Abstract: This chapter examines how our sense modalities interact in the perception of persistence. The chapter concentrates on two questions. The first concerns perceptual processing—do perceptual computations of object persistence ever integrate and compute over representations from more than one modality? It argues that this question should be answered affirmatively. The second question concerns perceptual experience—do experiences of object persistence ever exhibit a constitutively multisensory phenomenal character, or is the phenomenology of object persistence always uniquely associated with just one modality? The chapter argues that the available evidence underdetermines the answer to this question, but suggests ways it might be empirically resolved.

Keywords: multisensory perception, object perception, object persistence, multisensory experience, crossmodal interaction

Chapter 20

The multisensory perception of persistence

E. J. Green

Introduction

We perceive objects as persisting through time. As I watch my cat dash across the room, it is visually apparent that the animal I see now is the same one I saw a moment ago, even if he passes briefly behind a barrier. While discussion of the perception of persistence has focused mostly on vision, the capacity generalizes to other modalities. If I pass a pencil between my hands, I experience the object now in my left hand as the same one I earlier felt in my right. Nonetheless, while we can perceive objects as persisting *within* more than one modality, this does not settle how, if at all, the modalities *interact* in the perception of persistence. The present chapter examines this issue.

While many philosophers have considered crossmodal interactions in object perception, the primary focus has been on multimodal *binding*—that is, whether perception binds features apprehended through separate modalities to a single individual (Deroy, 2014; de Vignemont, 2014; O'Callaghan, 2014). However, ordinary episodes of object perception recruit many capacities beyond binding. The perception of objects as persisting over time is one example. Multisensory contributions to the perception of persistence are ripe for investigation.

This chapter explores two questions about the multisensory perception of persistence.[1] The first concerns perceptual processing. I ask whether perceptual computations of persistence are *multimodally penetrable*—roughly, whether they ever receive representations from multiple modalities as input. I argue that research on the audiovisual bounce effect makes a compelling case for multimodal penetration.

The second question concerns perceptual experience. I ask whether experiences of persistence ever exhibit *constitutively multisensory* phenomenal character—roughly, whether the phenomenal character associated with perceiving an object as persisting over time ever outstrips

[1] Caveats: I will limit my discussion to vision, audition, and touch, omitting the chemical senses. Nor will I address the apparent persistence of dynamic, temporally extended auditory streams like melodies or speech (but see Green, 2019a). The latter cases may differ from the perception of ordinary material objects in notable respects. According to O'Callaghan (2016) (although see Skrzypulec, 2020), sound streams seem to *perdure*: they seem to be composed of a series of temporal parts. Conversely, material objects seem to *endure*: my cat strikes me as wholly present at each moment I see him.

that which is uniquely associated with just one modality or another. The clearest examples would involve *multisensory diachronic grouping*—experiencing an object as persisting over time, even though it is alternately perceived through separate modalities. While there is suggestive evidence of this, the case remains inconclusive. I consider ways of resolving the issue.

Crossmodal processing of persistence

Certain perceptual processes receive inputs from multiple modalities. Consider multisensory cue combination. When two modalities produce conflicting estimates of a variable—for example, size or shape—our sensory systems often integrate these representations to form a single combined estimate of that variable. Typically, the output estimate is a reliability-weighted average of the unimodal estimates (van Dam *et al*., 2014).

The processes responsible for cue combination receive distinct visual and, say, haptic representations of a variable as input, and produce a combined estimate of that variable as output. When a process thus receives representations from multiple modalities as input, let us say it is *multimodally penetrable*.

A sensory process might be affected by processing in multiple modalities without being multimodally penetrable. Suppose an auditory representation of sound location is affected by the visual representation of the location of its source. And suppose that another process computes audible motion based on that auditory representation of sound location, but receives no further information as input. Then the motion computation is affected by both vision and audition, but is not multimodally penetrable, since its inputs would consist of auditory representations alone.

The multimodal penetration issue mirrors familiar discussions of the cognitive penetration of perception (Green, 2020; Macpherson, 2012; Pylyshyn, 1999). Several have argued that directness is a necessary condition for cognitive penetration. There is, on this view, an important distinction between the claim that a perceptual process is merely affected by cognition and the claim that a perceptual process receives cognitive representations as input and computes over them (Gross, 2017; Quilty-Dunn, 2020). In the latter case, perceptual processing is, in an important sense, *continuous* with cognition (Pylyshyn, 1999). A similar distinction applies to crossmodal effects. Some processes are merely affected by processing in multiple modalities; others receive representations from multiple modalities as input.

This section considers whether perceptual processes that compute object persistence are multimodally penetrable. I will focus on two cases that have received extensive investigation: crossmodal influences on apparent motion, and the audiovisual bounce effect.

Crossmodal influences on apparent motion

Suppose you see two objects flashed in alternation on opposite sides of a computer screen. With the proper timing and spacing, you will experience a single object moving back and forth. The problem of determining persistence through apparent motion is called the *correspondence problem*. When objects are perceptible at times T1 and T2, perception needs to determine correspondences (persistence relations) among them.

The direction of apparent motion in one modality can *capture* the direction of apparent motion in another. When subjects are required to judge the direction of an apparent motion stimulus in one modality, they are less accurate if an irrelevant stimulus is presented to another modality moving in the opposite direction (Soto-Faraco *et al*., 2004). Here I will focus on a case where auditory motion influences ambiguous visual motion.

Alink *et al*. (2012) showed participants a stimulus where several columns of dots were visible through an aperture. This was followed by a second frame in which the columns shifted some distance rightward. When the rightward shift was exactly half the distance between columns, subjects were equally likely to see the columns moving rightward or leftward. This display was presented alongside an auditory stimulus (four sounds in succession) that produced an impression of either rightward or leftward auditory motion. Critically, auditory apparent motion influenced reports of visual apparent motion. Ambiguous visual stimuli were more often reported as moving in the same direction as the auditory stimulus.

The perception of motion does not imply the perception of persistence.[2] However, supposing in this case that either the columns or the dots within them were seen as persisting between apparent motion frames, then audition affected the apparent persistence of visually perceived objects.

