nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-019-0377-4

# Perceptual straightening of natural videos

Olivier J. Hénaff[1,6*], Robbe L. T. Goris[2] and Eero P. Simoncelli[1,3,4,5]

**Many behaviors rely on predictions derived from recent visual input, but the temporal evolution of those inputs is generally complex and difficult to extrapolate. We propose that the visual system transforms these inputs to follow straighter temporal trajectories. To test this 'temporal straightening' hypothesis, we develop a methodology for estimating the curvature of an internal trajectory from human perceptual judgments. We use this to test three distinct predictions: natural sequences that are highly curved in the space of pixel intensities should be substantially straighter perceptually; in contrast, artificial sequences that are straight in the intensity domain should be more curved perceptually; finally, naturalistic sequences that are straight in the intensity domain should be relatively less curved. Perceptual data validate all three predictions, as do population models of the early visual system, providing evidence that the visual system specifically straightens natural videos, offering a solution for tasks that rely on prediction.**

It is generally believed that sensory systems transform their inputs into internal representations that efficiently and effectively capture information needed for current and future tasks[1]. For example, the retina removes redundant spatiotemporal structure from incoming light, such that it may be efficiently transmitted through the optic nerve[2–7]. Cortical area V1 transforms the retinal representation by extracting frequently occurring features such as edges[8], in support of efficient coding and discrimination of natural images[9–11]. Higher level visual areas in the ventral stream further transform this representation by separating objects' identity from viewing conditions, simplifying the task of object recognition[12].

Yet most visual tasks require more than simply analyzing static images, as they rely on predictions about future outcomes given past observations. The patterns of light arriving at the retina evolve according to complex, nonlinear dynamics that are difficult to extrapolate. As a result, we propose that the brain transforms the incoming stream of visual input to make it more predictable. Specifically, we propose that the nonlinear spatial representations of naturally occurring visual input are structured to straighten their temporal trajectories, enabling their prediction through linear extrapolation.

Temporal prediction has long been exploited by the video engineering community in building compression systems[13]. In neuroscience, a number of authors have proposed that temporal prediction could serve as a universal principle for the evolution, development and learning of visual representations[14–17]. Straightening of natural videos is consistent with current descriptions of the neural basis of object recognition, which propose that the brain 'untangles' trajectories that evolve according to changing viewing conditions[18].

To test the temporal straightening hypothesis, we developed a procedure for estimating the curvature (conversely, straightness) of the human perceptual representation of a sequence of images. By comparing this value to the curvature calculated from the pixel intensities of the image sequence, we tested three distinct predictions of our hypothesis. First, natural sequences that are curved in the intensity domain should be straighter perceptually. On the other hand, unnatural sequences (those that are unlikely to occur in the real world) need not be straightened and could even exhibit increased perceptual curvature. Finally, synthetic sequences that contain naturalistic changes (for example, shifts in luminance or contrast) should be relatively straight. We show that human perceptual capabilities are consistent with all three predictions. In addition, we show that simple nonlinear population models of the early visual system partially account for these behaviors, while deep convolution networks optimized for object recognition do not.

## Results

**Estimating perceptual curvature.** We gathered video sequences whose duration was roughly matched to the interval between successive saccades and estimated their curvature in the pixel-intensity and perceptual domains. Each frame of such a sequence can be represented in either domain as a point in a high-dimensional space (Fig. 1a). A natural measure of curvature for a sequence of points in either domain is the (unsigned) angle between consecutive segments and we summarize the curvature of a sequence using the average of these angles over the full sequence. This measure, known as discrete curvature, has the desirable property that it does not depend on the overall scale or units of the representation. It is zero only for straight (linear) sequences and increases as they become more curved. Intuitively, discrete curvature quantifies the dissimilarity between successive difference vectors and thus, the difficulty in linearly extrapolating the trajectory.

The curvature of sequences in the intensity domain can be measured directly, by computing the differences between successive frames in the high-dimensional space of pixel intensities, and the angles between them (Fig. 1b, Methods). Curvature in the perceptual domain is estimated from the discriminability (or perceptual distance) of all pairs of frames in a sequence, as measured from human subjects. Intuitively, the more curved a sequence is perceptually, the more the perceptual distance between a pair of frames (for example the first and third frames in Fig. 1b) should fall short of the summed intermediate distances (connecting the first, second and third frames). We measure the discriminability of a pair of frames by presenting them, on a given trial, as part of a sequence of three images in which the second is equal to the first or the last (an 'AXB' paradigm; Fig. 2a). By asking the observer to report which image is the unique one and measuring their performance over many trials, we arrive at an estimate of the perceptual distance between these two frames. Having obtained in this manner the distances between

[1]Center for Neural Science, New York University, New York, NY, USA. [2]Center for Perceptual Systems, University of Texas at Austin, Austin, TX, USA. [3]Howard Hughes Medical Institute, New York University, New York, NY, USA. [4]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA. [5]Department of Psychology, New York University, New York, NY, USA. [6]Present address: DeepMind, London, UK. *e-mail: henaff@google.com

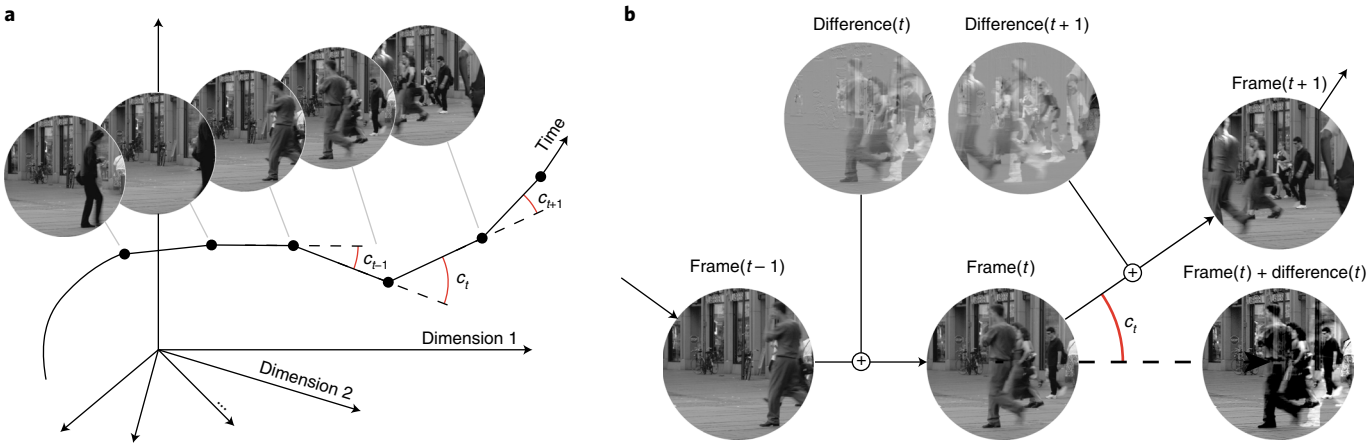


**Fig. 1 | Quantifying straightness of image sequences in the intensity and perceptual domains. a**, Visualization of a high-dimensional representation of a temporal sequence of images. We consider representations in two domains: the 'pixel-intensity' domain (axes correspond to pixel intensities in each frame) and the 'perceptual' domain (axes correspond to internal responses that underlie the perceptual judgments of human subjects). Each frame in the sequence corresponds to a point in the representational space. The discrete curvature at a given frame is equal to the angle between the segments connecting it to adjacent frames. We define the curvature of a sequence as the average of these angles. **b**, In the pixel-intensity domain, curvature can be calculated directly by computing the pixel-wise differences between successive frames and the angles between them. Note how this sequence of frames is curved in the intensity domain (difference images are dissimilar) but seems natural perceptually. In contrast, a linearly extrapolated frame in the intensity domain (bottom right) is perceptually unnatural.

all pairs of frames drawn from an 11-frame sequence (Fig. 2b), we search for a perceptual trajectory that accounts for these data, whose curvature we can then measure as in the intensity domain.

We start by proposing a candidate perceptual trajectory that might account for the pattern of discriminability we measured experimentally (chosen randomly; Fig. 2c, top left). Assuming (without loss of generality, see Methods) that our perceptual judgments are limited by additive Gaussian noise, the pairwise distances between points along the trajectory make a prediction regarding this pattern of discriminability[19] (Fig. 2c, top middle). This allows us to measure the discrepancy between the predicted and observed patterns of discriminability (Fig. 2c, top right). In this case, a straighter perceptual trajectory provides a better match to the data, suggesting that these data are more consistent with low perceptual curvature (Fig. 2c, bottom row). Given this, it is tempting to iteratively adjust this trajectory until arriving at the most likely one (similarly to nonlinear dimensionality reduction methods[20,21]) and reporting its curvature, computed as in the intensity domain. But this two-step method is plagued by estimation bias when used with the amounts of data available from our experiments (Supplementary Fig. 1a). As an alternative, we developed a data-efficient and nearly unbiased procedure for estimating the curvature that is most likely, by averaging over many plausible perceptual trajectories (Supplementary Fig. 1b, Methods). For visualization purposes, we show the perceptual trajectory whose length and curvature are equal to the average across plausible perceptual trajectories (Fig. 2c, bottom row).

