Predictive coding of visual motion in both monocular and binocular human visual processing

Elle van Heusden
Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne, Australia
Helmholtz Institute, Department of Experimental Psychology, Utrecht University, Utrecht, the Netherlands

Anthony M. Harris
Institute of Cognitive Neuroscience, University College London, London, UK
Queensland Brain Institute, The University of Queensland, Brisbane, Australia

Marta I. Garrido
Queensland Brain Institute, The University of Queensland, Brisbane, Australia
School of Mathematics and Physics, The University of Queensland, Brisbane, Australia
Centre for Advanced Imaging, The University of Queensland, Brisbane, Australia
Australian Research Council Centre of Excellence for Integrative Brain Function, The University of Queensland, Brisbane, Australia

Hinze Hogendoorn
Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne, Australia
Helmholtz Institute, Department of Experimental Psychology, Utrecht University, Utrecht, the Netherlands

Neural processing of sensory input in the brain takes time, and for that reason our awareness of visual events lags behind their actual occurrence. One way the brain might compensate to minimize the impact of the resulting delays is through extrapolation. Extrapolation mechanisms have been argued to underlie perceptual illusions in which moving and static stimuli are mislocalised relative to one another (such as the flash-lag and related effects). However, where in the visual hierarchy such extrapolation processes take place remains unknown. Here, we address this question by identifying monocular and binocular contributions to the flash-grab illusion. In this illusion, a brief target is flashed on a moving background that reverses direction. As a result, the perceived position of the target is shifted in the direction of the reversal. We show that the illusion is attenuated, but not eliminated, when the motion reversal and the target are presented dichoptically to separate eyes. This reveals extrapolation mechanisms at both monocular and binocular processing stages contribute to the illusion. We interpret the results in a hierarchical predictive coding framework, and argue that prediction errors in this framework manifest directly as perceptual illusions.

Introduction

Neural processing of sensory input in the brain takes time, and for that reason our awareness of visual events lags behind their actual occurrence. If the visual system did not somehow compensate for neural transmission delays, we would consistently mislocalize moving

Citation: van Heusden, E., Harris, A. M., Garrido, M. I., & Hogendoorn, H. (2019). Predictive coding of visual motion in both monocular and binocular human visual processing. *Journal of Vision*, *19*(1):3, 1–12, https://doi.org/10.1167/19.1.3.

https://doi.org/10.1167/19.1.3 Received July 27, 2018; published January 10, 2019 ISSN 1534-7362

objects behind their actual position. Nevertheless, human observers are typically very accurate at localizing moving objects, achieving near-zero lag when object trajectories are predictable (Brenner, Smeets, & de Lussanet, 1998). One explanation for how the brain might overcome its internal delay is through extrapolation: By exploiting knowledge about an object's past trajectory, the brain predicts its present position.

Although accurate interaction with moving objects could also be achieved by extrapolation in the motor system (e.g., Kerzel & Gegenfurtner, 2003), extrapolation mechanisms in the visual system have been hypothesized to underlie a class of visual illusions in which visual motion signals affect the perceived location of stationary objects. This includes the much-studied flash-lag effect, in which a moving object that is physically aligned with a stationary flash is perceived ahead of that flash (Nijhawan, 1994). In this interpretation, the brain extrapolates the position of the moving object along its expected trajectory to compensate for lag that would otherwise arise due to processing time. When the flash is presented aligned with the moving object, it is compared to the *extrapolated* position of the moving object, and hence appears to lag behind it.

In the years following Nijhawan's initial demonstrations of the flash-lag effect, numerous other motion-induced position shifts have been reported, including the flash-drag (Whitney & Cavanagh, 2000), flash-jump (Cai & Schlag, 2001), and flash-grab (Cavanagh & Anstis, 2013) effects. Although the underlying mechanisms have been hotly debated (e.g., Eagleman & Sejnowski, 2000; Krekelberg, 2000; Patel, Ogmen, Bedell, & Sampath, 2000; Whitney & Murakami, 1998), convergent evidence points to an important role for predictive extrapolation mechanisms in causing these effects (Nijhawan, 2008). For instance, animal neurophysiology studies have demonstrated the existence of predictive extrapolation mechanisms in the retinae of salamanders, mice, and rabbits (Berry, Brivanlou, Jordan, & Meister, 1999; Schwartz, Taylor, Fisher, & Harris, 2007), as well as in cat primary visual cortex (Jancke, Erlhagen, Schöner, & Dinse, 2004). In humans, it has been demonstrated that moving objects are extrapolated into regions of visual space where they could physically not be detected, such as the blind spot—ruling out explanations in terms of differential latencies (Maus & Nijhawan, 2008). Modeling studies have shown that a Bayesian model of perceived position that incorporates neural delays generates predictive position shifts such as seen in the flash-lag effect (Khoei, Masson, & Perrinet, 2017), and most recently an unsupervised predictive neural network exposed to natural video sequences (including motion) was found to have developed a pattern of response consistent with the flash-lag effect (Lotter, Kreiman, & Cox, 2018).

An important conceptual challenge to interpreting motion-induced position shifts as resulting from predictive mechanisms arises from the observation that the perceived position of a static event is biased primarily by motion presented *after* that event, rather than before it. This led Eagleman and Sejnowski (2000) to coin the term *post-diction*, as a temporal counterpart to *prediction*. In this original post-diction account, events were essentially back-dated in perception, rewriting recent perceptual history. Several years later, the same authors presented a refined version of this model, in which local velocity signals integrated over a brief period *after* an event interact with local position signals to bias its perceived position (Eagleman & Sejnowski, 2007). Much has been made of what seems like reverse causality in the postdiction account, and the apparent contrast with predictive extrapolation mechanisms. However, the predictive model and the postdictive motion-biasing model are mechanistically the same, differing only in the time-window during which motion signals are integrated. Eagleman and Sejnowski (2007) note that "Motion biasing will normally push objects closer to their true location in the world [. . .] by a clever method of updating signals that have become stale due to processing time" (p. 9), which is precisely what motion extrapolation also does (Nijhawan, 2008). Eagleman has more recently argued that prediction and postdiction cooperate to compensate for neural delays (Eagleman, 2008), simply because predictions by their nature sometimes do not come true, necessitating posthoc revisions to the timeline of experience. Viewed more broadly, prediction and postdiction are simply two halves of the same mechanism, split along the line separating past from future. However, in the context of sensory processing, this line is artificial: Due to neural delays, *all* cortical areas process information collected in the objective past. Although they seem polar opposites, predictive and postdictive accounts both push the representation of an object closer to its true location in the world at a given instant. Given the reality of neural processing delays, this means anticipating (i.e., predicting) the present.

