displays with 6 and 12 distracters but the difference increases dramatically for the largest number of distracters.
suggesting that display size is a major factor in determining the usefulness of the sound.

The results for the slow targets in Experiment 7 tell a different story than the results to the fast targets. Interestingly, RTs are slowest when locating a visual target accompanied by a congruent sound relative to the incongruent or sound absent conditions. Also, it appears that the largest difference between the congruent and incongruent sound conditions occurred for displays with 12 distracters. Perhaps for the smallest number of distracters, participants are relying on the visual stimulus alone. In a relatively uncluttered dynamic scene with only 6 distracters, the task of finding a visual target is relatively easy, therefore attending to a sound may not further facilitate this search performance. However, in a scene with 12 distracters, the task is more difficult and the presence of sound may have been used to guide the search. For the most cluttered scenes with the largest number of distracters, it may be that participants find it difficult to attend to more than just the visual stimuli when searching for the target and as a result perform best on the sound absent condition.

5.5 General Discussion

The results of Experiment 6 and 7 suggest that the presence of an auditory motion stream during a dynamic visual search task can hinder performance relative to the absence of such a sound. Curiously, it was found that for slow moving targets in particular, a directionally incongruent sound could reduce search latencies and improve accuracy relative to a congruent sound.

The contrasting performance found for fast and slow moving targets implies that the differences in target velocity across the experiments can be easily discriminated. As the participants’ ability to detect the motion direction in the
auditory signal had been pre-determined in a pilot study (which used the same sounds as were used in the present study), it can be assumed that this directional information was still discriminable in the audio-visual version of the present experiments. Lopez-Moliner & Soto-Faraco (Lopez-Moliner & Soto-Faraco, 2007) investigated how sensory sources containing different velocity information influenced each other. Participants were presented with directional sound (left to right or right to left) similar to that used in the experiments reported above. The velocity of the sounds were varied and a staircase procedure was used to determine which sounds corresponded to which visual objects. The results suggested that a shift in the perceived velocity of sound occurred as a function of the velocity of the concurrent visual stimulus. In essence, a fast moving target had the effect that sounds were perceived to be moving more quickly than they actually were. This visual dominance in motion perception may help to explain why incongruent sounds have been helpful in this study. Perhaps the auditory information is dominated by the visual signal resulting in a shift in the apparent speed or direction of the incongruent sounds to something more coherent within the scene, thus increasing their usefulness.

In order to understand the differences in performance across the fast and slow moving targets, we should consider the fact that the frequencies used for the congruent slow moving targets were considerably lower (350hz and below) and more difficult to localize than those used for congruent fast moving targets (see results of the pilot study in Appendix A). However, an incongruent sound could be any of the 6 alternate sounds used in the experiment. For example, for a slow horizontally moving target, a directionally incongruent sound may be a fast moving vertical sound. Thus, for slow moving visual targets, a higher frequency sound (i.e. fast) when present may be acting as an alerting mechanism resulting in improved search performance. This
alerting, caused by the inter-sensory conflict in incongruent trials may have resulted in faster target detection. Bertelson & Aschersleben (Bertelson & Aschersleben, 1998) described how a cognitive bias can occur whereby participants consciously or unconsciously adopt strategies based on their awareness of the conflicting nature of some trials. The performance on incongruent trials in this study may be an example of this in that participants may have been consciously or unconsciously aware of the incongruent trials in this study and adapted their response behaviour accordingly resulting in improved performance for these trials.

An alternative explanation for the overall improved performance for the trials in which the direction of the sound was incongruent to that of the visual target relative to congruent trials, could be provided by the distracters and their association with the incongruent sound. As was the design of the experiment, in a sound incongruent trial, the auditory stimulus, though not directionally congruent with the visual target, was nevertheless congruent with a number of distracters present in the scene. The number of distracters increases with display size, therefore the number of associated items which are actually moving in a direction which is the same as the 'incongruent' sound also increased. It may therefore be the case that when the sound direction is incongruent with the target, the distracters items which are moving in a direction compatible with the incongruent sound are grouped together by the sound. Since there was a benefit for the incongruent sound trials on search performance, this suggests that those items which were grouped were subsequently excluded from the search, allowing the participant to search within the remaining subset of items, including the target. In a study by Humphreys, Quinlan, & Riddoch (1989), it was reported that distracters which are homogenous (i.e. highly similar) can be grouped together, thus separating them from the visual target. Such grouping may have occurred for
distracter items that were moving in the same direction as the incongruent sound. With such a strategy, the target itself would consequently be excluded from the search process when the sound direction was congruent, thus leading to relatively inefficient search in this condition as was found.

McLeod et al. (1991) described how search for a target that is defined by movement and form is parallel. According to McLeod et al., a movement filter can direct attention to stimuli with common motion characteristics. In this case, it may be that sound is the common characteristic among the incongruent distracter items causing them to be grouped together. Research by McLeod et al. (1988) argues that we can attend to spatially dispersed perceptual groups rather than being limited to a small region of visual space. In the present context, this might suggest that our distractors could be anywhere within the scene and still be grouped together. Also, the effects of perceptual grouping in crossmodal dynamic capture as described by Sanabria, Soto-Faraco, & Spence (Sanabria, Soto-Faraco, & Spence, 2004), may help in understanding the integration of directional audio-visual motion streams. Whereas Sanabria et al. found effects of vision on the perceived direction of auditory motion, we may have uncovered an effect of a virtual visual stimulus on the perceived motion of an opposing auditory stream.

A number of studies have investigated the stages at which conflicting cross-sensory processes can occur (see Vroomen, Bertelson, & de Gelder, 2001). For example, in the case of the ventriloquist effect, it appears to occur at a pre-attentive stage. However, other studies such as Vroomen & de Gelder (2000) have reported that a high tone embedded in a series of low tones improved detection of a synchronously presented target. The abruptness of the tone increased the visibility of the target indicating that audiovisual interactions can occur at the level of perceptual
organisation. When determining the stages at which audio-visual sensory interaction occurs, the results are somewhat inconclusive and seem to depend on the stimuli and experimental procedure used. The processing of an inter-sensory conflict between audition and vision particularly in the incongruent trials could be occurring at an early sensory stage, which might explain the relatively fast reaction times where as the processing of the congruent trials may be occurring at the later decision making stage.

Although we did not find a multisensory benefit on visual search for the presentation of audio-visual stimuli, we cannot rule out the possibility that integration occurred during the sound present trials, particularly when considering the performance to the directionally incongruent stimuli. In their study of crossmodal dynamic capture, Soto-Faraco et al. (2004) estimated the magnitude of this effect to be roughly 50% meaning that for conflicting or incongruent trials, the sound direction was judged erroneously half of the time whereas it was near perfect for the congruent trials. Our results suggest that response time and accuracy to congruent trials fall below that of incongruent trials, particularly for slow moving targets. The Soto-Faraco et al. study found that an apparent auditory motion stream could be mislocated toward a concurrent light flash, causing a ventriloquist effect. It is possible that a similar effect occurred in the experiments described above whereby the direction of the incongruent auditory stimulus was somehow biased by the visual signal with the effect that the incongruent sound would not hinder the detection of the visual target. The evidence from both Experiment 6 and 7 suggests that the mechanism by which audio-visual information is processed and integrated in dynamic scenes differs from those of static scenes (see results section of Chapters 2 and 3). While a spatially co-located sound can benefit target detection for static events, it seems that spatial
congruency may operate differently for dynamic events, where conflicting audio-visual stimuli can aid in target detection.

