Evidence for joint encoding of motion and disparity in human visual perception

Peter Neri¹ ² and Dennis M. Levi²

¹. Institute of Medical Sciences, Aberdeen Medical School, Foresterhill, Aberdeen AB25 2ZD, UK
². School of Optometry and Helen Wills Neuroscience Institute, University of California at Berkeley, California CA 94720, USA

Correspondence: pn@white.stanford.edu

ABSTRACT
Electrophysiological recordings have established that motion and disparity signals are jointly encoded by subpopulations of neurons in visual cortex. However, the question of whether these neurons play a perceptual role has proven challenging and remains open. To answer this question we combined two powerful psychophysical techniques: perceptual adaptation and reverse correlation. Our results provide a detailed picture of how visual information about motion and disparity is processed by human observers, and how this processing is modified by prolonged sensory stimulation. We were able to isolate two perceptual components: a separable component, supported by separate motion and disparity signals; and an inseparable joint component, supported by motion and disparity signals that are concurrently represented at the level of the same neural mechanism. Both components are involved in the perception of stimuli containing motion and disparity information, in line with the known existence of corresponding neuronal subpopulations in visual cortex.

INTRODUCTION
We perceive individual objects as possessing multiple attributes: a red ball falls behind a brown fence. This perceptual representation combines four different attributes (color, shape, motion and depth), and provides an example of feature conjunction. Feature conjunction is essential to visual processing (Robertson, 2003). The conjunction of motion and depth through binocular disparity is particularly interesting, because these two features often co-vary in ways that reflect the three-dimensional structure of the environment. For example, if the observer moves to the right while maintaining fixation on a given point in space, objects further than fixation will move rightward on the retina while nearer objects will move leftward. The speed of moving objects or parts of a given object are often directly related to its three-dimensional structure, as is the case for a set of dots.
arranged on a rotating cylinder: as dots approach the edges of the cylinder, their retinal speed is reduced. Based on these and other similar considerations, it is reasonable to expect that the visual system of animals capable of stereoscopic processing would contain mechanisms that are jointly selective for both motion and disparity.

In line with this expectation, electrophysiological recordings from single neurons in monkey visual cortex have demonstrated that some of these neurons respond selectively to the combination of motion and disparity signals (Grunewald and Skoumbourdis, 2004; Anzai, Ohzawa & Freeman, 2001; DeAngelis et al., 1998; Pack et al., 2003; Roy et al., 1992). Although this property has been reported for some neurons in primary visual cortex (V1), it appears that most neurons in this area are not jointly selective for these two features (Pack et al., 2003; Read and Cumming, 2005a). Joint selectivity seems more common in the middle temporal (MT) area, a region of visual cortex generally believed to play a central role in both motion and disparity processing (DeAngelis et al., 1998). More electrophysiological research will be necessary to establish the exact regions of visual cortex where joint encoding is predominant, but there is no doubt that this property is represented in visual cortex within specific neuronal subpopulations.

From the neurophysiological standpoint, the general question of joint motion-disparity selectivity has been answered.

This question, however, remains open from a behavioural standpoint. The neurophysiological evidence mentioned above shows that visual cortex contains signals that are jointly selective for motion and disparity, but it does not tell us whether these signals are used by the brain to generate percepts and drive behavioural responses. This issue has been recently emphasized by Read and Cumming who demonstrated that Pulfrich-like phenomena, commonly regarded as providing unambiguous evidence that motion and disparity signals are combined in visual cortex, do not in fact require joint encoding (Read and Cumming, 2005b). These authors implemented some earlier considerations by Tyler (1974, 1977) within a physiologically plausible model consisting of neurons that are selective for either motion or disparity, but not both. Because motion and disparity signals are correlated in specific ways for Pulfrich stimuli, the model is able to exploit these correlations and mimic the human percept associated with Pulfrich-like phenomena (Read and Cumming, 2005b). There is ongoing controversy as to whether all aspects of this model are physiologically plausible and whether it is able to capture all variants of the Pulfrich effect, but at present its validity remains unchallenged.

As a consequence of these results, the behavioural issue of joint motion-disparity encoding has been re-opened and remains unresolved. The only method that can resolve this issue directly is visual psychophysics. Evidence from fMRI, for example, would be marginally relevant to the
question asked above, as it would only confirm in humans the electrophysiological evidence already obtained in monkeys. The question asked by Read and Cumming is strictly behavioural, and as such can only be answered by behavioural techniques. This question has direct implications for the design of the system under natural selection, as evolutionary pressure acts on behaviour. Without behavioural evidence, the significance of the neurophysiological evidence is greatly reduced: it then becomes unclear whether the neural signals mattered to the way in which the system was shaped by evolutionary pressure.

It may appear that a resolution to the problem should come from disparity-contingent motion aftereffects (Anstis, 1974; Nawrot and Blake, 1989; Verstraten et al., 1994; Sohn and Seiffert, 2006). These are typically demonstrated by presenting a long-lasting adaptor consisting of two transparent planes of dots. One plane contains Near dots moving (say) Downward, the other plane contains Far dots moving Upward. Following adaptation, a test stimulus containing no bias is presented: both Near and Far dots move in all directions. Despite the lack of bias, Near dots in the test stimulus appear to move Upward while Far dots may appear to move Downward. The effect varies across observers, some reporting an aftereffect only for the Near plane and others reporting little or no effect for either (Verstraten et al., 1994), but the general phenomenon has been documented by several studies and is relatively easy to experience. This phenomenon is not explained by Read and Cumming's model (see Discussion for details). However, most aspects of the associated phenomenology can be explained by other plausible models that rely on cognitive strategies and do not require joint encoding. An example from this class of models can be constructed by combining two successive stages, a disparity-selective stage followed by a motion-selective one. Although each stage by itself does not implement joint encoding, the final two-stage process does (we present an implementation of this model, as well as empirical evidence to support it).

Because of the difficulties associated with the question just discussed, we approached it with the view that the underlying perceptual process needed extensive and detailed characterization. We opted for a reverse correlation approach because this is a powerful tool for retrieving the full perceptual filter underlying a variety of perceptual processes (Eckstein and Ahumada, 2002; Neri and Levi, 2006). We combined this technique with visual adaptation, allowing us to derive both pre-adapted and post-adapted perceptual filters within a protocol analogous to the disparity-contingent motion aftereffect. Our results expose two perceptual components, a joint (inseparable) component and a separable component which does not jointly code motion and disparity. A simple computational model captures the main features of our data by assuming that these two components are affected differently by adaptation. Overall, our results are in agreement with existing
physiological evidence for two separate neuronal subpopulations whose response properties parallel those associated with the two perceptual components exposed by our data.

METHODS

Stimuli

These were generated by a VSG graphics card (CRS, Rochester, UK) and presented on a CRT monitor (Vision Master 17, Iiyama) at a viewing distance of 114 cm. To provide independent stimulation of the eyes, the graphics card was synchronized with ferroelectric (FE) stereogoggles at 100 Hz (50 Hz for each eye). Each static stereopair was rendered by 4 actual video frames (switching between right-eye / left-eye / right-eye / left-eye) before the next motion step. This makes it unlikely that the use of FE stereogoggles resulted in mispairing of disparity and horizontal motion signals. In addition, target and adapting stimuli contained vertical (Upward/Downward) motion. Consequently, the most relevant modulations in our data were centred on vertical motion (which is not confounded with horizontal disparity shifts and thus orthogonal to potential issues with using stereogoggles). Observers were instructed to maintain binocularly aligned fixation on a Nonius marker in the screen centre. The marker consisted of a central cross whose upper and left arms were shown to one eye, while lower and right arms were shown to the other eye. The cross was framed by a binocular square (25.4×25.4 arcmin). This marker was present at all times. Stimuli were confined to two rectangular regions (2×4 (W×H) deg each, centered 1.27 deg from fixation), one to the left and one to the right of fixation, and consisted of dots (100 on each side) that were half bright (74 cd/m²) and half dark (0 cd/m²) on a grey background (37 cd/m²) and measured 6.35 arcmin in diameter. Each dot moved at 2.2 deg/sec in a given direction and maintained a given disparity for the 240 ms duration of the stimulus. Motion direction and disparity could be controlled independently for each dot. Dots were wrapped around the rectangular regions when their motion path reached the edge. Their appearance was restricted to these regions in both eyes, eliminating potential monocular cues generated by position disparity. On one side of fixation, half the “signal” dots moved Downward and had a Near disparity of 5.08 arcmin while the remaining half moved Upward and had a Far disparity of -5.08 arcmin. We refer to this configuration as “target.” A cartoon three-dimensional representation of this stimulus is shown in Figure 1D where signal dots are indicated by asterisks. A representation in direction-disparity space is shown in Figure 1E where signal dots correspond to the two peaks (bright pixels) at the positions indicated by red crosses. On the other side of fixation motion direction and disparity were paired in the opposite way: half the signal dots moved Upward at the Near disparity, while the remaining half moved Downward at the Far disparity (“non-target”). The corresponding representation in direction-
disparity space is shown in Figure 1F, where the positions of signal dots are indicated by blue crosses. The target stimulus could appear on either side of fixation, randomly selected on every trial. The remaining “noise” dots could take any of 16 motion directions spanning the circle, and any of 17 disparity levels from –10.16 to +10.16 in steps of 1.27 arcmin (this range was selected based on previous reverse correlation experiments in the stereoscopic domain (Neri et al., 1999)). The corresponding modulations span the entire surface at Figure 1E and 1F. The stimulus presented every sixth trial (1st, 7th, 13th, ...) consisted of two full-signal targets (Figure 1A) on both sides of fixation. We refer to this stimulus as “adaptor.” The corresponding direction-disparity representation is shown in Figure 1B and 1C, where it can be seen that the only modulations are at the positions indicated by red crosses in Figure 1D to reflect the absence of noise dots. The percentage of signal dots was 60% (S1), 50% (S2), 16% (S3), 50% (S4) and 40% (S5). All observers were naïve except S3 (author PN).

