Testing a mechanism for temporal prediction in perceptual, neural, and machine representations

by

Olivier J. Hénaff

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Center for Neural Science

New York University

September 2018

Eero P. Simoncelli

To my parents, Ann and Patrick,

who shared their passion for discovery

Acknowledgements

I would first of all like to thank my advisor, Eero Simoncelli, for bringing me on the most exciting intellectual adventure I could have imagined, and for creating a thoughtful and nurturing environment in which I have thrived for the past five years. I have probably been irreversibly branded by his unique way of conducting scientific research, and for that I am extremely grateful.

This atmosphere of creativity and collaboration led to several fantastic relationships that have equally shaped my dissertation. I was very lucky to meet and build a collaboration with Robbe Goris during our shared time in Eero's lab, whose passion and focus have inspired large parts of this thesis. Prior to that, I had the wonderful opportunity to be taken under the wings of Neil Rabinowitz and Johannes Balle, who demonstrated that scientific excellence and quirky humor go hand in hand. I am also very grateful for the many exhilarating conversations I had with Joan Bruna during the early stages of my research. Finally, I would like to thank Yann LeCun for his initial guidance, and for inviting me to do a research internship that was the start of my graduate career.

None of the work presented in the third chapter would have been possible without the stellar collaboration of Yoon Bai, who collected the physiology data therein, and has been a continual source of inspiration and camaraderie. I am also very grateful to Natalie Pawlak, Rebecca Walton, and Lydia Cassard who provided invaluable assistance in collecting the psychophysical data presented in the second chapter. Most of the analyses in this thesis were enabled by the superhuman effort and support provided by Shenglong Wang, Robert Young, and Paul Fan. Finally, I would like to thank the valiant chefs of *Sidewalk Tacos*, Valentino, Miguel, Carlos and Eduardo, whose craft fueled much of the work in this thesis.

My time at NYU was brightened by fellow graduate students from several cohorts. Jonathan and Silvia in particular played key roles in my introduction to neuroscience, through long conversations over dosas in Washington Square Park. The members of the Laboratory for Computational Vision cemented that introduction, with Alex and Corey having a particular gift for lovely conversation about anything from science and politics to art and cinema, be it over espresso or on a ski lift.

Prior to and during my graduate studies, I have been a part of a stimulating and caring group of friends and family, to whom I owe everything. My parents, Ann and Patrick, have been an unwavering source of love and support, and continue to impress me with their boundless curiosity. My siblings, Elizabeth and Mikaël, have been and continue to be amazing role models, not least for their unending appetite for everything musical, scientific, and artistic. My good friends Léonard and Mandy have been a steady presence during my time in New York, and inspired me to take advantage of and contribute to all that the city has to offer. Last but not least, these five years have been uplifted and enchanted by my fortuitous encounter with Loreal, for which I will forever be grateful.

Preface

Chapter 3 is the result of a fruitful and thoroughly enjoyable collaboration with Yoon Bai, Julie Charlton, Ian Nauhaus, and Robbe Goris. Robbe Goris was also involved in much of chapters 2 and 4.

Abstract

Many decisions require that we take into account events that may occur far into the future. Yet natural scenes evolve according to complex, nonlinear dynamics that are difficult to extrapolate. We propose that the brain transforms visual input such that it follows straighter temporal trajectories, thereby enabling prediction through linear extrapolation. In this thesis, we test this 'temporal straightening' hypothesis in three different contexts: human psychophysics, primate physiology, and computational image synthesis.

We develop a methodology for estimating the curvature of internal trajectories from human perceptual judgments. We use this method 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, providing evidence that the visual system selectively straightens the temporal trajectories of natural image sequences.

If the visual hierarchy has learned to straighten the trajectories of natural videos, we would expect individual visual areas to contribute to perceptual straightening. We test this hypothesis by applying our curvature estimation methodology to population recordings from primary visual cortex, and find that the curvature of these neural trajectories is well predicted by our perceptual results. Finally, our hypothesis provides us with a framework for testing the metric properties of machine representations and their relationship to human vision. Together, these results point to the metric properties of natural videos as an effective way of identifying behaviorally relevant computation along the visual hierarchy.