5.5.1 Summary

The results of Experiments 6 & 7 suggest that in visual search for a moving target, the presence of a directional sound can incur a cost in performance. Contrary to expectation, search performance for visual targets accompanied by a directionally incongruent sound was relatively more efficient than when accompanied by a directionally congruent sound, particularly when the visual target was slow moving. These findings indicate fundamental differences in how cross-sensory stimuli are combined in dynamic scenes relative to static scenes. A better understanding of how multisensory benefits found in static audio-visual events translates to dynamic environments is necessary and requires further research.
6.0 An investigation of the role of temporally synchronous audition on visual search in dynamic scenes

Abstract

In this study we investigated whether sound can affect visual search performance when the target was a dynamic object embedded in a dynamic display. The task was to detect presence (and direction) or absence of a moving target. Auditory information was either temporally congruent or temporally incongruent with the intermittent appearance and disappearance of the visual target in the display (or sound was absent). A temporally congruent sound was one which shared the same onset and offset profile as the visual target. A temporally incongruent sound was one which did not share the same temporal profile as the visual target nor the profile of any of the distracter objects. The auditory signal used was a tone and therefore did not provide any top-down information regarding the identity of the visual target nor did it provide any spatial cue to the target’s location. We found that search times to visual targets which were simultaneously presented with a temporally congruent sound were slower than search times to visual targets presented with temporally incongruent sound. Furthermore, the search times to a visual target when no sound was present were faster than to either of the sound conditions. Our findings suggest that sound can affect visual search in dynamic displays and have important implications for our understanding of multisensory influences in target detection in complex dynamic scenes.
6.1 Introduction

Searching for a target in a cluttered scene can be a challenging task. It is even more demanding when the scene in which we are searching is continually changing and items within it continuously moving. In our environment, temporal synchrony across modalities can often provide information about causal relationships and thus reduce perceptual ambiguity. For example, if we see two cars collide and we hear a concurrent crash sound, we are inclined to think the two sensory instances are related (i.e. the sound and sight of the crash). Indeed it has been shown that sound can influence the perception of causal relationships between visually moving objects as has been shown in the Sekuler illusion (Sekuler, Sekuler, & Lau, 1997). Vroomen & de Gelder (2000) also found that the presence of a sound change which was simultaneous with the onset of a visual target improved the ability to detect that target in a rapidly presented sequence of visual distracters. In audiovisual perception, it is often argued that vision dominates audition for spatial tasks (Vroomen et al., 2001) whereas audition dominates vision for temporal tasks (Morein-Zamir et al., 2003; Shams, Kamitani, & Shimojo, 2000). The aim of the studies in this chapter was to investigate the role of temporally synchronous auditory information in a visual search task. More specifically, we wanted to look at the effect of a temporally congruent auditory stimulus on the ability to localise a moving visual target in a dynamic environment.

Principal theories of multisensory integration describe how information may be integrated across sensory modalities, thus enhancing the perception of our environment. However, the ability to attend to one stimulus to the exclusion of another is also of fundamental importance in our day-to-day cognitive functioning. As
stated, coincident auditory and visual information can be perceived to emanate from the same location or event. There is, however, a window of temporal integration for which this holds true. This window is generally thought to be around 25-50ms for integration of basic auditory and visual stimuli (see Zampini, Shore, & Spence, 2003) but it can be as large as 200ms for speech perception (van Wassenhove, Grant, & Poeppel, 2007). The order in which the auditory and visual stimuli occur can also be significant in our experience of a unified percept. Early work by Dixon & Spitz (1980) showed that audiovisual desynchrony is easier to detect when the auditory signal precedes the visual signal.

Other studies have shown that sound can guide our attention towards the location of visual objects within a static scene. These studies have, for the most part, found a benefit on localisation performance when the auditory and visual stimuli occur at the same location (McDonald & Ward, 2000; Perrott & Saberi, 1990; Spence & Driver, 1997). However, evidence has also emerged to suggest that even non-spatial auditory signals can also guide visual attention provided there is a temporal congruency between the two signals. For example, as demonstrated in the case of the ventriloquism effect, the perceived location of an auditory stimulus can be drawn to the location of a visual target (Spence & Driver, 2000). Vroomen and de Gelder (2000) demonstrated how, when synchronised with a visual target, a high tone which was embedded in a sequence of low tones could facilitate the detection of a visual target in a rapid serial visual search task when these cross-modal events were synchronised. In their experiment, visual targets were presented in a sequence of rapidly changing visual distracters. Similar results have suggested that the same is true of within modal auditory attention. For example, Dalton & Lavie (2004) found evidence for auditory attentional capture in that when searching sequences of sounds...
for a target sound (distinguished by frequency, intensity or duration), the presence of an irrelevant distracter sound reduced search performance for the target sound. However, when the same distracter sound synchronised with the presence of the target sound then search performance was facilitated. In the following experiment, we presented dynamic audio-visual stimuli where the sound was temporally synchronous in one condition and temporally asynchronous in another. Van der Burg et al. (2008) reported that the presentation of a temporally synchronous cue facilitated visual search performance relative to performance when no cue was presented. Participants searched for horizontal or vertical line segments among other distracter line segments oriented at +/- 22.5 degrees. An auditory tone (presented over headphones) was, when present, synchronized with the colour change of the target stimulus. For displays consisting of 24-48 items, a temporally synchronous auditory cue reduced search latencies by more than 1000ms. The above research suggests that we should find a benefit on search performance when the presence of the visual target is synchronised with the presence of an irrelevant auditory stimulus.

However, conflicting evidence has surfaced in recent years regarding the role of audiovisual temporal synchrony in speeded search tasks. For example, Fujisaki, Koene, Arnold, Johnston, & Nishida, (2006) showed that the detection of a visual target (singleton) that changed in synchrony with an auditory stimulus was gradually impaired as the number of unsynchronized visual distracters increased. This suggests that a serial search is occurring and performance on these tasks are not automatic or pre-attentive. In contrast to this finding, Van der Burg and colleagues found that an auditory ‘pip’, when synchronized with colour change of a visual singleton facilitated the speed of visual target detection regardless of the number of distracters present in the scene. In other words, the presence of a synchronised auditory stimulus allowed
the visual target to 'pop' out of the scene. This is evidence of a pre-attentive mechanism that occurs when audiovisual events are synchronized and facilitates the spatial localisation of visual objects.

For the most part, studies in the area of visual search and indeed audio-visual perception use sparse displays and simple stimuli. This approach has been important in demonstrating the occurrence of audiovisual interactions and how these can influence temporal and spatial perception. However, in order to further the ecological validity of these studies it is important to investigate whether these audiovisual interactions occur in more realistic environments. As the real world is rarely static, we looked at the role of audition in visual search within continuously changing dynamic environments.