**Procedure**

We collected data for two separate types of blocks. On pre-adapted blocks, the adaptor on the sixth trial lasted 240 ms (like the other trials). On post-adapted blocks, the adaptor on the 1st trial lasted 60 seconds and stimuli on every subsequent sixth trial lasted 6 seconds (top-up). Each stimulus presentation was followed by a 1-sec interval during which observers were required to press a button indicating the side of fixation containing the target (this response was required only on trials that did not contain an adaptor). At the end of this interval one of two sounds informed the observer of whether his/her response was correct/incorrect. If no response was recorded during the 1-sec time window, the corresponding trial was excluded from analysis (1.2 % of total) and no tone was produced. The task was explained to observers by showing them target and non-target in zero-noise conditions, so that they could clearly see the signal structure. We collected 1494/1260 (pre-adapted/post-adapted) trials for S1, 2324/2442 for S2, 2048/2016 for S3, 1704/1648 for S4 and 2533/2313 for S5 (2021/1936 on average), amounting to a total average of ~4K trials per observer. This figure may appear to compare poorly with the large numbers required by psychophysical reverse correlation (e.g. Neri et al. 1999). However the adaptation protocol increased the time required for testing by ~3×, translating our dataset into an extensive amount of data collection.

**Reverse correlation analysis**

If \( N_t(m,d) \) is the number of noise dots at motion direction \( m \) and disparity \( d \) for the target (t), and \( N_{nt}(m,d) \) is for the non-target (nt), \( \Delta N_c = N_t(m,d) - N_{nt}(m,d) \) where \( c=0 \) for trials on which the observer responded incorrectly and \( c=1 \) for correct-response trials. The final perceptual filter \( P \) is simply
\[ P(m,d) = E(\Delta N_1) - E(\Delta N_0) \] where \( E() \) is the average across trials (Abbey and Eckstein, 2002).

**Modeling (Fig. 8)**

Our model consists of three filters: 1) an attentional depth filter \( F_a \) obtained by collapsing the average pre-adapted perceptual filter (smoothed using a 2-pixel smoothing kernel and sign-inverted within the Upward region) across the direction of motion (Figure 8B); 2) a directional filter \( F_d \) obtained by collapsing the average pre-adapted perceptual filter (smoothed using a 2-pixel smoothing kernel) across the direction of disparity (Figure 8C) and 3) a joint motion-disparity filter \( F_j \) equal to the average differential filter smoothed using a 2-pixel smoothing kernel (Fig. 8D). The scalar output corresponding to the target was \( O_t = g_1 (I_t(m,d) \ast (F_a \times F_d)) + g_2 (I_t(m,d) \ast F_j) \) where \( I_t(m,d) \) is the direction-disparity configuration of the target on a given trial, \( \ast \) is cross-correlation and \( \times \) is pixel-by-pixel multiplication. Similarly \( O_{nt} \) for the non-target. The model responded correctly on that trial if \( O_t > O_{nt} \), incorrectly otherwise (a random response was selected when \( O_t = O_{nt} \)). The percentage of signal dots used for the model was 12%.

**Correlation with basic templates (Fig. 4)**

Four values were computed from the raw perceptual filter, corresponding to the average modulation within each of the four 3×3 square regions indicated by the dashed outlines in Fig. 2A-C. We refer to these values as \{↑Far, ↓Far, ↑Near, ↓Near\}. We then computed the correlation coefficient between this vector and \{1, -1, -1, 1\} to obtain abscissa values in Fig. 4, or \{-1, 1, -1, 1\} to obtain ordinate values.

**Modulation ratios (Fig. 5)**

Overall modulation within a region of the raw perceptual filter was defined as the sum of squared values within that region: \( \Sigma CI(m,d)^2 \). For example, we computed the modulation ratio between the Near and the Far regions: \( \Sigma CI(m,d>0)^2 / \Sigma CI(m,d<0)^2 \) (plotted in Fig. 5A as grey symbols), or the modulation ratio between the Downward and the Upward regions: \( \Sigma CI(\downarrow\pm\pi/2,d)^2 / \Sigma CI(\uparrow\pm\pi/2,d)^2 \) (plotted in Fig. 5B). We also computed ratios between quadrant regions of the perceptual filter, e.g. between Downward-Near and Downward-Far: \( \Sigma CI(\downarrow\pm\pi/2,d>0)^2 / \Sigma CI(\downarrow\pm\pi/2,d<0)^2 \) (red symbols in Fig. 5A).

**Separability index from singular value decomposition (Fig. 7)**

Singular value decomposition (SVD) was applied to the smoothed perceptual filter to compute the separability index \( s_2/s_1 \), the ratio between the second and the first singular values (we used
smoothed rather than raw surfaces for this analysis because SVD was very sensitive to the noisy modulations in our data). This index equals 1 for a fully inseparable surface and 0 for a fully separable one (Mazer et al., 2002; Peña and Konishi, 2001), and has already been used to assess joint encoding of motion and disparity in the physiological literature (Grunewald and Skoumbourdis, 2004).

**Calculation of ideal threshold**

If \( s \) is the number of signal dots in the target (and non-target), the number of dots at each motion-disparity location is Poisson-distributed with mean and variance \( \sigma_p^2 = (100 - s) / 272 \) (total number of dots (100) minus number of signal dots, divided by the total number of motion-disparity locations). \( d' \) is \( \frac{(R_t - R_{nt})}{\sigma_n} \) where the response to the target \( R_t \) for the ideal filter \( \{1, -1, -1, 1\} \) is simply \( s \) (average response to noise is 0), the response to the non-target \( R_{nt} \) is simply \(-s\), and the variance \( \sigma_n^2 \) for both is \( 4 \sigma_p^2 \). At threshold we set \( d' = 1 \), resulting in \( s = \sigma_p \). We substitute the formula for \( \sigma_p \) and solve the resulting quadratic equation to obtain \( s = 0.645 \). Because this is a non-integer value, it does not correspond to an actual stimulus outside motion-disparity space (number of dots must be integer).

**Procedure for averaging perceptual filters across observers**

There is no generally accepted procedure for generating an average perceptual filter from individual images for different observers. One approach is to combine trials from the different observers as if they referred to a single “aggregate” observer, and compute the corresponding perceptual filter (e.g. Neri and Levi, 2007). The resulting image, however, can be biased towards individual observers. If, for example, different observers collected significantly different numbers of trials, observers who collected more trials will impact the aggregate perceptual filter more than those who collected less trials. This is a potential issue for the present study, because some observers (S1 and S4) collected significantly fewer trials than the others.

Another issue is that, even for equal numbers of trials across observers, the threshold point may differ for different observers. Depending on the stimulus and experimental design, different threshold points may correspond to different noise intensities and the amplitude of the perceptual filter depends (among other factors) on noise intensity (Ahumada 2002). Consequently, observers who tolerate higher noise intensities at threshold will contribute noise samples with higher amplitudes to the aggregate perceptual filter, potentially affecting it more than observers who tolerate less noise. This is a potential issue for the present study, because the selected signal
percentages varied for different observers; in particular, the threshold signal percentage for S3 was $\frac{1}{3}$ of the average value for the remaining 4 naïve observers.

Finally, even for equal numbers of trials and threshold signal-to-noise ratios (SNR) across observers, there is a residual issue stemming from the fact that the amplitude of the perceptual filter depends on the observer's internal noise (Ahumada 2002). Observers with less internal noise will impact the aggregate perceptual filter more than observers with considerable internal noise. We cannot assess the potential impact of this issue for the present study, but the variability in performance across observers (different threshold SNR's and percentage correct values, see Fig. 9) indicates that it may play a role. For all these reasons, we normalized the perceptual filters for individual observers before averaging them (simple mean) to obtain the aggregate perceptual filters (shown in Fig. 2). Normalization consisted of dividing the two raw perceptual filters (before and after adaptation) for each observer by the mean square modulation across both images. This ensured that data for individual observers was scaled to comparable units of modulation before averaging them across observers (and that the relative amplitudes before and after adaptation were preserved). Average percent correct values (plotted in the main panel of Fig. 9) were computed by simply averaging the corresponding values across observers (shown in the smaller panels of Fig. 9).

**Computation of baseline noise perceptual filters**

For some of the analyses performed in this study it was necessary to compare the results obtained for the experimentally derived perceptual filters to baseline values expected from noise. We therefore computed baseline noise perceptual filters for each observer by resampling the corresponding dataset as if the observer responded randomly on every trial, and deriving the associated perceptual filter (together with the corresponding scalar metrics). This procedure was carried out separately for pre-/post-adapted filters and differential filters because differential filters are expected to be noisier. For this reason, differential baseline noise filters were obtained after subtracting two separate estimates as was done for the actual differential perceptual filters.