Contents

List of Figures

Chapter 1

Introduction

1.1 On the purpose of vision

The ascidian tadpole, having completed its journey along the sea floor and found a suitable place to spend the rest of its life, attaches itself to the nearest rock and proceeds to digest its rudimentary eye, visual system, and the rest of its cerebral vesicle (Cloney, 1978). Though somewhat extreme, the situation of this tadpole makes a compelling point: vision is only useful insofar as it leads to action. What elements of vision, then, do we need to interact with our environment?

In the mammalian nervous system, it takes on the order of 100 ms for changes in light intensity perceived by the retina to be transmitted to decision-making and motor areas. Were our visual system to report the current state of our environment, we would only perceive and react to past events, misestimating for example the position of a moving car or saber-tooth tiger by several meters. In order to overcome these sensory delays, we must therefore make predictions about the state of the world some time in the future. More importantly, many decisions (for example, whether to cross the road, which subway to take, or whether to go to graduate school) require that we take into account events that

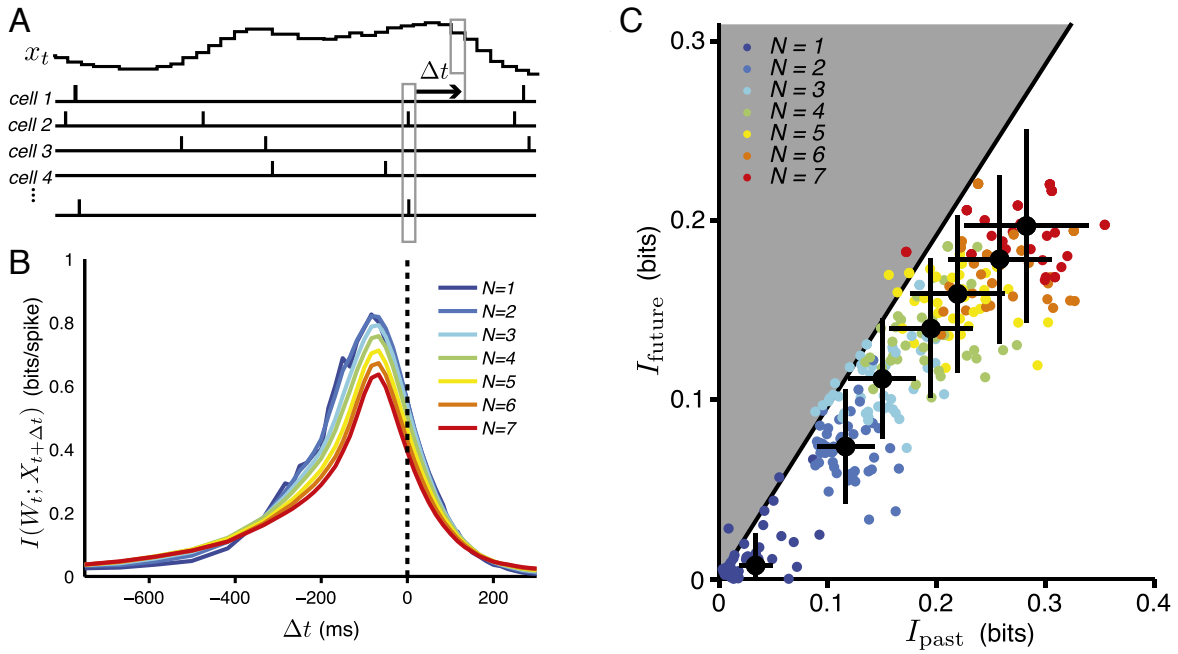


Figure 1.1 Information contained in the activity of a population of retinal ganglion cells about the position of a moving bar. (**a**) Mutual information is calculated between the pattern of neural activity at a given time bin and the position of the bar, at different time lags Δt. (**b**) This information is maximal around the response latency of the cells, but extends into the past and future. Different colors depict groups of neurons of varying sizes. (**c**) The information contained about the future is close to being as high as possible, given the amount of information contained about the past (the gray zone is physically unrealizable). Adapted from (Palmer et al., 2015)

.

may occur far into the future. How then, are we able to make predictions about various aspects of our environment at many different timescales?