Models of visual search can offer suggestions as to the nature of audiovisual processing in complex, dynamic scenes. The distinction between parallel and serial search has dominated research in visual attention for many years. Feature Integration Theory (FIT) as proposed by Treisman & Gelade (1980) indicates that features such as shape and colour are processed automatically without the use of attention provided the target is uniquely identifiable by one of these features. In the following study, the target objects used could not be defined by any single unique feature in that the targets shared features such as shape and colour with the distracters. However, the visual targets were identifiable through a unique combination of these features. As such, FIT would indicate that conjunction search was necessary where more than one feature such as the object's colour and shape needs to be used to identify it. This would mean integrating colour and shape for each object until the target is located. The resulting serial search performance would mean relatively long search times, with search time increasing the more distracters in the scene. However, as mentioned in the
previous chapters, the FIT did not account for the effect of auditory input on visual search performance. However, as suggested by the research discussed above, a temporally congruent auditory signal is likely to influence the detection of a visual target. Moreover, when audition is synchronised with the visual target, this sound may operate as a unique target feature thus affecting a more parallel search performance or even target pop out.

McLeod et al. (1988) found evidence to suggest that search is parallel when the target is defined by a conjunction of the features such as movement and form. This supports the suggestion that a dynamic visual target with a conjunction of features should pop-out from the scene. Jeremy Wolfe’s Guided Search model (Wolfe et al., 1989) further suggests that information from top-down and bottom-up processing of the stimulus is used to create a ranking of items in the scene in order of their attentional priority. While this model does not account for multisensory influences, there is a possibility that in the case of visual search when sound is present, the auditory signal feeds into the bottom up processing of the scene, ranking the visual target higher (when accompanied by sound) relative to the non-synchronised distracters and enabling more accurate and faster detection.

6.1.2 Outline of the Experiment

The following experiment investigated that role of audio temporal synchrony in the detection of a visual target within a dynamic scene. Specifically, we examined the effect of temporally synchronous, non-spatial auditory cues on the efficiency with which a visual target was located compared to temporally asynchronous sound or the absence of sound. The target was either a small red object or a large green object and could appear in any location in the scene. Target cars were intermittently occluded by
bridges randomly placed in the scene. In the temporally synchronous condition, the sound (a pure tone) was audible when the target was visible and inaudible when the target was occluded. In the temporally asynchronous condition, the presence of auditory stimulus and the visibility of the visual target did not share the same temporal profile. Our hypothesis was that for visual targets that shared the same temporal profile as the auditory stimulus, search performance would be improved. Performance was measured in terms of participant’s ability to quickly and accurately detect the presence or absence of a visual target.

6.2 Experiment 8

6.2.1 Method

Participants

An opportune sample of 24 participants (9 male and 15 female) who were attendees at a Science Gallery Exhibition in Trinity College took part in this study. Their ages ranged from 18 to 27 years with a mean age of 23 years. All reported normal or corrected to normal vision. Prior approval was obtained from the School of Psychology Ethics Committee. Following a briefing on the experimental protocol, participants provided written consent to take part in the study.

Stimuli

The stimuli used in this experiment were a set of dynamic visual scenes, in particular virtual images of traffic scenes. These scenes were developed using ‘OGRE’ (open source graphics engine) software as part of a collaboration with
colleagues from the Graphics and Visualisation Group, School of Computer Science and Statistics at Trinity College Dublin. Each scene depicted a set of static, horizontal virtual ‘roadways’ containing virtual ‘bridges’ and in which cars could move either left and right (i.e. on the horizontal axis only). Figure 6.1 provides an illustration of a typical scene used in this experiment. The scenes were then populated with a number of items, specifically cars, according to the design of the experimental conditions.

The criteria for scene generation was as follows: each scene consisted of a) four roads, and each road had two lanes for cars moving in two different directions, and b) 12 bridges which were randomly distributed over the four roads with a minimum requirement of at least 3 bridges per road. All of the car stimuli included in a scene were moving and when a car moved underneath a bridge, it was completely occluded from view. Moving cars were presented in a loop such that, for example, when a car moved off screen left, it would reappear on the same road on screen right and vice-versa. Car objects were randomly positioned within each scene. The onset of the stimulus was coincident with the onset of the motion of the cars in the scene. All moving objects in the scene moved at the same speed and all target objects appeared in the scene for the same duration.
Figure 6.1 An example of a static image taken from a dynamic scene depicting four roads with cars which were moving left and right. The vertical gray rectangles are "bridges" under which the cars were occluded. The target object (in this case a green car) in this scene is circled in red for illustration purposes only. The set size is eight in this example, meaning that there were a total of 8 cars in the scene.

There were four different types of moving car stimuli: small red, small green, large red, and large green cars. The large cars subtended approximately 1.27° (horizontal) x 0.5° (vertical) visual angle whereas the visual angle of the small cars subtended approximately 0.6° (horizontal) x 0.45° (vertical). Three set sizes were used across the scenes in the experiment of either 8, 16, or 32 items in the scene. Specifically, set size referred to the number of moving cars in a scene. The target car was either the large green or the small red car and was always present in the scene. Only one target appeared per scene. The auditory stimulus used was a 300hz pure tone and was delivered via headphones at a consistent and comfortable listening level.

Apparatus

The experiment was programmed using a Presentation™ (Neurobehavioral Systems®) script which ran on an Intel® Apple Macbook with 1GB RAM and an
external 20” CRT Monitor. The scenes were presented on a Mac monitor with a 1024 x 768 pixel resolution. Sennheiser HD headphones were used to transmit the audio signal. Correct responses and reaction time data were recorded using a standard PC keyboard. The participant was seated so that their head was positioned at about 60cm from the monitor and they were asked to maintain that position for the duration of the trial.

**Design**

The experiment was based on a fully factorial, 3x3 within subjects design with temporal congruency of the sound (congruent, incongruent or absent) and set size (8, 16, 32) as factors. In the sound congruency condition the proportion of trials within each of the congruent, incongruent and sound absent levels were according to the following ratio 2:1:3. Thus there were as many trials with an accompanying sound as trials without a sound, and when sound was present it was more likely to match the temporal profile of the target car.

Within the temporally congruent sound condition, the temporal profile of the sound was manipulated to match the same temporal profile as the visually moving target. This meant that the sound was continuously on while the car was visible but was rendered inaudible when the car moved under a bridge and was thus occluded from view. In this condition, the temporal profile of the sound matched the target car only and did not match any of the visual temporal profiles of the distracter cars. For the trials in the temporally incongruent sound condition, the sound was uninformative in that it neither matched the temporal profile of the visibility of the target object nor any of the distracter objects.
The trials were presented in separate blocks and blocked according to the target type (large red car or small green car) and block order was counterbalanced across participants. The target direction (left or right) was also counterbalanced. Within each block, all trials were randomly presented across participants. The dependent variables were response time and accuracy rates.

Procedure

Each participant began the experiment by completing a series of practice trials to ensure task familiarity. They were given feedback on their performance and had opportunity to pose any questions. Following the practice session, the participant were then presented with the test phase. Prior to the test block, the participant was presented with an image of the target car which they were required to locate in the scene. In this phase, a trial began with a central fixation cross which was presented for 500ms on which participants were required to focus. A dynamic scene was then presented for eight seconds. The task for the participant in this experiment was to indicate the direction of a moving target (i.e. either the large red car or small green car depending on the block) in the scene by pressing the “z” key to indicate ‘left’ and the “/” key to indicate ‘right’. Participants were requested to complete the task as quickly and as accurately as possible. Either a response or the end of the trial triggered the onset of the next trial. Participants could take a self-timed break between blocks. The experiment took approximately 15-20 minutes for each participant to complete.
6.2.2 Results

For each of the temporal congruency of the sound conditions, the mean reaction times and accuracy rates were calculated (see Figures 6.2 and 6.3 respectively). Performance in all conditions was greater than chance (50%). The average reaction time across all participants was 2,240ms (std. dev. was 460 ms). Based on the data presented in Figures 6.1 and 6.2, the first observation we can make is that sound had an effect on task performance. We then conducted separate 3 x 3 repeated measures ANOVAs on the correct responses and RTs with temporal congruency (3) and set size (3) as factors.