**Perceptual straightening of natural videos.** The primary prediction of the temporal straightening hypothesis is that natural image sequences that are curved in the intensity domain should be less curved in the perceptual domain. We measured the intensity-domain and perceptual-domain curvatures of 12 natural image sequences which differed in content (experiment 1; see Fig. 3a for three frames from an example sequence, Supplementary Figs. 2 and 3 for all sequences). To visualize our analysis and gain an intuition for the results, we projected the intensity-domain and perceptual-domain representations for a single sequence and observer onto the first two principal components (Fig. 3b). The trajectories are strikingly different. Trajectories of the first two components of this natural image sequence are highly curved in the intensity domain (curvature = 39°). Consistent with our hypothesis, the same sequence appears much straighter perceptually (curvature = 4°). This difference in curvature is not simply a byproduct of dimensionality reduction: in the high-dimensional intensity and perceptual domains, the difference was even more substantial (intensity-domain curvature = 99°, perceptual-domain curvature = 8°). Moreover, this curvature reduction is robust across sequences and observers (Fig. 3c, blue histogram; median difference in curvature = −23°; $P < 0.001$, two-tailed Wilcoxon signed-rank test).

Since our curvature estimates are obtained through a novel analysis method, we wanted to verify the reliability of those estimates. To that end, we simulated data obtained from model observers who were identical to our human observers in their ability to discriminate successive frames, lapse rates and number and distribution of trials. Crucially, however, we designed these model observers to base their responses on a perceptual representation whose curvature was matched to the pixel-domain curvature (Methods). When applied to these synthetic data, our analysis method found no reduction in curvature (Fig. 3c, gray downward histogram; median difference in curvature = 4°; $P = 0.99$, one-tailed test). Our estimation method is thus not inherently biased towards curvature reduction. Moreover, this analysis reveals that the average reduction in curvature estimated for our human observers is significantly greater than the variability in the estimates ($P < 0.001$, two-tailed test on the difference between human observers and simulated controls). These data provide clear supporting evidence for the notion that the human visual system straightens natural videos.

Although we motivated our curvature measurements with their connection to the accuracy of first-order linear extrapolation, we wanted to know whether perceptual straightening could also improve the accuracy of higher-order predictors. To that effect, we measured the performance of optimal second, third and fourth-order predictors (that is, that predict the next frame from a linear combination of the three, four or five previous frames) which can account for homogeneous curvature (Methods). These higher-order predictors are more accurate than first-order ones, but nevertheless exhibit significantly better performance on straighter perceptual trajectories than their curved intensity-domain counterparts (Supplementary Fig. 4). Indeed, we found prediction errors to

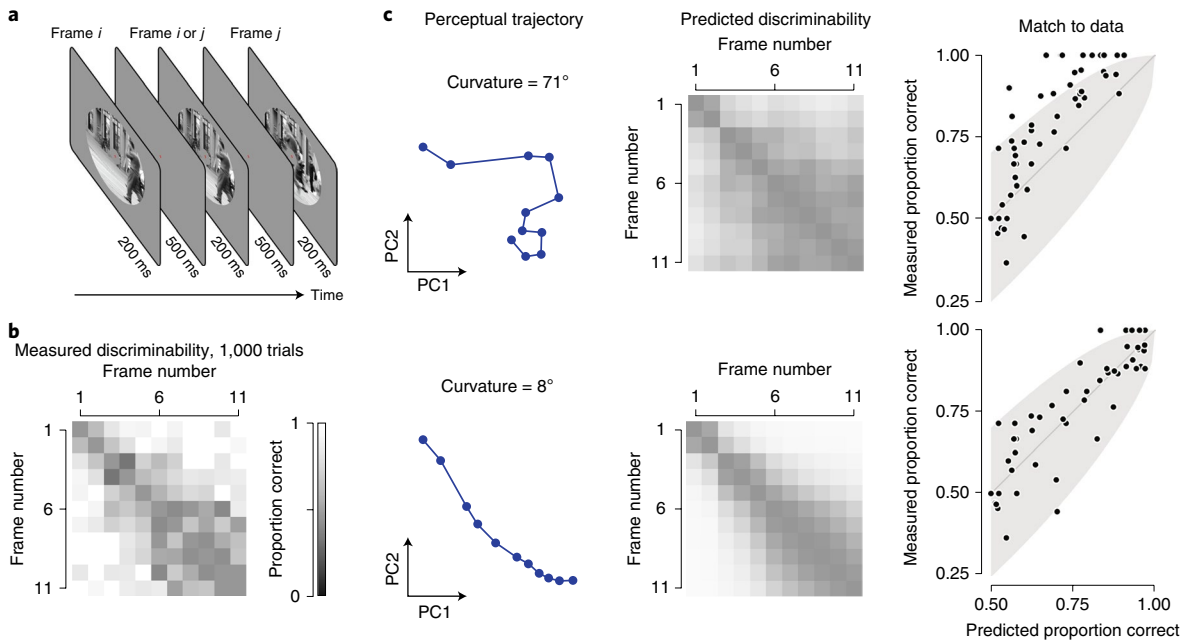


**Fig. 2 | Measuring perceptual straightness of image sequences. a**, Psychophysical AXB task. On each trial, observers viewed a sequence of three images. The first and the last are randomly selected frames from a given sequence; the middle one is identical to one of the other two. Observers indicated whether the first or the last image was the unique one. **b**, Performance of a single observer for all pairs of frames in a given sequence (total of 1,000 trials). Pixel brightness depicts proportion correct in discriminating the corresponding pair of frames (brighter indicates more discriminable). **c**, Inferring perceptual curvature from psychophysical data. Left: two-dimensional projections of ten-dimensional perceptual trajectories (*x*- and *y*-axes represent first and second principal components, respectively). Each point illustrates the centroid of a two-dimensional Gaussian distribution corresponding to the noisy perceptual representation of a frame in the sequence. Middle: pattern of performance on the AXB task predicted from the pairwise distances between points along the trajectory. Right: match between empirical and trajectory-predicted proportion correct (one point for each pair of frames), along with the 95% confidence interval expected from binomial variability (gray region). Top: a curved perceptual trajectory predicts a moderate increase in discriminability as frames are further separated in time, providing a poor match to the data. Bottom: a straighter perceptual trajectory predicts a faster increase in discriminability and provides a better match to the data.

be reduced relative to those in the intensity domain (Fig. 3d, blue histogram; median difference in third-order prediction error between perceptual domain and intensity domain = −21%, $P < 0.001$), as well as to those of simulated control observers with internal curvature matched to that of the intensity domain (Fig. 3d, gray histogram; median difference in third-order prediction error between humans and controls = −15%, $P < 0.001$).

**Perceptual distortion of artificial videos.** The straightening of curved natural videos exhibited by our subjects implies that their perceptual responses arise from a nontrivial transformation of their visual input. It does not by itself indicate that this transformation is specifically tailored for this purpose. It could be the case that most, or even all, sequences are straightened by the visual system, regardless of whether they could occur under natural conditions. But if, as we propose, temporal straightening targets sequences that occur naturally, then sequences that are unlikely to occur should not be straightened. On the contrary, these sequences are more likely to be distorted by the nonlinear hierarchical transformations of the visual system[22] and therefore exhibit increased perceptual curvature.

We tested this second prediction of the temporal straightening hypothesis by estimating the perceptual curvature of artificial image sequences that are strictly linear in the intensity domain (experiment 2). Specifically, we created synthetic sequences that fade from the initial to the final frame of each of the natural videos used in experiment 1. These sequences are straight (that is, they have zero curvature) in the intensity domain but unnatural in that, for example, the interpolated middle frames contain pixel-wise averages of two different images (see Fig. 4a for three frames from an example sequence, and Supplementary Figs. 2 and 3 for all sequences). We estimated the perceptual curvature of these sequences for the same observers that participated in experiment 1.

Consider the two-dimensional projections of the intensity- and perceptual-domain representations of a single observer (Fig. 4b). Consistent with our hypothesis, the perceptual-domain trajectory of this artificial sequence is much more curved than the intensity-domain trajectory (difference in curvature = 48°). This effect was just as prominent in the high-dimensional spaces (difference in curvature = 49°) and was consistent across all artificial sequences and observers (Fig. 4c, green histogram; median difference in curvature = 53°). Note that curvature is a positive-valued quantity and since the image-domain curvature of these sequences is zero, some increase in curvature is expected due to estimation error. To determine this baseline expectation, we simulated model observers that preserved image-domain curvature but were otherwise matched to our human observers. For these model observers, the median increase in curvature was 29° (Fig. 4c, gray histogram), significantly smaller than the 53° increase observed in our human subjects ($P < 0.001$).