I turn to another case where audition influences visual apparent motion.

Suppose you see two objects flashed in alternation, one on the right and one on the left. If the temporal spacing between flashes is constant, you typically experience a single object hopping back and forth. However, if spacing is uneven—if, say, the time from left to right flash is shorter than the time from right to left flash—the percept shifts. Instead, you typically see an object moving from left to right and then vanishing, followed by a new object appearing on the left and moving in the same direction (von Gruneau, 1986).

Freeman and Driver (2008) employed an apparent motion stimulus with equal temporal spacing between flashes. However, they played sounds that either slightly led or slightly lagged some of the flashes. For example, one sound could be played just after the left flash, and another just before the right flash. Because the perceived timing of sounds attracts the perceived timing

[2] Perceived motion *typically* involves perceived persistence (Paul, 2010; Scholl, 2007), but not always. Motion energy signals can induce an impression of movement without any perceptually differentiable object appearing to move (Cavanagh, 1992).

of visible events (Morein-Zamir *et al*., 2003), Freeman and Driver predicted that this display would induce a unidirectional percept of objects moving from one side to the other and vanishing. This prediction was confirmed through both subjective reports and adaptation after-effects. Thus, correspondence solutions for visually perceived stimuli were causally influenced by audition. In the presence of auditory input, subjects perceived an object *ceasing to persist*. Without auditory input, they would have perceived it as *continuing to persist*. Again, we have a crossmodal effect on perceived persistence.

Both of these cases exhibit causal influences of processing in one modality on persistence computations for stimuli apprehended through another modality. But do they show that these persistence computations are multimodally penetrable? I suggest not.

Consider the Freeman and Driver (2008) study. In this case, audition affects perceived persistence in visual apparent motion, but it plausibly does so *by way* of affecting the visual representation of temporal properties. Thus, the correspondence computation might have received only visual representations as input. It is just that some of these representations (representations of temporal features) were modified by audition.[3]

What about crossmodal capture? I suggest that this also need not involve multimodal penetration of persistence computations. It could be mediated by crossmodal influences on the allocation of attention, which in turn affects the perception of persistence.

Suppose you see four dots forming the vertices of an imaginary tilted square (Figure 20.1A). In the next frame, the dots have shifted, forming an imaginary square resting on its base (Figure 20.1B). If you see these frames in alternation, you can experience the dots moving either

[3] Another possibility is that audition affects not the perceived *timing* of visual stimuli, but their apparent *perceptual grouping* (Roseboom *et al*., 2013).

clockwise or counterclockwise. In displays of this sort, we have some control over the perceived direction of movement (Verstraten *et al*., 2000). A leading explanation is that the computations underlying apparent motion are sensitive to shifts in attention (Cavanagh, 1992; Xu *et al*., 2013). On the 'attentional pointer' hypothesis (Cavanagh *et al*., 2010), for example, the perception of persistence in high-level, object-based apparent motion sometimes results from shifting spatial attention. When you shift attention from one object's location in frame 1 to another object's location in frame 2, this '[links] the two locations together as the changing location of a single target' (Cavanagh *et al.*, 2010. p. 151). By shifting attention in one direction vs another, you can influence the perceived direction of motion in an ambiguous stimulus.

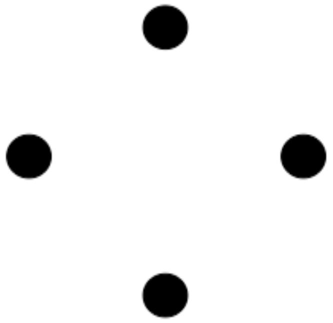

Figure 20.1A

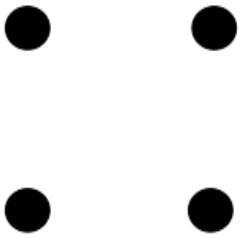

Figure 20.1B

Such attentional effects could be involved in crossmodal capture. There is independent evidence that the allocation of either endogenous (Driver and Spence, 2004) or exogenous (Spence and Driver, 1997) attention in one modality biases the allocation of attention in other modalities. Suppose that in Alink *et al*.'s (2012) ambiguous motion display, perceived motion direction was biased by subjects' sequential allocation of visual attention. If visual attention was first directed towards locations or dots on the left, and then to locations or dots on the right, this biased the system towards left–right motion. Suppose, further, that the concurrent unambiguous auditory motion stimulus attracted attention sequentially to the locations through which the sound appeared to move. If the direction of *auditory* attention biased the direction of *visual* attention, it may have also biased the perception of visual apparent motion, and thus visual persistence.[4]

Consistent with the hypothesis that crossmodal capture effects go via attention, attended motion streams in one modality are more effective in capturing apparent motion in another modality (Oruc *et al*., 2008). There is also evidence that louder, more attention-grabbing auditory stimuli exert stronger capture effects on tactile motion (Occelli *et al*., 2009). Finally, there is evidence that attention-grabbing static sounds can bias the perception of an ambiguous visual apparent motion stimulus, even without any auditory motion (McBeath *et al*., 2019).[5]

If crossmodal capture effects are mediated by attention, this raises the possibility that crossmodal capture does not involve multimodal penetration. Instead, persistence computations within a modality might receive *attention-weighted unimodal representations* as input. Thus, the computation of persistence in visual apparent motion might receive purely visual representations of objects, properties, and locations, alongside information about the amount of attention allocated to each. Processing in other modalities could influence the allocation of attention, and

[4] This story is even simpler if there is a single multimodal system of spatial attention, since there would be no need for a mediating influence of auditory attention on visual attention.

[5] McBeath *et al*. created a visual stimulus that could be seen as either starting on the left and moving rightward or starting on the right and moving leftward. They found that static auditory cues presented on the left promoted percepts of rightward visual motion originating on the left, and vice versa for auditory cues presented on the right.

thus modify these inputs to visual persistence computations, but auditory and haptic representations would not themselves constitute inputs to those computations.

Thus, while extra-modal factors *affect* perceived persistence through apparent motion within a modality, it is unclear whether the relevant persistence computations literally integrate and compute over representations from multiple modalities. I note that there is a striking analogy here with the literature on cognitive penetration, where it is often disputed whether purported cognitive effects on perception are direct or rather mediated by attention.

