Disparity-based stereomotion detectors are poorly suited to track 2D motion

Marina Zannoli
Université Paris Descartes, Sorbonne Paris Cité, Paris, France
Laboratoire Psychologie de la Perception, Paris, France

John Cass
School of Psychology, University of Western Sydney, Sydney, New South Wales, Australia

David Alais
School of Psychology, University of Sydney, Sydney, New South Wales, Australia

Pascal Mamassian
Université Paris Descartes, Sorbonne Paris Cité, Paris, France
Laboratoire Psychologie de la Perception, Paris, France

A study was conducted to examine the time required to process lateral motion and motion-in-depth for luminance- and disparity-defined stimuli. In a $2 \times 2$ design, visual stimuli oscillated sinusoidally in either 2D (moving left to right at a constant disparity of 9 arcmin) or 3D (looming and receding in depth between 6 and 12 arcmin) and were defined either purely by disparity (change of disparity over time [CDOT]) or by a combination of disparity and luminance (providing CDOT and interocular velocity differences [IOVD]). Visual stimuli were accompanied by an amplitude-modulated auditory tone that oscillated at the same rate and whose phase was varied to find the latency producing synchronous perception of the auditory and visual oscillations. In separate sessions, oscillations of 0.7 and 1.4 Hz were compared. For the combined CDOT + IOVD stimuli (disparity and luminance [DL] conditions), audiovisual synchrony required a 50 ms auditory lag, regardless of whether the motion was 2D or 3D. For the CDOT-only stimuli (disparity-only [DO] conditions), we found that a similar lag (~60 ms) was needed to produce synchrony for the 3D motion condition. However, when the CDOT-only stimuli oscillated along a 2D path, the auditory lags required for audiovisual synchrony were much longer: 170 ms for the 0.7 Hz condition, and 90 ms for the 1.4 Hz condition. These results suggest that stereomotion detectors based on CDOT are well suited to tracking 3D motion, but are poorly suited to tracking 2D motion.

Keywords: stereomotion, stereopsis, motion, audio-visual integration


Introduction

To estimate the depth order relationships between objects, the visual system relies on multiple cues to depth. One such cue is binocular disparity. Stereopsis refers to the perception of depth derived from disparities between the two eyes’ retinal images. Whilst static objects may be defined purely by their interocular spatial correlations, objects undergoing motion-in-depth are defined by correlations that co-occur across space and time. There are two main cues to extract motion-in-depth. The visual system can extract binocular disparity of an object relative to the fixation plane and track how this information varies over time (change of disparity over time [CDOT]). Another possibility is to combine the velocity of an object extracted from each monocular image (interocular velocity difference [IOVD]). Since the seminal work of Rashbass and Westheimer (1961), extensive psychophysical work has been done to understand the nature and the relative utility of these two cues that are often present redundantly in natural scenes (Harris, Nefs, & Grafton, 2008; Nefs & Harris, 2010). A majority of the studies conducted in the last decade have focused on understanding how monocular motion signals are combined to detect motion-in-depth (Cumming & Parker, 1994) and to discriminate the speed of motion-in-depth (Brooks, 2002; Brooks & Stone, 2004; Harris & Watamaniuk, 1995; Rokers, Czuba, Cormack, & Huk, 2011).
Less interest has been shown in understanding the mechanisms underlying the tracking of changes in disparity over time. Harris, McKee, and Watamaniuk (1998) showed that the detection of motion-in-depth was more affected by disparity noise than was the detection of lateral motion, and they suggested that the detection of 3D motion was carried out by static disparity mechanisms rather than specific mechanisms sensitive to the change of disparity over time. Using the same paradigm, Harris and Sumnall (2000) showed that there was no effect of the viewing distance on the detection of 3D and 2D motion, suggesting that motion-in-depth detection is based on relative and not absolute signals.