Although this increase in curvature probably makes these sequences less predictable for a first-order linear extrapolator, they could remain predictable for a higher-order one. Hence, we measured the accuracy of second-, third- and fourth-order predictors, but nevertheless found them all to be reduced in the perceptual domain compared to the intensity domain (Supplementary Fig. 4; Fig. 4d, green histogram; median difference in third-order prediction error between perceptual and intensity domains = 66%, $P < 0.001$) and to simulated controls who did not increase the curvature of these sequences (Fig. 4d, gray histogram; median difference in third-order prediction error between humans and controls = 61%, $P < 0.001$).

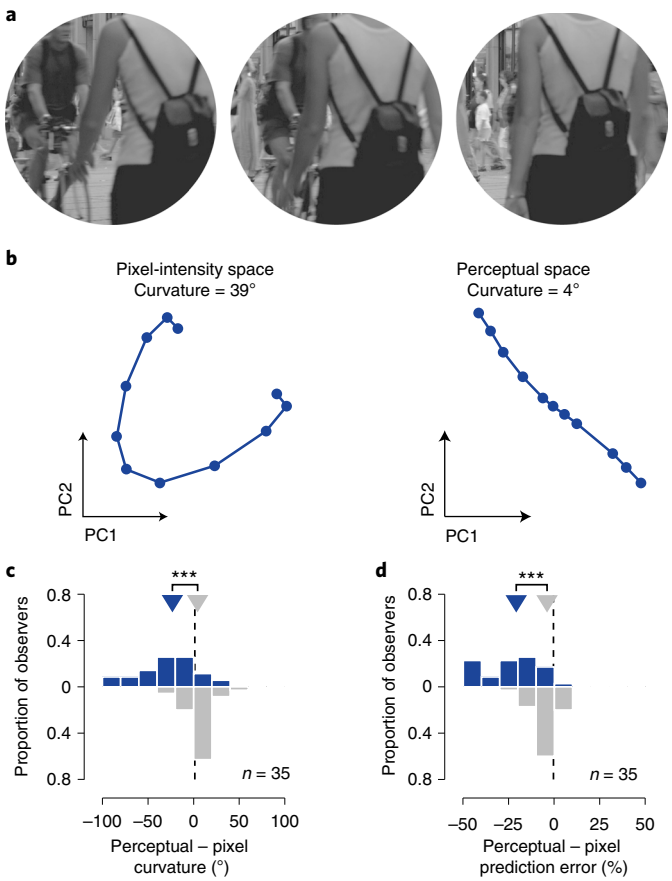


**Fig. 3 | Curvature reduction for natural image sequences. a**, Initial, middle and final frames of an example sequence (a person walking in front of a cyclist). **b**, Two-dimensional projections of an example sequence in the intensity domain (left) and in the inferred perceptual domain (right). Each point represents a frame. **c**, Difference in curvature between the intensity and perceptual domains, for 12 natural image sequences and 18 observers ($n = 35$ sequence–observer pairs total). Blue histogram, perceptual curvature estimated from human subject data (median = −23°, interquartile range (IQR) = 38°). Gray histogram, perceptual curvature estimated from data simulated from model (control) observers whose perceptual curvature is matched to the intensity-domain curvature, with all other parameters matched to those of the human observers (median = 4°, IQR = 12°). Triangles indicate the median of each distribution. **d**, Difference in third-order prediction error between the intensity and perceptual domains (human observers, median = −21%, IQR = 24%; simulated controls, median = −4%, IQR = 7%). Same layout as **c**. ***$P < 0.001$, two-tailed Wilcoxon signed-rank test.

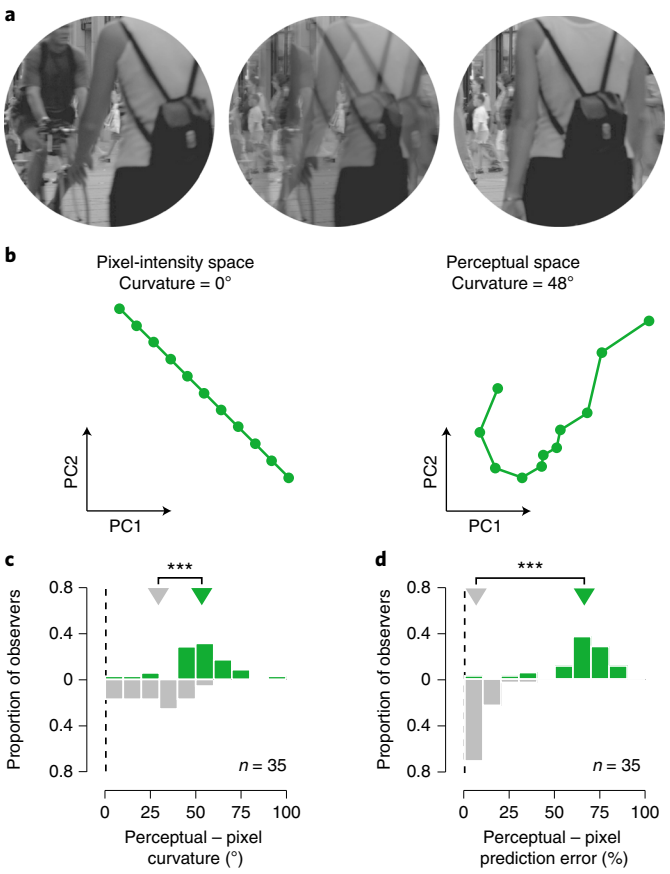


**Fig. 4 | Curvature increase for artificial image sequences. a**, Initial, middle and final frames of one such image sequence. Initial and final frames are identical to those of the corresponding natural sequence (Fig. 3a), whereas intermediate frames are generated by linearly interpolating (fading) between the initial and final frames. **b**, Low-dimensional projections of an example sequence in the intensity domain (left) and in the inferred perceptual domain (right). **c**, Difference in curvature between the intensity and perceptual domains, for 12 artificial image sequences and 18 observers ($n = 35$ sequence–observer pairs total). Green histogram, perceptual curvature estimated from human subject data (median = 53°, IQR = 14°). Gray histogram, perceptual curvature estimated from data simulated from model observers whose perceptual curvature is matched to the intensity-domain curvature (in this case, zero), with all other parameters matched to those of the human observers (median = 29°, IQR = 29°). Triangles indicate the median of each distribution. **d**, Difference in third-order prediction error between the intensity and perceptual domains (human observers, median = 66%, IQR = 14%; simulated controls, median = 6%, IQR = 9%). Same layout as **c**. ***$P < 0.001$, two-tailed Wilcoxon signed-rank test.

**Curvature preservation of naturalistic videos.** We interpreted the outcome of experiment 2 as evidence that the nonlinear computations underlying perceptual straightening target natural sequences and exhibit the opposite effect on straight artificial sequences. But it could be the case that all videos that are straight in the pixel-intensity domain yield highly curved perceptual-domain representations, regardless of whether they are natural or artificial.

To resolve this ambiguity, we characterized the perceptual-domain curvature of a new set of synthetic sequences, that are straight in the intensity domain but mimic natural transformations. Specifically, we constructed sequences by gradually and monotonically changing the contrast of the initial frame over time (Fig. 5a). These sequences have zero curvature in the intensity domain by construction and are more natural than the sequences of experiment 2 because they approximate changes in scene visibility (for example due to the onset of fog). The temporal straightening hypothesis predicts that these sequences should be much straighter than their unnatural counterparts. Figure 5b shows the low-dimensional projection of a perceptual-domain trajectory, for an example observer in experiment 3. In this case, the perceptual distortion of the contrast-varying sequence is minimal (difference in low-dimensional curvature = 9°; difference in high-dimensional curvature = 18°). We found this result to be consistent across all sequences and observers. Specifically, although these sequences evoked a significant increase in curvature (median difference between human observers and simulated controls = 12°, $P = 0.009$), this increase was significantly smaller than that of the artificial sequences in experiment 2 (Fig. 5c; difference in median curvature = 22°, $P = 0.003$). Similarly, these sequences were significantly