A natural approach for describing how the brain makes predictions is to posit that the information encoded in the activity of populations of neurons is relevant to future states of the world, and mutual information provides a means of quantifying this relationship (Shannon, 1948). This quantity has been shown to be limited by the information contained about the past: one cannot predict future events without measuring past ones with some amount of precision (Tishby et al., 1999). Applying this framework to populations of

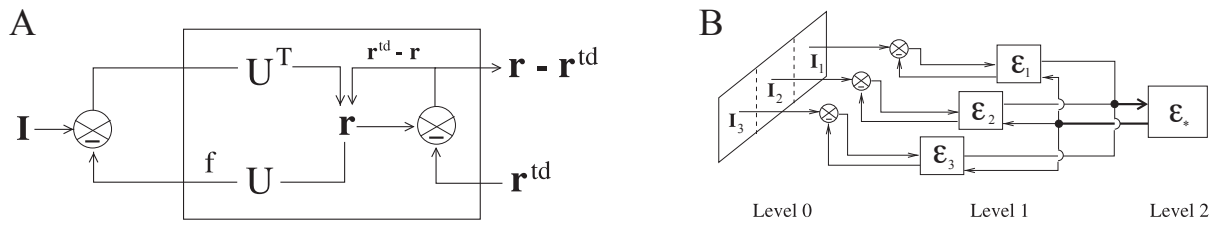


Figure 1.2 The predictive coding scheme. (**a**) Information is sent to subsequent stages by subtracting predictions from feed-forward input. (**b**) Such information is aggregated across multiple spatial locations (or time-points) to form more abstract predictions. Adapted from (Rao & Ballard, 1999)

retinal ganglion cells, Palmer et al., (2015) find that their activity is not only predictive of future environmental states (**Fig. 1.1)** but optimally so, given the amount of information they carry about the past (Chechik et al., 2005). This is compelling evidence for the idea that the activity of these neurons carries information about future stimuli, but does not speak to how this information could be exploited. Furthermore, the generality of their approach often makes their measurements impractical as soon as the dimensionality of the problem (number of neurons or stimulus dimensions) becomes even moderately high, limiting its applicability to small populations of neurons that represent simple features in the environment.

At the other extreme, theories of predictive coding commit to a specific functional form, in which predictions from higher areas are subtracted from the feed-forward activity coming from lower areas (**Fig. 1.2)**. While this scheme makes very direct experimental predictions, and successfully accounts for some aspects of early cortical physiology (Rao & Ballard, 1999), its subtractive comparison between predictions and observations could be overly restrictive. In particular, this computation fully specifies what should be represented at higher stages, even though the functional properties of mid- to high-level areas remain an open problem in visual neuroscience.

1.2 The temporal straightening hypothesis

In this thesis, we introduce a theory for temporal prediction that resides at an intermediate level of abstraction. By describing precisely how predictions are to be made, we aim for a theory that is more tractable and testable than the information theoretic one. On the other hand, by remaining agnostic about what is represented by neural populations, we are able to apply it to representations whose functional properties are not fully known, from human psychophysics to artificial neural networks and primate physiology. Specifically, we propose that internal representations of the environment are structured to enable prediction of their own future states, through linear extrapolation. Since the pioneering work of Kalman, (1960) linear prediction has been widely adopted in many engineering problems, including navigation and tracking. Despite non-linear extensions of this work (Jazwinski, 1970; Julier & Uhlmann, 1997; Ghahramani & Roweis, 1999; Haykin, 2001; Krishnan et al., 2015), prediction of high-dimensional, non-linear signals remains an open problem. We propose that biological organisms leverage their non-linear processing of visual scenes to facilitate the task of prediction. In particular, if this processing were structured so as to straighten the temporal evolution of a scene, then linear extrapolation would be sufficient to accurately predict future states of the environment (**Fig. 1.3)**.