The audiovisual bounce effect

I turn to another candidate case of multimodal penetration. Suppose two objects start on opposite sides of a computer screen and approach each other until they overlap. Afterward, two objects emerge following the original motion trajectories (Figure 20.2). Two percepts are possible: the objects can appear either to *stream past* each other or to *bounce off* each other. Typically, streaming percepts dominate. However, if a sound is played at the moment of overlap, the pattern reverses, with bouncing perceived on nearly 80% of trials (Sekuler *et al*., 1997).

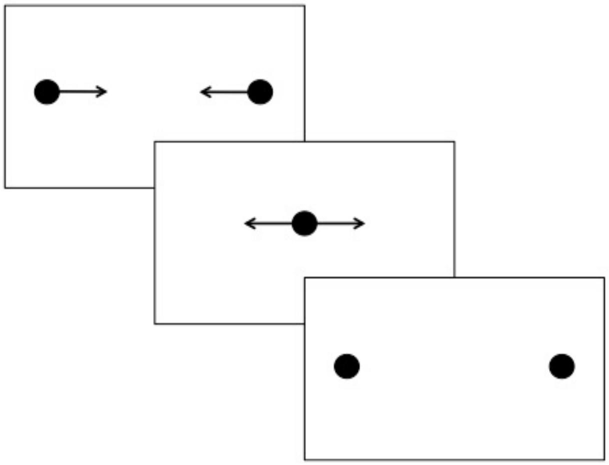

Figure 20.2. Bouncing/streaming display.

The objects at the beginning of the bouncing/streaming display can be seen as persisting in two incompatible ways, and audition helps settle which interpretation dominates. But is this a case where persistence computations are multimodally penetrated? It depends on what the computations receive as input. We can distinguish two models.

First, it is possible that the audiovisual bounce effect is mediated by auditory influences on visual attention. Perhaps the sound attracts attention away from the motion event, causing the visual system to 'miss' the moment of complete overlap between the objects. Accordingly, the visual system non-veridically represents what Scholl and Nakayama (2004) call an 'illusory crescent'—a slice of one object that is never fully covered by the other. The attentional diversion reduces the perceived overlap between the objects, and in the absence of full overlap, bouncing becomes a more likely interpretation than streaming. Meyerhoff and Scholl (2018) found that perceivers are indeed more likely to perceive illusory crescents in the bouncing/streaming display on sound-present trials than on sound-absent trials. They suggest that attentional diversion causes illusory crescents, which promote bouncing percepts (2018, pp. 92–3).[6] Call this the *attentional model*. This model is consistent with (but does not entail) the absence of multimodal penetration. Perhaps the inputs to visual persistence computations are exhausted by visual representations of motion trajectories and spatial relations. By diverting attention, audition modifies the visual representation of spatial relations (viz. whether a crescent is represented), and so modifies those inputs.

[6] One concern with the attentional model is that it is unclear why diverting processing resources away from the overlapping objects would lead the system to infer an illusory crescent, instead of simply representing the objects as continuing their original trajectories. (Thanks to Casey O'Callaghan for this point.)

Second, it is possible that the inputs to persistence computations include representations of relations between events apprehended through separate modalities. When a sound occurs at the moment of overlap, perhaps the perceptual system determines that the visible event *caused* the sound.[7] This information is passed to the persistence computation, which determines that the objects bounced rather than streamed. After all, bouncing normally causes sound, while passing normally does not. Call this the *causal model*. On this model, persistence computations receive both visual and auditory representations as input. The representation of the causal relation delivered to the computation contains representations from both modalities as constituents (*visible event V caused audible event A*).[8]

Some evidence appears to support the attentional model. Watanabe and Shimojo (1998) found that bounce percepts are promoted not only by sounds, but also by visible flashes, which would also be expected to distract attention from the bouncing/streaming event. Further studies suggest that other abrupt perceptible events—also known as 'transients'—also facilitate bouncing. A brief tactile pulse promotes bouncing and also induces illusory crescents (Meyerhoff *et al*., 2018; Watanabe and Shimojo, 2005). Meyerhoff and Suzuki (2018) found that sudden *offsets* of sound have a similar result. Meyerhoff *et al*. (2018, p. 2236) suggest that the attentional model might explain the generality of these effects: '[I]t is possible that any coinciding transient distracts attention from the bouncing/streaming display. [This] might result

[7] The proposal that perceptual systems represent causal relations raises many interesting issues, such as whether sensory cues to causation are innate, learnt, or both. Unfortunately, I cannot address these issues here.

[8] Alternatively, one might hold that some non-causal relation between visual and auditory events is represented and drives the bounce effect. This would also be a form of multimodal penetration.

in missing the central frame which in turn induces illusory crescents as well as bouncing impressions.' Such evidence seems to favour the attentional model over the causal model. Collisions normally produce sounds but rarely produce visible flashes, tactile vibrations, or sound offsets. It is unclear why the perceptual system would opt for a causal interpretation in the latter cases.

However, other evidence favours the causal model. First, not all sounds promote bouncing with equal effectiveness. Real-world impact sounds possess a characteristic amplitude profile: abrupt 'attack' followed by gradual decay. Grassi and Casco (2009) generated artificial sounds that could be either impact-consistent or impact-inconsistent in this respect. One sound had a loud onset and decayed gradually, while the other had a gradual onset and abrupt decay. Critically, the impact-consistent sounds generated more bounce percepts than impact-inconsistent sounds. Furthermore, Grassi and Casco (2010) found that billiard ball collision sounds generate a stronger bounce effect than water drops or fireworks, even though all three sounds are equally attention-grabbing. Together, these findings suggest that the processes responsible for distinguishing bouncing from streaming *are* sensitive to whether the transient is a viable effect of the motion event.