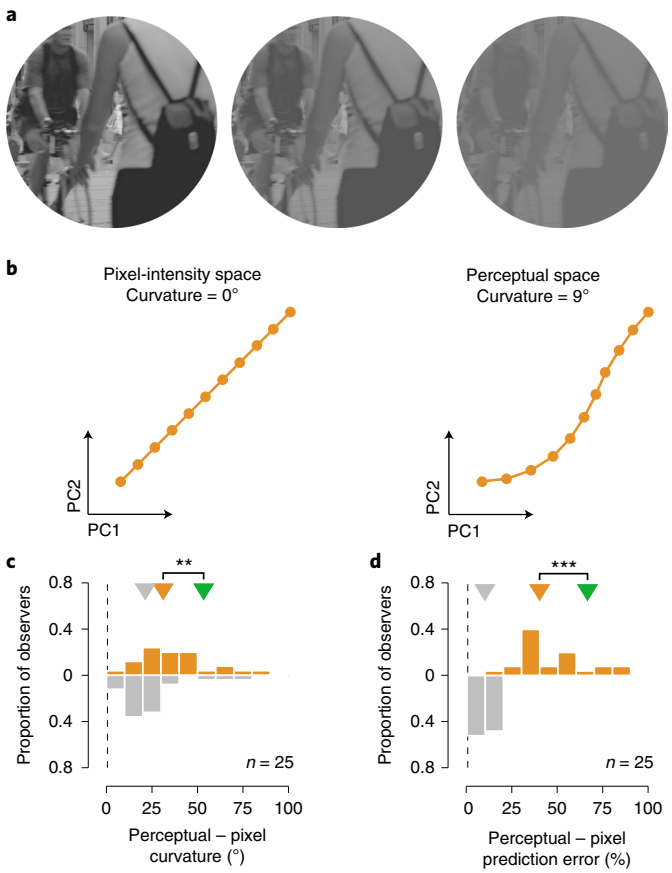


**Fig. 5 | Curvature conservation for naturalistic, intensity-linear image sequences. a**, Initial, middle and final frames of one such sequence. Initial frame is identical to those in Figs. 3 and 4, and the rest are generated by gradually reducing contrast. **b**, Low-dimensional projections of an example sequence in the intensity domain (left) and in the inferred perceptual domain (right). **c**, Difference in curvature between the intensity and perceptual domains, for nine natural, intensity-linear image sequences and nine observers ($n = 25$ sequence–observer pairs total). Yellow histogram, perceptual curvature estimated from human subject data (median = 31°, IQR = 26°). Gray histogram, perceptual curvature estimated from data simulated from model observers whose perceptual curvature is matched to the intensity-domain curvature (in this case, zero), with all other parameters matched to those of the human observers (median = 21°, IQR = 14°). Triangles indicate the median of each distribution. Green triangle is copied from Fig. 4c, showing that naturalistic sequences are significantly less curved than their artificial counterparts. $^{**}P = 0.003$, two-tailed Mann–Whitney *U*-test. **d**, Difference in third-order prediction error between the intensity and perceptual domains (human observers, median = 40%, IQR = 29%; simulated controls, median = 10%, IQR = 8%). Same layout as **c**. $^{***}P < 0.001$, two-tailed Mann–Whitney *U*-test.

more predictable than their unnatural counterparts (Fig. 5d; difference in median third-order prediction error = 26%, $P < 0.001$). Hence, consistent with the temporal straightening hypothesis, we conclude that the human visual system is able to largely preserve the linearity and predictability of straight, naturalistic videos.

**Computational basis of perceptual straightening.** Finally, we asked how perceptual straightening could arise from the underlying neural activity of the visual system. In particular, if straightening is a fundamental goal of visual processing, we might expect that each successive transformation throughout the visual hierarchy could serve to further reduce the curvature of natural videos. To probe this hypothesis, we examined responses of a two-stage model that mimics the nonlinear functional properties of the early visual system (Fig. 6a). The first stage is comprised of center-surround filtering followed by local luminance and contrast gain control operations, capturing the primary nonlinear transformations performed by the retina and lateral geniculate nucleus[23,24]. The second stage further transforms this representation using a set of oriented filters whose responses are squared and combined over phase, capturing the nonlinear behavior of complex cells in primary visual cortex (area V1)[25]. We constructed response trajectories by applying the model to each video frame independently and evaluated the curvature of these trajectories directly. Both model stages induced systematic changes in curvature that were consistent with the changes we measured perceptually. Specifically, both stages straightened natural image sequences, distorted unnatural ones and preserved the linearity of naturalistic 'contrast' sequences (Fig. 6b). The first stage of the model likely straightens natural videos by providing robustness to local fluctuations in luminance and contrast[23,26], whereas the second provides further straightening through the position- and phase-invariance properties of the 'energy model' for complex cells[25,27]. The straightening induced by both stages is substantial, although still less than the perceptual straightening observed in the human subjects.

We also tested the straightening capabilities of artificial neural networks constructed from many stages of rectified linear filters. Such models have shown impressive capabilities when optimized for object recognition[28] and have been proposed as candidate models of biological vision[29–31]. We hypothesized that the ability of these networks to 'untangle' image manifolds associated with object categories[32] might also extend to straightening of natural videos. To test this, we evaluated the changes in curvature induced by each stage of the AlexNet architecture[33] (Fig. 6c). Unlike the simple biological models and our human subjects however, we found that this model did not straighten any natural videos (Fig. 6d). We tested several other current deep neural network architectures used for image classification[34–37] and found that all of them increased the curvature of natural videos (Supplementary Fig. 5). In principle, we would expect these networks to be capable of approximating the nonlinear transformations of the two-stage biological model (local gain control and energy), which exhibit substantial straightening. We thus conclude that optimizing such networks for static object recognition fails to endow them with the nonlinear temporal straightening capabilities found in the human visual system.

## Discussion

We have introduced the temporal straightening hypothesis, which provides a normative explanation for the structure of sensory representations. We developed a methodology for estimating perceptual curvature and provided behavioral evidence for three distinct predictions of the hypothesis. Our results demonstrate that the visual system nonlinearly transforms its inputs such that naturally occurring temporal image transformations give rise to straighter trajectories in perceptual space than in the input space. We also find that synthetic, behaviorally irrelevant sequences that are straight in the intensity domain are distorted by the visual system, breaking their perceptual contiguity. Nonetheless, we found that the visual system is able to largely preserve the linearity of naturalistic, intensity-linear sequences.

To design an experimental test of temporal straightening, we assumed a restricted form of the hypothesis in which linear predictability over time is achieved through nonlinear spatial processing of visual input. Moreover, by measuring curvature between successive frames in a sequence, we have restricted our tests of straightening to a specific timescale, corresponding to the interval between frames. Despite these restrictions, we have found human perceptual capabilities (and simple models of the early visual system) to behave as predicted. How would these results generalize to the straightening

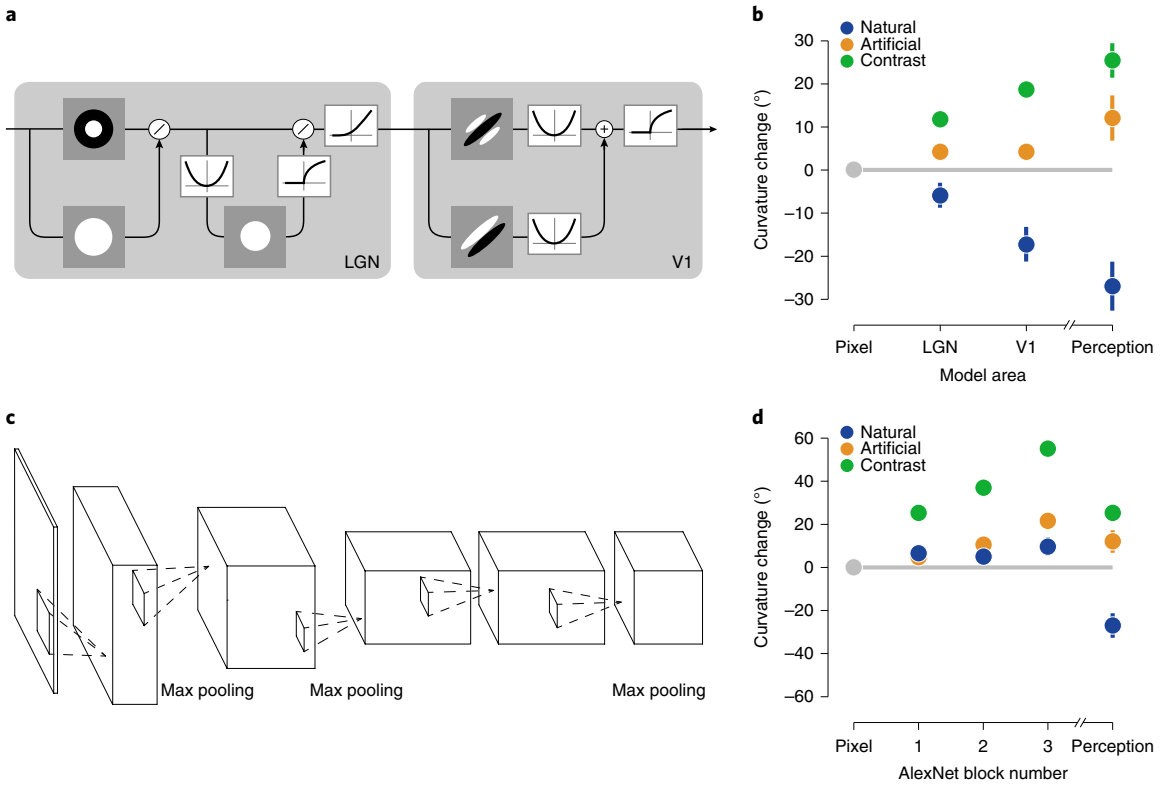


**Fig. 6 | Changes in curvature induced by models of the visual system. a**, Two-stage cascade model describing computations found in the retina, lateral geniculate nucleus (LGN) and V1. The first stage performs bandpass filtering (gray boxes contain icons representing spatial filters), followed by luminance and contrast gain control. The second stage decomposes the output of the previous stage with an oriented, multiscale linear transform and measures the local energy in each of its sub-bands (only one sub-band shown). **b**, Change in curvature induced by these computations for natural, artificial and contrast sequences. Each stage in the model incrementally contributes to the changes in the curvature found perceptually (circles indicate the median across sequences, error bars show the 68% confidence interval, $n=12$ sequences for the natural and artificial stimuli, $n=9$ sequences for the naturalistic 'contrast' stimuli). We report the perceptual data as the median difference in curvature between human observers and simulated controls, to correct for any estimation bias. **c**, Multistage neural network (AlexNet) trained for object recognition. **d**, Change in curvature induced by this network, for the same sequences (circles indicate the median across sequences, error bars where visible show the 68% confidence interval; $n=12$ sequences for the natural and artificial stimuli, $n=9$ sequences for the naturalistic 'contrast' stimuli). Despite strong performance on object classification, this model does not straighten natural sequences.

