Acta Psychologica 178 (2017) 41–47

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

Crossmodal attentional control sets between vision and audition☆

Frank Mast[a], Christian Frings[a,*], Charles Spence[b]

[a] University of Trier, Department of Psychology, D-54286, Germany
[b] Crossmodal Research Laboratory, Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, United Kingdom

ARTICLE INFO

Keywords:
Contingent capture
Crossmodal attention
Multisensory
Vision
Audition

ABSTRACT

The interplay between top-down and bottom-up factors in attentional selection has been a topic of extensive research and controversy amongst scientists over the past two decades. According to the influential contingent capture hypothesis, a *visual* stimulus needs to match the feature(s) implemented into the current attentional control sets in order to be automatically selected. Recently, however, evidence has been presented that attentional control sets affect not only visual but also crossmodal selection. The aim of the present study was therefore to establish contingent capture as a general principle of multisensory selection. A non-spatial interference task with bimodal (visual and auditory) distractors and bimodal targets was used. The target and the distractors were presented in close temporal succession. In order to perform the task correctly, the participants only had to process a predefined target feature in either of the two modalities (e.g., colour when vision was the primary modality). Note that the additional crossmodal stimulation (e.g., a specific sound when hearing was the secondary modality) was not relevant for the selection of the correct response. Nevertheless, larger interference effects were observed when the distractor matched both the stimulus of the primary as well as the secondary modality and this pattern was even stronger if vision was the primary modality than if audition was the primary modality. These results are therefore in line with the crossmodal contingent capture hypothesis. Both visual and auditory early processing seem to be affected by top-down control sets even beyond the spatial dimension.

1. Introduction

During every waking moment, we are flooded by a vast amount of sensory information as a result of the many ongoing events in our environment. Due to our limited cognitive capacities, however, only a minority of this information can be processed consciously. Thus, there is a constant competition amongst the different stimuli in order to be selected for further processing (Pashler, 1998). One of the central roles for attention here is thought to be to bias the processing of early incoming perceptual information. According to prominent theories of *visual* attention (e.g., the Feature Integration Theory, Treisman & Gelade, 1980; the Guided Search Model, Wolfe, 1994, 2007; the Theory of Visual Attention, Bundesen, 1990), selection is affected by both, top-down and bottom-up mechanisms. In everyday life, however, visual input is rarely processed in isolation, but rather together with information from the other senses, in order to enhance behavioural efficiency. For example, when searching for a friend at a lively party, it seems plausible to search the visual scene for his face but also to listen out for the booming sound of his voice. It is reasonable, therefore, to assume that top-down mechanisms (e.g., knowledge about the appearance and voice of one's friend) not only play an important role in unisensory but also in crossmodal attentional selection (that is selection between different senses) as well as in the selection and integration of multisensory events (see e.g., Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010; Tang, Wu, & Shen, 2016, for reviews). In this paper we therefore enhance and generalize the idea of contingent capture to crossmodal selection as research on this particular attention phenomenon has only recently started to be examined in multisensory contexts (Mast, Frings, & Spence, 2015; Matusz & Eimer, 2013).

One paradigm that has frequently been used in research on visual selective attention in order to dissociate top-down from bottom-up mechanisms is the exogenous spatial cuing task (see Posner, 1980; Yantis & Jonides, 1984). Over the last couple of decades, researchers have investigated the exogenous control of crossmodal spatial attention for all possible combinations of visual, auditory, and tactile cue and target stimuli (Spence & Driver, 1997; Spence, Nicholls, Gillespie, & Driver, 1998).

Originally, the assumption was that certain stimulus events, no

☆ The research reported in this article was supported by a grant of the Deutsche Forschungsgemeinschaft to Christian Frings and Charles Spence (FR 2133/5-1).
* Corresponding author at: Trier University, Faculty 1, Department of Psychology, Germany.
E-mail address: chfrings@uni-trier.de (C. Frings).

http://dx.doi.org/10.1016/j.actpsy.2017.05.011
Received 5 September 2016; Received in revised form 3 March 2017; Accepted 25 May 2017

matter whether they were unisensory or multisensory, had the potential to capture attention and reflect a purely stimulus-driven mechanism of selective attention (e.g., Jonides, 1980; Theeuwes, 1991; Yantis & Jonides, 1984). This notion was challenged by Folk and colleagues' (1992; see also Folk, Remington, & Wright, 1994) notion of *contingent* capture. According to the latter account, participants can set up attentional control sets for a specific task-relevant feature. As a consequence of attentional control sets, only those stimuli that match the current attentional control set have the potential to automatically capture a participant's spatial attention. When, for example, the task involves localizing a red target stimulus, participants are assumed to set-up their attentional control sets for the colour 'red' (see also Ansorge & Becker, 2014; Goller & Ansorge, 2015). Accordingly, a stimulus (cue or target) needs to be red in order to be selected automatically. One might think of attentional control sets as an abstract inner representation of the searched-for target stimuli. Since Folk, Remington, and Johnston (1992) published their influential first study, *contingent attentional capture* has been replicated in various studies (see Awh, Belopolsky, & Theeuwes, 2012; Burnham, 2007; Theeuwes, 2010, for reviews; but see also Lamy & Kristjánsson, 2013).