Another study provides particularly strong evidence against a pure attentional model. Adams and Grove (2018) reasoned that if transients induce the bounce effect simply because they distract attention, then their effect should be stronger when they occur farther from the moving objects, since this diverts attention farther from the overlap event. However, in the case of visible flashes, precisely the *opposite* occurred. Visible flashes produced more bounce responses when they occurred at the same location as the objects' overlap than when they occurred on the opposite side of the screen. This result conflicts with the attentional model's predictions but comports well with the causal model. Perhaps the perceptual system is wired to treat abrupt transients at or near the location of the overlap as indicating a causal transaction has occurred.

So there is evidence supporting both the attentional and causal models. What could break the impasse? I suggest that both models capture aspects of the perceptual computation of persistence in bouncing/streaming displays.

Recall that Grassi and Casco (2009) found that impact-consistent sounds generate more bounce percepts than impact-inconsistent sounds. However, they also discovered an important distinction *within* the class of impact-inconsistent sounds. In one experiment, they examined impact-inconsistent sounds that had the same average intensity as impact-consistent sounds: the latter sounds began at 87 dB and softened to 47 dB, while the former sounds did the opposite. In this case, impact-inconsistent sounds produced *no* bounce effect—bouncing percepts were no more frequent than in the silent display. However, an interpretive difficulty here is that when a sound is softer at onset, it may also be less attention-grabbing. Thus, a proponent of the attentional model could reply that impact-inconsistent sounds were simply less distracting. To address this concern, Grassi and Casco ran another experiment where impact-inconsistent sounds had the same intensity as impact-consistent sounds at onset, but gradually got louder rather than decaying. In this case, the impact-inconsistent sound *did* induce a bounce effect (nearly 50% bounce responses vs 20% in the silent display), but the effect was nonetheless weaker than that produced by impact-consistent sounds (nearly 80% bouncing).

A hybrid model could explain these results. When sounds have a louder onset, they are more attention-grabbing (Grassi and Casco 2009, exp. 3). Consistent with the attentional model, this attentional distraction promotes illusory crescents, which bias the system towards bouncing.

Beyond this, however, the persistence computation also consults information about whether the sound is a credible effect of the motion event. When the motion event is represented as causing the sound, this information is sent as input to the persistence computation, which is then even more likely to opt for bouncing. On this view, attentional and causal factors are mutually reinforcing. Similar remarks apply to the Adams and Grove (2018) study. Although visible flashes near the overlapping objects resulted in a stronger bounce effect, the effect was not eliminated with distant flashes. The hybrid model could claim that while distant flashes were less likely to be represented as caused by the visible motion event, they effectively diverted attention and induced illusory crescents. Finally, the hybrid model can explain why abrupt transients with no obvious causal connection to the motion event (e.g., brief silences) also promote bouncing: they still draw attention.

If this analysis is correct, then the answer to our first question—whether persistence computations are multimodally penetrable—is yes. At a computational level, the perception of persistence is sometimes a multimodal operation. In determining how objects apprehended through one modality persist through time, the perceptual system sometimes consults representations from other modalities. Such representations constitute inputs to persistence computations.

Multisensory experience of persistence

So far I have discussed crossmodal interactions in the subpersonal computation of persistence. This section turns to the question of how crossmodal interactions might reshape our perceptual experience of persistence. I ask whether such experiences are ever *constitutively multisensory*.

Preliminaries

Some perceptual capacities are multisensory at the level of perceptual processing, but not at the level of experience. In the McGurk effect, our experience of an uttered syllable arises from computations that combine visual information about mouth movement with auditory information about acoustic features, but arguably the experience of the syllable is purely auditory (McGurk and MacDonald, 1976). Likewise, although persistence computations sometimes integrate information from multiple modalities, the experiences that result from these interactions might not be multisensory in any deep respect. In the audiovisual bounce effect, the experience of the objects bouncing may be purely visual despite issuing from crossmodal interactions.

The present question is whether perceptual experiences of persistence are ever *constitutively multisensory*. To say that an experience is constitutively multisensory is, roughly, to say that its phenomenal character cannot be fully factored into components, each of which is uniquely associated with just one modality. However, applying this principle presupposes some conception of when an aspect of phenomenal character is 'associated' with one modality vs another. Clarifying this idea is notoriously difficult (Macpherson, 2011). In recent years, Casey O'Callaghan (2015, 2019) has offered the most refined approach to the issue.

On O'Callaghan's account, an aspect of phenomenal character is associated with a modality when it could be instantiated by a *corresponding mere experience of that modality*. A *mere* experience of modality A is an experience that belongs to A, and not to any other modality. Thus: '[A] merely visual experience is visual but not auditory, tactual, olfactory, or gustatory. To get a fix on this, consider the other sense organs as blocked or anesthetized' (O'Callaghan, 2015). So let us say a merely visual experience is an experience of the sort that is normally produced, absent sensory malfunctions, when other sense organs receive no stimulation. A *corresponding* mere experience of a modality on an occasion is 'a perceptual experience merely of that modality under equivalent stimulation' (2015). Suppose that a subject actually undergoes a perceptual experience owing to stimulation S1 of her retina and S2 of her cochlea. Then a corresponding merely visual experience is one that the subject could have undergone on that occasion had she received only S1, and not S2. Call this conception of constitutively multisensory experience the *unisensory correspondence model*.

One concern should be flagged. O'Callaghan proposes that a corresponding mere experience of a modality is one that the subject could have had under *equivalent* stimulation. However, Wadle (2020) argues that this rule yields implausible consequences. Our experiences of uttered syllables during the McGurk effect are plausibly purely auditory, not multisensory, even though they arise from processes that integrate visual and auditory information. However, the unisensory correspondence model appears to deem such experiences constitutively multisensory. Consider someone who receives auditory input consistent with /ba/ and visual input consistent with /ga/, but experiences /da/ instead. If they had received only the equivalent auditory input, they would have experienced /ba/, not /da/. Thus, the model implies that the actual phenomenal character of experiencing /da/ is constitutively multisensory. Thus, the unisensory correspondence model has trouble distinguishing constitutively multisensory experiences from unisensory experiences that are only producible through subpersonal crossmodal interaction.