Interestingly, this predictive rationale fits neatly with a more recent computational model aiming to unify position and motion perception (Kwon, Tadin, & Knill, 2015), which applied a Bayesian approach to another motion-position illusion (motion-induced position shifts; De Valois & De Valois, 1991; Ramachandran & Anstis, 1990). Kwon et al. advocate a model in which motion and position judgments mutually interact to make optimal inferences about the generative causes underlying sensory signals. The model is implemented as a Kalman filter, and therefore

the represented position at a given time explicitly depends only on velocity signals integrated before that time. However, the precise time-window over which signals are integrated relative to objective external time was not the focus of the model, nor does it invalidate the mechanistic similarities it shares with the account proposed earlier by Eagleman and Sejnowski (2007). Indeed, Eagleman and Sejnowski (2000) note that postdiction is commonplace in engineering, where it is simply known as smoothing. Most importantly, the model's core feature—that it causes a position signal to be shifted in the direction of a motion signal—is the same: It is a predictive mechanism that causes anticipatory activation at the object's future position.

A more recent illusion, the flash-grab effect, has provided the opportunity to study how predictive motion extrapolation mechanisms behave when motion vectors abruptly change, such that the anticipated future does not come true (Cavanagh & Anstis, 2013). In this illusion, a target is briefly flashed on a moving background as the motion unexpectedly reverses direction, which results in the perceived position of the flash being shifted in the direction of the second motion sequence. Although neither Eagleman and Sejnowski (2007) nor Kwon et al. (2015) made reference to the flash-grab effect, the flash-grab effect can also be readily explained by the same mechanism. Figure 1 shows schematically how this would work. A moving object is represented at a given level of processing with a certain delay. It is possible to compensate for that delay by using information about the object's velocity to extrapolate the true position of the object at that instant (e.g., Krekelberg & Lappe, 2001; Nijhawan, 1994; see Figure 1A). When the object reverses direction, it again takes time to detect the reversal, during which the object's position continues to be extrapolated beyond the reversal point (Figure 1B). When sensory information about the object's actual trajectory then becomes available, the represented position must rapidly shift from the predicted trajectory to the new trajectory. This rapid shift in represented position equates to a brief spike in velocity (Figure 1C, upper plot). Importantly, the key features in Figure 1 are not hypothetical: The overshoot in represented position and subsequent acceleration to intercept the new trajectory exactly mirror population codes reported in the mouse and salamander retina for such reversing stimuli (Schwartz et al., 2007). Although such mechanisms have not yet been directly demonstrated in the brain itself, because prediction error signals arise already in the retina, they are passed on to the rest of the visual processing hierarchy even if they would not be calculated there. The mechanism proposed by Eagleman and Sejnowski (2007) then predicts that any stationary object flashed at the reversal point would interact with this motion signal and be mislocalized. This is the effect we know as the flash-grab effect (Cavanagh & Anstis, 2013). Importantly, the magnitude of this mislocalization would be a direct reflection of how far into the future the neural representation of the moving object has been extrapolated. As is evident in Figure 1A through C, the longer the processing delay, the further into the (local) future a representation must be extrapolated in order to compensate, and the longer it would take before a violation of that extrapolation is detected in that brain area. This would yield a stronger velocity spike as the area "catches up," and a bigger flash-grab effect.

Consistent with this interpretation of the flash-grab effect as resulting from failed prediction, we recently demonstrated that the same neural mechanisms that cause the location of the target in the flash-grab effect to be misperceived also influence saccades aimed at that target (Van Heusden, Rolfs, Cavanagh, & Hogendoorn, 2018). Most importantly, we showed that the degree of saccade error increased with increasing saccadic latency. This indicates that the visuomotor system was extrapolating the (stationary) target's position as if it was actually moving, confirming that mislocalization in this illusion results from an extrapolation process. It can be readily appreciated from Figure 1 that if processing delays are larger (for whatever reason), then the original trajectory will be extrapolated further into the future, the object will move even further along its actual trajectory, and the total position error will be greater. The transient peak in velocity as the system adjusts will therefore also be greater, in turn leading to a larger flash-grab effect.

Where in the visual hierarchy the neural mechanisms responsible for extrapolation operate is still unknown. In animals, predictive neural mechanisms have been identified at multiple levels of the visual system, including the retina (Berry et al., 1999; Hosoya, Baccus, & Meister, 2005; Schwartz et al., 2007), lateral geniculate nucleus (Sillito, Jones, Gerstein, & West, 1994), primary visual cortex (Jancke et al., 2004), and V4 (Sundberg, Fallah, & Reynolds, 2006). In humans, we recently demonstrated that the visual brain predicts the position of a moving object using an EEG classification paradigm (Hogendoorn & Burkitt, 2018a). This study revealed that for an object in apparent motion, the neural representation of the object's position is preactivated when the object moves along a predictable trajectory. However, this was only true for neural representations evoked around 130 ms after stimulus presentation, whereas the latency of earlier neural position representations was not modulated by prediction. In contrast, a previous EEG study of the flash-grab effect revealed that the target's illusory position was represented in the EEG signal as early as 81 ms poststimulus (Hogendoorn, Verstraten, & Cavanagh, 2015). These two studies therefore seem to