To complement these behavioral results, two fMRI studies have proposed that motion-in-depth is computed by specific neurons in the visual cortex. First, Likova and Tyler (2007) recorded bold activation in the dorsal stream after presentation of motion-in-depth stimuli containing only CDOT information and discovered a visual area, anterior to hMT+ (previously found to be sensitive to motion and disparity information; Maunsell & Van Essen, 1983) exclusively sensitive to changes of disparity over time. Later, Rokers, Cormack, and Huk (2009) found that area hMT+ was sensitive to both CDOT and IOVD types of information. In addition, they reported specific activation of V3A and the lateral occipital complex (LO) after presentation of stimuli containing only CDOT information. Taken together, these results suggest that: (a) changes of disparity over time are mediated by cortical mechanisms separate from those associated with the processing of static disparity signals; and (b) that these CDOT-selective mechanisms are associated with the perception of motion-in-depth.

At the same time, another line of work focused on understanding the interactions between the processing of motion and binocular disparity. Maunsell and Van Essen (1983) were the first to report the existence of cells sensitive to both binocular disparity and fronto-parallel motion in macaque MT, suggesting that lateral motion is treated separately for different depth planes. This was confirmed by more recent psychophysical work on motion transparency. For example, Hibbard and Bradshaw (1999) and Snowden and Rossiter (1999) measured thresholds for the identification of the direction of motion for stimuli in which signal and noise elements were given various disparities, and they found that performance was substantially better when signal and noise had different disparities. Similarly, Edwards and Greenwood (2005) and Greenwood and Edwards (2006) showed that observers are able to detect a larger number of transparent motion directions when they are carried by signals that are distributed across distinct depth planes.

Even though the processing of motion and binocular disparity seem to share common cortical resources, their underlying mechanisms have different spatial and temporal resolutions. In the stereo domain, both temporal and spatial resolution have been found to be worse than for lateral motion (Norcia & Tyler, 1984; Regan & Beverley, 1973; Tyler, 1971). It has also been shown that differences in temporal resolution and time of processing can be found within the stereo system itself. For example, Julesz (1960) observed that depth from random-dot stereograms (RDSs) took more time than for stimuli with monocular segmentation information. However, more recently Uttal, David, and Welke (1994) reported that observers were above 50% performance when asked to recognize a 3D shape on a RDS presented for 1 ms. Both studies showed an effect of practice on the latency of stereopsis from RDSs.

The aim of the present study was to investigate how long it takes the visual system to process changes in direction of motion, comparing stimuli oscillating at a constant (nonzero) disparity over time (2D motion in the fronto-parallel plane) with stimuli oscillating through varying disparities over time (3D motion-in-depth). Performance in these lateral motion and motion-in-depth conditions are compared for stimuli with and without the contribution of monocular segmenting information. In this way, we compare the temporal resolution of the luminance and disparity-defined motion systems for detecting changes in the direction of moving stimuli in a 2D or 3D context. We do this using a method introduced by Moutoussis and Zeki (1997) to study relative processing latencies between different stimulus attributes. Their stimulus consisted of a pattern of colored squares that oscillated in position (up/down) and color (red/green) following a square-wave pattern. By shifting the phase of the color alternation relative to motion until they both appeared to change synchronously, they showed that color changes were perceived 70–80 ms before motion changes. They then verified that this was a fixed offset by testing different frequencies of color/motion change.

We employ an analogous paradigm to investigate the processing latencies for changes in perceived direction of luminance- and disparity-defined motion within and across depth planes. As this method involves continuously cycling stimuli, it has an important advantage over other paradigms using brief stimuli because the phase-lag required for perceptual synchrony can be assumed to reflect a pure latency difference and not to include the time needed to fuse the two eyes’ images and compute the disparity map, which would happen only once, at stimulus onset. Because our measure is free of a time-to-fuse component, it allows us to compare the optimal latency for synchrony with the same measure obtained in the luminance domain, and for lateral motion versus motion-in-depth. Several
studies using visual objects defined by luminance have reported that the auditory event must be presented 30–40 ms after the visual stimulus to perceive audiovisual synchrony (Lewald & Guski, 2003). However, little is known about the time required to compute audiovisual simultaneity for disparity-defined visual stimuli.

**Method**

**Participants**

Five observers (four naïve and one author) with normal or corrected-to-normal vision were recruited from the laboratory. All had experience in psychophysical observation and had normal stereo acuity and hearing.