One response to this problem is to relax the requirement that corresponding unimodal experiences be producible under *equivalent* stimulation. Perhaps they need only be producible under some *appropriately similar* stimulation. Even if the McGurk subject could not have experienced /da/, given only the acoustic input she, in fact, received, there is a related acoustic stimulus that could have produced an experience with that phenomenal character without concurrent inputs to other sense organs. It is unclear just what it means for two conditions of stimulation to be 'appropriately similar', and certain construals raise further puzzles (Wadle, 2020, pp. 14–15). However, as I will explain below, my application of the unisensory correspondence model does not hinge on how exactly we construe the notion of appropriately similar stimulation, so we need not settle the issue here.[9]

Multisensory diachronic grouping

We can distinguish the phenomenal character associated with an object's apparent *persistence* from that associated with its other properties. Suppose you see a red ball pass behind a barrier and then emerge. You have a strong impression that the emerging object is the *same one* that passed behind the barrier. This impression of persistence plausibly possesses a distinctive phenomenal character separate from that associated with, say, the ball's apparent colour and shape. Studies suggest that by slightly increasing the interval before an object emerges, we can

[9] For those doubtful that any version of the unisensory-correspondence model can succeed, a weaker principle would suffice for my purposes. Specifically: if someone perceptually experiences a relation between two objects, or two time slices of an object, that are wholly apprehended through separate modalities, then the phenomenal character associated with experiencing this relation is constitutively multisensory. Suppose you experience a visible event as *simultaneous* with an audible event. Then the experience of simultaneity is neither uniquely visual nor uniquely auditory, but constitutively audiovisual (O'Callaghan, 2017, pp. 166–7). Likewise, if you perceive an object as persisting from T1 to T2, but it is perceived wholly through vision at T1 and wholly through audition at T2, then the experience of persistence is neither uniquely visual nor uniquely auditory, but constitutively audiovisual.

remove the impression of persistence while leaving the phenomenology of other properties unchanged (Flombaum and Scholl, 2006).

My present interest concerns whether this distinctive phenomenology of persistence is ever constitutively multisensory. Given O'Callaghan's unisensory-correspondence criterion, the issue becomes more precise: consider the phenomenal character of persistence exhibited by an experience of an object moving through the scene. Could that very phenomenal character—abstracting away from the phenomenology associated with other features—be instantiated by a corresponding experience merely of a single modality? If it could not, then it is constitutively multisensory.

What could establish constitutively multisensory phenomenology of persistence? Let us say *multisensory diachronic grouping* occurs if an object is perceived wholly through one modality at time T1, then wholly through a separate modality at time T2, but is experienced as persisting from T1 to T2. For example, you might see a ball roll behind a barrier at T1, then feel the ball behind the barrier at T2. If you experience the ball as persisting from T1 to T2, then your experience exhibits multisensory diachronic grouping. This would establish constitutively multisensory experience of persistence. No corresponding unimodal experience could share the phenomenal character associated with perceiving the ball as persisting from T1 to T2. Any experience produced under equivalent visual stimulation, but without haptic stimulation, would omit the object at T2, and vice versa for haptic stimulation without vision. This holds even if we relax the requirement of equivalent unimodal stimulation. No remotely similar pattern of visual stimulation could produce an experience of the object as persisting at T2.

The section 'Crossmodal influences on apparent motion', p. XXX examined cases where apparent motion in one modality influences apparent motion in another modality. While these are not candidate instances of multisensory diachronic grouping, other crossmodal stimuli are—namely, displays where stimuli are alternately presented to different modalities and the subject is asked whether they perceive motion between them.

Consider visual–tactile apparent motion. Harrar *et al*. (2008) had subjects place their index fingers in a pair of cups that could emit brief taps to the skin. Both cups had LED displays mounted on top that emitted brief flashes. This set-up allowed comparison of visual–visual, tactile–tactile, and visual–tactile apparent motion sequences by presenting flashes and/or taps in rapid succession. Subjects rated the quality of apparent motion across various inter-stimulus distances and temporal intervals. Crucially, for inter-stimulus delays of 200–300 ms, the quality of visual–tactile apparent motion was equivalent to visual–visual and tactile–tactile motion.

There is debate about whether reports of crossmodal apparent motion are genuinely perceptual (O'Callaghan, 2017), or instead result from post-perceptual inference or compliance with task demands (Spence, 2015; Spence and Bayne, 2014). I believe the non-perceptual hypothesis is unlikely in the present case. Suppose that subjects reported experiencing visual–tactile motion simply because they believed that the purpose of the experiment was to document its existence. This account provides no obvious explanation of why the rated quality of visual–tactile motion was systematically governed by the temporal interval between stimuli. Further support for the perceptual account derives from evidence that crossmodal apparent motion has the propensity to capture unimodal auditory apparent motion (Jiang and Chen, 2013).

Sceptics about crossmodal apparent motion have observed that in another study, Huddleston *et al*. (2008) tested for the presence of audiovisual apparent motion and found negative results. Subjects were shown an array of two loudspeakers and two LED displays arranged in a circle. Lights and white noise bursts occurred in either clockwise and counterclockwise order. The authors found that while subjects could reliably distinguish the implied direction of audiovisual motion, they reported no genuine percepts of audiovisual motion. However, another study investigated audiovisual apparent motion, with more promising results.

Kluss *et al*. (2012) had subjects sit in a chair surrounded by loudspeakers arranged in a semicircle, with an LED attached to each speaker cone (Figure 20.3). They compared three conditions: a *unimodal–auditory* condition where, say, the speakers at −54°, −18°, 18°, and 54° emitted noise bursts in rapid succession; a *coherent–bimodal* condition, where the sounds were presented in the same order, but flashes were spatio-temporally interpolated between the sounds; and an *incoherent–bimodal* condition, where the sounds were presented in the same order, but flashes were temporally interpolated between them at random locations. Relative to the other conditions, subjects reported perceiving continuous motion at longer temporal intervals between sounds in the coherent–bimodal condition. It is as though flashes at the appropriate spatial locations helped 'fill the gaps' between sounds, promoting a percept of seamless motion. The authors take this to support an 'amodal interpolation subsystem accepting unimodal activity from both the auditory and the visual domain' (2012, p. 64; cp. Stiles *et al*., 2018).