of continuous streams of images and over different timescales? We used sequences with sampling rates roughly matched to the integration times of photoreceptors (30 frames per second or less), whose response would probably be similar to those under static presentation. However, we might expect downstream areas that process visual input over longer timescales to respond differently (for example, direction-selective neurons in areas V1 or MT). A more general psychophysical protocol, in which perceptual representations are evaluated within their recent temporal context, seems necessary to test straightening at these longer timescales.

Our temporal straightening hypothesis provides a specific instance, as well as an augmentation, of the efficient coding hypothesis—one of the most widely discussed and successful theories of early sensory processing[1]. Efficient coding posits that sensory representations are structured to preserve information in natural signals, while reducing redundancy and minimizing the use of neural resources (for example, cells and spikes), a goal that is especially relevant for early sensory areas that are separated from cortex by a communication bottleneck. Temporal straightening (and more generally, prediction) offers a specific form of coding efficiency, given that predictable signals can be coded via small residual errors[13]. Beyond this bottleneck, however, coding efficiency may no longer suffice to fully explain the form or specifics of sensory processes[38–41]. Rather, as sensory information propagates through the brain, it is combined with experience (memory), goals, desires and other internal states that govern behavioral relevance and probably play an important role in specifying which information is processed and which is discarded. Temporal prediction offers a potential unification, by augmenting coding efficiency with a universal goal that is essential for a large class of behaviorally relevant tasks[14,17,42].

Temporal straightening offers a simple and readily testable instantiation of the temporal prediction hypothesis: assuming a first-order extrapolation model, the curvature of a sequence is equivalent to its predictability. Although we found that straightening improved the accuracy of a number of higher-order predictors as well, we do not know whether human observers base their predictions on a model from this class. Therefore, a new psychophysical paradigm that measures the ability of human observers to predict future frames of a sequence seems essential to characterize the mechanism by which we extrapolate observations.

Temporal straightening also bears similarity to the 'untangling hypothesis' that has been proposed as a normative explanation for the visual representations underlying object recognition capabilities[18]. Specifically, the fundamental difficulty of object recognition lies in constructing representations that vary substantially across object categories, while being unaffected by the substantial variability in visual appearance that arises from changes in viewing conditions and configuration. This hypothesis posits that the goal of the visual system is to produce a representation of objects that can be linearly decoded, which requires that variation due to viewing

conditions be confined to a low-dimensional subspace. The straightening hypothesis is more restrictive in that it seeks to contain the representation of individual videos in one-dimensional subspaces but also more general in that it does not rely on the definition or categorization of objects. It could therefore provide a practical means of learning such an untangled representation in an unsupervised manner, one of the most important open problems in machine learning[28]. Indeed, if the brain were to learn to straighten image sequences that evolve according to changes in viewpoint or lighting (as is the case for most natural sequences), the resulting representation would restrict these ephemeral fluctuations to a low-dimensional subspace, while preserving object-persistent information. Moreover, this objective could enable the untangling of more complex naturally occurring image variations, such as the motion of articulated or flexible objects and materials.

Our hypothesis is defined with regard to an unspecified internal perceptual representation, which presumably corresponds to the activity of some collection of neurons within the visual system. Although the perceptual measurements we report offer no direct indication as to where these neurons reside, our computational modeling suggests that straightening might emerge through the incremental transformations achieved by successive stages of visual processing, in line with current descriptions of the emergence of feature and object selectivity in the ventral stream[29,43,44]. The curvature estimation methodology we have developed is agnostic to the particular form of experimental measurement and we have begun to explore its application to physiological data (spiking responses recorded with multi-electrodes[45]) to directly evaluate the curvature of neural population representations in different visual areas.

Our findings also provide a new means of evaluating the adequacy of models of biological visual systems. A number of recent studies have examined the appropriateness of learned artificial neural network representations as models for biological perception[24,29,30]. Since they fail to straighten the timecourse of natural videos, these models cannot provide a complete account of biological vision. Our model of early visual processing is able to account for a significant portion of the straightening properties found in humans and it could be that downstream computations could also be identified by their effect on curvature. Going further, these models can be more stringently tested by asking which sequences are the straightest in their representational space (in technical terms, the 'geodesics' of the representation). We have developed a computational method to generate such geodesic sequences[46] and we have used the perceptual straightness of these sequences as a measure for comparing candidate models of the human visual system[47].

Although we have stated and tested the straightening hypothesis in terms of fixed response properties of the visual system, we can view it more generally as a means of adapting sensory representations to the properties of natural temporal inputs. This suggests, for example, that temporal straightening might play a role in perceptual learning. If so, it should be possible to induce perceptual straightening of arbitrary sequences through repeated or prolonged exposure. There is already some evidence in support of this: a series of studies showed that consecutive presentation of pairs of images at different positions[48] or scales[49] can change the invariance properties of single cells in visual area IT, as well as the robustness of perceptual discriminability in human observers[50]. But a more direct test of this idea, over more general input sequences, is warranted.

Finally, even if the straightening hypothesis proves consistent with physiological measurements, this does not answer the question of whether it is sufficiently powerful to serve as an objective for structuring representations throughout the visual system. To test this, one would need to simulate a system that learns to temporally straighten the content of natural videos and examine its similarity to biological systems. Our understanding of vision, whether biological or machine-based, has progressed furthest with regard to the simple representations that occur in early stages of hierarchical processing. In these, principles of coding efficiency have been useful, both in testing for the efficiency of visual representations and in showing that they can be learned by maximizing efficiency in the representation of naturally occurring stimuli[1,38,39]. Temporal straightening, as a task-relevant generalization of efficient coding, holds promise to fulfill an analogous role in downstream visual areas.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41593-019-0377-4.

Received: 15 December 2017; Accepted: 7 March 2019;
Published online: 29 April 2019

## References

1. Barlow, H. B. Possible principles underlying the transformation of sensory messages. *Sensory Communication* (ed. Rosenblith, W.) 217–234 (M.I.T. Press, 1961).
2. Atick, J. J. & Redlich, A. N. Towards a theory of early visual processing. *Neural Comput.* **320**, 1–13 (1990).
3. van Hateren, J. H. A theory of maximizing sensory information. *Biol. Cybern.* **68**, 23–29 (1992).
4. Meister, M., Lagnado, L. & Baylor, D. A. Concerted signaling by retinal ganglion cells. *Science* **270**, 1207–1210 (1995).
5. Balasubramanian, V. & Berry, M. J. A test of metabolically efficient coding in the retina. *Network* **13**, 531–552 (2002).
6. Puchalla, J. L., Schneidman, E., Harris, R. A. & Berry, M. J. Redundancy in the population code of the retina. *Neuron* **46**, 493–504 (2005).
7. Doi, E. et al. Efficient coding of spatial information in the primate retina. *J. Neurosci.* **32**, 16256–16264 (2012).
8. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154 (1962).
9. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
10. Bell, A. J. & Sejnowski, T. J. The 'independent components' of natural scenes are edge filters. *Vision Res.* **37**, 3327–3338 (1997).
11. Goris, R. L. T., Simoncelli, E. P. & Movshon, J. A. Origin and function of tuning diversity in macaque visual cortex. *Neuron* **88**, 819–831 (2015).
12. Rust, N. C. & DiCarlo, J. J. Selectivity and tolerance ('invariance') both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci.* **30**, 12978–12995 (2010).
13. Le Gall, D. MPEG: a video compression standard for multimedia applications. *Commun. ACM* **34**, 46–58 (1991).
14. Tishby, N., Pereira, F. C. & Bialek, W. The information bottleneck method. *In Proc. Allerton Conference on Communication, Control and Computing* **37**, 368–377 (1999).
15. Wiskott, L. & Sejnowski, T. J. Slow feature analysis: unsupervised learning of invariances. *Neural Comput.* **14**, 715–70 (2002).
16. Richthofer, S. & Wiskott, L. Predictable feature analysis. In *Proceedings IEEE 1fourth International Conference on Machine Learning and Applications* (2016).
17. Palmer, S. E., Marre, O., Berry, M. J. & Bialek, W. Predictive information in a sensory population. *Proc. Natl Acad. Sci. USA* **112**, 6908–13 (2015).
18. DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends Cogn. Sci.* **11**, 333–341 (2007).
19. Noreen, D. L. Optimal decision rules for some common psychophysical paradigms. *Proc. of the Symposium in Applied Mathematics of the American Mathematical Society and the Society for Industrial and Applied Mathematics* **13**, 237–279 (1981).
20. Tenenbaum, J. B., De Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–23 (2000).
21. Roweis, S. T. & Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–6 (2000).
22. Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J. & Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. *Advances in Neural Information Processing Systems* **29**, 3360–3368 (2016).
23. Mante, V., Bonin, V. & Carandini, M. Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli. *Neuron* **58**, 625–638 (2008).
24. Berardino, A., Ballé, J., Laparra, V. & Simoncelli, E. P. Eigen-distortions of hierarchical representations. *Advances in Neural Information Processing Systems* **30**, 3530–3539 (2017).

25. Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2**, 284 (1985).
26. Carandini, M. & Heeger, D. J. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2012).
27. Mallat, S. Group invariant scattering.*Commun. Pur. Appl. Math.* **65**, 1331–1398 (2012).
28. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
29. Yamins, D. L. K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
30. Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
31. Tacchetti, A., Isik, L. & Poggio, T. Invariant recognition drives neural representations of action sequences. *PLoS Comput. Biol.* **13**, e1005859 (2017).
32. Hong, H., Yamins, D. L. K., Majaj, N. J. & Dicarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19**, 613–22 (2016).
33. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* **25**, 1–9 (2012).
34. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *In Proc. International Conference on Learning Representations* **3**, 1–14 (2015).
35. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *In Proc. International Conference on Machine Learning* **7**, 1–9 (2015).
36. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *In Proc. Conference on Computer Vision and Pattern Recognition* **29**, 770–778 (2016).
37. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. *In Proc. Conference on Computer Vision and Pattern Recognition* **30**, 2261–2269 (2017).
38. Simoncelli, E. P. & Olshausen, B. A. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).
39. Barlow, H. Redundancy reduction revisited. *Network* **12**, 241–253 (2001).
40. Machens, C. K., Gollisch, T., Kolesnikova, O. & Herz, A. V. M. Testing the efficiency of sensory coding with optimal stimulus ensembles. *Neuron* **47**, 447–456 (2005).
41. Geisler, W. S. Visual perception and the statistical properties of natural scenes. *Annu. Rev. Psychol.* **59**, 167–192 (2008).
42. Bialek, W., De Ruyter Van Steveninck, R. R. & Tishby, N. Efficient representation as a design principle for neural coding and computation. In *Proc. International Symposium on Information Theory*, 659–663 (2006).
43. Fukushima, K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernet.* **36**, 193–202 (1980).
44. Serre, T., Oliva, A. & Poggio, T. A feedforward architecture accounts for rapid categorization. *Proc. Natl Acad. Sci. USA* **104**, 6424–6429 (2007).
45. Bai, Y., et al. Neural straightening of natural videos in macaque primary visual cortex. *Soc. Neurosci. Abstr.* 485.07 (2018).
46. Hénaff, O. J. & Simoncelli, E. P. Geodesics of learned representations. *In Proc. International Conferenceon Learning Representations* **4**, 1–10 (2016).
47. Hénaff, O.J., Goris, R.L.T. & Simoncelli, O.J. Perceptual evaluation of artificial visual recognition systems using geodesics. *Cosyne Abstr.* II-72 (2016).
48. Li, N. & DiCarlo, J. J. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science* **321**, 1502–1507 (2008).
49. Li, N. & DiCarlo, J. J. Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron* **67**, 1062–1075 (2010).
50. Cox, D. D., Meier, P., Oertelt, N. & DiCarlo, J. J. 'Breaking' position-invariant object recognition. *Nat. Neurosci.* **8**, 1145–1147 (2005).


## Acknowledgements

We thank S. Palmer and J. Salisbury for making available the video sequences in their Chicago Motion Database. We are also grateful to Y. Bai for helpful comments on the manuscript. This work was supported by the Howard Hughes Medical Institute (O.J.H, R.L.T.G., E.P.S).

## Author contributions

O.J.H., R.L.T.G. and E.P.S. conceived the project and designed the experiments. O.J.H. designed the analysis and performed the experiments. O.J.H, R.L.T.G. and E.P.S. wrote the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41593-019-0377-4.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Correspondence and requests for materials** should be addressed to O.J.H.

**Journal peer review information:** *Nature Neuroscience* thanks Konrad Kording and other anonymous reviewer(s) for their contribution to the peer review of this work.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Methods

**Stimuli.** We measured the perceptual curvature of 12 natural image sequences that are representative of the diversity found in real videos (experiment 1; Supplementary Figs. 2 and 3, blue path). To constrain our choice of sequences, we established a list of attributes that distinguish natural videos. These pertain to the content of the videos (discrete, isolated objects versus dense textures) as well as the types of motion and transformations over time (camera motion, rigid object motion and flexible/articulated object motion). Supplementary Table 1 indicates the diversity of the chosen set.

We obtained eight of these ('water', 'prairie', 'egomotion', 'ice', 'bees', 'carnegie-dam', 'leaves-wind', 'chironomus') from the Chicago Motion Database (https://cmd.rcc.uchicago.edu), one from a feature film ('Dogville', Lions Gate Entertainment, 2003) and two from the LIVE Video Quality Database[51,52] ('smile' and 'walking'). The last ('boats') was generated by translating a single image over time. For most sequences, we used consecutive frames at the sampling rate of the original videos (30 frames per second). However, because curvature can only be resolved when successive frames are sufficiently discriminable, we temporally down-sampled videos with little variation. As a result, each 11-frame sequence lasted anywhere from 92 ms to 1,650 ms in real time (on average, about 300 ms) but each contained roughly the same average change (measured perceptually) from one frame to the next. All video frames had a spatial resolution of 512 × 512 pixels.

For each sequence, we collected data from three to four observers (41 sequence–observer pairs in total). Because we required perceptual trajectories of sufficient length (when measured in terms of $d'$) for curvature estimation, we excluded data from observers with unusually low average discriminability (specifically, those whose proportion of correct answers did not exceed 0.7), leaving two to four observers per sequence and 35 sequence–observer pairs.

Each observer was also shown an artificial sequence that faded linearly between the first and the last frame of the corresponding natural sequence (experiment 2; Supplementary Figs. 2 and 3, green path). Finally, nine of these observers also viewed nine natural, intensity-linear sequences that were generated by manipulating the contrast of a single frame (experiment 3; 25 sequence–observer pairs in total). Four such sequences varied the contrast of the first frame of the 'water', 'walking', 'bees' and 'boats' sequences, respectively, from 50% to 100% in linear steps. The five others varied the contrast of the first frame of the 'prairie', 'walking', 'smile', 'bees' and 'egomotion' sequences, respectively, from 10% to 100% in logarithmic steps.

**Experimental paradigm.** We tested 18 observers (6 females, 12 males; ages 19–30 yr) with normal or corrected-to-normal vision. Protocols for selection of observers and experimental procedures were approved by the human subjects committee of New York University and all subjects signed an approved consent form. One observer was an author (O.J.H.); all others were naive as to the purposes of the experiments.

Each trial followed an AXB paradigm. Three images were shown in sequence, the first and the last being a randomly chosen pair of frames from the sequence, the middle one being identical to one of the other two. Observers were asked to indicate with a button-press whether the first or last image was unique and were given feedback after each trial. Sequences obtained from the Chicago Motion Database were luminance calibrated and we showed their intensities on a calibrated display. For the others, intensities were transformed with pixel-wise power (gamma) function, whose exponent we obtained from the camera meta-data (ICC profile), to approximate the original luminance. Images were presented for 200 ms, with a 500 ms inter-stimulus interval, in an annulus whose inner and outer radii were equal to 2 and 12 degrees, respectively. Subjects were instructed to fixate on a small cross in the center of the annulus. Their eye position was monitored with an EyeLink 2000 eye-tracker and a warning signal indicated when their gaze deviated from the cross by more than 1.5 degrees. Trials for which eye position deviated by more than 2 degrees were discarded. Trials were grouped into blocks of 40, in which observers were presented with images from a single sequence. Blocks presenting natural and artificial stimuli were interleaved within a session, alleviating the need for separate experimental groups and blind data collection. Each observer performed 1,000 trials for each sequence on which they were tested, resulting in an average of 18 trials for each pairwise comparison.