**RESULTS**

**Perceptual filters before and after adaptation**

We derived motion-disparity perceptual filters by reverse-correlating the observers' responses with noisy stimuli that varied across both dimensions (Figure 1D-F), both before and after exposure to a long-lasting adaptor (Figure 1A-C). Figure 2A shows the motion-disparity perceptual filter before adaptation (disparity is on the x axis, motion on the y axis). For an ideal observer, this surface would contain two positive peaks at the motion-disparity locations indicated by red crosses (target
signal) and two negative peaks at the locations indicated by blue crosses (non-target signal). Although far from ideal, the surface shows a visible positive modulation at the level of the bottom-right cross (Near plane and Downward motion). This modulation is very noisy, as is the overall filter, so we applied a smoothing Gaussian kernel to the surface to reduce high-frequency noise. This procedure requires the choice of a value for the standard deviation of the Gaussian kernel. For the purpose of guiding this choice, we estimated the rough extent of the relevant modulations by fitting an unconstrained two-dimensional Gaussian to the positive peak in the surface (yellow contours in Figure 2). We then chose a Gaussian smoothing kernel that would generate a similar structure in the smoothed surface, shown in Figure 2D (more specifically we selected an isotropic standard deviation of 1 surface pixel, which is smaller than the smallest standard deviation of the unconstrained fit whose lower bound was 1.16 pixels; we also present data for a standard deviation of 2 surface pixels at Figures 3, 6, 7 and 8). The same smoothing kernel was used to generate Figure 2E from Figure 2B, the motion-disparity perceptual filter resulting from data that was collected while observers were adapted to a long-lasting stimulus that consisted of two positive peaks at the motion-disparity locations indicated by red crosses (see Figure 1 and METHODS). We note at this stage that the bulk of our analysis (to be presented below) was performed on the raw, not on the smoothed data. Smoothed data is presented here mainly to facilitate visual inspection of the surfaces. Smoothing is the norm in noise image classification studies (Beard and Ahumada, 1998; Tadin et al., 2005). It is best avoided when possible (Neri and Levi, 2007), but it was necessary for visual inspection of the data presented here. In addition, all statistical tests were carried out on metrics extracted from data for individual observers, not on average data, so the procedure we chose for averaging data across observers (see METHODS) is irrelevant to the p values we report.

Both before and after adaptation (Figure 2D and Figure 2E respectively) the structure of the perceptual filters is clearly far from ideal: Downward motion is weighed positively and Upward motion is weighed negatively across the whole disparity range from Near to Far (compare profiles for slices at different disparity values in Fig. 2G-H), inconsistent with signal motion (shown by crosses) which inverts at Far disparities. Moreover the amplitude of the filter is large at Near disparities but smaller at Far disparities (compare red and blue profiles in Fig. 2H), inconsistent with signal modulation which is equal at both. Although this Near/Far imbalance is not visibly pronounced in the average pre-adapted filter (Fig. 2D), the disparity profile obtained by averaging across motion (after inverting Upward values) shows that it is present here too (see magenta profile at Fig. 8B). This asymmetry in the data suggests that observers may have been preferentially allocating attention to the Near region of the stimulus. After this attentional selection stage, the stimulus may have been processed by a simple directional filter preferring Downward motion. As
we demonstrate in the modelling section, this simple scheme accounts reasonably well for the shape of these surfaces. We refer to this model as the “separable component.”

In order to expose the specific effect of adaptation on the perceptual filters, we simply subtracted the pre-adapted filter from the post-adapted one. The result is plotted in Figure 2C (raw) and 2F (smoothed). Figure 2F exposes a crossed structure that was not immediately visible in either the pre-adapted or the post-adapted filter: directional preference inverts between Near and Far (compare red and blue profiles in Figure 2I). More specifically, there are now two clear positive modulations: a bright peak at Downward-Near and a second positive peak of comparable amplitude at Upward-Far. There are also two negative peaks, one at Downward-Far and one of smaller amplitude in the vicinity of Upward-Near. This structure resembles the target signal (indicated by red crosses) and (although to a lesser extent) the non-target signal (shown by blue crosses). It is therefore also similar to the structure expected of an ideal matched template. As we demonstrate below, it bears the signature of joint motion-disparity encoding and is to a significant extent inseparable in the two dimensions of motion and disparity (in this respect it differs from the ideal template which is separable, see below for more details). We refer to this structure as the “joint component.”

Fig. 3 shows perceptual filters for individual observers. We found a significant degree of variability across observers, making it difficult to draw conclusions from simply inspecting individual perceptual filters. For this reason we performed additional analyses that captured a variety of aspects of these surfaces, and quantified each aspect using a single value for each perceptual filter (scalar metrics). This made it then possible to perform simple population statistics in the form of t-tests, and confirm or reject specific hypotheses about the overall shape of the perceptual filters.

Assessment of filters using scalar metrics I: Correlation with basic templates

As discussed in the previous section, two main modulation patterns appear evident from inspection of the average perceptual filters in Fig. 2: one in which directional tuning remains similar across disparity, which we refer to as the “uncrossed” pattern, and one in which it inverts between Near and Far, which we refer to as the “crossed” pattern. If we consider only the four locations in motion-disparity space targeted by signal dots (shown by crosses) and refer to them in the order Upward-Far / Downward-Far / Upward-Near / Downward-Near, the uncrossed pattern can be described as -1 / +1 / -1 / +1 or -/+/-/+ for simplicity, while the crossed pattern is +/+-/-/+ (these are represented by the surfaces indexed Y and X respectively in Fig. 4).

Fig. 4 plots correlation coefficients between the four regions in the raw perceptual filter
indicated by dashed outlines in Fig. 2A-C and each of these two patterns on y and x axes respectively. Perceptual filters before adaptation correlate well with the uncrossed pattern for all five observers (open black symbols), but not with the crossed one. More specifically, the correlation coefficients for the uncrossed pattern are significantly different from 0 (two-tailed t-test returns $p < 10^{-4}$) but not for the crossed pattern ($p = 0.49$). After adaptation (red) both patterns generate correlation coefficients that differ significantly from 0 ($p < 0.05$ for uncrossed and $p < 0.002$ for crossed). Finally, the differential perceptual filters (After- Before, blue symbols) only correlate with the crossed pattern ($p < 0.002$), but not with the uncrossed one ($p = 0.65$). All perceptual filters contain structure because they fall outside the range expected for baseline noise (indicated by the grey circular region at the origin; see METHODS for how we computed baseline noise perceptual filters). We also confirmed these results using partial correlation coefficients. More specifically, the partial correlation coefficients between pre-adapted filters and the uncrossed pattern while controlling for the crossed pattern are significantly different from 0 (two-tailed t-test returns $p < 10^{-4}$), but not between filters and the crossed pattern while controlling for the uncrossed one ($p = 0.36$). After adaptation both patterns generate partial correlation coefficients that differ significantly from 0 ($p < 0.04$ for uncrossed while controlling for crossed and $p < 10^{-3}$ for crossed while controlling for uncrossed). The differential perceptual filters (After- Before) only show partial correlation with the crossed pattern while controlling for the uncrossed one ($p < 10^{-3}$), but not with the uncrossed one while controlling for the crossed one ($p = 0.92$).

The results in Fig. 4 prompt two immediate considerations. First, it is puzzling that perceptual filters before adaptation (black symbols) correlate so poorly with the crossed structure, because this structure also reflects the signal to be detected. In the context of a simple linear amplifier model, these filters would be very poor at discriminating target from non-target in our task; human observers instead performed above chance before (as well as after) adaptation (see Fig. 9). We do not have a ready explanation for this potential inconsistency, but it may reflect the presence of significant nonlinear processing that is not captured by the perceptual filters alone. Second, the good correlation of the post-adapted perceptual filters (red symbols) with both crossed and uncrossed patterns may simply indicate that these filters conformed to the structure indicated by Z in Fig. 4. Correlation coefficients for this structure map to the yellow cross, which overlaps with the average post-adapted perceptual filter (solid red symbol).

We believe the structure in Z only provides a coarse account of the post-adapted perceptual filter: in Fig. 2H, for example, the Far region of the filter contains two small positive modulations at both Upward and Downward directions (blue profile); these would not be detected by the analysis presented in Fig. 4. The major challenge posed by this coarse account is that the transition from Y to
Z can be obtained by simply shifting the disparity profile towards the Near region (see next section for a more detailed discussion of this hypothesis), without any need for joint encoding. We therefore proceed to characterize the perceptual filters in more detail, in order to exclude this simple hypothesis and demonstrate the existence of a joint component in our data.

Assessment of filters using scalar metrics II: Modulation ratios
As briefly mentioned above, from Fig. 4 it would appear that the data can be captured by the single and simple hypothesis that, following adaptation, the disparity profile was shifted towards the Near region of the perceptual filter, thus leading from structure Y in Fig. 4 to structure Z. Structure X would then trivially result from taking the difference. In the Discussion we consider plausible scenarios that may produce such an effect.

This potential explanation, which we refer to in the following as the “disparity-shift” hypothesis, makes a series of testable predictions. In this section we present results that do not conform to these predictions. In the process of testing the hypothesis we performed a variety of analyses and statistical tests, which could be broadly grouped under one of two scenarios: either “the hypothesis predicted effect x, but we did not observe it”, or “the hypothesis predicted no effect, but we observed a statistically significant effect.” Of these we present only results from the latter class of analyses as they are clearly more compelling. In particular, we tested the following prediction: if the transition from the pre-adapted to the post-adapted filter simply involved a change in the disparity profile, we expect that the modulation ratio between Near/Far should be similar for Upward and Downward directions (i.e. it would affect all motion directions equally). A related prediction is that the Downward/Upward modulation ratio, which should be equal for Near and Far if the surfaces are separable, should remain unaffected by this transition and be the same for pre-adapted and post-adapted perceptual filters. These predictions are straightforward, but we decided to confirm them directly for the noise statistics used in our experiments by simulating the disparity-shift hypothesis in our model (oval outlines in Fig. 5). More specifically, we used a model with only the separable component and simulated a shift in the disparity profile that would generate modulation ratios of comparable size to those observed experimentally. All predictions were confirmed by these simulations.