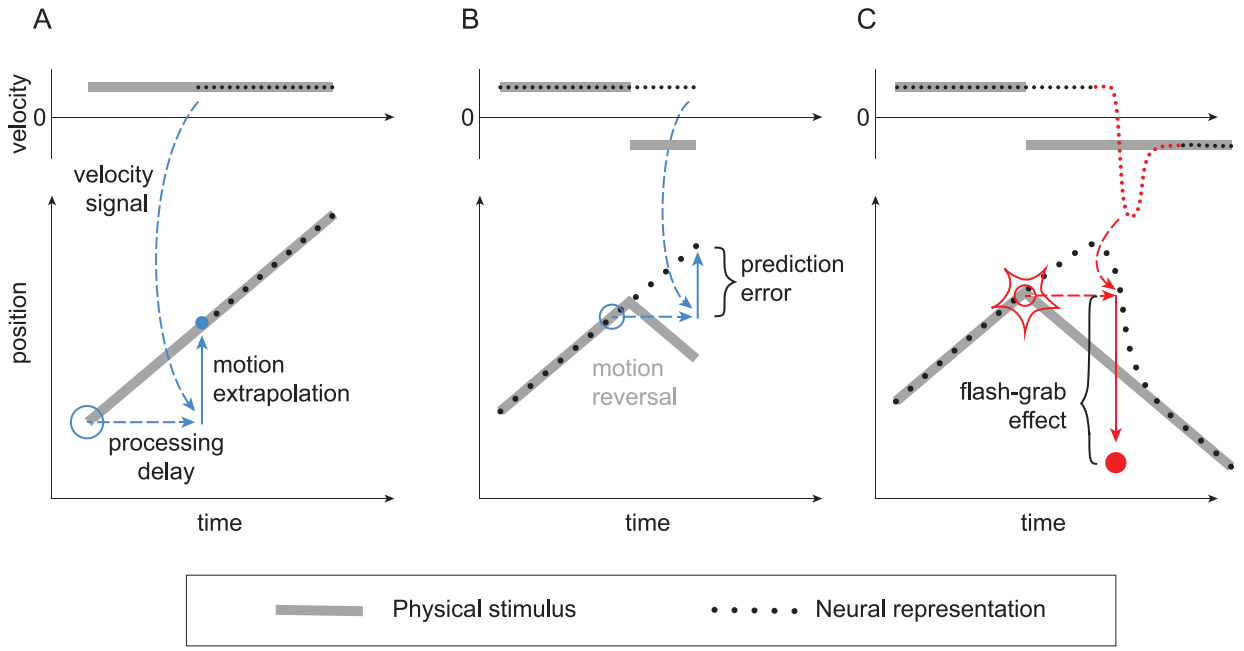

Figure 1. Schematic illustration of extrapolation in the flash-grab effect. In each panel, the lower plot shows position as a function of time, and the upper plot shows velocity as a function of time. Solid gray traces indicate the properties of the physical stimulus as presented on the screen, and dotted black traces indicate (predictive) neural representations of the same stimulus, as demonstrated empirically in the retina by Schwartz et al. (2007). (A) In order to accurately localize a moving object despite neural transmission delays, the visual system uses concurrent velocity signals to extrapolate the real-time position of the object (blue lines). (B) When an object unexpectedly reverses direction, at any given level of representation, some time elapses before the reversal is detected. During that time, the object will continue to be (erroneously) extrapolated into positions where it is never presented, creating a prediction error. (C) As the represented position shifts from the predicted trajectory to a new trajectory, the rapid shift in represented position creates a brief spike in the represented velocity (dotted red trace). If a (stationary) flash is presented at the same time as the reversal, then the position of the flash will interact with the (large) transient velocity signal and be mislocalised, resulting in the flash-grab effect (red lines).

reveal extrapolation processes at different stages in the visual hierarchy. Hence, the limited evidence from humans is consistent with the evidence from animal neurophysiology.

With the cortical EEG reflecting the target's extrapolated position already at about 80 ms post-stimulus, the question is, where along the route from retina to cortex does this extrapolation take place? The vast majority of visual information reaches the cortex through the retino-geniculo-cortical pathway, passing from retina to the lateral geniculate nucleus of the thalamus before flowing on to the primary visual cortex (V1). Although there are alternative pathways to the cortex (a point to which we return in the discussion), given the severely limited timeframe, it is likely that the visual extrapolation mechanisms responsible for the flash-grab effect operate along the geniculate pathway. This function would parallel the predictive mechanisms in the retina, LGN, and V1 revealed in animals, but it remains unknown whether (and if so, which of) these areas similarly carry out extrapolation in the human visual system.

In order to answer this question, here we make use of the fact that visual information from the two eyes does not converge until V1. Neurons that carry information either from the retina to LGN, or from LGN to V1, carry information from only one eye, with the first binocular neurons in the visual pathway located in V1 itself (Parker, 2007). We use the flash-grab effect, and employ dichoptic presentation to separate the different components of the flash-grab stimulus across the two eyes. In so doing, we prevent those components of the stimulus interacting at an early (monocular) stage of the visual hierarchy. We manipulate which components of the flash-grab stimulus sequence (motion prior to the flash, the flash itself, and the motion following the flash) are presented to which eye, and measure the strength of the resulting illusion. If the shift in the perceived position of a target presented in one eye is reduced when the flanking motion sequence is presented to the opposite eye, then this would be evidence that extrapolation mechanisms operate already at monocular stages of processing.

Here, we show that this is the case. The flash-grab effect is indeed attenuated when the flash is not presented in the same eye as both the preceding and the subsequent motion sequences. This makes a strong case for the existence of neural extrapolation mechanisms in early, monocular stages in the visual hierarchy. The fact that the illusion was not entirely eliminated in these conditions indicates that extrapolation also occurs in later binocular areas. Altogether, the results therefore point towards extrapolation computations being carried out at multiple stages of the early visual pathway.

Methods

Observers

Twenty observers performed the experiment. All observers had normal or corrected-to-normal vision and gave informed consent prior to participating. The experiment was conducted in accordance with the Declaration of Helsinki and was approved by the ethics committee of the Melbourne School of Psychological Sciences. Data from three observers were excluded from the analysis because target detection was lower than 50%. The remaining observers successfully detected an average of 88% of targets across conditions.

Apparatus

Stimuli were presented on an ASUS ROG Swift PG258Q monitor running at 100 Hz with a resolution of 1920 × 1080 pixels, controlled by a Dell Precision computer. The experiment was presented using MATLAB (MathWorks, Natick, MA) and Psychtoolbox 3.0.8 extensions (Brainard, 1997). A mirror-stereoscope set-up (including chin rest) was placed 50 cm away from the screen.

Stimulus

All stimuli were presented on a gray background. The stimuli consisted of two annuli (presented one to each eye), composed of 16 patches, which showed an alternating black and white pattern (see Figure 1A) and rotated at an angular velocity of 200° per second. The annuli had inner and outer radii of 4.3 and 6.1 degrees of visual angle (dva) respectively. The two annuli were viewed through a mirror-stereoscope and fused into a single percept (Figure 1B). A black square was presented around both annuli throughout the experiment to assist in maintaining binocular fusion. The square was 9.3 dva wide, drawn with a linewidth of 0.8 dva. Fixation dots were presented in the center of both annuli (diameter: 0.6 dva). To give the observers some reference as to where they perceived the target, and to aid binocular fusion and avoid torsional eye movements, a white line was presented on the vertical meridian just below both annuli (width: 0.06 dva; height: 0.5 dva; 3.4 dva from fixation).