Response times

We found a main effect of temporal congruency \([F (2, 48) = 7.103, p < 0.01]\) which suggested that participants performed fastest when the audio-visual stimuli were temporally incongruent. A Tukey post-hoc comparison revealed that response times to locating the target in the temporally congruent sound were significantly slower than those in the temporally incongruent sound \((p<0.005)\) and significantly faster than those to the sound absent condition \((p<0.001)\). Response times to the targets in the temporally incongruent sound were also significantly faster than those in the sound absent condition \((p<0.01)\). We found a main effect of set size \([F (2, 48) = 25.702, p < 0.001]\). A Tukey post-hoc analysis revealed a significant increase in response times between set size 16 and 32 \((p<0.01)\). No significant interactions were found.
Correct Responses

As shown in Figure 6.3, we found a main effect of temporal congruency \( [F_2, 48] = 13.335, p < 0.001 \) in favour of the sound present conditions. A Tukey post-hoc comparison revealed no difference in accuracy between the congruent and incongruent sound factors. Performance was, however, significantly better in the congruent condition than in the sound absent condition \( (p<0.01) \) and in the incongruent than in the sound absent condition \( (p<0.05) \). There was a main effect of set size \( [F_2, 48] = 7.103, p < 0.01 \). A post-hoc analysis (Tukey) showed no difference in accuracy performance between set size 8 and 16 but revealed a significant decrement in performance from set size 8 to 32 \( (p<0.001) \) and from set size 16 to 32 \( (p<0.001) \). The interaction between congruency and set size was not significant \( [F_2, 48] = 0.879, p = 0.479 \).
6.2.3 Discussion

In this experiment it was found that participants performed better on a dynamic visual search task with the presence of a 300hz auditory tone. However, when this sound was temporally congruent with the onset and offset of the visual target (as it appeared and disappeared under a virtual bridge), it did not offer a benefit on accuracy performance over temporally incongruent sound. Moreover, the response time results suggest the surprising result that search times were faster when the sound was incongruent with the temporal profile of the target (or even any of the distracters) than when the sound matched this profile. In general, however, the presence of the sound, informative for locating the target or otherwise, improved search times and reduced errors relative to when no sound was presented. While the findings support the research in aurally aided visual perception, they pose new questions. In particular, it is not clear why the spatial localisation of a target in a high-level dynamic scene does not benefit from the presence of sound when temporal profiles are matched.
There may be a number of explanations. First, the literature informs us that the auditory system codes temporal information more precisely than the visual system (e.g. Navarra, Hartcher-O'Brien, Piazza, & Spence, 2009). It could be the case that a temporal shift is occurring (as in the case of temporal ventriloquism) and participants are adapting to the asynchronous audiovisual signals. The presence of temporal "aftereffects", as in the case of audiovisual speech perception, (see e.g. Vatakis, Navarra, Soto-Faraco, & Spence, 2007) suggest that the mechanisms integrating information from our senses are somewhat flexible in terms of reducing temporal disparities and hence, optimizing the perception of the events around us (Navarra et al. 2009). In our experiment, for the sound present trials, the auditory stimulus had two potential temporal profiles: one that was identical to that of the visual target (temporally congruent) and one that was not (temporally incongruent). If we consider a trial where the auditory stimulus was temporally incongruent with the visual target, the sound is not in sync with the complete disappearance and reappearance of the car. However, there is a small time lag between when the sound goes off and when the target is occluded/visible. Perhaps it is the case that this temporal disparity is being reduced and hence the temporally incongruent sound is of use to the participant in their search. If such audiovisual temporal adaptation occurs as is the case with temporal ventriloquism, this may imply a shift in the processing of the visual signal toward the auditory signal and may be evidence of the flexibility of the speed at which the auditory signal is processed. In static, low-level visual scenes (i.e. containing basic shapes) this temporal disparity becomes perceivable after about 300ms (Jack & Thurlow, 1973). It may be possible that this temporal disparity is larger in a dynamic environment, although this has yet to be established.
The temporal co-occurrence of the auditory stimulus and visual target in this experiment is relatively rare overall (33%), but occurs in twice as many trials as the audio-visual asynchronous trials. The purpose of this was to increase the likelihood of the congruent sound being useful. Nevertheless, we found no evidence that the temporally aligned auditory cue facilitated search performance relative to a sound cue which was not aligned to the target. Instead, the findings suggest a generalised benefit for the presence of a sound stimulus on performance relative to when no sound was available. It is possible, therefore, that sound in this experiment, was acting as a general attentional cue which increased arousal thus improving performance generally when it was available. What is not clear, however, is how such a mechanism could explain the distinction in performance between the congruent and incongruent conditions.

In general, it could be argued that our findings are not at conflict with those of Van der Burg et al. (2008) who showed clearly that temporally relevant but non-spatial auditory signals can improve visual search for a static target which appears and disappears in the same spatial location. However, it should be noted that in the Van der Burg et al. experiments, when not paired with the visual target, the tone used was paired with a distracter object instead with the result that search performance was less efficient. Also, it should be noted that the scenes used differed dramatically from those used in the current study. For example, Van der Burg’s scenes were dynamic in that they used colour and luminance variation not object motion as was used in the present study. To our knowledge, aurally aided visual search had not previously been investigated using more meaningful or high-level dynamic scenes.

The results in this study enable us to look deeper at the nature of search in high-level dynamic scenes. Treisman’s FIT may partially account for these results. If
participants are conducting a conjunctive search task, Treisman’s theory states that this would involve serial processing of the visual items in the scene. Our results are in accordance with this prediction since the slope in accuracy rates decreases and response times increases as the number of distracters increases. According to the results of McLeod et al. (McLeod et al., 1988), they found that when searching for a moving target, the search process segregates the stimuli into moving and stationary items, and search for the target is subsequently conducted on the moving group of stimuli only. This implies that a parallel search is more likely for targets defined by a conjunction of movement and form. If we apply this theory to the present study, this kind of parallel search would be near impossible as both the target and distracters are both moving and therefore intermittently present in the scene. In other words, while some objects (target/distracter) are occluded by the virtual bridges in the scene, others are visible and the reverse is then the case as the scene evolves.

Because of the fact that search performance for a moving target was not facilitated by a temporally congruent sound but more by a temporally incongruent sound, it is difficult to account for this finding in terms of one of the main theories of visual search, namely Wolfe’s guided search theory. In any case, search performance was always above chance, with or without the presence of sound, therefore it is likely that certain visual features were used to guide search. For example, Guided Search model indicates that a combination of bottom up and top down processing is used to rank the stimuli in a scene in terms of attentional priority and based on this ranking the target is located. An activation map is used in which the level of activation at a location reflects the likelihood that the location contains a target. Areas in the scene in which the target is occluded, such as the virtual bridges, may be low activation locations whereas roadways would be highly activated. If the target is the small red
car, a peak in the activation map would occur for red cars (large and small) and the search could then take place within that subgroup. In this kind of search, search times are longer when distracters share one or more features with the target stimuli and increase again as the number of distracters increase.