This particular mechanism for prediction, 'temporal straightening', is closely related to two other theories for sensory processing. On one hand, efficient coding posits that neural representations are structured so as to preserve the information in natural signals, while reducing their redundancy (Barlow, 1961). Temporal straightening offers a specific form of coding efficiency, given that predictable signals can be coded via small residual errors, a feature that is well-known to the video and audio compression communities. In

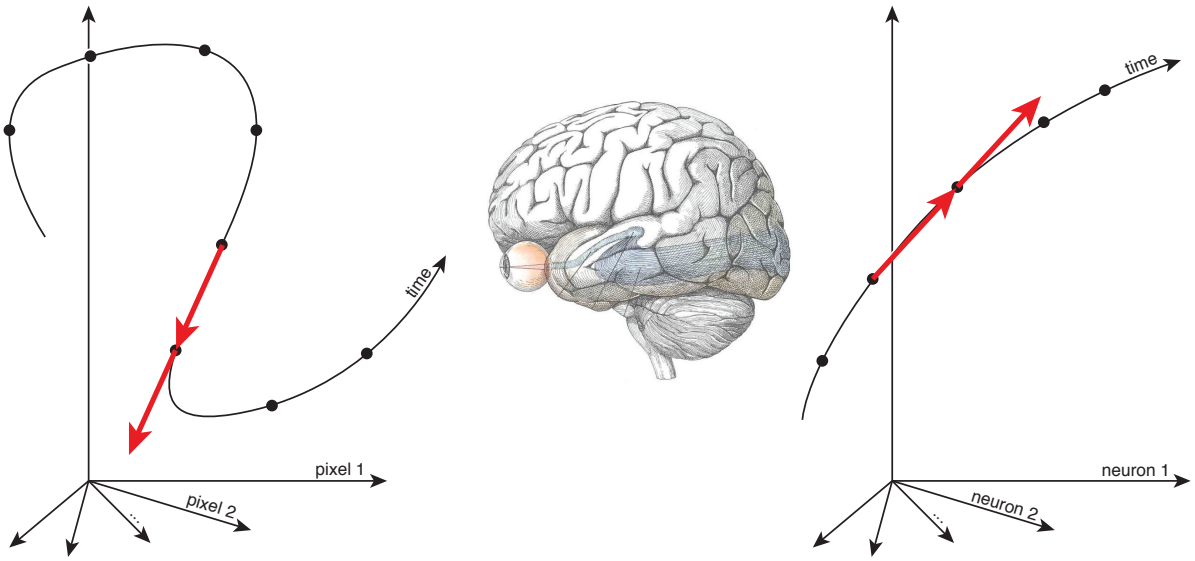


Figure 1.3 The temporal straightening hypothesis. Because the temporal trajectories of natural videos are highly curved in the domain of pixel-intensities (left), linear extrapolation fails to predict future stimuli. We propose that the visual system transforms its inputs into a new representation (right) whose evolution is much straighter, enabling prediction through linear extrapolation.

contrast to this bottom-up criterion, goal-directed learning proposes that the visual system is designed to support a range of visual tasks such object recognition. This theory has been successful in explaining the selectivity and invariances of higher visual areas (Hung et al., 2005; Rust & DiCarlo, 2010), and allowed for a quantitative comparison between artificial neural networks whose sole purpose is classifying objects and these same areas (Yamins et al., 2014). The similarity between these representations was later explained by the fact that both biological and artificial representations 'untangle' the image manifolds that correspond to particular objects, enabling classification with simple, linear operations (**Fig. 1.4;** DiCarlo & Cox 2007; Hong et al. 2016). And yet the ways in which these systems learn such untangled representations are very different. Artificial neural networks are generally trained with massive amounts of supervision, reproducing human annotations