**Curvature: definition and notation.** The curvature at a given node of a sequence is defined as the angle between the segments connecting it to adjacent nodes and can be computed directly when we have access to the high-dimensional locations of each node. If $x = \{x_t\}_{t=0,\dots,T}$ are such locations (for example a sequence of vectors, each containing the pixel luminances of a frame of a video or model responses to a frame; $T = 10$ in our experiments), we can define a sequence of normalized displacement vectors $\{\hat{v}_t\}_{t=1,\dots,T}$:

$$v_t = x_t - x_{t-1}$$
$$\hat{v}_t = \frac{v_t}{\|v_t\|}$$

and the curvature at a given node $t$ is simply the angle between two such vectors, which can be computed from their dot product:

$$c_t = \arccos(\hat{v}_t \cdot \hat{v}_{t+1})$$

The global curvature of the sequence $\hat{c}$ is the average of the local curvatures over time, in degrees.

**Modeling: observer model.** We wish to infer the perceptual curvature of a sequence from the discriminability of pairs of frames. To this end, we formulate an observer model that assigns a location in a $D$-dimensional perceptual space to every frame in the sequence and explains the observed discriminability as arising from distances in that space. Since $T$ dimensions are sufficient to render any pattern of pairwise distances amongst $T+1$ points, we can choose $D = T = 10$ without loss of generality. Let $x = \{x_t\}_{t=0,\dots,T}$ be the perceptual locations associated with the frames in a video, for a given subject. For a given pair of frames $(i,j)$ we describe the subject's $n_{ij}$ correct and $m_{ij}$ incorrect responses in terms of the probability of being correct $p_{ij}$ with a binomial likelihood

$$\mathbb{P}(n_{ij}, m_{ij} | p_{ij}) = \binom{n_{ij}+m_{ij}}{n_{ij}} p_{ij}^{n_{ij}} (1-p_{ij})^{m_{ij}}$$

The probability of a correct response $p_{ij}$ is a linear combination of the probability of successfully performing the AXB task $p_{ij}^{\mathrm{AXB}}$ and the probability of successfully guessing the correct answer ($\frac{1}{2}$) weighted by the probability of guessing $p^{\mathrm{G}} = 2\lambda$, where $\lambda$ is known as the lapse rate[53]:

$$\begin{aligned} p_{ij} &= (1-p^{\mathrm{G}}) p_{ij}^{\mathrm{AXB}} + \frac{p^{\mathrm{G}}}{2} \\ &= (1-2\lambda) p_{ij}^{\mathrm{AXB}} + \lambda \end{aligned}$$

The discriminability of two frames does not determine their relative locations in perceptual space or the shape of their associated noise distributions. In particular, we are free to choose noise distributions for each frame and have them determine their relative locations. To apply the same definition of curvature we use in the intensity domain (which assumes Euclidean geometry in the dot product), we assume that task performance is limited by additive Gaussian noise. The probability of successfully performing the task may then be expressed using standard methods from signal detection theory[19]:

$$p_{ij}^{\mathrm{AXB}} = \Phi\left(\frac{d_{ij}}{\sqrt{2}}\right)\Phi\left(\frac{d_{ij}}{2}\right) + \Phi\left(-\frac{d_{ij}}{\sqrt{2}}\right)\Phi\left(-\frac{d_{ij}}{2}\right)$$

where $d_{ij} = \|x_i - x_j\|$ is the Euclidean distance between the perceptual locations of frames $i$ and $j$ and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal.

**Modeling: perceptual curvature estimation.** Intuitively, a natural procedure for estimating perceptual curvature consists of maximizing the likelihood of the perceptual locations $\{x_t\}_{t=0,\dots,T}$ of each frame given the entire dataset of correct and incorrect responses of an observer $(n, m) = \{n_{ij}, m_{ij}\}_{0 \le i,j \le T}$ and then computing the curvature of this trajectory. Although this is accurate in the limit of large amounts of data, for our experimental data (1,000 trials per sequence and per observer) it is prone to substantial biases, consistently preferring curvature values that are closer to 90° (the most probable configuration of random vectors in a high-dimensional space; Supplementary Fig. 1a). Instead, we perform a direct maximum-likelihood estimate of the curvature, by parameterizing the trajectory in terms of its curvature and marginalizing over the perceptual locations of individual frames. Although more complex, this procedure is substantially more robust than the greedy two-step process described above and is nearly unbiased (Supplementary Fig. 1b).

To develop this estimation method, we need to parameterize the trajectory in terms of its local (and global) curvature. As for pixel-domain curvature, we first express the frame vectors in terms of displacement vectors $\{v_t\}_{t=1,\dots,T}$, which are factored into distances and normalized displacements:

$$x_t = x_{t-1} + v_t$$
$$v_t = d_t \hat{v}_t$$

Since our objective is invariant to global translation, we choose $x_0 = 0$.

Next, we define the normalized displacement recursively as a function of the curvature at each node $c_t$ and the direction of curvature $\hat{a}_t$:

$$\hat{v}_t = \cos(c_t)\hat{v}_{t-1} + \sin(c_t)\hat{a}_t$$

where $\hat{a}_t$ is a unit vector orthogonal to the previous displacement vector $\hat{v}_{t-1}$, thereby ensuring that the curvature at node $t$ is equal to $c_t$. Since the objective is invariant to rotations, we choose $\hat{v}_1$ to lie in the direction of the first coordinate axis.

This polar parameterization can express the same set of trajectories as the initial Cartesian parameterization but allows us to directly estimate the global curvature while marginalizing over local variables. Specifically, we define a prior

probability over local curvatures that is Gaussian and centered around the global curvature $c^*$. Given a set of local curvatures, the optimal estimate of the global curvature is simply the average local curvature, consistent with our previous definition. We also define similar priors over local distances, directions and the lapse rate, by introducing a set of Gaussian-distributed auxiliary variables that are mapped through nonlinear functions:

$$\begin{aligned} d_t &= f_d(z_t^d) & z_t^d &\sim \mathcal{N}\left(f_d^{-1}(d^*), \sigma_d^2\right) \\ c_t &= z_t^c & z_t^c &\sim \mathcal{N}(c^*,\ \sigma_c^2) \\ \hat{a}_t &= f_a(z_t^a) & z_t^a &\sim \mathcal{N}(0,\ \Sigma_a) \\ \lambda &= f_\lambda(z^\lambda) & z^\lambda &\sim \mathcal{N}(0,\ 1) \end{aligned}$$

where $f_d$ is a smooth rectifying function, $f_a$ ensures that $\hat{a}_t$ is of unit length and orthogonal to $v_{t-1}$ and $f_\lambda(z) = \lambda_{\max}\Phi(z)$ effectively places a uniform prior on the lapse rate (we choose $\lambda_{\max} = 0.06$ as in ref. [53]). Here $\Sigma_a$ is diagonal and controls the effective dimensionality and aspect-ratio of the trajectory.

Define $\theta = \{d^*, c^*, \sigma_d, \sigma_c, \Sigma_a\}$ as the set of parameters governing random variables $z = \{z_t^d, z_t^c, z_t^a, z\}$. Direct curvature estimation amounts to maximizing the likelihood of these parameters to best account for the data, a form of empirical Bayes estimation. Computing the (log) likelihood of these parameters requires marginalizing over local variables, a high-dimensional integral that is intractable in practice:

$$\log p_\theta(n, m) = \log \int p(n, m|z)\ p_\theta(z)$$

Fortunately, variational methods provide a tractable lower bound on the likelihood[54]. Given an approximate Gaussian posterior $q_\phi(z|n, m)$, we replace the intractable integral with an analytical one:

$$\begin{aligned} \log p_\theta(n,m) &= \log \int \frac{q_\phi(z \mid n,m)}{q_\phi(z \mid n,m)} p(n,m|z)\, p_\theta(z)\, dz \\ &= \log \mathbb{E}_{q_\phi(z|n,m)}\left[\frac{p(n,m \mid z)\, p_\theta(z)}{q_\phi(z \mid n,m)}\right] \\ &\geq \mathbb{E}_{q_\phi(z|n,m)}\left[\log \frac{p(n,m \mid z)\, p_\theta(z)}{q_\phi(z \mid n,m)}\right] \\ &\geq \mathbb{E}_{q_\phi(z|n,m)}[\log p(n,m|z)] - D_{KL}\left(q_\phi(z|n,m) \middle\| p_\theta(z)\right) \end{aligned}$$

In practice, we optimize this lower bound simultaneously with respect to the global parameters $\theta$ and those of the approximate posterior $\phi$, using a stochastic gradient descent algorithm[55].

The example trajectories and curvature values in Figs. 3b, 4b and 5b are the result of this optimization procedure. Specifically, the optimal parameters $\theta^*$ contain our estimate of the trajectory's global length, curvature and shape, whereas the optimal parameters $\phi^*$ contain estimates of the trajectory's local distances, curvature and direction. When describing population data in Figs. 3c, 4c and 5c we further reduce the variance of our curvature estimates by reporting the mean of 100 bootstrapped samples.