**Stimulus presentation**

The stereograms were presented on a CRT monitor (ViewSonic 21”, resolution of 1280 × 960, refresh rate of 85 Hz [ViewSonic, Walnut, CA]) using a modified Wheatstone stereoscope at a simulated distance of 57 cm. Each eye viewed one horizontal half of the CRT screen. A chin rest was used to stabilize the observer’s head and to control the viewing distance. The display was the only source of light and the stereoscope was calibrated geometrically to account for each participant’s interocular distance. The auditory stimuli were presented binaurally through headphones.

**Stimuli**

**Visual stimuli**

Stimuli were generated using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). To avoid any coherent 2D motion signals in the monocular images of our disparity-defined stereomotion stimuli, we used dynamic random-dot stereograms (DRDS). In DRDSs, the stereogram is rebuilt on each video frame using a new pattern of random noise. Disparity was achieved by adding horizontal offsets in interocular disparity to a small portion of the left- and right-eye images. Stereomotion was then obtained by smoothly changing the value of the disparity offsets from frame to frame. This way, stereomotion in our stimuli was entirely defined by changes of binocular disparity over time.

The background consisted of a 3.1° × 3.1° square of dynamic random noise (mean luminance 40 cd/m²; one-pixel resolution; refreshed every frame). Visual stimuli (see Figure 1) consisted of fifteen randomly distributed squares (0.6° × 0.6°) defined either by disparity and luminance (DL) conditions or by disparity only (DO) conditions. In the DO condition, the squares were composed of dynamic random noise and were distinguished from the background only by disparity (and therefore visible only binocularly), whereas in the DL condition, they were black squares of 5 cd/m² luminance (and therefore monocularly visible), which were also disparate relative to the background. In the DL conditions, both CDOT and IOVD cues were available, while only the CDOT cue was present in the DO conditions. Each square in the set consistently moved from left to right in opposite directions in each eye between 6 and 12 arcmin, producing a percept of motion-in-depth (3D motion), or from left to right (6 arcmin amplitude) at a constant disparity of 9 arcmin, producing a percept of lateral displacement at a pedestal depth (2D motion).

The background was surrounded by a vergence-stabilization frame consisting of multiple luminance-defined squares (0.20° × 0.20°; gray: 40 cd/m² and white: 80 cd/m²) presented on a black background (5 cd/m²). Black nonius lines were presented at the center of the display (see Figure 1).

**Auditory stimuli**

The auditory stimulus was a 500 Hz sine-wave (44.1 kHz sample rate; mono) whose envelope (amplitude) was modulated between 0 and 70 dB at the same frequency as the modulations in visual motion direction. The sound was presented binaurally through headphones.

**Audiovisual modulations**

The audiovisual stimulus was presented for 2 s. In order to test whether the optimal latencies measured were dependant on the phase of the motion (in degrees) or whether they reflected absolute latencies (in ms), the experiment was replicated for two different frequency values: 0.7 (equivalent speed: 0.14°/s) and 1.4 Hz (equivalent speed: 0.28°/s). These values were chosen in order to maximize sensitivity to CDOT (Czuba, Rokers, Huk, & Cormack, 2010; Shioiri, Nakajima, Kakehi, & Yaguchi, 2008).

Both modulations were sinusoidal. The phase of the auditory modulation relative to the visual modulation varied between 0° and 345° in steps of 15°. A random phase was added to all modulations (see Figure 2).

**Procedure**

The experiment was divided into two sessions. In the first session, the audiovisual modulations were at a frequency of 0.7 Hz. In the second session, the frequency
was doubled to 1.4 Hz. For each session, the four conditions of DO/DL * 2D/3D motion were presented in separate blocks. The four blocks were presented in a random order. Each block contained a total of 192 trials (eight repetitions for each of the 24 auditory phases).

Participants were asked to match the direction of motion with the amplitude modulation. In the 3D motion conditions, participants were asked to press one key of a keyboard if the maximum auditory amplitude was synchronized with the squares being at their perceptually “farthest” position and another key if the maximum of the sound amplitude was synchronized with the squares being at their perceptually “closest” position. In the 2D motion conditions, participants used the left and right arrows to respectively indicate audio-visual synchrony between the maximum amplitude of the tone and left-most and right-most points in the 2D motion trajectory.