The neural basis of audio-visual spatial orienting in visual search is still in its infancy. However, FMRI studies have implicated the posterior parietal cortex in visual search and visuo-spatial orienting (Dent, Lestou, & Humphreys, 2010). Studies by Ashbridge, Walsh & Cowey (1997) used transcranial magnetic stimulation on participants while they were performing pop-out or conjunction search tasks. Their findings implicated the sub-region of the right parietal lobe as being important for conjunction search but not for pre-attentive pop-out. The superior temporal sulcus (STS) has been found to be an area sensitive to audiovisual synchrony with meaningful stimuli (Calvert, Campbell, & Brammer, 2000; van Atteveldt, Formisano, Goebel, & Blomert, 2007) and non-meaningful stimuli (Noesselt et al., 2007). While brain regions used during visual search have been identified, the neural basis of audiovisual spatial and temporal synchrony still requires further research.

6.4 Summary

The findings in this study suggest that when searching for a moving visual target, in a dynamic scene, the presence of a sound which shares the same temporal profile as the target, does not benefit search performance relative to a sound with a random temporal profile. On the contrary, a temporally incongruent sound was associated with faster search times than a temporally congruent sound. However, the presence of sound in general, facilitated search performance over the absence of
sound. Our results suggest that temporal audition plays a different role in searching for a moving visual target in a complex dynamic scene than when searching for a target in a static display (Van der Burg et al. 2008). Further study is clearly required to understand the integration of auditory stimuli in dynamic high-level scenes.
7.0 General Discussion

The current chapter will summarise the findings of each preceding experimental chapter. The findings reported in each chapter will be considered in relation to the wider theoretical implications. Questions which have arisen from the present research, and future directions, will also be discussed.

7.1 Summary of Results

In Chapter 2 the effects of semantically characteristic and spatially congruent sounds on visual search were investigated within the context of static visual scenes. Similar to previous findings reported by Iordanescu et al., (2008), the results of Experiment 1 suggested that semantically informative sounds improved performance in a visual search task. However, the hemifield in which the sound occurred seemed to have no affect on performance. Consequently, the auditory and visual stimuli, once semantically related, did not need to be spatially co-located for task performance to be facilitated. To further test this idea, in Experiment 2 we increased the difficulty of the task, requiring the participant to search for any one of a possible six targets per block of trials (where previously they were required to search for one target per block). Interestingly, it was found that when task difficulty was increased, participants then relied on the spatial location of the sound rather than the semantic information for detecting the presence or absence of the target. These effects occurred even though participants in both experiments were explicitly instructed to ignore the sound.

The results of the experiments reported in Chapter 3 extended the findings in the previous chapter. A similar search task was used in which participants searched
for a visual target embedded in a visual scene. Experiment 3 tested the effects of the spatial location of a meaningful sound on visual target detection in a complex visual scene. Size and spatial location of the visual target and the relative location of the sound source were manipulated. Adding to an existing body of research in crossmodal spatial perception (see Spence, 2010 for a review), it was found that the presence of an auditory stimulus which was spatially congruent (i.e. one which occurred in the same hemifield) with the visual target benefitted the detection of a visual target. The effect occurred for both response times and accuracy. Specifically, it was found that small targets located in the periphery benefitted more from the presence of a spatially congruent sound than large targets irrespective of their location.

The experiments reported in Chapter 4 examined the role of dynamic task-irrelevant information on attentional deployment using an attentional cuing task. In Experiment 4, participants judged the orientation of a spatially pre-cued visual target while a task irrelevant, dynamic distracter stimulus was presented at the opposite side of space and competed for spatial attention. Distracters included gender specific walking figures. First we found that task performance was best at the cued locations relative to uncued locations. However, when the cue was invalid, then the dynamic distracter reduced performance particularly when this stimulus depicted very feminine or very masculine walker. In Experiment 5, the distracter contained auditory (i.e. sound of footsteps) as well as visual information. Here, better performance was found when the pre-cue was invalid. These results suggest a response bias in which participants could predict the spatial location of the target as it always occurred in the opposite hemifield to the audiovisual distracter.
Chapter 5 examined the effects of a directionally congruent sound on visual search performance in dynamic scenes. We expected to find a crossmodal facilitation for detecting visual targets that had a directionally congruent sound simultaneously presented. However, we failed to find evidence that a spatially congruent, dynamic sound benefitted search performance. Instead, the findings of Experiment 6 demonstrated that a spatially incongruent sound benefitted performance relative to other sound conditions. Therefore, while the presence of the sound aided target detection, it did so in an unpredictable manner. In Experiment 7, the relative probabilities of the sound condition were altered. Even when the proportion of sound congruent trials was increased relative to incongruent trials, a benefit for directionally congruent sound was still not found. The decreased response times for a spatially co-located sound with the visual target which were found in Chapter 3 were not found here. These findings suggested that there may be fundamental differences in how multisensory stimuli are integrated between static (see Chapter 3) and dynamic scenes.

The aim of Experiment 8 reported in Chapter 6 was to examine the role of a temporally synchronous sound in the search for a visual target in dynamic scenes. In a speeded detection task, participants were required to indicate the direction of a moving target object (car) in a cluttered, continually changing environment. The visual target was presented with either a temporally synchronous sound (i.e. which shared the same onset and offset profile as the appearance and disappearance of the visual target respectively), or a temporally asynchronous sound (i.e. one which mismatched the appearance and disappearance of the visual target). It was found that
search times to the visual targets when a simultaneously presented temporally synchronous sound was presented were slower than when a temporally asynchronous sound was presented. Furthermore, it was found that search times to a visual target when no sound was present were faster than to either the temporally synchronous and asynchronous sound conditions. The findings suggest that the presence of non-spatial auditory signals can impair search for a visual target in dynamic displays and have implications for our understanding of multisensory influences in complex dynamic scenes.

7.2 Implications of findings

The five experimental chapters in this thesis can be divided into three themes; the role of spatially informative audition on attentional deployment in static scenes, the role of spatially and temporally informative audition on attentional deployment in dynamic scenes, and the role of task irrelevant or semantically informative dynamic stimuli on attentional deployment. Following a general discussion of crossmodal integration in visual search, these three themes will be discussed in the context of the findings and the implications for the field of research.

7.2.1 The integration of audition and vision in visual search

Traditionally, it has been found that a stimulus which occurs in the auditory modality that is temporally synchronous (e.g. Van der Burg et al. 2008) or spatially co-located (e.g. Spence & Driver, 1997) with the visual target can be integrated to enhance performance in visuo-spatial search tasks. Additionally, the semantic content
of a stimulus has been shown to influence these interactions (e.g. Iordanescu et al. 2008), by improving performance when a stimulus is semantically consistent, and impairing performance when a stimulus does not correspond or is distracting to the observer. Chapters 2 & 3 have found consistent effects for static audio-visual events. However, findings for complex dynamic scenes have uncovered some conflicting evidence. The experiments reported in Chapters 5 & 6 have found evidence that conflicting sensory stimuli can result in improved performance on visuo-spatial tasks. A possible interpretation of these results is that for directionally incongruent stimuli, perceptual grouping may account for the improvements in performance while for temporally asynchronous stimuli, it is thought that visual dominance or temporal adaptation could explain the unexpected findings.