Fig. 5A plots Near/Far modulation ratios across the whole directional range (grey symbols), and shows that this ratio is larger for post-adapted as opposed to pre-adapted perceptual filters (data points lie above the unity line; two-tailed paired t-test for Before versus After returns p<0.05). This result is consistent with the hypothesis that the disparity profile shifted to emphasize the Near region after adaptation. However, the disparity-shift hypothesis predicts that this effect should be
similar for Upward and Downward directions; instead, differential analysis of the two directions reveals that the effect is only due to modulations within the Downward region (red symbols, p < 0.02), but not within the Upward region (blue symbols fail to show any difference between Before and After adaptation, p = 0.77). Furthermore, there was a striking correlation across observers between modulation ratios Before and After adaptation but only for the Downward direction (red symbols lie along the best-fit line shown in Fig. 5A with a correlation coefficient close to 1). These results cannot be attributed to an overall lack of structure in the Upward region of the perceptual filters, because the modulation ratio between Downward and Upward was close to 1 both before and after adaptation (grey symbols in Fig. 5B). We tested the prediction from the disparity-shift hypothesis more stringently by directly comparing the differential effect of adaptation on Upward and Downward modulation ratios, i.e. we subtracted pre-adapted modulation ratios from post-adapted modulation ratios and compared the difference between Upward and Downward. This analysis is further remote from the raw data, so the comparison was very close to significant on a paired t-test but did not survive the 0.05 threshold (p=0.052). There was also a marginal effect (survived a one-tailed (but not two-tailed) t-test at p<0.05) of Near/Far ratios being larger within the Downward than the Upward regions of post-adapted perceptual filters (red symbols are shifted upwards compared with blue symbols), but not of pre-adapted perceptual filters (p=0.9). This result is not compatible with either the prediction from the disparity-shift hypothesis or with the assumption of separability of the post-adapted perceptual filters, but some of the tests were only marginally significant so we analyzed the data further (see below).

When Downward/Upward ratios are computed across the whole disparity range they differ Before and After adaptation (grey symbols in Fig. 5B lie above the unity line; two-tailed paired t-test for Before versus After returns p <0.02). This already violates the prediction from the disparity-shift hypothesis that the ratio should remain unchanged, but the effect is relatively small (~15% increase in average ratio between Before and After). When the same analysis is performed separately within the Near and the Far regions of the perceptual filters, it becomes apparent that the prediction of equality holds true within the Far region (blue symbols, p = 0.45) but not within the Near region (red, p< 0.05). More specifically, Downward modulations were significantly larger than Upward modulations within the Near region of post-adapted perceptual filters (ordinate values for red symbols are greater than 1, two-tailed t-test returns p < 0.05), and Downward/Upward ratios were on average ~40% larger than before adaptation. These results are inconsistent with the disparity-shift hypothesis. In addition, there was a marginal effect (survived a one-tailed (but not two-tailed) t-test at p<0.05) of Downward/Upward ratios being larger within the Near than the Far regions of post-adapted perceptual filters (red symbols are shifted upwards compared with blue
symbols), but not of pre-adapted perceptual filters (p=0.7). Similarly to the analysis for Near/Far ratios, this result is not compatible with the assumption of separability of the post-adapted perceptual filters. We note that for Upward/Downward modulation ratios it is not necessary to compare directly Near and Far as we did for Upward and Downward in relation to Near/Far modulation ratios, because the prediction from the disparity-shift hypothesis is not only that Upward/Downward modulation ratios should not differ between Near and Far regions, but also that they should not differ before and after adaptation, so it is sufficient to demonstrate that data for either one (Near or Far) lies above the unity line (red symbols in Fig. 5B). Finally we note that the data plotted in Fig. 5B can, in principle, be obtained directly from that plotted in Fig. 5A (i.e. there is some degree of redundancy between the two panels).

**Assessment of filters using scalar metrics III: Separability index**

The disparity-shift hypothesis can be formulated in more general terms as follows. If the perceptual filter before adaptation is a separable surface of the type \( P(m,d) = g(m) \times f_1(d) \), where \( g \) is the motion profile and \( f_1 \) the disparity profile, the perceptual filter after adaptation becomes \( g(m) \times f_2(d) \), where \( f_2 \) is the new disparity profile after adaptation. The differential perceptual filter \( g(m) \times [f_2(d)-f_1(d)] \) is still a separable function of motion and disparity. In the absence of noise, the prediction is therefore that the differential perceptual filter, as well as the two original filters, should all be separable. This prediction is not straightforward in the presence of noise, an issue we address below.

As a preliminary observation, we note that the average differential perceptual filter (Fig. 2F) presents a modulation pattern that is not entirely compatible with a separable surface. The motion profile at 0 disparity, shown by the black curve in Fig. 2I, presents two positive modulations centred on \( \leftarrow \) and \( \rightarrow \). This raises the possibility that the differential filter may contain an oriented motion-disparity modulation of the general class that has been proposed to operate in visual plenoptic space (Adelson and Bergen, 1991), similar to the oriented space-time receptive field underlying human motion perception (Burr et al., 1986). The circularity of directional space (i.e. the top and bottom rows of Fig. 2F are connected) combined with simple symmetry considerations (i.e. we expect that \( \leftarrow \) and \( \rightarrow \) were perceptually equivalent for our task as it involved \( \downarrow \) and \( \uparrow \) signals) implies that such an oriented structure would have to appear in the form of a triangle modulation connecting the red crosses on the surface. It is possible to see hints of this structure in Fig. 6, where the differential perceptual filter (smoothed using a 2-pixel standard deviation for the smoothing kernel) has been replicated three times (to conform to the circularity of directional space) and modulations above
average are tinted in red, while those below are tinted in blue. It can be seen that the red-tinted modulation follows the triangle structure described above (indicated by yellow line).

One important caveat must be considered in interpreting Fig. 6. For the crossed structure +/- -/+ +, no amount of smoothing will generate the triangle modulation shown in Fig. 6. However, if for example the positive peaks are larger than the negative peaks, i.e. ++/-/-/++, then heavy smoothing will blur the positive peaks into oriented streaks not dissimilar from those shown in Fig. 6. It is therefore possible that the oriented structure shown in Fig. 6 is mainly a smoothing artefact. A few elements argue against this possibility, such as the comparable magnitude of positive and negative modulations in the differential perceptual filter (see Fig. 2I), or the equally large amplitude of the positive modulation at ← (see black profile in Fig. 2I), but a cautious view must be taken. The relevant observation at this stage is that, whether the true underlying structure is oriented or simply characterized by larger positive than negative peaks, both patterns would display some degree of inseparability (the extreme case of a crossed structure with positive peaks much larger than the negative peaks, i.e. +/0/0/+, is fully inseparable). This is not the case for the crossed structure +/- -/+ +, which is fully separable and therefore potentially explained by the disparity-shift hypothesis.

Fig. 7A plots both first (x axis) and second (y axis) singular values obtained by applying SVD to the smoothed perceptual filter (smoothing at 1-pixel standard deviation). Data for a fully separable surface would plot to the x axis: the second singular value would be close to 0. Data for a fully inseparable surface would plot to the unity line (black): the second singular value would be equal to the first singular value. Separability can therefore be assessed by taking the ratio between the second and the first singular values; this separability index is plotted in Fig. 7B, where we show that it is closer to 1 (more inseparable) for the differential perceptual filter compared to the post-adapted and pre-adapted perceptual filters. Before focusing on ratios, however, it is important to examine the singular values individually as their magnitude allows us to exclude that the results simply reflect the presence of noise.

As shown by the green symbols in Fig. 7A, baseline noise perceptual filters (see METHODS for how they were computed) plot to a region that is very close to the unity line; indeed, they fall along the magenta line defined by a noiseless oriented Gabor. In terms of separability index (ratio between first and second singular values), a noisy surface can therefore produce values that are comparable to those associated with a clearly oriented structure. This is a potential problem for interpreting the data from the differential perceptual filters (blue symbols), as they fall along the same line. However, the magnitude of the singular values is larger than expected from noise alone. More specifically, the second singular value is significantly larger than noise (p<0.05) for a one-tailed paired t-test (a one-tailed test, rather than two-tailed, is justified here because singular values
are expected to be greater or equal to baseline noise). This is also the case for pre-adapted perceptual filters (p<0.05), but not for the post-adapted ones (p=0.07), and none of the filters survives a t-test for the first singular value being greater than noise. Overall, this analysis indicates that singular values for the differential perceptual filters are larger than baseline noise, and in any case not closer to noise than those associated with the post-adapted and pre-adapted filters. We therefore conclude that a comparison of separability indices between different classes of perceptual filters is unlikely to reflect different degrees of noisiness associated with the filters. We now turn to this comparison.