Procedure

On each trial, observers viewed a rotating annulus for 1,000, 1,100, 1,200, 1,300, 1,400, or 1,500 ms (from now on referred to as the first motion sequence). During the very last frame of the motion sequence, a target (a red circle with a diameter of 0.6 dva) was presented at one of three possible target locations: 160°, 180°, or 200° polar angle offset from the top of the annulus, for a single frame (10 ms). Next, the direction of motion reversed and the annulus continued to rotate in the opposite direction for 400 ms, after which it gradually started to turn gray. The annulus was fully gray 100 ms later (these 500 ms are from now on referred to as the second motion sequence). This was done to ensure that participants were not distracted by the segments of the annulus when giving their response. At the end of each trial, observers used a mouse to report the position where they perceived the target. An image of the target was drawn at the cursor location for both eyes, and moved with the mouse cursor across the screen. When observers did not perceive the target, they were instructed to click at the location of the fixation dot.

Experimental design

Although observers perceived the same series of events on every trial, we used a mirror stereoscope to manipulate the information presented to each eye across five different conditions (Figure 1C). (a) In the Binocular condition, all the information (first motion sequence, the target, and the second motion sequence) was presented to both eyes; (b) in the monocular condition, all information was presented to one eye only; (c) in the interocular condition, both the first and second motion sequence were presented to one eye, while the target was presented to the other eye; (d) in the Before Reversal condition, the first motion sequence and the target were presented to one eye, while the second motion sequence was presented to the other eye; and (e) lastly, in the After Reversal condition, the first motion sequence was presented to one eye, while the target and the second motion sequence were presented to other eye. The first motion sequence (and

all consecutive events) occurred in the left and right eye with equal probability. The experiment consisted of nine blocks, with 110 trials in each block. All conditions were randomly interleaved within each block. On 10% of the trials, no target was presented. These trials served as catch-trials (1.8% of which were wrongfully reported).

Results

Observers viewed a flash-grab sequence (motion-flash-motion) in one of five conditions, in each case reporting the perceived position of the flashed target (Figure 2). The strength of the illusion was calculated as the polar angle between the reported position of the target and the target's real position, with errors in the direction of the second motion sequence (i.e., post-reversal) taken as positive. Mean illusion strength in each condition is plotted in Figure 3, with and without baseline-correcting for variability in the mean strength of the illusion across observers. All statistical analyses were carried out on the nonbaselined data.

First, one-sample t tests revealed that mislocalization was evident in all conditions (all $p < 0.001$). A repeated measures analysis of variance subsequently revealed a highly significant effect of condition, $F(4, 64) = 10.3$, $p = 1.7 \times 10^{-6}$, partial $\eta^2 = 0.391$. To further interpret the results, we made planned comparisons between the conditions using paired-samples t tests. Each condition was compared to the baseline Binocular condition, and additional planned pairwise comparisons were made between the three split conditions. The Monocular condition did not differ from the Binocular condition, $t(16) = -1.15$, $p = 0.26$, Cohen's $d = -0.28$. Conversely, mislocalization was significantly reduced in the Interocular, $t(16) = -3.43$, $p = 0.003$, Cohen's $d = -0.83$; Before Reversal, $t(16) = -4.67$, $p = 0.0003$, Cohen's $d = -1.13$; and After Reversal, $t(16) = -2.27$, $p = 0.037$, Cohen's $d = -0.55$, conditions. Finally, mislocalization did not differ significantly between the Interocular and After Reversal conditions, $t(16) = -1.0$, $p = 0.31$, Cohen's $d = -0.25$, although the Before Reversal condition was significantly reduced relative to the After Reversal condition, $t(16) = -3.09$, $p = 0.007$, Cohen's $d = -0.75$, and there was a trend suggesting that the Before Reversal condition might also produce less mislocalization than the Interocular condition, $t(16) = -1.98$, $p = 0.066$, Cohen's $d = -0.75$.

The strength of the illusion was maximal when the entire stimulus sequence was presented to either one or both eyes. Importantly, the strength of the illusion was reduced when the motion and the flash were presented in separate eyes (*Interocular* condition) as well as when the first and second motion sequences were presented to separate eyes (*Before Reversal* and *After Reversal* conditions). Maximal reduction (the weakest illusion) was evident in the Before Reversal condition in which the flash was presented to the eye that received the first motion sequence.

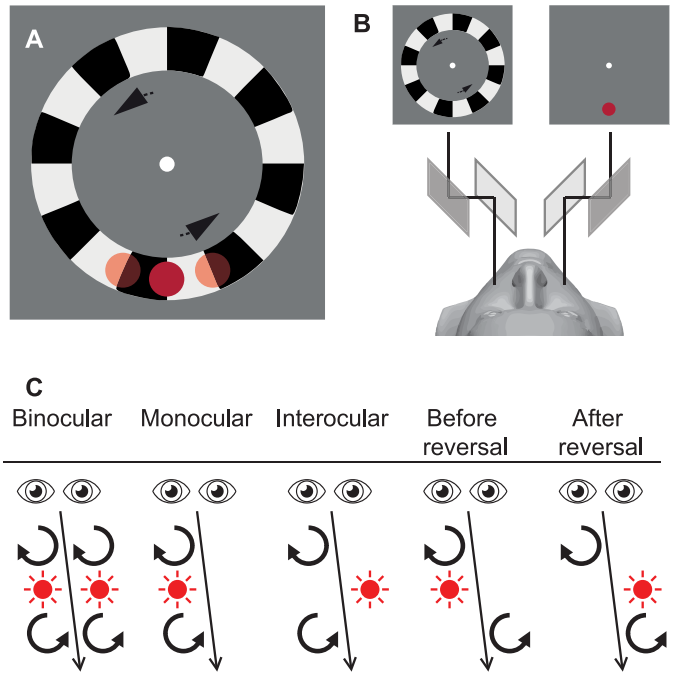

Figure 2. Stimulus and procedure. (A) Observers ($N = 17$) viewed a flash-grab sequence consisting of a rotating annulus that unexpectedly reversed its direction of motion. At the reversal, a red target disc was presented at one of three possible target locations, and observers reported the perceived location of the target after the trial using a mouse. (B) Using a mirror stereoscope, we manipulated the information presented to each eye. (C) Stimuli were presented in five different conditions. Binocular condition: All information is presented to both eyes. Monocular condition: All information is presented to one eye. Interocular condition: The moving annulus is presented to one eye, while the target is presented to the other eye. Before reversal condition: The first motion sequence and the target are presented to one eye, after which the annulus is presented to the other eye (rotating in the opposite direction). After reversal condition: The first motion sequence is presented to one eye, after which both the annulus (rotating in the opposite direction) and the target are presented in the other eye.