Overall, the present studies have found evidence indicating that audition and vision interact within complex static and dynamic scenes. For static events, our findings extend previous findings by Perrott et al. (1996) who demonstrated that audition could aid in the localisation of a visual target. Whilst Perrott and colleagues used basic low-level stimuli, the present studies used natural scenes and semantically meaningful sounds (Chapters 2 & 3). The realistic nature of the static stimuli used in Chapters 2 & 3 allowed for greater generalisation of these and previous findings to real world scenarios. For dynamic scenes, investigating the role of temporal synchronous sounds in visual search showed greater accuracy for the presence of sound but did not uncover differences in accuracy for synchronous versus asynchronous targets. Though the findings in Chapter 6 differ from those of Van der Burg et al. (2008), who found a decrease in visual search latencies for a temporally
synchronous sound relative to a temporally asynchronous sound, this may be explained by methodological differences across the two studies. For example, the van der Burg et al. (2008) study demonstrated how non-spatial auditory signals could guide visual attention to a stationary singleton, the colour of which changed in synchrony with an auditory ‘pip’ sound. The experiment described in Chapter 6 incorporated visual targets (and distracters) moving horizontally across the entire scene. Also, the scenes were continually updated whereas those used by Van der Burg et al. (2008) were not.

It is apparent from the present findings regarding audition in dynamic scenes that the brain does not integrate all sensory information that is temporally synchronous. Instead, as suggested by Posner & Driver (Posner & Driver, 1992), mechanisms exist for attending to relevant stimuli while ignoring irrelevant or distracting information. Until now, these mechanisms have been almost exclusively tested using static audiovisual events. Interestingly, it was found that even when multisensory stimuli are temporally synchronous but occurring at different spatial locations, integration can still occur, resulting in performance benefits (see Experiments 1, 6 & 7 in Chapters 2 & 5).

As the perception of our environment depends largely on the combination of critical information from different sensory inputs, intuitively we know that pairing corresponding information from multiple modalities can be helpful when searching for items. The findings reported in Chapter 2 (Experiment 1) regarding semantically characteristic sounds and visual search support those of Laurienti, Kraft, Maldjian, Burdette, & Wallace (2004), who found that semantic congruence is a critical factor.
in behavioural performance when multisensory stimuli are used. It seems that integrating temporally coincident but semantically inconsistent information from the auditory and visual modality can impair performance in search tasks, as was found in Experiment 1 reported in Chapter 2. These results extend those of Iordanescu et al. (2008) & (2010) who demonstrated that characteristic sounds could facilitate visual search in a target discrimination task. In contrast to the Iordanescu et al. 2008 study, Experiments 1 and 2 in Chapter 2 incorporated images of real objects embedded in naturalistic scenes rather than more simple graphical representations. Also, the latter was a target discrimination task whereas the task in Experiment 1 and 2 required searching a naturalistic visual scene to detect the presence or absence of a target. Additionally, these studies manipulated the spatial location of the auditory stimulus and tested the interaction between spatial location and its semantic relevance in relation to the visual target. The results of Experiment 2 suggest that the effect of a characteristic sound on visual search was task dependent. It was found that a sound-induced search strategy occurred whereby, depending on the nature of the task (across both Experiments 1 and 2 reported in Chapter 2), participants were either guided by the semantic information or its spatial location of the sound but not both types of information.

Spence, Nicholls, & Driver (2001) reported that while the effects are small, participants generally respond quicker and more accurately when targets appeared in the expected modality rather than in an unexpected modality. Similar cuing effects were found in Experiment 4 reported in Chapter 4 whereby performance improved for visual targets which appeared in pre-cued locations. However, no evidence was found
of such cueing effects when distracters were audio-visual. Studies have reported impaired performance on visuo-spatial tasks when crossmodal distracters are presented (e.g. Fan, Flombaum, McCandliss, Thomas, & Posner, 2003; Stein, Meredith, Huneycutt, & McDade, 1989). However, we found an improvement in task performance when a highly salient multisensory distracter was simultaneously presented. The findings of Experiment 5 suggest this was an example of crossmodal attentional capture caused by the audiovisual distracter stimulus. This resulted in performance benefits at uncued locations which it was argued was due to a response strategy based on the participant predicting the location of the target as it always occurred at the opposite side of space to the auditory and visual distracter.

Throughout this thesis, a large portion of the experiments were designed to investigate the role of audition on visual search performance. When interpreting the findings in this context, the main models of visual search (i.e. Feature Integration Theory and Guided Search) do not account for multisensory influences. As has been clearly demonstrated throughout findings in this thesis, sound can impact on the perception of visual events. Consequently, I would argue that these models should be revisited and updated to incorporate crossmodal effects. This would further increase the usefulness of such models outside of laboratory conditions and help to account for crossmodal effects.

7.2.2 Role of spatially congruent sound in searching static scenes

A spatially co-located sound has previously been shown to improve accuracy and reduce search latencies for visual targets in static scenes (e.g. Perrott et al., 1991;
Spence & Driver, 1997). To date, the majority of studies in this area have failed to test the effects of sound on target detection in more complex high-level visual scenes. The experiments reported in Chapters 2 and 3, embedded target objects within a variety of visual scenes and tested observers’ ability to locate the target using a speeded search task. It was found that a spatially co-located sound does aid target detection in static visual scenes. In particular, it was found in Experiment 3 that a spatially co-located sound was most beneficial for targets located in the periphery. This finding appeared to be robust even considering the rough spatial alignment of the auditory and visual stimulus (i.e. a congruent sound was one which occurred in the same hemifield as the visual target). These findings also confirmed that target size affected search performance in static visual scenes. While large targets were found quicker, there was an interaction between target size and target location as mentioned. It is clear from the results of these experiments, that the presentation of spatially congruent auditory stimuli results in a facilitation of visual target detection at that location.

The implications of these findings are that spatially coincident sound can enhance perception of a visual target in a static high-level visual scene. This confirms that the results of previous studies demonstrating crossmodal spatial orienting in basic low-level environments hold true for complex visual scenes.

Naturally, explaining such crossmodal shifts in attention neurologically will require further investigation. However, the integration observed behaviourally may relate to the converging of multisensory neurons in the superior colliculus as described in early work by Stein et al. (1989). Single unit recordings in rhesus monkeys have attempted to uncover the neural basis of such attentional shifts (e.g.
Ignashchenkova, Dicke, Haarmeier, & Thier, 2004). Such studies have found evidence of activation for neurons in the intermediate layer of the superior colliculus also responsible for the preparation of saccades.

7.2.3 The role of directionally congruent and temporally synchronous sound in visual search in dynamic scenes

The findings from Experiment 6, 7 and 8 reported in Chapters 5 and 6 have revealed significant multisensory interactions in the perception of auditory and visual motion. It was found that in some cases, conflicting sensory information can result in improved performance on visuo-spatial tasks. These findings support evidence from earlier studies by Anstis (1973) and Zapparoli & Reatto (1969) whereby directional information in one sensory modality can bias the perceived direction of information in another modality. Such biasing effects were found to exist across variations in target direction and speed. This may be attributed to a Cross-modal Dynamic Capture (CDC) whereby sounds can appear to move in the same direction as a synchronized visual moving object, despite these actually moving in opposite directions. Similar results have been found by Soto-Faraco et al., (2002) & (2004). These results were consistently observed in Experiments 6 and 7 reported in Chapter 5. The experiments were carried out using artificially induced inter-sensory conflict scenarios and the findings have implications for the multisensory binding of auditory and visual events in virtual and potentially natural environments. As found by Valjamae & Soto-Faraco (2008), the combination of sensory processes in such scenes appears to occur at an early perceptual stage rather than a decisional one. Neurological evidence has also
found such integration to occur early (Giard & Peronnet, 1999; Talsma, Doty, & Woldorff, 2007).