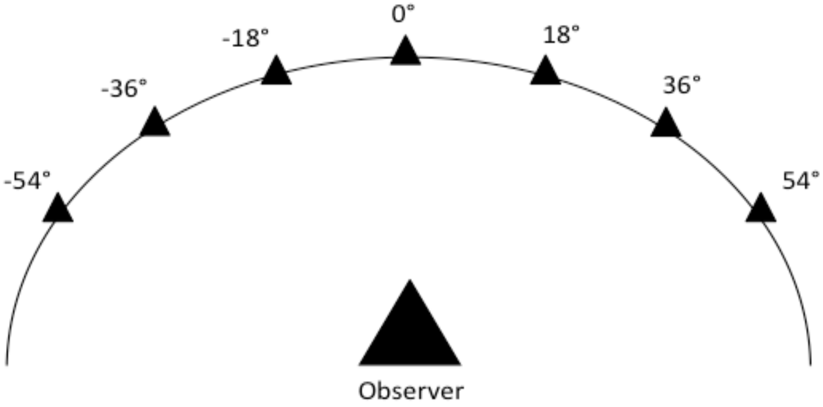

Figure 20.3. Design of the Kluss *et al.* (2012) study.

Suppose, then, that studies of crossmodal apparent motion are unearthing a real perceptual phenomenon. Do they establish constitutively multisensory experience of persistence? I argue that, at present, they do not. While crossmodal apparent motion involves constitutively multisensory experiences of *some* kind, it is unclear whether these are experiences of persistence, or of some other relation.

Consider Harrar *et al.*'s (2008) study of visual–tactile apparent motion. I argued that the systematic dependence of quality ratings on inter-stimulus interval suggests that subjects were reporting their perceptual experiences at face value. At longer intervals, the subjects experienced the stimuli as standing in some relation—a relation they did not experience at shorter intervals. Consider the phenomenology associated with perceiving this relation. On the unisensory-correspondence model, this phenomenology is constitutively multisensory if it is not shared by any corresponding experience merely of a single modality. Plausibly this is true. Any merely visual experience would omit the tap, while any merely tactual experience would omit the flash. And if one of the relata is not experienced at all, then neither is the relation between them. Thus, experiences of crossmodal apparent motion have constitutively multisensory phenomenology.

But do they have constitutively multisensory phenomenology of *persistence*? When subjects see a flash and then feel a tap, do they experience the tap as a *continuation* of the flash? While this may seem like a reasonable interpretation, subjects' self-reports in the Harrar *et al.*'s (2008) study suggest otherwise. Thus (p. 810):

> [S]ubjects in the visuotactile condition reported perceiving some type of multimodal apparent motion, but they often described it as being 'more causal' than the unimodal apparent motion. Our participants mainly interpreted their perception like a switch flicking on a light or like a cannon firing that was felt on

> one hand and then the flash from the landing explosive was seen on the other hand.

As I have argued elsewhere (Green, 2019b), such reports raise the possibility that in crossmodal apparent motion, subjects do not experience persistence at all, but rather a causal relation between distinct objects or events. Compare Michotte-style causal perception displays (Kominsky *et al*., 2017), wherein one disc approaches another and contacts it, after which the second disc begins moving in the same direction. Here subjects report a strong impression that the first disc *launched* the second. But while they arguably experience the discs as causally related, they do not experience them as stages of a single persisting object. Crossmodal apparent motion may fit the same mould.[10]

How could we settle whether crossmodal apparent motion involves experience of persistence? One option would be to apply paradigms thought to reveal the perception of persistence in unimodal cases. One such paradigm is the *object-reviewing* task (Kahneman *et al*., 1992): subjects perceive a pair of objects, and features briefly appear on the objects before vanishing. Then the objects move to new locations, and a feature appears on one of them. Subjects' task is to report whether the feature matches either of those encountered earlier. Typically, responses are faster when there is a match, but faster still when the feature appears in the object in which it initially appeared—an *object-specific preview benefit* (OSPB). The standard explanation of the OSPB is that the perceptual system maintains object

[10] This story could also apply to audiovisual apparent motion in Kluss *et al*. (2012). Rather than experiencing a single moving object, perhaps subjects experienced an uninterrupted causal chain linking the lights and sounds, and simply reported this as continuous motion for lack of a better description.

representations—*object files*—over time and stores information about an object's currently and recently perceived features in its file (Green and Quilty-Dunn, 2021). When an object's feature matches information already in its file, responses to the feature are speeded. To perceptually represent an object as persisting from time T1 to T2 just is to maintain a file that sustains reference to the object from T1 to T2.

Suppose this story is correct. Then a natural idea would be to examine object-reviewing in a crossmodal context. Suppose an object is first seen, then moves behind an occluder where it can only be touched. And suppose that the features to be compared are perceptible both visually and haptically (e.g., a simple shape or oriented line segment). If an OSPB emerges in reidentifying these features, then the perceptual system must have deemed the object to persist from one time to the next despite being perceived wholly through separate modalities.

Notably, there is evidence that OSPBs are not modality-specific. Jordan *et al*. (2010) found that when subjects saw a picture of a telephone appear within an object, there was an OSPB for reidentifying the feature by its audible ring (see also Zmigrod *et al*., 2009). However, while Jordan *et al*.'s experiment provides evidence for multisensory *binding*, it does not directly bear on the question of multisensory *persistence*, since the objects (square-shaped wireframes) to which the features were bound could have been selected and reidentified through vision alone. To address the persistence issue, we need a case where the *objects*, and not merely the features bound to them, are picked out and then reidentified using separate modalities.

Unfortunately, there are problems with using object-file maintenance as a guide to the experience of persistence. Some evidence has been taken to show that object files sometimes diverge from the experience of persisting objecthood (Mitroff *et al*., 2005). In general, however, the perceptual experience of persistence is at least reliably yoked to the maintenance of object files. Odic *et al*. (2012) examined a variety of subtly different apparent motion displays biased towards different correspondence solutions and found that object-file maintenance, as indexed by the OSPB, did systematically track the conscious experience of persistence in apparent motion. Divergence between object files and perceptually experienced persistence is probably the exception, not the rule.