Recent studies have indicated that attentional control sets can also be compiled of multiple features from different sensory modalities, namely from vision and touch (Mast et al., 2015; see also Matusz & Eimer, 2013). Mast and his colleagues combined a non-spatial visual response compatibility task with additional (response irrelevant) tactile stimulation. Each trial consisted of two visual stimuli that were presented from the same location in close temporal succession. The participants were instructed to try and ignore the identity of the first stimulus (the distractor) and to respond to the identity of the second stimulus (the target). In response compatible trials, the distractor was mapped on to the same response as the subsequent target. In the response incompatible trials, by contrast, the distractor was mapped on to the opposing response instead. In order to examine the compilation of crossmodal attentional control sets, the visual primary task was combined with additional tactile stimulation. That is, the visual target stimulus was always accompanied by a simultaneously-presented additional tactile stimulus. It is important to stress that the tactile stimulation itself was not mapped on to either response. The co-occurrence of the visual target and the tactile stimulus was assumed to result in participants establishing a bimodal attentional control set (incorporating both visual and tactile components). Intriguingly, bimodal distractors caused more pronounced interference effects than unimodal distractors. It was argued that the difference in the size of the interference effects was due to differences in the feature-overlap between the features of the distractor (unimodal vs. bimodal) and the features implemented into the participants' top-down sets. Therefore, the results suggest multisensory top-down sets having both visual and tactile features.

The aim of the present study was therefore to further support the *crossmodal* contingent capture hypothesis and to underline the importance of contingent capture in selection in general. Studying crossmodal attentional control sets across the different senses and in different experimental paradigms, is important since the interplay between top-down and bottom-up mechanisms has been found to differ between the different senses modalities (e.g., see Bundesen, Kyllingsbaek, Houmann, & Jensen, 1997; Moray, 1959, for differences in the potential of one's own name to capture attention depending on whether the name is spoken or written). What is more, the top-down influence on automatic distractor processing has been found to vary as a function of the modalities combined in a given task (e.g., see Mast, Frings, & Spence, 2014, where participants were able to ignore tactile distractor information when attending to a visual target but not vice versa). Therefore, we examined the compilation of audiovisual attentional control sets in a non-spatial interference task that was derived from the typical contingent capture task (see Matusz & Eimer, 2013, for a crossmodal exogenous spatial cuing task). All of the published studies that previously examined crossmodal contingent capture (Mast et al., 2015; Matusz & Eimer, 2013) combined a *visual* primary task with additional crossmodal (auditory or tactile) stimulation. Thus, we went beyond the previous research in this area by analysing audio-visual contingent capture effects and further by varying whether vision or audition was the response-relevant dimension. Note that previous research on audio-visual integration/selection have shown differences in dependence of whether vision or audition was the task-relevant modality (e.g., Thelen, Matusz, & Murray, 2014; van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008; Yuval-Greenberg & Deouell, 2009). Thus, our study can answer the question whether crossmodal contingent capture affects selection also if participants respond to non-visual targets and hence reflects a modality-unspecific attention mechanism.

1.1. Overview of the present study

Following Mast et al. (2014, 2015), a non-spatial response compatibility task was used. In Experiment 1, the presentation of the visual target was always accompanied by a sound. It can be argued that participants set-up their attentional control sets for a visual and an auditory feature. On the one hand, the visual feature (i.e., colour) should be implemented into the top-down sets because its identity indicates the response that should be executed (the response feature). On the other hand, the auditory feature should be implemented into the participant's attentional control set because it indicates the presence of the target stimulus (the selection feature). While the targets were always accompanied by the same auditory stimulus, the distractors were combined with either a target congruent sound or else with a target incongruent sound (see Fig. 1). Note that the distractor sound was not correlated with the identity of the subsequent visual target (i.e., it was non-predictive).

In Experiment 2, the relevant modality was audition while a particular visual feature (a coloured circle) always accompanied the targets but only half of the distractors displayed this particular target feature. Once again, the accompanying feature was not correlated with the response feature.

In both experiments, the strength of attentional capture effects for the distractors was assumed to vary as a function of the feature-overlap between the features of the distractor and the features of the attentional control sets. That is, more pronounced attentional capture effects were expected as the feature overlap between the distractor and the multisensory top-down set increases.

2. Experiment 1

As outlined above, Experiment 1 combined stimuli from both vision (red and green circles) and audition (with either 200 or 700 Hz pure tones). In line with the previous research (e.g., Mast et al., 2015; Matusz & Eimer, 2013), the hypothesis was that the participants compile their top-down sets for both the visual and auditory features. Thus, the participants' top-down sets should contain at least two crossmodal features; colour as the response feature (given that colour indicates the correct response) and the pitch of the target sound as the selection feature (a feature that indicates the presence of the target). The featural-overlap of the distractor and the participants' top-down sets is assumed to vary as a function of the congruency of the auditory stimulus. The proposed differences in feature overlap between the distractor and the top-down set should be reflected in less pronounced attentional capture effects for the auditory incongruent condition as compared to the auditory congruent condition (see Fig. 2 for an explanation).

2.1. Methods

2.1.1. Participants

Twenty-one students (3 male; mean age of 22 years ranging from 19