Fig. 7B plots separability indices for differential perceptual filters (y axis) versus corresponding indices for pre-adapted (black) or post-adapted (red) perceptual filters, for two smoothing levels (indicated by symbol size and insets). It is clear that the former are generally greater than the latter (points lie above unity line), and only the former fall within the region defined by an oriented noiseless Gabor wavelet (magenta line). More specifically, differential perceptual filters (smoothing level 1-pixel standard deviation) show significantly larger indices than post-adapted filters (two-tailed paired t-test returns p<0.05) and this effect is marginally significant in the comparison with pre-adapted filters (survives a one-tailed paired t-test at p<0.05). For the larger smoothing kernel (large symbols) the effect is significant (two-tailed paired t-test returns p<0.05) for the comparison with pre-adapted filters (but not post-adapted). Overall, differential perceptual filters display a larger degree of inseparability than both pre- and post-adapted perceptual filters.

As well as being inconsistent with the disparity-shift hypothesis, this analysis emphasizes the highly constrained nature of the requirements for joint encoding that we are imposing on our data (mainly reflected in the metrics used in FIG. 5-7). These requirements are far more restrictive than in related single-unit studies. For example, at the level of individual neurons it is sufficient to show that the motion-disparity surface is multiplicative rather than additive in nature (Grunewald and Skoumbourdis, 2004). This criterion is too lax for our purposes, because in some circumstances multiplicative separable filters can be implemented by plausible models where motion and disparity are not coded jointly (a demonstration is provided by the separable component in our model, discussed below). More generally, these differences stem from the challenging issue of translating joint encoding from the neuronal to the perceptual level (see Discussion).

Modeling
As anticipated in the previous section, a plausible explanation for the results in Figure 2 is that two processing components are at play, a separable and an inseparable joint one. The separable component is the product of a two-step process whereby Near information in the stimulus is first
selectively attended; the output from this attentional depth filtering stage is then processed by a directionally selective filter preferring Downward motion. Central to the present discussion, the initial depth-selective stage is not motion-selective, and the subsequent motion-selective stage is not depth-selective. In this sense the process does not reflect joint encoding by the same neural mechanism. In contrast, the joint component relies on mechanisms that are jointly selective for both motion and disparity, and is thus able to generate an inseparable motion-disparity surface like that in Figures 2F and 6.

To demonstrate our interpretation of the results, we constructed a simple model that incorporates these two components and we challenged it with the same stimuli used with human observers. The separable component is implemented by the two filters in Figure 8B and 8C (directly obtained from smoothed data). The joint component is implemented by the filter in Figure 8D (equal to the differential filter estimated experimentally). The outputs from the two components are summed only after their gains are separately modulated by the multiplicative factors g₁ and g₂ (see METHODS). As shown in Figure 8E-G, we were able to capture the essential experimental results by assuming that the ratio between g₁ and g₂ was different before and after adaptation (pixel-by-pixel correlation coefficients between experimental and simulated surface were 0.97 for pre-adapted, 0.95 for post-adapted and 0.97 for differential filters). In other words, we are assuming that adaptation affected separable and joint components differentially, allowing us to resolve them. For example, adaptation may have reduced the firing rate of neurons involved in the separable component by a factor of 20, while reducing the firing rate of neurons supporting the joint component by a factor of 2 (so that the ratio g₁/g₂ changes by 10×).

This model is also able to capture other aspects of the data. For example, it violates the predictions of the disparity-shift model like the real data (Fig. 5). Although the absolute values of the Near/Far modulation ratios differ significantly enough from the human data that it resulted impractical to plot them in Fig. 5A, they (as well as the Downward/Upward ratios) show the same trends. Fig. 9 shows that the model is also able to capture the difference in performance between adapted and non-adapted conditions (open symbols) at least qualitatively.

DISCUSSION

Logic of the experiments and brief summary of the results

Our goal was to expose a perceptual structure that would display joint encoding of motion and disparity. For this purpose we designed a task that required observers to rely on both attributes (see Fig. 1E-F) and exposed them to a long-lasting adaptor consisting of a joint motion-disparity pattern (Fig. 1B). Our reasoning was as follows: if there exists a perceptual component that supports joint
encoding, this component would be selectively targeted by the joint adaptor and become visible in our post-adapted data. By combining the powerful psychophysical techniques of selective adaptation and reverse correlation, we were able to extract this component (Fig. 2F) and show that it conforms to the joint structure induced by the adaptor (Fig. 4). We therefore confirmed our hypothesis that there exists a joint perceptual component, and that this component was exposed by selective adaptation to joint stimulus statistics.

However there is a plausible competing hypothesis which we refer to as the disparity-shift hypothesis, according to which the joint pattern exposed by the adaptor does not reflect selective adaptation of a joint perceptual component, but rather results from a shift in the disparity profile whereby observers rely more on the Near region of the stimulus following adaptation. We were able to rule out this hypothesis with additional data analysis (Fig. 5), and therefore conclude that the joint structure exposed by adaptation is not accidental. This conclusion is based entirely on raw data (analyses in Figs. 4-5).

We then attempted to provide further evidence to support our conclusions by demonstrating that the structure induced by adaptation is not only joint but also inseparable in the dimensions of motion and disparity (Figs. 6-7). This analysis was based on smoothed data and is very sensitive to measurement noise, so its purpose is mainly confirmatory rather than critical to this study. Our main conclusions are independent of this analysis.

Finally, we instantiated a simple model to illustrate that the main assumption underlying our interpretation of the data is that joint and separable components are affected differently by adaptation (Fig. 8). We address this assumption in more detail later in the Discussion.

The thorny issue of defining joint encoding psychophysically
It is not entirely obvious what criteria should be met by a process in order for it to qualify as “joint encoding.” For example, if the requirement for joint encoding is simply that a process must register, at some level, information about both motion and disparity, then the discrimination task in our experiments requires joint encoding by design. Target and non-target are statistically indistinguishable when only information about one dimension is used. This is a consequence of the fact that both target and non-target contain (on average) the same number of Upward-moving, Downward-moving, Far and Near dots. It is the way in which Upward/Downward and Far/Near are paired that defines target and non-target, but this information is lost when the surface distributions in Figure 1E (target) and 1F (non-target) are collapsed across either motion or disparity. That observers could perform well above chance in the discrimination task is, by itself, a clear indication that they must have been using stimulus information about both motion and disparity (see Fig. 9 for
observer performance levels).

The above definition of joint encoding is clearly too lax. For example, a two-stage process like the one illustrated in Figure 8B-C would satisfy it. Although each stage on its own is only registering information about one dimension, the combined process displays some degree of joint encoding in relation to our task. This process is able to discriminate target from non-target above chance, because the cross-correlation between the resulting filter and the stimulus generates larger values for target stimuli as opposed to non-target stimuli. Nevertheless, this process does not conform to the notion of joint encoding that would satisfy a physiologist: if one were to record from the neural circuitry that may support the process in Figure 8B-C, it is plausible that no individual neuron would be found to carry information about both motion and disparity. Similarly, the model developed by Read and Cumming (2005b) would be able to discriminate target from non-target in our experiments, because the two stimuli generate different correlation patterns for the firing of motion-encoding and disparity-encoding populations. However their model does not qualify as a form of joint encoding; rather it was intended as a counter example to joint encoding which could accomplish the same outcome (Read and Cumming, 2005b).

It appears sensible to start from a definition that is satisfactory from a physiological standpoint, and then work out a psychophysical equivalent. A suitable physiological definition could simply be that a system displays joint encoding when at least some of its constituent neurons are simultaneously selective for both motion and disparity (as mentioned in the Introduction, such neurons have been found in visual cortex so their existence is not in dispute). To translate this definition into psychophysical terms we recur to the method of adaptation, also known as the “psychophysicist's microelectrode” (Frisby, 1979) because the perceptual effects of adaptation are generally regarded as reflecting the properties of mechanisms that are as close to individual neurons as is psychophysically possible. Using this tool, the requirement for joint encoding would state that a process must display adaptive properties reflecting the co-registration of information about both motion and disparity. This more restrictive definition excludes Read and Cumming's model (Read and Cumming, 2005b), because disparity-contingent motion aftereffects require that the visual system represents the specific combination of motion and disparity signals that were presented during the adapting phase. This statement is not obvious, and we therefore discuss it in detail in the following two paragraphs.

During the adapting phase of a disparity-contingent motion aftereffect, Read and Cumming's model would correctly perceive the joint motion-disparity configuration of the adaptor because, similar to Pulfrich stimuli, specific combinations of motion and disparity signals occur in a temporally correlated fashion. Although the adaptor contains an equal number of Downward-
moving and Upward-moving dots as well as an equal number of Far and Near dots, Near dots always move Downward and Far dots always move Upward. At a given point in the stimulus both combinations would appear the same number of times across the whole duration of the adaptor, but they would do so at different times. Suppose there are four neurons with overlapping receptive fields centred on this point: a Far-preferring, a Near-preferring, a Downward-preferring and an Upward-preferring neuron. This population of 4 neurons does not represent motion and disparity jointly. However, the activity across time of the Near-preferring neuron is positively correlated with the activity of the Downward-preferring neuron, because every time a Near dot appears at that point in the stimulus it also moves Downward. The same is true for the Far-preferring and the Upward-prefering neurons. A read-out mechanism that relies on the temporal correlation in activity between neurons would pick out these pairings and correctly report that the stimulus contained Downward-moving Near dots and Upward-moving Far dots. This, combined with the intrinsic correlations of Pulfrich stimuli (Tyler, 1974), allows Read and Cumming's model to explain Pulfrich-like phenomena without the need for joint encoding (Read and Cumming, 2005b).