Chapter 6 investigated the effect of temporally synchronous and asynchronous sounds in searching for a dynamic visual target. While a number of studies have demonstrated performance benefits when sound and vision are spatially correlated (e.g. Bolia et al., 1999; McDonald & Ward, 2000), few have demonstrated performance benefits for the presence of non-spatial auditory stimuli. Though, a substantial improvement in accuracy performance was found for the presence of sound, there was no benefit in terms of accuracy for a temporally synchronous sound over a temporally asynchronous sound. However, similar to findings in Chapter 5, there was a reduction in search latencies for visual targets that had conflicting auditory information simultaneously presented. In 2008, Van der Burg et al. (2008) found search benefits for visual targets with synchronously presented sounds which were spatially uninformative. However, a fundamental difference between the stimuli used in Experiment 8 and that used by Van der Burg (2008) is that in the latter, the target (and distracters) were continually moving left and right on the horizontal axis. In the Van der Burg study, the stimuli was stationary but had a dynamic component which was a colour change that corresponded to a pure tone. Thus, while the results of both studies have demonstrated benefits for non spatial auditory stimuli, the nature of the stimuli differed and this may account for the contrasting results.
7.2.4 The role of meaningful, non meaningful and task irrelevant stimuli in the deployment of visuo-spatial attention

Our findings in Experiment 1 and 2 in Chapter 2 suggest that semantically relevant sounds can facilitate target detection in visual search but this effect seems to only occur when there is a high likelihood of the sound being helpful. In Experiment 1, participants searched for one target per block. Thus only one sound was semantically relevant and others could be ignored. In Experiment 2, participants searched all 6 targets and hence all sounds were potentially useful in locating the target. The findings of Experiment 1 extend those of Iordanescu et al., (2008) who found that a characteristic sound facilitated visual search, by embedding objects within visual scenes and also manipulating the spatial location of the target. Initially, it seems that top-down semantic information benefited search regardless of spatial location. However, in Experiment 2, as task difficulty was increased and the likelihood of the sound being relevant was decreased, participants were more inclined to rely on auditory spatial rather than semantic information when localising the target. In Experiment 1 participants searched one target within a series of scenes and subsequently (in Experiment 2) they searched for 1 of 6 targets. To be efficient at such a task, it was necessary to memorise the images of 6 targets as well as their corresponding sounds. As mentioned, in the latter study, a shift in performance was noticed with increased task difficulty where the spatial location of the sound appeared to guide the participant’s attention instead of its characteristic relevance. This finding concurs with work by Belke, Humphreys, Watson, Meyer, & Telling (2008) who found that top-down semantic knowledge in visual search could be modulated by
cognitive load. Semantically consistent audiovisual pairings have been associated with integration effects in the posterior superior temporal sulcus (pSTS) and additionally in the superior temporal gyrus in humans (Hein et al., 2007). They reported that semantic congruency and object familiarity were important factors for the integration of sounds and images to occur. The findings in chapter 2 may act as a foundation to further explore the neural correlates of object-related audiovisual integration.

The presentation of semantically congruent stimuli will not always capture attention and can improve task performance for uncued targets (as found in Experiment 5 in Chapter 4). Participants performed better at a spatial cueing task when an auditory and visual distracter were simultaneously presented. Clearly, this was an unexpected result and did not occur when a visual distracter was presented. However, it could perhaps be explained by a response bias whereby the participant learned that the target was always located in the opposite hemifield to the sound. Thus, the sound cued the observer to the location of the visual target instead of the visual cue which was presented before each trial. Further investigation is required to better understand the nature of such effects. A follow up study which manipulated the spatial location of the auditory stimulus (so it did not always appear opposite the target) would prevent the sound from cuing the location of the target. Also, it would be of interest to run a version of this experiment which used an auditory distracter only (Experiment 4 uses a visual distracter and Experiment 5 uses an audiovisual distracter). This would help to ascertain if the results found in Experiment 5 could be
attributed to the sound cue only or if they were caused by the multisensory nature of the distracter.

7.3 Theoretical implication of the research findings

Despite a large body of literature devoted to understanding visual search behaviour, a consensus on the precise nature of mechanisms underlying selective attention has yet to be reached. As has been demonstrated in the experiments throughout this thesis, search efficiency is dependent on a number of factors. Traditional models of visual search such as Treisman's Feature Integration Theory (FIT) (Treisman & Gelade, 1980) describe how elements defined by a single feature (e.g. colour or orientation) will 'pop out' of a scene, while items defined by a conjunction of features usually requires a serial search
Currently, only visual feature conjunctions are incorporated into FIT. A multisensory implementation of Treisman's Feature Integration Theory (as depicted in Figure 7.0) could offer an explanation as to how we find objects within a real multisensory environment. Such an implementation would mean that feature information from auditory and visual inputs are integrated and processed within the attentional spotlight to form an efficient search. As found in Experiment 3 (Chapter 3), an auditory spatial cue can guide attention to the location of a peripheral visual target. In such a model,
the auditory signal would cause a rapid shift in the location of the attentional spotlight toward the location of the target object, resulting in a faster more efficient search. Depending on the quality of the auditory signal, this may eliminate the need for serial search. In an example where semantic auditory and visual information is consistent, the features from both modalities, are integrated automatically and the target ‘pops out’ (similar to what may be occurring in Experiment 1, Chapter 2) regardless of the number of distracters.

This model could also be adapted to include some of the fundamental principles of Guided Search Theory (Wolfe, 1989). Wolfe described attentional deployment in visual search as being determined by peaks in activation generated by the processing of basic features. An activation map was thus generated for which the highest activation peaks were attended to first. This ranking of items according to attentional priority could be incorporated into a multisensory model of Guided search whereby auditory and visual stimuli which are spatially and/or semantically consistent would be ranked above stimuli with inconsistent sensory information.

7.4 Practical implications of the research findings

The studies reported in this thesis were designed to address a wider objective to develop a realistic virtual replica of Dublin city as part of the ‘Metropolis’ project. Specifically, the aim was to conduct a comprehensive set of studies into human perception of computer generated, virtual motion with a focus on multi-sensory input. This was achieved by investigating the effects of auditory spatial location and temporal synchrony within the context of static natural scenes and dynamic traffic scenes. The work has highlighted factors affecting localisation of objects within busy
scenes such as the spatial location, temporal asynchrony and the number of distracting elements present within a scene. Also, it was found that semantic auditory information played a role in the perception and localisation of objects in visual scenes. Furthermore, the findings in this thesis have highlighted a need for careful consideration when presenting dynamic sound that is to be perceptually valid in a virtual environment.