Results

Figure 3 shows the averaged response curves for the five participants. The proportion of “near” (for the 3D motion condition) or “left” (for the 2D motion condition) responses is plotted as a function of auditory latency. A negative latency represents an auditory signal lag while a positive latency codes for a visual lag. When an audio-visual phase lag of zero is applied to the auditory signal, the maximum amplitude (70 dB) is synchronized with the maximum visual disparity value (12 arcmin), or with the left position for the 2D motion condition. If perception of the auditory and visual changes occurred with no differential latency, we would expect the maximum of the response curve to peak at a value of 0 ms. If this maximum deviates from 0 ms, it suggests that visual and auditory information are perceived at different times. We fitted logit functions (Mamassian & Wallace, 2010; see Figure 3) to the distributions and extracted slopes for the four conditions. For each latency \( h \) the probability \( p \) to perceive the maximum of auditory amplitude synchronized with the “left” or “near” position (depending on the motion condition) of the visual stimuli is characterized by the following logit model:

\[
\text{logit}(p) = \ln \left( \frac{p}{1-p} \right) = \gamma - \beta_\theta |\theta - \theta_0| \pi
\]

where \( \theta_0 \) is the optimal latency, \( \beta_\theta \) represents the strength of the effect of latency on the proportion of
Figure 3. Results of the experiment. This plot shows the proportion of "left" or "near" responses as a function of the latency. The data is pooled across the five participants. The visual stimuli were defined by 3D or 2D motion (blue or red) and by DO or by DL (light or dark). A logit function was fitted to the data from the four experimental conditions.

"left" or "near" responses, and γ is a constant. In this equation, \(|x|_\pi = \text{acos}(\text{cos}(x))\). The parameter \(\beta_0\) shows how sensitive an observer is for small variations of latency (its unit is in ms\(^{-1}\) when latencies are expressed in ms). Figure 4 shows the group mean slopes as a function of the latencies extracted from the best-fitting logit functions for the four conditions at each oscillation rate.

A repeated-measures ANOVA was run on the mean latency and the slope with the type of stimuli (DO vs. DL) and the type of motion (2D vs. 3D) as within-subject variables for the two frequency values sessions (0.7 and 1.4 Hz). The ANOVA for the 0.7 Hz session revealed a significant effect of the type of stimuli, \(F(1, 3) = 14.8, p < 0.01\); for the slope and \(F(1, 3) = 13.6, p < 0.05\) for the latency; the type of motion, \(F(1, 3) = 12.0, p < 0.05\) for the slope and \(F(1, 3) = 9.96, p < 0.05\) for the latency; and a significant interaction (type of stimuli \(\times\) type of motion) effect, \(F(3, 1) = 90.7, p < 0.01\) for the slope and \(F(3, 1) = 14.8, p < 0.05\) for the latency. For the 1.4 Hz session, the ANOVA revealed a significant effect of the type of stimuli, \(F(1, 3) = 66.2, p < 0.01\), and the type of motion, \(F(1, 3) = 35.5, p < 0.01\), but no significant interaction effect, \(F(3, 1) = 6.58, p = 0.06\) for the slope. For the latency measure in the 1.4 Hz session, the ANOVA revealed no significant effect of the type of stimuli, \(F(1, 3) = 3.86, p = 0.12\), a significant effect of the type of motion, \(F(1, 3) = 19.3, p < 0.05\), and a significant interaction effect, \(F(3, 1) = 9.74, p < 0.05\).

To further investigate the effects found in the ANOVAs, we tested multiple comparisons with Tukey least-significant difference corrections. In the 0.7 Hz session, for the slope and latency measures, we found no difference between the DO-3D, DL-2D, and DL-3D conditions and a significant difference between the DO-2D condition and the three other conditions. In the 1.4 Hz session, for the slope measures, we found no difference between the DO-3D, DL-2D, and DL-3D conditions and a significant difference between the DO-2D condition and the three other conditions. For the latency measure, only the comparison between the DO-3D and DO-2D conditions was significant.

A casual exploration of Figure 4 suggests a potential relationship between latency and slope: small latencies (i.e., low bias) are linked to large slopes (i.e., high sensitivity). However, with only eight conditions (and a clear outlier), this apparent relationship should be taken with caution.