Discussion

We have previously demonstrated that the flash-grab effect (Cavanagh & Anstis, 2013) involves neural extrapolation mechanisms that operate very early on in the visual pathway (Hogendoorn et al., 2015; Van

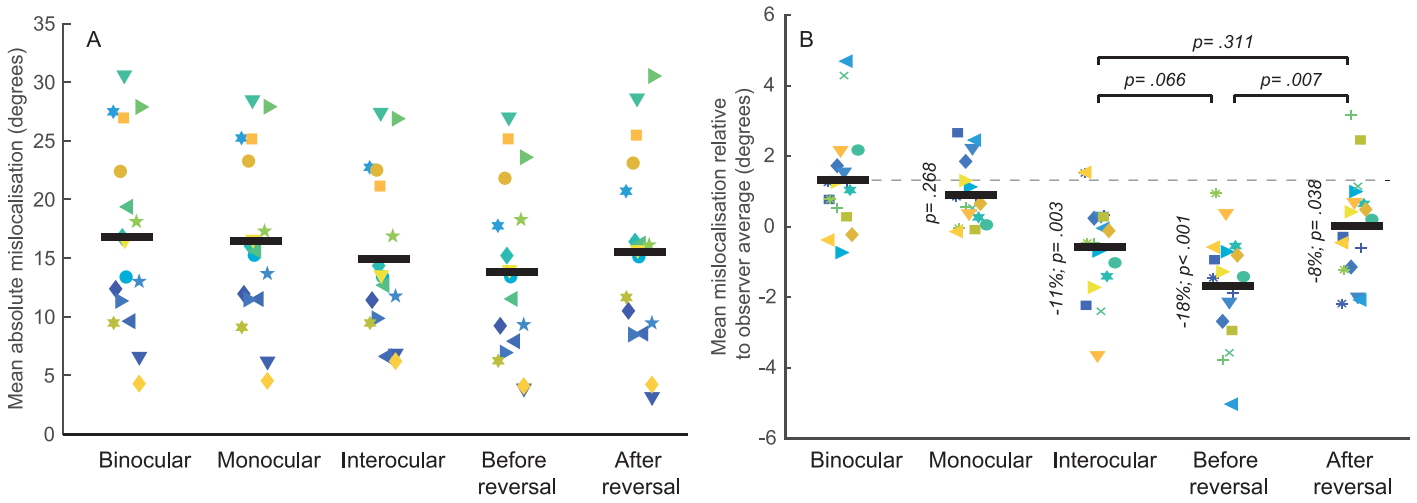

Figure 3. Results. (A) Mean mislocalization for individual observers in each of the five dichoptic presentation conditions. Each combination of marker shape and color represents an individual observer. Solid black lines indicate means across observers. (B) The same data after baseline-correction (subtracting the overall mean of each observer from each of the conditions for that observer). This reveals that although observers vary widely in the magnitude of the illusion, they all demonstrate a comparable pattern of illusion strength: In the three conditions in which the complete stimulus sequence was not presented to the same eye (Interocular, Before Reversal, and After Reversal conditions), the strength of the illusion was significantly attenuated. Statistical comparisons with paired-sample *t* tests are unaffected by the baseline correction and are illustrated only in Panel B for clarity. Vertical text indicates comparisons against the Binocular condition.

Heusden et al., 2018). Here, we used dichoptic presentation to further narrow down when and where these mechanisms operate. Because visual input from the two eyes does not converge until primary visual cortex, separating the components of the flash-grab effect allowed us to discriminate whether motion extrapolation takes place at early, monocular stages (possibly retinal or subcortical), or at later binocular stages (V1 and/or beyond). The results reveal that dichoptic presentation attenuates, but does not eliminate the illusion, indicating that extrapolation mechanisms operate in both monocular and binocular visual processing.

This finding is important for two reasons. Firstly, it is consistent with animal work showing predictive mechanisms in monocular parts of the early visual pathway, including the retina (Berry et al., 1999; Hosoya et al., 2005; Schwartz et al., 2007) and LGN (Sillito et al., 1994). The present results reflect a similar involvement of predictive mechanisms in humans at both early, monocular stages (e.g., retina or LGN) and later, binocular stages (e.g., V1 and beyond; Hubel & Wiesel, 1965, 1968).

Secondly, both monocular and binocular neural populations contributed to the effect. This indicates that extrapolation occurs at multiple hierarchical (rather than single) processing stages. Extrapolation, or prediction, at multiple stages of sensory processing is the central property of hierarchical predictive coding, an influential theoretical and computational account of neural sensory processing (Huang & Rao, 2011; Rao & Ballard, 1999). In this model, successive layers of neurons "predict" their own input through feedback connections to earlier layers, feeding only the prediction error forward to higher layers (Rao & Ballard, 1999).

For example, in the visual system, a high-level neuron might represent a Gabor patch at a given position, spatial frequency, and orientation, and "predict" the local luminance of the lower level neurons (with smaller, simpler receptive fields) that project to it. The lower level neuron receiving the prediction then essentially compares the "prediction" to its input, and only feeds forward the deviations from that prediction—i.e., any properties of the stimulus not captured, or predicted, by the activity of the high-level neuron representing the Gabor. Conceptually, such a hierarchy would converge on patterns of connectivity and activation that minimize total prediction error in the system. This would minimize metabolic requirements of sensory signaling, while optimizing information-theoretic properties of the network (a principle that has been dubbed the Free Energy Principle; Friston, 2005, 2010).