Despite asymmetric correlations in the activity of the 4 neurons, the overall activity of each neuron is the same. This is a consequence of the balanced structure of the adaptor, which delivers (on average) an equal number of signals at Far, Near, Downward and Upward. The result of this uniform activity pattern is that all 4 neurons should adapt to the same extent, creating no bias in the system. In other words, an adaptive version of Read and Cumming's model would not preserve any trace of the correlations that drove it in the first place. Indeed, it cannot do so by its very definition of separable encoding: any trace of correlations in the adaptor (e.g. Downward + Near) would need to be stored at the level of some representation within the model, and this representation would need to be joint. Because it cannot be jointly adaptive, Read and Cumming's model does not predict any subsequent bias in perceiving an unbiased test stimulus (for which no specific correlations exist), but this bias is perceived by human observers during the disparity-contingent motion aftereffect (Sohn and Seiffert, 2006). We conclude that Read and Cumming's model is not able to account for disparity-contingent motion aftereffects, and would not predict any difference between post-adapted and pre-adapted filters in our experiments.

On the other hand, the two-stage process in Figure 8B-C can explain various aspects of disparity-contingent aftereffects. The first stage (supported by an attentional depth filter) silences signals from the Far region of the adaptor, reducing it to a stimulus containing predominantly Downward-moving dots. The second stage is supported by directionally selective neurons which would undergo motion adaptation, leading to an aftereffect. The two-stage model appears therefore able to provide a reasonable account of disparity-contingent aftereffects (see later in Discussion for
more details), but we believe it fails to satisfy the above requirement for joint encoding because its adaptive properties are not strictly joint. This failure is exposed by the detailed perceptual filters that can be derived using psychophysical reverse correlation, as we have done in this study and as we discuss further in the Discussion.

It may be argued that the two-stage process in figure 8B-C is actually an example of joint encoding rather than separable encoding: although it is true that a hypothetical single-unit experiment aimed at targeting this process would not find units simultaneously selective for both motion and disparity within the process itself, joint selectivity may be recorded from the downstream neurons which combine the signals generated by the process after they are passed on to the read-out stage. This criticism does not apply for a number of reasons. We first note that it relies on a specific implementation of the read-out process via a grandmother-cell-like final layer. Very little is known about the read-out process that converts early sensory signals into the final percept, so this specific implementation is only one of many possibilities. Secondly, although the combined signals do display some degree of joint selectivity in the context of our task and stimuli, this selectivity is very weak in that the signals do not change directional preference across depth but simply attenuate from Near to Far. Most importantly, this attenuation is the result of an attentional filter (Figure 8B), and is therefore dependent upon the cognitive state of the animal. If the animal deploys attention uniformly across depth (or does not deploy it actively to a specific depth region as may be the case under anaesthesia during acute recording), the signals generated by the process in Figure 8B-C cease to display any degree of joint encoding. In other words, the very weak joint selectivity displayed by the process in Figure 8B-C is not an intrinsic property of individual neurons within the circuitry, but rather depends on the cognitive state of the animal. In a typical single-unit experiment where the recorded neuron is probed using a single plane of dots, and the disparity and direction of the dots is varied from trial to trial, the animal would attend equally to the plane regardless of its disparity (or would not modify its attentional state at all if anesthetized). This situation would map to a separable receptive field, selective for direction of motion but not for disparity (like the one shown in Figure 8C). The same logic applies to any attentional depth profile that is symmetric around zero disparity (e.g. one that peaks at zero and decreases for both Near and Far). In conclusion, the process represented in Figure 8B-C does not strictly qualify as a form of joint encoding at the neuronal level.

Potential role of learning and attentional strategies

Before proceeding further in discussing our data, we must consider perceptual learning as a serious potential confound in the interpretation of our results. Some aspects of the transition from pre-
adapted to post-adapted filter (Figure 2D and 2E respectively) may be interpreted as resulting from sharpening of sensory tuning. For example, the positive peak at Near-Downward (bottom-right region of the surface) is sharper across the disparity dimension following adaptation, as demonstrated by the yellow contours in Figures 2A-B (we were not able to detect significant changes in directional tuning, see directional profiles in Fig. 2G-H). Li, Levi and Klein have reported that perceptual filters derived using psychophysical reverse correlation become more sharply tuned with learning (Li et al., 2004). If learning played a role in our experiments, it may have been the source of at least some of the effects we report here.

Although learning was certainly taking place in our experiments, we can exclude that it played a role in the comparison between pre-adapted and post-adapted filters. This is because learning should have affected both equally, since the two conditions were alternated during data collection. Although we could not mix the two conditions within the same block, observers collected a few blocks without adaptation, then a few blocks in the presence of adaptation, and alternated in this fashion. This protocol ensured that if learning was taking place it would affect the two conditions equally. This is indeed what we observed, as shown in Figure 9: percentage correct improved similarly for post-adapted (red) and pre-adapted (black) trials throughout data collection. Except for a vertical shift of ~7%, the two curves overlap within ±1 standard error of the mean (compare region within red dashed lines with grey shaded region).

In addition, we can further exclude a learning-based explanation of our results because it would predict a negative modulation at Upward-Far in the differential filter (whereas we observed a positive modulation, see upper-left region in Figure 2F). This prediction results from the fact that the pre-adapted filter shows a negative modulation at Upward-Far (see blue trace in Fig. 2G). Sharpening would lead to a more pronounced negative peak within this region of the filter, generating a negative modulation in the differential filter in the same way that it generates a positive modulation at Downward-Near corresponding to the positive peak in the pre-adapted filter within this region of direction-disparity space.

A related concern is that the transition from Fig. 2D to Fig. 2E may reflect a simple change in attentional strategy: it may be hypothesized that, following long-lasting exposure to the adaptor, observers are able to deploy their attentional resources more sharply to the Near region of the stimulus. There is no obvious reason to hypothesize that such an attentional shift may have occurred in our experiments, but this hypothesis belongs to a more general class of explanations which we referred to as the disparity-shift hypothesis in the Results section. According to this class of explanations, the joint structure exposed by the differential perceptual filter in Fig. 2F would not reflect a separate component, but would simply result from subtracting two separable perceptual
filters that only differ in the distribution of the disparity profile. Because we wish to demonstrate that our adaptation protocol exposed a separate joint component in Fig. 2F, it is essential to rule out this alternative class of explanations.

A simple change in the shape of the disparity profile makes specific predictions for how the perceptual filter should differ between pre- and post-adapted conditions. For example, it predicts that the same degree of disparity shift should be observed for all motion directions (the perceptual filters are assumed to be separable). We demonstrated that this prediction is violated by the raw data (Fig. 5A). A related prediction is that the directional tuning profile may be scaled, but should not change shape. Again, this prediction was directly violated by the raw data (Fig. 5B). Finally, the disparity-shift hypothesis predicts that the differential perceptual filter, like the pre- and post-adapted filters, should all be separable to similar extents. The results of both the analysis based on modulations ratios (Fig. 5) and that based on separability indices (Fig. 7) were inconsistent with this prediction. We conclude that the disparity-shift hypothesis and related explanations cannot account for our results. Furthermore, we conclude that the transition from pre- to post-adapted perceptual filters, captured by the differential filter in Fig. 2F, bears the signature of joint encoding in the form of a crossed structure induced by the adaptor (Fig. 4) which is also significantly inseparable in the dimensions of disparity and motion (Fig. 6-7).

Attentional shift of both motion and disparity
The disparity-shift hypothesis may be augmented by a concomitant change in the motion profile, both acting on the separable component in our model without need for the joint component. However this hypothesis amounts to a form of joint encoding: whatever mechanism (possibly attentional) may be altering the disparity and the motion profile at the same time, this mechanism would need to represent the two dimensions simultaneously. In addition, this explanation fails to explain the effect of reduced separability we measured for the post-adapted perceptual filters (Fig. 5), as it predicts that pre- and post-adapted perceptual filters should be equally separable (however it could account for the inseparability of the differential filter, so the analysis presented in Fig. 6-7 is not useful for excluding this hypothesis). We conclude that our data is unlikely to be explained by a simple joint motion-disparity shift hypothesis; even if a model could be formulated along the lines of this hypothesis to explain our data, we question whether this model should be defined as an example of separable or joint encoding, largely reducing its utility for the present discussion.

Performance was better following adaptation
Figure 9 shows that performance was generally improved by adaptation, with the exception of S3
and S5 whose performance was not modified (compare solid circles before (black) and after (red) adaptation). This is an interesting result, in that we expect that observers were at the limit of their performance in the conditions of our experiments (we used a 2AFC task with feedback) and it is not obvious that adaptation would result in better performance, as most often it reduces performance for the adapted parameters (e.g. Blakemore and Campbell (1969) and the vast literature that followed). On the other hand, this effect is consistent with the overall structure of the perceptual filters as assessed by their correlation with the ideal template in Fig. 4: pre-adapted perceptual filters showed a poor correlation with this template (x values around 0 for black symbols), whereas post-adapted perceptual filters were significantly correlated with it (x values are positive for red symbols). This effect is also captured by our model (open circles in Figure 9).

We had to reduce the stimulus SNR significantly in order to bring model performance within human range (simulated number of signal dots was 28% of average human value). This is not surprising, as human performance is affected by internal noise (Burgess and Colborne, 1988) whereas our model is noiseless. Internal noise, however, is expected to result in similar reductions of pre- and post-adaptation performance leaving the comparison relatively unaffected (although the extent of the reduction depends on how internal noise is modelled). For this reason, we conclude that our model correctly captures the differential effect of performance before and after adaptation. This conclusion is not compromised by the noiseless nature of the model.