I suggest that while studies of object-file maintenance do not offer an *experimentum crucis* regarding whether crossmodal apparent motion involves the experience of persistence, they offer a key source of evidence. Objects in crossmodal apparent motion are experienced as 'linked' in some way—the question is what sort of link they are experienced as exhibiting. So suppose an OSPB is produced during crossmodal apparent motion, indicating that an object-file is maintained throughout. Then we would have evidence that the perceptual system is representing persistence between objects apprehended through separate modalities. A prima facie plausible hypothesis would be that the associated perceptual experience represents persistence as well. If so, that experience would be constitutively multisensory.

Summing up: it remains an open question whether we enjoy constitutively multisensory experiences of persistence. This section has described the type of phenomenon (multisensory diachronic grouping) that would establish this convincingly. However, while crossmodal apparent motion supplies suggestive evidence for multisensory diachronic grouping, alternative interpretations remain viable. The issue is not whether crossmodal apparent motion is *perceptual*, but whether it involves the perception of *persistence*.

Conclusion

This chapter has explored multisensory interactions in the perception of persistence. This issue can be framed at the level of either perceptual processing or perceptual experience. I have argued that research on the audiovisual bounce effect offers compelling evidence that certain perceptual persistence computations are directly penetrated by representations from more than one modality. I then considered the issue of whether perceptual experiences of persistence are ever constitutively multisensory. I argued that the available evidence does not settle this question, but that it is not beyond empirical resolution.

Acknowledgements

Thanks to Tyler Brooke-Wilson, Alex Byrne, Casey O'Callaghan, and an anonymous reviewer for helpful comments that greatly improved this chapter.

References

Adams, K. L. and Grove, P. M. (2018). The effect of transient location on the resolution of bistable visual and audiovisual motion sequences. *Perception*, 47(9), 927–42.

Alink, A., Euler, F., Galeano, E., Krugliak, A., Singer, W., and Kohler, A. (2012). Auditory motion capturing ambiguous visual motion. *Frontiers in Psychology*, 2(391), 1–8.

Cavanagh, P. (1992). Attention-based motion perception. *Science*, 257(5076), 1563–5.

Cavanagh, P., Hunt, A. R., Afraz, A., and Rolfs, M. (2010). Visual stability based on remapping of attention pointers. *Trends in Cognitive Sciences*, 14(4), 147–53.

Deroy, O. (2014). The unity assumption and the many unities of consciousness. In *Sensory Integration and the Unity of Consciousness*, edited by D. J. Bennett and C. Hill, 105–24. Cambridge, MA: MIT Press.

de Vignemont, F. (2014). Multimodal unity and multimodal binding. In *Sensory Integration and the Unity of Consciousness*, edited by D. J. Bennett and C. Hill, 125–50. Cambridge, MA: MIT Press.

Driver, J. and Spence, C. (2004). Crossmodal spatial attention: evidence from human performance. In *Crossmodal Space and Crossmodal Attention*, edited by C. Spence and J. Driver, 179–220. Oxford: Oxford University Press.

Flombaum, J. I. and Scholl, B. J. (2006). A temporal same-object advantage in the tunnel effect: facilitated change detection for persisting objects. *Journal of Experimental Psychology: Human Perception and Performance*, 32(4), 840–53.

Freeman, E. and Driver, J. (2008). Direction of visual apparent motion driven solely by timing of a static sound. *Current Biology*, 18(16), 1262–6.

Fulkerson, M. (2014). Rethinking the senses and their interactions: the case for sensory pluralism. *Frontiers in Psychology*, 5, 1426.

Grassi, M. and Casco, C. (2009). Audiovisual bounce-inducing effect: attention alone does not explain why the discs are bouncing. *Journal of Experimental Psychology: Human Perception and Performance*, 35(1), 235–43.

Grassi, M. and Casco, C. (2010). Audiovisual bounce-inducing effect: when sound congruence affects grouping in vision. *Attention, Perception, and Psychophysics*, 72(2), 378–86.

Green, E. J. (2019a) A theory of perceptual objects. *Philosophy and Phenomenological Research*, 99(3), 663–93.

Green, E. J. (2019b). Binding and differentiation in multisensory object perception. *Synthese*, doi: https://doi.org/10.1007/s11229-019-02351-1

Green, E. J. (2020). The perception–cognition border: a case for architectural division. *Philosophical Review*, 129(3), 323–93.

Green, E. J. and Quilty-Dunn, J. (2021) What is an object file? *British Journal for the Philosophy of Science*, 72(3), 665-699.

Gross, S. (2017). Cognitive penetration and attention. *Frontiers in Psychology*, 8, 1–12.

Harrar, V., Winter, R., and Harris, L. R. (2008). Visuotactile apparent motion. *Perception and Psychophysics*, 70(5), 807–17.

Huddleston, W. E., Lewis, J. W., Phinney, R. E., and DeYoe, E. A. (2008). Auditory and visual attention-based apparent motion share functional parallels. *Perception and Psychophysics*, 70(7), 1207–16.

Jiang, Y. and Chen, L. (2013). Mutual influences of intermodal visual/tactile apparent motion and auditory motion with uncrossed and crossed arms. *Multisensory Research*, 26(1–2), 19–51.

Jordan, K. E., Clark, K., and Mitroff, S. R. (2010). See an object, hear an object file: object correspondence transcends sensory modality. *Visual Cognition*, 18(4), 492–503.

Kahneman, D., Treisman, A., and Gibbs, B. J. (1992). The reviewing of object files: object-specific integration of information. *Cognitive Psychology*, 24, 175–219.

Kluss, T., Schult, N., Schill, K., Fahle, M., and Zetzsche, C. (2012). Investigating the in-between: multisensory integration of auditory and visual motion streams. *Seeing and Perceiving*, 25, 45–69.

Kominsky, J. F., Strickland, B., Wertz, A. E., Elsner, C., Wynn, K., and Keil, F. C. (2017). Categories and constraints in causal perception. *Psychological Science*, 28(11), 1649–62.