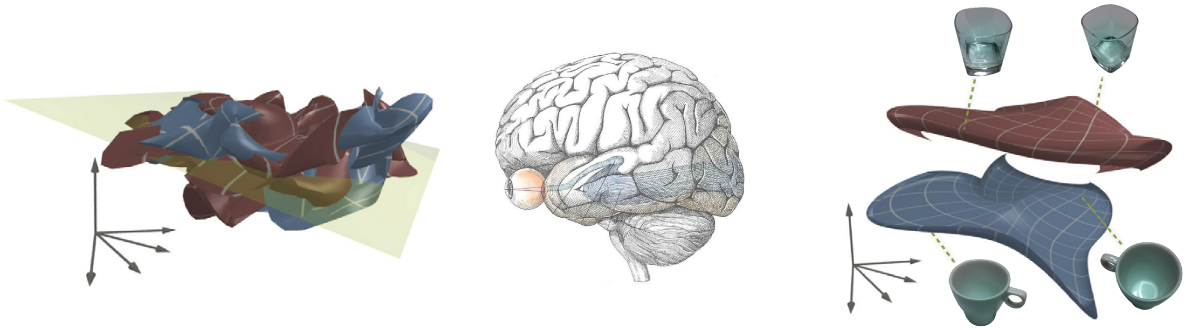


Figure 1.4 The untangling hypothesis. Because the manifolds representing two different objects (in this case, a water glass and an espresso cup) under different viewpoints and lighting conditions are hopelessly intertwined in the domain on pixel-intensities (left), a linear classifier would be hard-pressed to distinguish them. The untangling hypothesis proposes that the visual system solves this problem by representing these objects such that changes in viewing conditions are confined to a quasi-linear subspace (right). Adapted from (DiCarlo & Cox, 2007).

over millions of trials (Deng et al., 2009). Biological systems on the other hand are thought to be able to learn about their environment in a largely unsupervised manner, with humans being able to generalize from a single instance of a new category (Lake et al., 2015). Temporal straightening of natural videos could be a way of achieving such an unsupervised understanding of the environment. Indeed, objects tend to persist in a scene over relatively long timescales, and the trajectories of natural videos are therefore contained in individual object manifolds. Straightening these trajectories would then require straightening that portion of the object manifold, and doing so for many videos might result in the object untangling thought to be necessary for real-world vision.

Finally, an alternative hypothesis that explains the difference in sample efficiency between current artificial systems and biological agents is that the latter could have an innate understanding of objects and the physics that govern them (Battaglia et al., 2013). Indeed, a considerable body of literature has shown that even young infants expect objects to

persist over time, interact in specific ways, and belong to distinct categories (Spelke, 1990; Tenenbaum et al., 2011). Despite its success in modeling behavior, a neural implementation for this rich prior knowledge remains elusive. Object untangling (DiCarlo & Cox, 2007), by confining the representations of individual objects to distinct subspaces, could be a way of representing this structure, and temporal straightening a way of learning it.

In summary, if the brain indeed performs some form of temporal straightening, it may provide not only a mechanism for how we make long-term predictions, but also an explanation of how we have come to understand the world through experience and how to enable machines to learn in the same way. This thesis describes a series of studies that test the straightening hypothesis in different contexts.

1.3 Structure of the thesis

Chapter 2 evaluates the curvature of image sequences through human perceptual judgments. Consistent with the straightening hypothesis, we find the curvature of natural videos to be reduced perceptually. This straightening appears to be specific to natural sequences, as artificial ones that are linear in the image-domain see their curvature increased. Nevertheless, the curvature of naturalistic sequences that are linear in the image-domain is preserved. Finally, we find that all three perceptual effects can be accounted for by a hierarchical model of computations in the early visual system, but not by a generic artificial network, suggesting that the visual system has learned to straighten the trajectories of natural videos.

Chapter 3 provides a direct test of the hypothesis that early visual areas contribute to perceptual straightening. Consistent with our behavioral results, we find the neural trajectories of natural videos to be straightened in primary visual cortex, while artificial ones are

distorted. Moreover, changes in curvature found in primary visual cortex are predictive of the changes found perceptually. These results confirm that perceptual straightening could arise through successive computations in the visual hierarchy.

Chapter 4 uses the straightening of image sequences as a framework for evaluating the perceptual relevance of machine representations. We develop a procedure for synthesizing image sequences that are of minimal length according to a representation. Measuring the perceptual length of these 'geodesic' sequences allows us to efficiently assess their match to human vision. Moreover, visualizing these sequences exposes their selectivities and invariances, and suggests a way of improving their biological relevance.