The 2D and 3D motion conditions for DL stimuli were similar in terms of optimal latency for perceived synchrony (mean auditory lag: 43 and 37 ms for the 0.7 Hz condition and 59 and 48 ms for the 1.4 Hz condition for the 2D and 3D motion conditions, respectively). Surprisingly, even though stereopsis is often thought to be a slow process, we found the optimal latency for DO-3D motion stimuli was only slightly longer (mean auditory lag: 55 ms and 64 ms for the 0.7 Hz and 1.4 Hz conditions, respectively). However, when participants had to judge synchrony for the DO-2D motion stimuli, it led to larger latencies (170 and 90 ms for the 0.7 Hz and 1.4 Hz conditions). In addition, in the DO-2D motion, the slope of the distribution was substantially shallower than in the three other conditions for the two frequency conditions, suggesting that the task was much harder (see Figures 3 and 4).

We found a similar pattern of results in the two experiments (similar latencies and slopes for three conditions and longer latency and shallower slope in the DO-2D motion condition). Latencies in the two experiments are equivalent in terms of absolute
latencies except for the DO-2D motion condition. In this condition, the latency was divided by two in the 1.4 Hz experiment compared to the 0.7 Hz experiment.

**Discussion**

To sum up, we measured the latencies required to perceptually align an auditory modulation with an oscillating visual motion. We compared lateral (2D) motion and motion-in-depth (3D), for motion tokens defined either by DL, or by DO. Of these four conditions, we found a similar optimal latency for audiovisual synchrony in three conditions: 2D and 3D motion for the luminance stimuli and 3D motion for the DO condition all produced latencies in the range of 50–60 ms. The exception was in the DO-2D condition where the latency was up to three times larger than for the three other conditions. In addition, the slope of the distribution of synchrony judgments for this particular experimental condition was much shallower than for the other three conditions. Together, these results indicate that the stereomotion system is able to detect changes in the direction of motion as rapidly and precisely as the luminance system can detect direction changes, provided the signal contains changes in depth. Even though this result might seem at odds with previous observations by Tyler (1971), we think that it is hazardous to compare our results with Tyler’s because of several empirical differences. Tyler measured movement sensitivity, so he used small motion amplitudes that were difficult to perceive. We measured the optimal latency for the perception of synchrony between sound and visual motion. For this purpose, we used stimuli in which displayed motion (2D or 3D) was suprathreshold. Another difference is that our visual stimuli moved around a disparity pedestal, whereas Tyler’s were around zero disparity.

The second main result of our study is, however, that when disparities do not vary across time, as in lateral motion at a fixed nonzero disparity, the stereomotion system is very sluggish.

According to the *continuity* rule stated by Marr and Poggio (1976), smooth modulations of disparity over time are easier to detect than abrupt changes. Let us consider a limited area adjacent to the edge of one of the squares present in the visual stimulus on the DO-2D condition. Through this small window, the edge of the square is successively present or absent creating abrupt changes of disparity. If the stereo system relied on such transient information to compute the 2D motion in this stimulus, the task would be much harder than for the other stimulus configurations, leading to degraded performances. We ran a control experiment to test whether the performance obtained in the DO-2D condition could be explained by the temporal integration of square (on/off) modulations of disparity in a limited area of the stimulus. The same participants ran two separate sessions similar to the ones from the main experiment. For both frequency conditions, the mean slope from the control condition was significantly steeper than for the DO-2D condition, \( t(4) = 3.1, p < 0.05 \) for the 0.7 Hz condition and \( t(4) = 3.53, p < 0.05 \) for the 1.4 Hz condition, suggesting that the task was easier in the control condition. Therefore, degraded performances in the DO-2D condition cannot be accounted for by local integration of square modulations of disparity.

In a pilot experiment run on one author, we also tested whether introducing a small amount of 3D motion (1.2 and 2.4 arcmin) would result in a reduction of latency and an increase in slope. We found that slopes and latencies in these two conditions were similar than in the 2D motion condition.