Importantly, the "predictions" in current models of predictive coding are predictive only in the hierarchical sense, but not in the temporal sense of predicting *future* activity (Bastos et al., 2012; Spratling, 2012, 2017); but see Friston (2005) for a discussion of predictive coding in time, and Garrido, Kilner, Stephan, and Friston (2009) for an empirical demonstration applied to expectation and mismatch negativity. However, neural transmission delays mean that for any time-variant input (such as visual motion), prediction errors at a given neural population are minimized by the higher

area predicting that population's future, rather than current, input. In a toy example, Area 1 sends visual information about the position of a moving object to Area 2, which in turn sends a "prediction" back to Area 1. That prediction is compared with input in Area 1 and any mismatch error is recursively minimized by adjusting the feedback signal to *Area 1* to line up with its *input at the time the signal arrives there* (for details, see Hogendoorn & Burkitt, 2018b). Minimizing error therefore requires compensating for the delays incurred in both feed-forward and feed-back signaling. In the case of visual motion, compensating for these delays can be achieved by extrapolation: Simply multiplying the instantaneous velocity of an object by the expected delay (feed-forward and feedback) yields a *spatiotemporal* prediction which is predictive in both the hierarchical and temporal sense. Indeed, several authors have proposed neural mechanisms for motion prediction, based on adaptation (Erlhagen, 2003) and Bayes-optimal motion-position estimation (Khoei et al., 2017; Kwon et al., 2015). However, these have not been related to the broader hierarchical predictive coding framework.

Because delays are incurred between each successive stage in the processing hierarchy, extrapolation must similarly occur at each stage if total prediction error is to be minimized. Extrapolation at each stage would require information about rate of change (i.e., velocity) at each stage. This is consistent with known properties of the early visual system: In lower vertebrates, velocity is extracted already in the retina (Amthor & Grzywacz, 1993), and although the proportion of direction-selective retinal ganglion cells in higher vertebrates is reduced (Bach & Hoffmann, 2000), in these animals direction-selective cells have been reported in the lateral geniculate nucleus (Niell, 2013). V1 itself of course also represents velocity (Hubel & Wiesel, 1962). Our finding that both monocular and binocular stages in the visual processing hierarchy carry out extrapolation is therefore anatomically plausible, and consistent with a version of hierarchical predictive coding that takes into account neural transmission delays.

Resulting as it does from an unexpected reversal of a moving background pattern, the flash-grab illusion is thought to occur due to a violation of expected motion (Cavanagh & Anstis, 2013; Hogendoorn et al., 2015). In the predictive coding framework, this amounts to a prediction error: A higher level area extrapolates the position of the moving background, but by the time that predictive signal arrives at the lower level area, the stimulus has reversed and the prediction (having been extrapolated in the initial direction, as indicated schematically in Figure 1) is very far from the new input. The resulting prediction error means that the represented position subsequently shifts rapidly over time, yielding a spike in velocity in the direction of the new motion (Figure 1C). As per the mechanism proposed by Eagleman and Sejnowski (2007), and principally consistent with Kwon et al. (2015), this velocity signal biases the perceived position of the (stationary) target that is briefly flashed superimposed on the background. In this sense, the flash-grab effect can be thought of as a direct reflection of prediction error. Two things about this interpretation remain to be elucidated. Firstly, even when no flash is presented, the reversal point of the sector edge on which the flash would otherwise be presented still undershoots the true physical point. Conversely, the neural representation of the edge, as measured in the retina (Schwartz et al., 2007) and proposed here to underlie the flash-grab effect, does not (Figure 1). The fate of these neural representations that do not reach awareness remains to be elucidated. In a similar vein, this sequence of neural representations generates a spike in the velocity signal (Figure 1C). We argue here that this causes a concurrent flash to be mislocalized, but perhaps this velocity spike also has other perceptual consequences. One interpretation could be that this velocity signal actually masks the final section of the position signal that represents the overshoot, comparable to the mechanisms proposed for saccadic suppression during eye movements (Ibbotson & Cloherty, 2009; Ibbotson, Crowder, Cloherty, Price, & Mustari, 2008), but this remains to be further explored.

The pattern of illusion strength in the three split conditions gives some further insight into the mechanisms that are likely to play a role. In the *Binocular* and *Monocular* conditions, in which the entire stimulus sequence is presented within a single eye, the violation of the background's motion direction can be detected at an early monocular stage. The prediction error therefore arises early, and is available to influence the neural representation of the target's position at both monocular (since the target is presented in the same eye) and binocular stages. This results in maximal prediction error (evident as maximal illusion strength, Figure 3). In the Interocular condition, the violation can be detected monocularly, but with the target being presented to the other eye, the target's representation can only be influenced when the monocular channels converge at a later binocular stage, thereby reducing the magnitude of the illusion. In the two other split conditions (Before Reversal and After Reversal), because the two motion sequences are presented to different eyes, the violation is only detected at a later binocular stage. Because receptive fields are generally larger further down the hierarchy, such that the discrepant extrapolated and actual positions of the target are more likely to fall within the same cell's receptive field as one looks further down the hierarchy, a given violation might be expected to yield a smaller error further down the hierarchy. Consistent with this,

in our results, the illusion is attenuated in the two split conditions (paired-samples t test of the two split conditions averaged together vs. binocular; $t(16) = 4.3$, $p < 0.001$, Cohen's $d = 0.96$. Finally, the illusion is more strongly reduced in the Before Reversal than After Reversal condition, $t(16) = 3.1$, $p = 0.007$, Cohen's $d = 0.75$. This is a consequence of the so-called Frohlich effect (Kerzel, 2010), in which the onset position of a moving object is shifted in the direction of that object's subsequent trajectory. In the After Reversal condition, the second motion sequence does not violate a monocular motion prediction per se, but it still generates a monocular error signal due to the motion onset. In the Before Reversal condition, the same prediction error arises in the opposite eye to the target. Because this can only influence the target's position at the later binocular stage, this again leads to a smaller mislocalization illusion.

The flash-grab effect has alternatively been explained in terms of trajectory shortening (Cavanagh & Anstis, 2013). It has been reported that the perceived trajectory of an object that reverses its direction is shortened (Sinico, Parovel, Casco, & Anstis, 2009), and that the perceived shift in the endpoint of the trajectory is linked to the position shift induced by the flash-grab effect (Cavanagh & Anstis, 2013). The trajectory-shortening explanation is formulated at a more abstract, computational level of description and therefore cannot offer any insight about how the effects should vary under monocular, binocular, or dichoptic presentation. Cavanagh and Anstis (2013) argued that the flash-grab effect critically depends on attention, which might be taken as corresponding to a late, presumably binocular neural locus. As the available information at these stages would not be affected by dichoptic presentation, this interpretation in itself therefore does not provide a parsimonious explanation why the illusion would be attenuated in these conditions.