Finally, it is interesting that observers were overall remarkably inefficient at performing the task. The threshold percentage of signal dots for the ideal observer (see METHODS) translates into an average human efficiency of ~1%. The reverse correlation analysis indicates that this low efficiency results from relying largely on highly suboptimal filters that are separable in motion-disparity space (Fig. 2D-E and black symbols in Fig. 4A). Considering that subjects were receiving trial-by-trial feedback, and that it is known from the physiology that there exist subpopulations of neurons in MT that could be used to implement a much more efficient filter for the task in our experiments, it is surprising that observers were not able to deliver higher performance. This suggests that access to the joint subpopulations may not be easily achieved by the visual system, and that the two-stage strategy implemented by the separable component in our model may play an obligatory role in the processing of motion and disparity.

**Cognitive impenetrability of the differential filter**

The differential filter in Figure 2F exposes aspects of visual processing that are not explicitly accessible to the observer. By this we mean that it would not be possible for the observer to consciously generate the modulations reported in Figure 2F (Von der Malsburg, 1999). These
modulations are obtained by computing a noise average from thousands of trials collected at one time, then computing a separate noise average from thousands of trials collected at a different time, and by taking the difference between the two. In order for the observer to have cognitive control over the final difference, he/she would need to be able to monitor the motion-disparity noise distributions for the two conditions separately, and be able to predict the outcome of the average difference so as to guide his/her responses accordingly. This is highly unlikely.

The cognitive impenetrability of the differential filter may be one of the factors that confers this operator the ability to expose the joint motion-disparity component otherwise masked by the separable component. The separable component is to a large extent under explicit control of the observer, in the sense that the observer may choose to pay attention to the Far stimulus as opposed to the Near stimulus if he/she wishes to do so (however observers naturally attend to the front plane as this is perceptually more salient (Andersen, 1990)). This change in attentional strategy would significantly modify the shape of the pre-adapted and post-adapted filters in Figures 2D-E: the largest modulations would appear in the Far region of the surfaces, rather than the Near region. We speculate that the joint component is not under the same degree of control by the observer, because it relies on targeted monitoring of a specific neuronal subpopulation and cannot be described as a cognitive strategy similar to the one defining the separable component. As a consequence of this difference in cognitive penetrability, the two components may be differentially affected by adaptation, which is the main assumption underlying our model (Figure 8). This in turn would confer the differential filter the ability to expose the joint component, as demonstrated by our simulations (Figure 8G).

Is this assumption at least partly justified? Previous research has shown that attention has profound effects on adaptation to motion (Alais and Blake, 1999) as well as to disparity (Rose et al., 2003). More specifically, adaptation correlates positively with attention: the more a visual stimulus is attended, the more it undergoes adaptation (Alais, 2005). Although it is not immediately obvious how these findings may translate in the context of the experiments described here because none of them looked at both motion and disparity jointly, their general conclusions lend support to our assumption. In our model only the separable component is under attentional control. From these previous studies it may therefore be expected that this component would be more greatly affected than the joint component during the adapting phase.

The considerations just made are largely speculative, and there may well be other potential interpretations of the differential effect of adaptation on joint and separable components. We emphasize that our model does not rely on the notion of cognitive impenetrability or related concepts. We used these concepts to offer a partial justification for the differential effect of
adaptation on $g_1$ and $g_2$, but the main objective of our model is to demonstrate that it is possible to capture our data using a simple and plausible circuit such as that shown in Figure 8A-D consisting of two separate components.

Psychophysical and physiological plausibility of the separable component

The separable component in our model consists of two steps: an attentional depth filter followed by a directional filter (Figure 8B and 8C respectively), both separable in the dimensions of motion direction and disparity. This model component is consistent with previous psychophysical studies. Most relevant here are the studies by Nakayama and Silverman (1986) and Sohn and Seiffert (2006).

Nakayama and Silverman (1986) showed that reaction times can be independent of set size for certain types of conjunction search. In particular, they studied the conjunction of disparity and orientation. Their interpretation of the lack of a set-size effect was based on a two-stage model: elements are first segregated across disparity by selecting only one depth level (e.g. Near); the selected elements are subsequently analysed for orientation to identify the odd element in this dimension. The conjunction task is thus converted into a single-feature (pop-out) task by directing attention to only one depth level, resulting in no set-size effect for reaction times. This model is almost identical to our separable component, except we used motion direction instead of orientation. The separable component in our model therefore has established precedents in the perceptual literature. In line with the present discussion, these previous authors have recognized that this serial two-stage process does not qualify as a form of joint encoding.

Sohn and Seiffert (2006) characterized disparity-contingent motion aftereffects in detail with an extensive set of quantitative measurements on different stimulus configurations. For the present discussion the relevant configuration consisted of an adaptor that was very similar to the one used in our experiments, followed by a test stimulus consisting of either dots moving in random directions within the depth planes defined by the adaptor (their Experiment 1) or consisting of dots moving orthogonally to the directions defined by the adaptor (their Experiment 2). In Experiment 1, observers were asked to indicate the direction of either Near or Far surface in the test and the aftereffect was assessed using the percentage of times that the unbiased test pattern was seen to move in one direction rather than the opposite one. Sohn and Seiffert found that motion aftereffects often occurred in opposite directions for Near and Far surfaces, consistent with previous reports of the disparity-contingent motion aftereffects (Verstraten et al., 1994), but their method did not provide a transparent measure of the intensity of the aftereffect.

To overcome this limitation, in Experiment 2 the authors measured intensity in the form of
directional deflection of the test. Their results were entirely consistent with the separable component in our model, in that they observed a regular motion aftereffect for the Near plane, but virtually no effect for the Far plane (their Fig. 4h). When there was an effect, it was often in the direction expected for the Near plane, exactly as predicted by the separable component at Fig. 8B-C. Furthermore, Sohn and Seiffert interpreted their results using the same framework we used for the separable component in our model. More specifically, these authors speculated that “attentional allocation in three-dimensional space may be responsible” for the variability they observed among observers, and that some observers “might have been attending more to the front surface during adaptation” (quoted from their page 126).

Even though it appears that the separable component was playing an important role in Sohn and Seiffert's experiments, we do not actually believe this was the only factor that affected their measurements as our own data demonstrates the presence of a joint component and we think this component played an important role in their experiments. We therefore believe that their measurements did reflect joint motion-disparity processing. However, the above considerations highlight the fact that the perceptual reports of apparent motion assessed by Sohn and Seiffert are not sufficiently detailed for a close inspection of the issue of joint encoding. Together with Nakayama and Silverman's results, these observations add plausibility to the separable (attentional) component in our model.

*Psychophysical and physiological plausibility of the joint component*

The joint component in our model consists of a single filter (Figure 8D), inseparable in the dimensions of motion direction and disparity. This model component is consistent with previous physiological (Grunewald and Skoumbourdis, 2004; Anzai, Ohzawa and Freeman, 2001; DeAngelis et al., 1998; Pack et al., 2003; Roy et al., 1992) and psychophysical (Anstis, 1974; Nawrot and Blake, 1989; Verstraten et al., 1994; Sohn and Seiffert, 2006) studies.

Although there are indications that some neurons in visual cortex may implement motion-disparity filters of the type shown in Figure 8D, there is no conclusive evidence to support this notion. Some MT neurons respond selectively to the rotation of 3D cylinders (Bradley et al., 1995; Dodd et al., 2001), which may indicate that they prefer one direction of motion at a Near disparity and the opposite direction of motion at a Far disparity, somewhat like the filter in Figure 8D. However, such neurons may be selective for only one of the two motion-disparity combinations and still display directional selectivity for the rotation of the 3D cylinder. There is extensive evidence for this type of selectivity whereby neurons respond to one specific combination of motion and disparity (Grunewald and Skoumbourdis, 2004; DeAngelis et al., 1998), and microstimulation
experiments in rhesus monkeys have shown that these neurons can drive behaviour (DeAngelis and Newsome, 2004). Based on this evidence, the separable filter in our model is best interpreted as deriving from the combination of multiple neuronal subpopulations, e.g. one selective for Near disparity and Downward motion, another one selective for Far disparity and Upward motion, and possibly others selective for different combinations of these two attributes.

The joint component in our model is based on existing physiological findings, and is also consistent with previous psychophysical studies, even though those studies did not expose it as clearly. As discussed previously, Pulfrich-like effects are all consistent with a joint component but can also be explained without one (Read and Cumming, 2005b). Disparity-contingent motion aftereffects are consistent with a joint component (Sohn and Seiffert, 2006), but previous studies of this class of phenomena have not characterized the underlying mechanisms in sufficient detail to exclude explanations solely based on a separable component similar to the one we incorporated in our model. It is not obvious how our results may relate to this class of aftereffects (see next section for a more detailed discussion), but both lines of research converge on the existence of a joint component. Overall, previous psychophysical findings provided several indications of joint motion-disparity encoding but did not rule out the possibility that their origin may have been separable. Our study was specifically designed to target this issue.

**Relationship between the differential filter and the motion aftereffect**

Because this is the first time in the literature that motion adaptation and psychophysical reverse correlation are combined, there are no clear expectations as to how adaptation should affect the shape of the behavioural sensory filters derived using these techniques. It may be reasonable to expect that the changes would be restricted mainly to the adapted regions of motion-disparity space, as we observed in our data (Figure 2), but a priori it is not straightforward that they would consist of positive rather than negative modulations: the latter scenario may actually appear more likely based on the accepted notion that adaptation reduces neuronal gain (Mather et al., 1998).