Macpherson, F. (2011). Cross-modal experiences. *Proceedings of the Aristotelian Society,* 111(3), 429–68.

Macpherson, F. (2012). Cognitive penetration of colour experience: rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research*, 84(1), 24–62.

McBeath, M. K., Addie, J. D. and Krynen, R. C. (2019). Auditory capture of visual apparent motion, both laterally and looming. *Acta Psychologica*, 193, 105–12.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–8.

Meyerhoff, H. S., Merz, S., and Frings, C. (2018). Tactile stimulation disambiguates the perception of visual motion paths. *Psychonomic Bulletin and Review*, 25(6), 2231–7.

Meyerhoff, H. S. and Scholl, B. J. (2018). Auditory-induced bouncing is a perceptual (rather than a cognitive) phenomenon: evidence from illusory crescents. *Cognition*, 170, 88–94.

Meyerhoff, H. S. and Suzuki, S. (2018). Beep, be-, or–ep: the impact of auditory transients on perceived bouncing/streaming. *Journal of Experimental Psychology: Human Perception and Performance*, 44(12), 1995–2004.

Mitroff, S. R., Scholl, B. J., and Wynn, K. (2005). The relationship between object files and conscious perception. *Cognition*, 96(1), 67–92.

Morein-Zamir, S., Soto-Faraco, S., and Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloquism. *Cognitive Brain Research*, 17(1), 154–63.

O'Callaghan, C. (2014). Intermodal binding awareness. In *Sensory Integration and the Unity of Consciousness*, edited by D. Bennett and C. Hill, 73–103. Cambridge, MA: MIT Press.

O'Callaghan, C. (2015). The multisensory character of perception. *Journal of Philosophy*, 112(10), 551–69.

O'Callaghan, C. (2016). Objects for multisensory perception. *Philosophical Studies*, *173*, pp. 1269–1289.

O'Callaghan, C. (2017). Grades of multisensory awareness. *Mind and Language*, 32(2), 55–81.

O'Callaghan, C. (2019). *A Multisensory Philosophy of Perception*. Oxford: Oxford University Press.

Occelli, V., Spence, C., and Zampini, M. (2009). The effect of sound intensity on the audiotactile crossmodal dynamic capture effect. *Experimental Brain Research*, 193(3), 409–19.

Odic, D., Roth, O., and Flombaum, J. I. (2012). The relationship between apparent motion and object files. *Visual Cognition*, 20(9), 1052–81.

Oruc, I., Sinnett, S., Bischof, W. F., Soto-Faraco, S., Lock, K., and Kingstone, A. (2008). The effect of attention on the illusory capture of motion in bimodal stimuli. *Brain Research*, 1242, 200–8.

Paul, L. A. (2010). Temporal experience. *Journal of Philosophy*, 107(7), 333–59.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, 22(3), 341–65.

Quilty-Dunn, J. (2020). Attention and encapsulation. *Mind and Language*, 35(3), 335–49.

Roseboom, W., Kawabe, T., and Nishida, S. Y. (2013). Direction of visual apparent motion driven by perceptual organization of cross-modal signals. *Journal of Vision*, 13(1):6.1–13.

Scholl, B. J. (2007). Object persistence in philosophy and psychology. *Mind and Language*, 22(5), 563–91.

Scholl, B. J. and Nakayama, K. (2004). Illusory causal crescents: misperceived spatial relations due to perceived causality. *Perception*, 33(4), 455–69.

Sekuler, R., Sekuler, A. B., and Lau, R. (1997). Sound alters visual motion perception. *Nature*, 385, 308.

Skrzypulec, B., 2020. Visual endurance and auditory perdurance. *Erkenntnis*, *85*, pp. 467-488.

Soto-Faraco, S., Spence, C., and Kingstone, A. (2004). Congruency effects between auditory and tactile motion: extending the phenomenon of cross-modal dynamic capture. *Cognitive, Affective, and Behavioral Neuroscience*, 4(2), 208–17.

Spence, C. (2015). Cross-modal perceptual organization. In *The Oxford Handbook of Perceptual Organization*, edited by J. Wagemans, 639–54. Oxford: Oxford University Press.

Spence, C. and Bayne, T. (2014). Is consciousness multisensory? In *Perception and Its Modalities*, edited by D. Stokes, M. Matthen, and S. Biggs, 95–132. Oxford: Oxford University Press.

Spence, C. and Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. *Perception and Psychophysics*, 59(1), 1–22.

Stiles, N. R., Li, M., Levitan, C. A., Kamitani, Y., and Shimojo, S. (2018). What you saw is what you will hear: two new illusions with audiovisual postdictive effects. *PLoS One*, 13(10), e0204217.

van Dam, L. C. J., Parise, C. V., and Ernst, M. O. (2014). Modeling multisensory integration. In *Sensory Integration and the Unity of Consciousness*, edited by D. J. Bennett and C. Hill, 209–29. Cambridge, MA: MIT Press.

Verstraten, F. A., Cavanagh, P., and Labianca, A. T. (2000). Limits of attentive tracking reveal temporal properties of attention. *Vision Research*, 40(26), 3651–64.

von Grünau, M. W. (1986). A motion aftereffect for long-range troboscopic apparent motion. *Perception and Psychophysics*, 40(1), 31–8.

Wadle, D. C. (2020). Sensory modalities and novel features of perceptual experiences. *Synthese*, doi: https://doi.org/10.1007/s11229-020-02689-x

Watanabe, K. and Shimojo, S. (1998). Attentional modulation in perception of visual motion events. *Perception*, 27(9), 1041–54.

Watanabe, K. and Shimojo, S. (2005). Crossmodal attention in event perception. In *Neurobiology of Attention*, edited by L. Itti, G. Rees, and J. Tsotsos, 538–43. New York, NY: Elsevier Academic Press.

Xu, Y., Suzuki, S., and Franconeri, S. L. (2013). Shifting selection may control apparent motion. *Psychological Science*, 24(7), 1368–70.

Zmigrod, S., Spapé, M., and Hommel, B. (2009). Intermodal event files: integrating features across vision, audition, taction, and action. *Psychological Research PRPF*, 73(5), 674–84.