7.5 Limitations and directions for future research

On consideration of the studies reported in this thesis, some limitations and potential directions for future research are apparent. Although a foundation has been laid in the investigation of the effects of audition on attentional deployment in complex scenes, there remains an opportunity to extend these findings by investigating the effects of sounds which are both meaningful and spatially informative in high-level dynamic environments. The majority of research investigating the perceptual effects of directionally informative audio-visual stimuli have used basic setups incorporating flashes and beeps (e.g. Soto-Faraco, Lyons, Gazzaniga, Spence & Kingstone, 2002). Major developments in computer graphics in the last decade have brought about the generation of three-dimensional virtual environments in which sensory processes can be investigated.

Throughout this thesis, the scenes used were restricted to two-dimensions. The reason for this was partially due to the expertise and resources necessary to create three-dimensional environments but also to investigate the role of audition in visuo-spatial tasks within simple static and dynamic scenes with a view to progressing to more complex environments. For instance, the use of advanced apparatuses such as
head mounted displays or large screen immersive displays could offer greater scope
presenting realistic auditory and visual content for testing purposes in future studies.

In any study pertaining to attentional deployment, it is important to ensure that
the findings are applicable outside of the laboratory. To ensure ecological validity,
experimental findings should be faithful to real world perceptual processes and care
should be taken when generalising from results on laboratory tasks to real world
scenarios. A difficulty often encountered in studies of auditory and visual spatial
attention is spatial coincidence. Previous related studies ensured spatial alignment of
auditory visual stimuli by placing LEDs on the loudspeakers (e.g. Spence & Driver,
1997). In complex virtual environments, particularly with dynamic information, the
spatial alignment of auditory and visual stimuli presents more of a challenge. Conflict
scenarios, where two modalities provide incongruent information is a useful means of
studying attentional deployment and sensory conflicts do occur in the natural
environment. However, auditory cues are typically presented from peripheral
locations when a visual stimulus is usually presented on computer monitors. Such
spatial discrepancies between cue and target for trials where stimuli are presented as
spatially congruent, may determine whether or not a spatial cuing effect is observed.
Simulating perceptual auditory experiences from the natural environment in the
laboratory can be achieved through the use of complex multiple speaker setups. Also,
the use of a Head Related Transfer Function (HRTF) which incorporates a type of
filtering based on the shape of the listener’s ear is an effective method of presenting
spatially informative sound. These methods should be considered in future studies
incorporating dynamic auditory stimuli.
Our experiments (6, 7 & 8) which were designed to investigate the role of audition within dynamic environments, used pure tones exclusively. While this stimuli was considered to be the most appropriate for the initial examination of congruent auditory effects within these environments, it may also be of interest to examine the use of semantically meaningful sounds in dynamic scenes. It would be of particular interest to explore the influence of top-down processing in a continually changing environment as it was found that these effects were dependent on task difficulty for static events (see Chapter 2).

In light of the large body of literature on crossmodal spatial-cuing effects for static events (e.g. Bolia et al., 1999; McDonald & Ward, 2000; Spence & Driver, 1997) it could be useful to explore other types of dynamic stimuli and their effects on spatial attention. Further examination of multisensory stimuli as a means of capturing spatial attention in a cuing task (as in Chapter 4) may reveal interesting results.

7.6 Synopsi

The experiments reported in this thesis contribute to an emerging body of research in the area of crossmodal spatial attention. The findings suggest that audition plays a crucial role in the deployment of attention within both visual search and spatial-cuing paradigms. The spatial location and semantic relevance of a sound can affect performance in a visuo-spatial task. As discussed in Chapter 5, the stage at which the integration of auditory and visual signals occur during these tasks seems to depend on the nature of the stimuli and the paradigm used. These findings highlight the complexity underlying the mechanisms which guide attention to objects in the environment. Furthermore, these studies have uncovered apparent differences in the
effects of audition on visual search in continually changing environments as compared with static scenes.

Further investigation examining the effects of conflicting sensory stimuli on search within complex dynamic environments is required to gain a deeper understanding of how audition and vision integrate to become a coherent percept in such environments. Investigating the role of semantically relevant sound in dynamic environments would also be a logical progression from this work.

In conclusion, the study of crossmodal spatial attention within complex environments is a growing area of research with many valuable implications for multisensory environments that are adopted for therapeutic purposes (e.g. treatment of phobias) and are of particular interest to the entertainment industry (film and games). It is hoped that the findings in these studies will make a worthwhile contribution to what is a very rewarding area of research.
References

A


B


10.1038/nn1333

164


2004-10560-012 [pii]


10.1080/02643294.2010.497727


E


nn874 [pii]


10.1098/rspb.2005.3327

10.1080/17470210600831119


I


J


K


10.1016/j.brainres.2008.01.070


10.1038/nature08103


152694799 [pii]


10.1016/j.plrev.2010.04.006

175


10.1016/j.actpsy.2008.08.004

10.1016/j.neuroimage.2007.03.065
10.1037/0096-1523.34.5.1053

10.1016/j.neuropsychologia.2006.01.001


10.1037/0096-1523.32.4.1063


10.1523/JNEUROSCI.0559-09.2009

178


Y


Z


Appendix A

Preparatory study to test participants ability to determine correctly the directional information in the sounds used in Chapter 5.

Method

Participants

8 participants (7 female with a mean age of 28 years) who were staff and students at the Trinity College Institute of Neuroscience took part in this experiment. All reported normal or corrected to normal vision. Prior approval was obtained from the School of Psychology Ethics Committee. Following a briefing on the experimental protocol, participants provided written consent to take part in the study.

Stimuli

The auditory stimulus used was a sine wave presented at varying frequencies and was generated using NCH Tone Generator software. Auditory motion was simulated by dynamically manipulating the interaural level difference between the visual target’s start and end point. There were four sound directions (left, right, up & down). A detailed description of the directions and corresponding tones used can be found in Table 2.0. The sound intensity level of the auditory stimuli was kept constant across all stimuli and all participants, at an average of 12dB SPL.
Table 3.0 describes the frequencies used for congruent horizontally and vertically moving targets. Horizontally moving targets had a continuous tone that was manipulated using interaural level difference (ILD) whereas vertically moving targets used a frequency shift as described above.

### Apparatus

The experiment was programmed using a Presentation™ (Neurobehavioral Systems®) script which ran on an Intel dual core PC. Sennheiser HD202 headphones were used to transmit the auditory information. Correct responses and reaction time data were recorded using a standard PC keyboard. Testing took place in a laboratory at Trinity College Institute of Neuroscience.

### Design and Procedure

Each sound was presented 16 times. This led to a total of 128 trials. For each participant, trials were presented in a randomised order. The dependent variables were response time and accuracy. In the experiment, a trial began with a central fixation cross which was presented for 500ms. A sound was then presented for 8 seconds or until the participant responded. The task for the participant was to indicate the direction in which the sound was moving by pressing the corresponding key (‘’ = up, ‘/’ = down, z = left, x = right). A subsequent trial occurred only when the participant had responded to the
previous trial. Both accuracy and response times were recorded. The experiment took approximately 10-15 minutes for each participant to complete.

**Results**

As shown in Figure 8a, the mean accuracy performance across all participants was 68.45% which was above chance level (25%). Thus the ability to accurately identify the direction of the auditory motion stimuli was ensured. The mean reaction time across all participants was 4719ms (see Figure 8b).

![Graph showing mean accuracy performance across all participants.](image)

**Figure 8.** Plot showing the a) mean correct responses and b) mean reaction times to indicating the direction of the sound. The results are shown for each speed (fast/slow) and direction (up, down, left and right).