There are two important reasons why these predictions are hazardous in the absence of quantitative modeling. First, recent recordings from MT neurons have demonstrated more complex and unexpected effects than simple gain reduction following adaptation (Kohn and Movshon, 2004); some of these effects would be more consistent with positive rather than negative modulations of gain. Second, even if adaptation is assumed to reduce overall gain, the amplitude of the filters returned by reverse correlation does not transparently reflect the absolute gain of the underlying neural mechanism due to the nature of this methodology (Ahumada, 2002). This is demonstrated by our two-component model, which is consistent with gain reduction of both
separable and joint components. Because the final filter does not depend on the absolute gain of the two components, but on their ratio \( \frac{g_1}{g_2} \), the model is able to replicate the positive modulations in the differential filter.

A related question is how the percept of the disparity-contingent motion aftereffect may relate to the differential filter we measured in our experiments. At first it may seem that the differential filter in Figure 2 is inconsistent with the phenomenology of the aftereffect, because it contains positive modulations corresponding to the adaptor. It may be expected that these modulations should lead to an aftereffect in the direction of the adaptor, opposite to what is observed phenomenologically. However, the link between changes in neuronal firing and the perceptual aftereffect is not fully understood at present (Kohn, 2007). A correct interpretation of how our data relates to the phenomenological aspects of motion adaptation would demand a more sophisticated understanding of the motion aftereffect than is currently available. Notwithstanding this limitation, we note that the analysis presented here does not depend on this relationship. During the piloting stages of this research we informally verified that our adapting protocol induced disparity-contingent motion aftereffects, but this was not necessary nor is it necessary to clarify the nature of this phenomenon in order to interpret the perceptual filters we derived at FIG. 2-3 and the subsequent analyses. Adaptation is not necessarily defined in relation to the presence of a perceptual aftereffect (certainly not in the physiological literature). In relation to the logic of our experiments, all that matters is that prolonged exposure to biased stimulus statistics causes some measurable recalibration in the response characteristics of the system (viewed as a detector attempting to identify a signal at a given SNR; no specific phenomenology need be associated with this detector). We were able to measure statistically significant effects of this recalibration at the level of the perceptual filters derived using reverse correlation. Whether and how these effects may relate to the aftereffect does not impact our conclusions.

GRANTS Supported by the NIH (RO1EY01728) and the Royal Society of London (URF to PN).


Read JCA, Cumming BG. All Pulfrich-like illusions can be explained without joint encoding of motion and disparity. *J Vis* 5: 901-927, 2005b.


FIGURE CAPTIONS

FIG. 1. Stimuli and procedure. The adaptor (A-C) consisted of a Near plane containing Downward-moving dots and a Far plane containing Upward-moving dots. Two such adaptors were presented, one to the right (B) and one to the left (C) of Nonius fixation. A is an attempt to capture the 3D appearance of the adaptor. B and C show the corresponding configuration in motion-disparity space, where the intensity of each pixel reflects the number of dots that moved in a given direction (indicated by the y axis) and were at a given disparity (indicated by the x axis). The adaptor was followed by a test stimulus (D-F). In the test, one side of fixation contained a target (shown in D) and the other side contained a non-target. For the target (red), a percentage of the dots (signal dots indicated by asterisks) moved like the dots in the adaptor, while remaining dots (noise dots) could take any motion direction and disparity. For the non-target (blue), signal dots were paired in the opposite way: Near dots moved Upward and Far dots moved Downward. The corresponding motion-disparity representations are shown in Fig. 1E-F, where the target is shown to the right of fixation (observer indicated by eyeballs) but the actual side of presentation in the experiments was randomized. Observers pressed one of two buttons to indicate which side of fixation contained the target. A and D are not faithful representations of the stimuli; B-C and E-F are (E-F show actual samples from a trial used with observer S3). Notice that target and non-target are statistically indistinguishable when collapsed across either motion or disparity.

FIG. 2. Average perceptual filters derived using reverse correlation. A and B show raw perceptual filters obtained before (A) and after (B) adaptation; C shows the differential filter (B minus A). Red crosses indicate signal locations for the target, blue crosses for the non-target (see Fig. 1). Thick yellow contours show half-height profiles of best-fitting two-dimensional Gaussian functions; thin contours show 95% confidence intervals. D-F show surfaces obtained after smoothing A-C (using a circular Gaussian kernel with standard deviation equal to 1 surface pixels). A-E are plotted to the grey scale shown in the upper left legend; F is plotted to the scale shown by the bottom right legend. G-I show slices along direction of motion averaged across the disparity range indicated by the corresponding dashed rectangular regions in D-F (red for Near, blue for Far, black for Nonius). Error bars show ±1 s.e.m. Units are arbitrary because averaging was performed after observer-by-observer normalization (see METHODS).
FIG. 3. Perceptual filters for individual observers (S1-S5). Plotting conventions are similar to FIG. 2. Smoothing was performed using a circular Gaussian kernel with standard deviation equal to 2 surface pixels.

FIG. 4. Correlation with basic templates. For each raw perceptual filter we computed the 4 values corresponding to the 3×3 square regions shown by dashed contours in Fig. 2 A-C (each value corresponding to the average within the square region) and correlated these values with the template shown in X (coefficients plotted on the abscissa) or in Y (coefficients plotted on the ordinate). Open symbols refer to individual observers (circle S1, square S2, diamond S3, left-pointing triangle S4, right-pointing triangle S5), solid symbols to the average (from Fig. 2A-C). Black symbols refer to pre-adapted, red to post-adapted, blue to differential filters. Error bars show ±1 s.e.m. Grey circular region centred on origin shows standard deviation of the range spanned by baseline noise (see METHODS).

FIG. 5. Modulation ratios (see METHODS for how they were computed). A plots Near/Far ratios Before (abscissa) and After (ordinate) adaptation for the entire directional range (grey), only for the Downward region (red) or only for the Upward region (blue). B plots Downward/Upward ratios Before (abscissa) and After (ordinate) adaptation for the entire disparity range (grey), only for the Near region (red) or only for the Far region (blue). Open symbols refer to individual observers (circle S1, square S2, diamond S3, left-pointing triangle S4, right-pointing triangle S5), solid symbols to the average. Error bars show ±1 s.e.m. Oval contours show mean ±1 standard deviation of the ranges spanned by simulations of the disparity-shift hypothesis (green and yellow corresponding to blue and red data respectively).

FIG. 6. Inseparable structure of the differential filter. Surface plots a triple replication of the differential filter (smoothed using a circular Gaussian kernel with standard deviation equal to 2 surface pixels) across direction of motion. Regions above mean are tinted in red, those below in blue. Yellow line shows triangle structure expected for an oriented filter.

FIG. 7. Analysis based on singular value decomposition. A plots both first (abscissa) and second (ordinate) singular values for pre-adapted (black), post-adapted (red) and differential (blue) filters. Green symbols show values for baseline noise (see METHODS), for both pre-/post-adapted filters (large symbols) and differential filters (small symbols). Magenta line shows values spanned by the
oriented Gabor surface shown in the inset (values scale with Gabor amplitude). B plots separability
indices (ratio between second and first singular values) for differential filters (After-Before on
ordinate) versus either pre-adapted (black) or post-adapted (red) filters (abscissa). Symbol size
reflects standard deviation of smoothing kernel (shown in insets) applied to the filters before
computing indices: small for 1-pixel and large for 2-pixel standard deviation. Magenta line as in A.
Error bars show ±1 s.e.m.

FIG. 8. Model and simulations. The input stimulus (A) was filtered by two separate components: a
separable component consisting of an attentional depth filter (B) followed by a directionally
selective filter (C), both derived from averaging 2-pixel smoothed experimental data (profiles are
shown in magenta), and a joint component consisting of an inseparable motion-disparity filter (D)
obtained directly from the differential filter shown in Fig. 6. The outputs from these two
components were combined (+) after their gain was controlled separately by $g_1$ and $g_2$ (see
METHODS). This model was challenged with the same stimuli used in the psychophysical
experiments, resulting in the perceptual filters shown in E-G corresponding to panels D-F in Figure
2. The effect of adaptation was simulated as a $10 \times$ change in the $g_1/g_2$ ratio.

FIG. 9. Learning effects. We divided each observer's dataset into 6 sequential epochs containing an
equal number of trials and computed the percentage of correct responses for each epoch. The
average across observers is shown in the large panel, individual observer data in the smaller panels.
Black data before adaptation, red data after adaptation. Shaded areas show ±1 s.e.m. Solid circles
show average across all epochs. Open circles show model performance. Dashed red lines show
where red data would fall after subtracting 6.6%. Error bars are smaller than symbols when not
visible. Dotted lines indicate 75% correct performance level.
Inseparable
Separable
First singular value (s1)
Second singular value (s2)
A Noise
After
Before
After−Before

s2/s1 Before or After

s2/s1 Before or After
Motion direction Before

\[ g_1/g_2 = 50 \]

Nonius +5 near

Motion direction After

\[ g_1/g_2 = 5 \]

Nonius +5 near

Motion direction After − Before

Disparity (arcmin)

\[ \frac{1}{25} - \frac{1}{5} \]

# noise dots

\[ \frac{1}{25} - \frac{1}{5} \]

A

B

Disparity (arcmin)

C

D

Output

\[ g_1 \]

\[ g_2 \]

Attentional depth filter

Directional filter

Joint motion–disparity filter

E

F

G

Motion direction

far −5 Nonius +5 near

Motion direction

Motion direction

Motion direction

Disparity (arcmin)