The proposition that motion and position signals interact already in monocular channels is consistent with a recent report studying motion-induced position shifts (the illusory displacement of the envelope of a Gabor patch when its carrier wave is moving; De Valois & De Valois, 1991). Hisakata, Hayashi, and Murakami (2016) observed that this illusory displacement is observed even when the carrier and the envelope are presented at widely divergent disparities, suggesting a disparity-insensitive monocular mechanism. They further showed that (as previously reported by Anstis, 1989) when illusory displacements are induced in the two eyes, the resulting illusory disparity yields an illusory depth percept, further supporting the involvement of motion-position interactions at monocular processing stages.

One limitation of the current study is that the three dichoptic conditions require binocular fusion, whereas the monocular and binocular presentations did not. Nevertheless, we do not believe this is a significant confound for a number of reasons. Firstly, a binocularly presented fixation point and large, high-contrast squares around the stimulus were presented to assist with binocular fusion, and during debrief none of the observers reported any difficulty with fusion. Furthermore, difficulty with binocular fusion mostly occurs when conflicting high-contrast stimuli are presented to each eye, which was never the case in our stimulus sequences. Instead, individual stimulus components were presented to one eye in the absence of any contrast energy in the other eye, a situation which does not negatively affect fusion (Alais & Blake, 2004). Finally, it is not clear how problems with binocular fusion would explain the observed differences between the different dichoptic conditions. Nevertheless, we cannot entirely rule out the possibility that differences in binocular fusion may have played a role in our experimental conditions.

In sum, we have used a dichoptic version of the flash-grab effect to study the monocular and binocular contributions to motion extrapolation in human visual motion perception. The results reveal that extrapolation mechanisms operate at both monocular and binocular processing stages—a finding that is consistent with an extension of the hierarchical predictive coding framework that accounts for neural transmission delays. The results further suggest that prediction errors in this framework can manifest directly as perceptual illusions.

Keywords: motion extrapolation, prediction, predictive coding

Acknowledgments

EvH and HH were supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP180102268). MG acknowledges a University of Queensland Fellowship (2016000071).

Commercial relationships: none.
Corresponding author: Hinze Hogendoorn.
Email: hhogendoorn@unimelb.edu.au.
Address: Melbourne School of Psychological Sciences, The University of Melbourne, Parkville VIC, Australia.

References

Alais, D., & Blake, R. (2004). *Binocular rivalry*. Cambridge, MA: MIT Press.

Amthor, F. R., & Grzywacz, N. M. (1993). Directional selectivity in vertebrate retinal ganglion cells. *Reviews of Oculomotor Research*, *5*, 79–100. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8420563

Anstis, S. M. (1989). Kinetic edges become displaced, segregated, and invisible. In D. Lam (Ed.), *Neural mechanisms of visual perception* (Vol. 2, pp. 247–260). Houston, TX: Gulf Publishing.

Bach, M., & Hoffmann, M. B. (2000). Visual motion detection in man is governed by non-retinal mechanisms. *Vision Research*, *40*(18), 2379–2385, https://doi.org/10.1016/S0042-6989(00)00106-1.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012, November 21). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711. NIH Public Access, https://doi.org/10.1016/j.neuron.2012.10.038.

Berry, M. J., Brivanlou, I. H., Jordan, T. A., & Meister, M. (1999, March 25). Anticipation of moving stimuli by the retina. *Nature*, *398*(6725), 334–338, https://doi.org/10.1038/18678.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9176952

Brenner, E., Smeets, J. B., & de Lussanet, M. H. (1998). Hitting moving targets. Continuous control of the acceleration of the hand on the basis of the target's velocity. *Experimental Brain Research*, *122*(4), 467–474. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9827866

Cai, R. H., & Schlag, J. (2001). Asynchronous feature binding and the flash-lag illusion. *Investigative Ophthalmology and Visual Science*, *42*, S711.

Cavanagh, P., & Anstis, S. M. (2013). The flash grab effect. *Vision Research*, *91*, 8–20, https://doi.org/10.1016/j.visres.2013.07.007.

De Valois, R. L., & De Valois, K. K. (1991). Vernier acuity with stationary moving Gabors. *Vision Research*, *31*(9), 1619–1626, https://doi.org/10.1016/0042-6989(91)90138-U.

Eagleman, D. M. (2008, April 14). Prediction and postdiction: Two frameworks with the goal of delay compensation. *Behavioral and Brain Sciences*, *31*(2), 205–206. https://doi.org/10.1017/S0140525X08003889.

Eagleman, D. M., & Sejnowski, T. J. (2000, March 17). Motion integration and postdiction in visual awareness. *Science*, *287*(5460), 2036–2038. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10720334

Eagleman, D. M., & Sejnowski, T. J. (2007). Motion signals bias localization judgments: A unified explanation for the flash-lag, flash-drag, flash-jump, and Frohlich illusions. *Journal of Vision*, *7*(4):3, 1–12, https://doi.org/10.1167/7.4.3. [PubMed] [Article]

Erlhagen, W. (2003). Internal models for visual perception. *Biological Cybernetics*, *88*(5), 409–417, https://doi.org/10.1007/s00422-002-0387-1.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836, https://doi.org/10.1098/rstb.2005.1622.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138, https://doi.org/10.1038/nrn2787.

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *120*(3), 453–463, https://doi.org/10.1016/j.clinph.2008.11.029.

Hisakata, R., Hayashi, D., & Murakami, I. (2016). Motion-induced position shift in stereoscopic and dichoptic viewing. *Journal of Vision*, *16*(13) 3, 1–13, https://doi.org/10.1167/16.13.3. [PubMed] [Article]

Hogendoorn, H., & Burkitt, A. N. (2018a). Predictive coding of visual object position ahead of moving objects revealed by time-resolved EEG decoding. *NeuroImage*, *171*, 55–61, https://doi.org/10.1016/j.neuroimage.2017.12.063.

Hogendoorn, H., & Burkitt, A. N. (2018b). Predictive coding with neural transmission delays: A real-time temporal alignment hypothesis. *BioRxiv*, https://doi.org/10.1101/453183.

Hogendoorn, H., Verstraten, F. A. J., & Cavanagh, P. (2015). Strikingly rapid neural basis of motion-induced position shifts revealed by high temporal-resolution EEG pattern classification. *Vision Research*, *113*(Part A), 1–10, https://doi.org/10.1016/j.visres.2015.05.005.

Hosoya, T., Baccus, S. A., & Meister, M. (2005, July 7). Dynamic predictive coding by the retina. *Nature*, *436*(7047), 71–77, https://doi.org/10.1038/nature03689.

Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580–593, https://doi.org/10.1002/wcs.142.