It is of particular interest to compare the two DO conditions. In these two conditions, the moving squares sustained the same amplitude of motion. In the 3D motion condition, the direction of motion (laterally) was in antiphase in one eye compared to the other, while the direction of motion was identical in the two eyes in the 2D motion condition. Therefore, 2D and 3D motion conditions differed only in the direction of lateral displacement across the eyes. It is likely that this difference is responsible for the optimal latency and performance differences between these two conditions.
Implications of the optimal latency reports

The method employed in the present study has been used in several psychophysical works to assess the timing of processing different perceptual attributes. Following Moutoussis and Zeki’s (1997a, 1997b) work on color and motion, Zeki and Bartels (1998) argued that the activity of neurons in a given system is sufficient to elicit a conscious experience of the attribute that is being processed. Therefore, the optimal latency for the perception of synchrony between two attributes directly reflects the timing of processing of these two attributes. Moutoussis and Zeki (1997a, 1997b) found that color information is perceived 60–80 ms before motion information. Stone et al. (2001) measured the point of subjective simultaneity between a light and a sound using a method similar to Moutoussis and Zeki (1997a, 1997b) and found that this optimal latency measure was observer-specific (ranging from −21 ms to +150 ms of auditory lag) and stable. Our results add to these previous observations by showing that the disparity system takes longer to process lateral motion than motion-in-depth.

Implications of the performance measures

The slopes extracted from the logit function fitted to the raw data add to the optimal latency reports and suggest that not only does the disparity system take longer to process changes in the direction of lateral motion than motion-in-depth, but also that it is less efficient at doing so. While it appears from our results that there exists a specific system dedicated to extract motion-in-depth from changes of disparity over time, lateral motion must be inferred from a series of snapshots when moving objects are defined only by disparity.

This result is in contradiction with a basic assumption of a majority of the physiological and computational models of stereopsis. Most cooperative models of stereopsis rely on two fundamental rules first proposed by Marr and Poggio (1976). The uniqueness rule states that “each item from each image may be assigned at most one disparity value” and the continuity rule states that “disparity varies smoothly almost everywhere” as a consequence of the cohesiveness of matter, except at the boundaries of objects. A recent physiological study demonstrated the existence of local competitive and distant cooperative interactions in the primary visual cortex of the macaque, via lateral connections (Samonds, Potetz, & Lee, 2009). These interactions improve disparity sensitivity of binocular neurons over time. These authors suggest that local competition could be the neural substrate of the uniqueness rule, while distant cooperation would favor the detection of similar disparities and therefore implement the continuity rule. These horizontal connections should favor the detection of similar disparities in adjacent positions of the visual field and thus support the processing of 2D motion. It is possible that, in our stimuli, the lack of sensitivity to lateral motion is due to the implementation of the continuity rule. Because it is based on distant lateral connections, it can be hypothesized that this computation is slow.

The discrepancy between performances for 2D and 3D motion for our DO stimuli also has interesting implications in terms of predictive coding. It has been hypothesized that to reduce redundancy, the brain transmits only the unpredicted portions of the sensory input. This information is then combined with a predictive signal, boosting compatible inputs and discarding unlikely ones to reduce detection thresholds (Huang & Rao, 2011; Rao & Ballard, 1999; Srinivasan, Laughlin, & Dubs, 1982). Predictive coding has proven an adequate description of certain aspects of motion perception. For example, Roach, McGraw, and Johnston (2011) showed that a motion signal induces a prediction about the aspect and position of a forward stimulus and that this prediction is combined with the future representation of this stimulus. Our results suggest that the visual system might be more efficient in predicting the variations in depth than in lateral position of an object when it is defined only by binocular disparity.

Conclusion

In the present study, we measured optimal latencies for the perception of synchrony between moving visual stimuli and amplitude modulating sounds. We found that binocular vision is able to efficiently track variations in the direction of motion when these changes are variations in disparity/depth. However, we were surprised to find that this same system dedicated to process binocular vision seems to be poorly suited to track frontoparallel 2D motion. By using visual objects defined only by their binocular disparity, we were able to control for the level of processing required to compute audio-visual integration. Because disparity information is not available before early visual cortical areas, the optimal latencies measured in this study cannot result from early multimodal feedforward integration at a subcortical level.

Acknowledgments

We would like to thank Shin’ya Nishida for discussions and acknowledge financial support from
the French ANR (grant ANR-10-BLAN-1910) and from the Ministère de l’Enseignement et de la Recherche.

Commercial relationships: none.
Corresponding author: Marina Zannoli.
Email: marinazannoli@gmail.com.
Address: Université Paris Descartes, Sorbonne Paris Cité, Paris, France.

References