**Modeling: evaluating curvature estimates with simulated observers.** After fitting our model to a given observer's data, we are left with a distribution over the perceptual trajectory's parameters. The mean of this distribution determines a trajectory whose curvature is equal to our estimate of the human observer's perceptual curvature (which we report in the results). If we replace its local curvature values with those of the intensity-domain trajectory, we arrive at the perceptual-domain trajectory of a simulated observer that is identical to the human observer but whose internal curvature is identical to the intensity-domain curvature. We then simulate a new dataset of responses from this observer, with the same number and distribution of trials as the original one. Fitting the model to this simulated dataset yields a perceptual-domain curvature estimate which reflects the null hypothesis for this set of sequences and observers. By comparing the distribution of perceptual-domain curvature for human observers to that of their simulated counterparts, we can assess whether perceptual-domain curvature differs significantly from intensity-domain curvature. This null distribution also provides a means of evaluating the bias and variance of our estimation procedure.

**Modeling: predictability of intensity domain and perceptual trajectories.** Having inferred a perceptual trajectory and its curvature for each image sequence and observer, we wanted to know whether straighter trajectories were more predictable than more curved ones. This is likely the case for first-order linear extrapolation but need not be for higher-order extrapolators. Having observed the frames $x_t$, $x_{t-1}$, and so on, a $K$th-order extrapolator predicts the next frame $x_{t+1}$ to be:

$$\hat{x}_{t+1} = \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_K x_{t-K}$$

in which the weights $\beta = (\beta_0, \ldots, \beta_K)$ can be fit using least-squares regression:

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \sum_t \left\| x_{t+1} - \sum_{k=0}^{K} \beta_k x_{t-k} \right\|_2^2$$

In Figs. 3d, 4d and 5d we report the error of such a predictor as a percentage of the average step size, for each sequence and observer's perceptual trajectory $x$:

$$\hat{e}[x] = 100^* \sqrt{\frac{\sum_t \|x_t - \hat{x}_t\|_2^2}{\sum_t \|x_t - x_{t-1}\|_2^2}}$$

As a control, we also applied this analysis to the perceptual trajectories of simulated observers described in the previous section.

**Statistical tests.** Unless specified otherwise, all statistical testing used a two-tailed Wilcoxon signed-rank test, typically for comparing curvature (or prediction error) in the intensity and perceptual domains or between human observers and simulated controls. The only exceptions are when ensuring that our curvature estimation methodology is not biased towards curvature reduction (we used a one-tailed test) and when comparing artificial and naturalistic 'contrast' sequences (we used a Mann–Whitney $U$-test).

Since the simulation process is inherently variable, we also compared the median change in curvature for human observers to the distribution of median change in curvature across simulated control populations, which yielded similar results (Supplementary Fig. 6). Specifically, the simulated populations we present in Figs. 3–5 show a median change in curvature which is equal to the median of the distribution across simulations. Moreover, for experiments 1 and 2 this distribution is concentrated around the median, such that the median change in curvature observed in human observers is significantly greater than for all simulated populations ($P < 0.001$, two-tailed Z-test; Supplementary Fig. 6, left and middle). For experiment 3, the median change in curvature shown by human observers relative to simulated controls is significantly smaller than in experiment 2 ($P = 0.02$; Supplementary Fig. 6, right). For these statistical tests, the distribution of median change in curvature across simulated control populations was assumed to be normal but this was not formally tested. As such, the simulated populations presented in Figs. 3–5 are representative of the distributions across simulated controls.

**Modeling: curvature in hierarchical models.** We constructed a two-stage model of early visual processing by cascading a model of retinal processing and one of primary visual cortex. The retinal model composes spatial center-surround filtering, luminance and contrast gain control and a rectifying nonlinearity[24]. The model of primary visual cortex uses a set of multi-scale, oriented and band-pass filters (a 'steerable pyramid'[56]), followed by squaring and summing over quadrature pairs to mimic the action of complex cells[25]. We used six scales and four orientations, excluding the high- and low-pass residual bands. The retinal model was optimized to match foveal perceptual discriminability judgments of human observers[24] but our images were presented at 2 degree eccentricity. To approximate the loss of visual acuity in the parafovea[57], we spatially down-sampled images by a factor of 2 using a Lanczos filter before presenting them to our model. Our results were robust to the precise choice of resolution (down-sampling by factors of 1, 2, 4 or 8 all give qualitatively similar results). We computed model response vectors for each frame in a sequence and measured the curvature of this sequence of responses.

We also evaluated the curvature of the same sequences as represented in each layer of a deep convolutional neural network (known as 'AlexNet') trained for object recognition[33]. The network contains a sequence of five rectified convolutional layers, with max pooling after the first, second and fifth layers. The convolutional layers have 64, 192, 384, 256 and 256 filters of size 11, 5, 3, 3 and 3 pixels respectively. We obtained the pre-trained model from the PyTorch Model Zoo, with corresponding Top-5 error rate on the ImageNet test set of 21%. We also tested more recent architectures (the 19-layer VGG model[34] with and without batch normalization[35], a 152-layer Residual Network[36] and a 121-layer Dense Network[37], with test errors of 8%, 9%, 6% and 8%, respectively) and obtained similar results. In all of these, we report curvature in the pooling layers, but obtained similar results in intermediate ones.

**Sample size and replication.** Our simulation and recovery analysis allowed us to estimate the variability in our curvature estimation method. This analysis revealed that six sequences and three observers per sequence would be sufficient to detect a curvature reduction of 20°. We collected these data which were the basis of our initial submission. For our final submission, we replicated our original findings with a new set of six sequences and three observers per sequence. We have presented the combination of both datasets.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The data supporting the findings of this study are available from the corresponding author on reasonable request.

## Code availability

The code used to analyze the data of this study is available from the corresponding author on reasonable request.

## References


51. Seshadrinathan, K., Soundararajan, R., Bovik, A. C. & Cormack, L. K. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Process.* **19**, 1427–1441 (2010).
52. Seshadrinathan, K., Soundararajan, R., Bovik, A. C. & Cormack, L. K. A subjective study to evaluate video quality assessment algorithms. In *SPIE Proceedings Human Vision and Electronic Imaging*, 1–10 (2010).
53. Wichmann, F. A. & Hill, N. J. The psychometric function: I. Fitting, sampling, and goodness of fit. *Percept. Psychophys.* **63**, 1293–1313 (2001).
54. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. Introduction to variational methods for graphical models. *Mach. Learn.* **37**, 183–233 (1999).
55. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *In Proc. International Conference on Learning Representations* **2**, 1–14 (2014).
56. Simoncelli, E. P. & Freeman, W. T. in *Proceedings second IEEE., International Conference on Image Processing*, 444–447 (1995).
57. Green, D. G. Regional variations in the visual acuity for interference fringes on the retina. *J. Physiol.* **207**, 351–6 (1970).

Corresponding author(s): Olivier Henaff (henaff@cns.nyu.edu)

Last updated by author(s): Feb 17, 2019

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☐ | ☒ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | We used the MGL library for stimulus presentation. |
| Data analysis | We used the Torch7 library for all data analyses. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data supporting the findings of this study are available from the corresponding author on reasonable request.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | Our simulation and recovery analysis allowed us to estimate the variability in our curvature estimation method. This analysis revealed that 6 sequences and 3 observers per sequence would be sufficient to detect a curvature reduction of 20 degrees. We collected these data which were the basis of our initial submission. For our final submission, we replicated our original findings with a new set of 6 sequences and 3 observers per sequence. The manuscript presents the combination of both datasets. |
| Data exclusions | We were unable to reliably estimate perceptual curvature from observers with very low accuracy. For that reason, we excluded data from observers whose proportion of correct responses was below 70%. This criterion was determined in simulation, before analyzing human human data. |
| Replication | For this revised version, we collected a second dataset (identical in size to the original) from an independent set of sequences. The results from this new dataset were perfectly consistent with our original findings. The manuscript presents the combination of both datasets. |
| Randomization | Experiments 1 and 2 were interleaved within a session for each sequence and observer, alleviating the need for group allocation and blinded data collection. A random subset of these observers also participated in experiment 3. |
| Blinding | Experiments 1 and 2 were interleaved within a session for each sequence and observer, alleviating the need for group allocation and blinded data collection. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology |
| ☒ | ☐ Animals and other organisms |
| ☐ | ☒ Human research participants |
| ☒ | ☐ Clinical data |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Human research participants

Policy information about studies involving human research participants

| | |
|---|---|
| Population characteristics | We tested 18 observers (6 females, 12 males; ages 19-30) with normal or corrected-to-normal vision. |
| Recruitment | We advertised the study with fliers in the common areas of the department, as well as a department-wide e-mail. All volunteers participated in the study. Given the nature of our measurements (visual discrimination of natural video frames), we do not expect any self-selection bias to impact our results. |
| Ethics oversight | New York University human subjects committee |

Note that full information on the approval of the study protocol must also be provided in the manuscript.