Hubel, D. H., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14449617

Hubel, D. H., & Wiesel, T. (1965). Binocultar interaction in striate cortex of kittens reared with artificial squint. *Journal of Neurophysiology*, *28*(6), 1041–1059, https://doi.org/10.1152/jn.1965.28.6.1041.

Hubel, D. H., & Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, *195*(1), 215–243, https://doi.org/papers://47831562-1F78-4B52-B52E-78BF7F97A700/Paper/p352.

Ibbotson, M. R., & Cloherty, S. L. (2009). Visual perception: Saccadic omission—suppression or temporal masking? *Current Biology*, *19*(12), R493–R496, https://doi.org/10.1016/j.cub.2009.05.010.

Ibbotson, M. R., Crowder, N. A., Cloherty, S. L., Price, N. S. C., & Mustari, M. J. (2008). Saccadic modulation of neural responses: Possible roles in saccadic suppression, enhancement, and time compression. *Journal of Neuroscience*, *28*(43), 10952–10960, https://doi.org/10.1523/JNEUROSCI.3950-08.2008.

Jancke, D., Erlhagen, W., Schöner, G., & Dinse, H. R. (2004). Shorter latencies for motion trajectories than for flashes in population responses of cat primary visual cortex. *Journal of Physiology*, *556*(3), 971–982, https://doi.org/10.1113/jphysiol.2003.058941.

Kerzel, D. (2010). The Fröhlich effect: Past and present. In R. Nijhawan & B. Khurana (Eds.), *Space and time in perception and action* (pp. 321–337). Cambridge, UK: Cambridge University Press, https://doi.org/10.1017/CBO9780511750540.019.

Kerzel, D., & Gegenfurtner, K. R. (2003). Neuronal processing delays are compensated in the sensorimotor branch of the visual system. *Current Biology*, *13*(22), 1975–1978, https://doi.org/10.1016/j.cub.2003.10.054.

Khoei, M. A., Masson, G. S., & Perrinet, L. U. (2017). The flash-lag effect as a motion-based predictive shift. *PLOS Computational Biology*, *13*(1), e1005068, https://doi.org/10.1371/journal.pcbi.1005068.

Krekelberg, B. (2000, August 18). The position of moving objects. *Science*, *289*(5482), 1107a–1107, https://doi.org/10.1126/science.289.5482.1107a.

Krekelberg, B., & Lappe, M. (2001). Neuronal latencies and the position of moving objects. *Trends in Neurosciences*, *24*(6), 335–339, https://doi.org/10.1016/S0166-2236(00)01795-1.

Kwon, O.-S., Tadin, D., & Knill, D. C. (2015). Unifying account of visual motion and position perception. *Proceedings of the National Academy of Sciences, USA*, *112*(26), 8142–8147, https://doi.org/10.1073/pnas.1500361112.

Lotter, W., Kreiman, G., & Cox, D. (2018). A neural network trained to predict future video frames mimics critical properties of biological neuronal responses and perception, 1–18. Retrieved from http://arxiv.org/abs/1805.10734

Maus, G. W., & Nijhawan, R. (2008). Motion extrapolation into the blind spot. *Psychological Science*, *19*(11), 1087–1091, https://doi.org/10.1111/j.1467-9280.2008.02205.x.

Niell, C. M. (2013). Vision: More than expected in the early visual system. *Current Biology*, *23*(16), R681–R684, https://doi.org/10.1016/J.CUB.2013.07.049.

Nijhawan, R. (1994, July 28). Motion extrapolation in catching. *Nature*, *370*(6487), 256–257, https://doi.org/10.1038/370256b0.

Nijhawan, R. (2008). Visual prediction: Psychophysics and neurophysiology of compensation for time delays. *Behavioral and Brain Sciences*, *31*(2), 179–239, https://doi.org/10.1017/S0140525X08003804.

Parker, A. J. (2007). Binocular depth perception and the cerebral cortex. *Nature Reviews Neuroscience*, *8*(5), 379–391, https://doi.org/10.1038/nrn2131.

Patel, S. S., Ogmen, H., Bedell, H. E., & Sampath, V. (2000, November 10). Flash-lag effect: Differential latency, not postdiction. *Science*, *290*(5494),1051. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11184992.

Ramachandran, V. S., & Anstis, S. M. (1990). Illusory displacement of equiluminous kinetic edges. *Perception*, *19*(5), 611–616, https://doi.org/10.1068/p190611.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87, https://doi.org/10.1038/4580.

Schwartz, G., Taylor, S., Fisher, C., Harris, R., & Berry, M. J., II. (2007). Synchronized firing among retinal ganglion cells signals motion reversal. *Neuron*, *55*(6), 958–969, https://doi.org/10.1016/j.neuron.2007.07.042.

Sillito, A. M., Jones, H. E., Gerstein, G. L., & West, D. C. (1994). Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex. *Nature*, *369*(6480), 479–482, https://doi.org/10.1038/369479a0.

Sinico, M., Parovel, G., Casco, C., & Anstis, S. (2009). Perceived shrinkage of motion paths. *Journal of Experimental Psychology: Human Perception and*

Performance, *35*(4), 948–957, https://doi.org/10.1037/a0014257.

Spratling, M. W. (2012). Predictive coding accounts for V1 response properties recorded using reverse correlation. *Biological Cybernetics*, *106*(1), 37–49, https://doi.org/10.1007/s00422-012-0477-7.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, *112*, 92–97, https://doi.org/10.1016/j.bandc.2015.11.003.

Sundberg, K. A., Fallah, M., & Reynolds, J. H. (2006). A motion-dependent distortion of retinotopy in area V4. *Neuron*, *49*(3), 447–457, https://doi.org/10.1016/j.neuron.2005.12.023.

Van Heusden, E., Rolfs, M., Cavanagh, P., & Hogendoorn, H. (2018). Motion extrapolation for eye movements predicts perceived motion-induced position shifts. *Journal of Neuroscience*, *38*(38), 8243–8250, https://doi.org/10.1523/JNEUROSCI.0736-18.2018.

Whitney, D., & Cavanagh, P. (2000). Motion distorts visual space: Shifting the perceived position of remote stationary objects. *Nature Neuroscience*, *3*(9), 954–959, https://doi.org/10.1038/78878.

Whitney, D., & Murakami, I. (1998). Latency difference, not spatial extrapolation. *Nature Neuroscience*, *1*(8), 656–657, https://doi.org/10.1038/3659.