Chapter 2

Perceptual straightening of natural videos

2.1 Introduction

Our first test of the temporal straightening hypothesis is perceptual. If our internal representation of a natural video follows a straighter trajectory than in the light-intensity domain, the pairwise perceptual distances (or discriminability) of pairs of frames should reflect that geometry. Specifically, the straighter a sequence of frames, the more distances should be additive, i.e. the more the distance between two end-frames should be equal to the sum of distances between intermediate frames. As a result, if we can estimate the perceptual distances between pairs of frames from a video, we should be able to combine them into an estimate of its curvature.

We developed a novel 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 in-

tensity 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 will likely exhibit increased perceptual curvature. Finally, synthetic sequences that contain naturalistic changes 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 are able to account for these behaviors, while deep convolution networks optimized for object recognition are not.

2.2 Quantifying straightness with discrete curvature

To test the temporal straightening hypothesis, 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. 2.1)**. 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.

Calculating intensity-domain curvature

We measure the curvature of sequences in the intensity domain directly, by computing the differences between successive frames in the high-dimensional space of pixels, and the

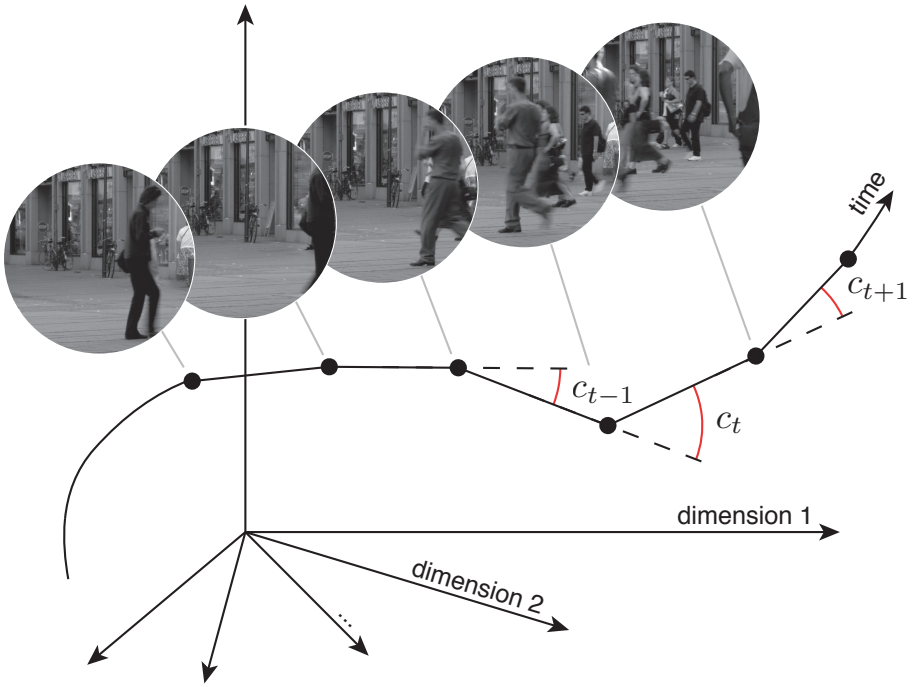


Figure 2.1 Quantifying straightness of image sequences in the intensity and perceptual domains. We consider representations in two domains: the 'pixel-intensity' domain (axes correspond to pixel intensities in each frame), and the 'perceptual' domain (axes correspond to responses of a set of neurons 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.

angles between them (**Fig. 2.2**). Specifically, if $x = \{x_t\}_{t=0,\ldots,T}$ are vectors containing the pixel luminances of a sequence of frames in a video (T = 10 in our experiments), we can define a sequence of normalized displacement vectors $\{\hat{v}_t\}_{t=1,\ldots,T}$:

$$v_t = x_t - x_{t-1}$$

$$\hat{v}_t = \frac{v_t}{\|v_t\|}$$

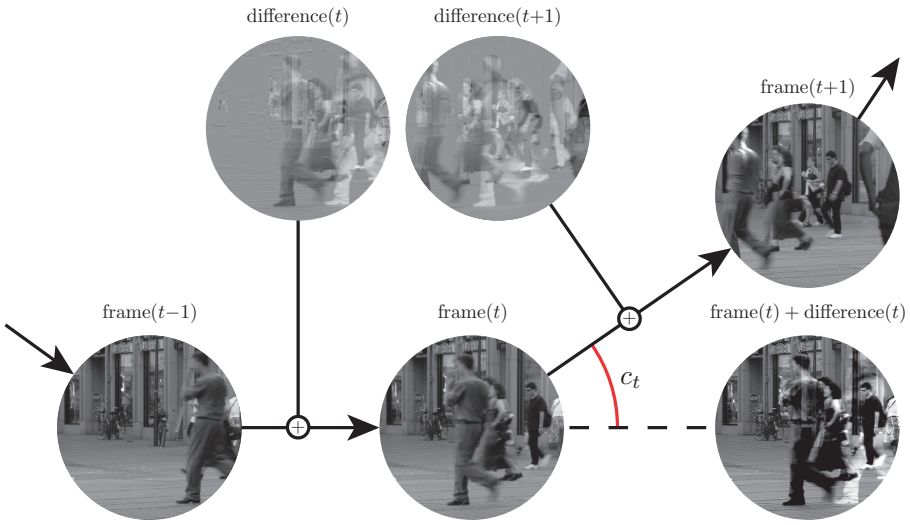


Figure 2.2 In the pixel-intensity domain, curvature can be calculated directly by computing the pixel-wise differences between successive frames, and the angles between them. Note how this sequence of frames is curved in the intensity domain (difference images are dissimilar) but seems natural perceptually. In contrast, a linear extrapolation in the intensity domain appears to be highly unnatural.

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 these local curvatures, in degrees.

2.3 Estimating perceptual curvature

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

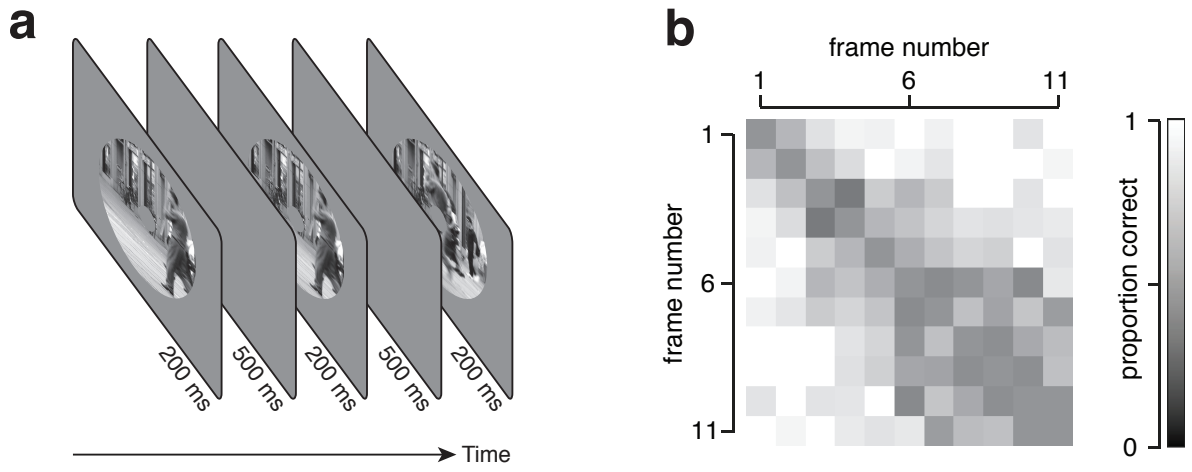


Figure 2.3 Measuring the discriminability of pairs of frames from a sequence. (**a**) Psychophysical 'AXB' task. On each trial, observers viewed a sequence of three images. The first and the last are randomly selected frames from a given sequence, and the middle one is identical to one of the other two. 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). The brightness of each pixel depicts the proportion correct in discriminating a pair of frames (brighter indicates more discriminable).

pair of frames (e.g. the 1st and 3rd frames in **Fig. 2.2)** should fall short of the summed intermediate distances (connecting the 1st, 2nd and 3rd 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. 2.3a**). 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 all pairs of frames drawn from an 11-frame sequence (**Fig. 2.3b**), 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. 2.4,** top

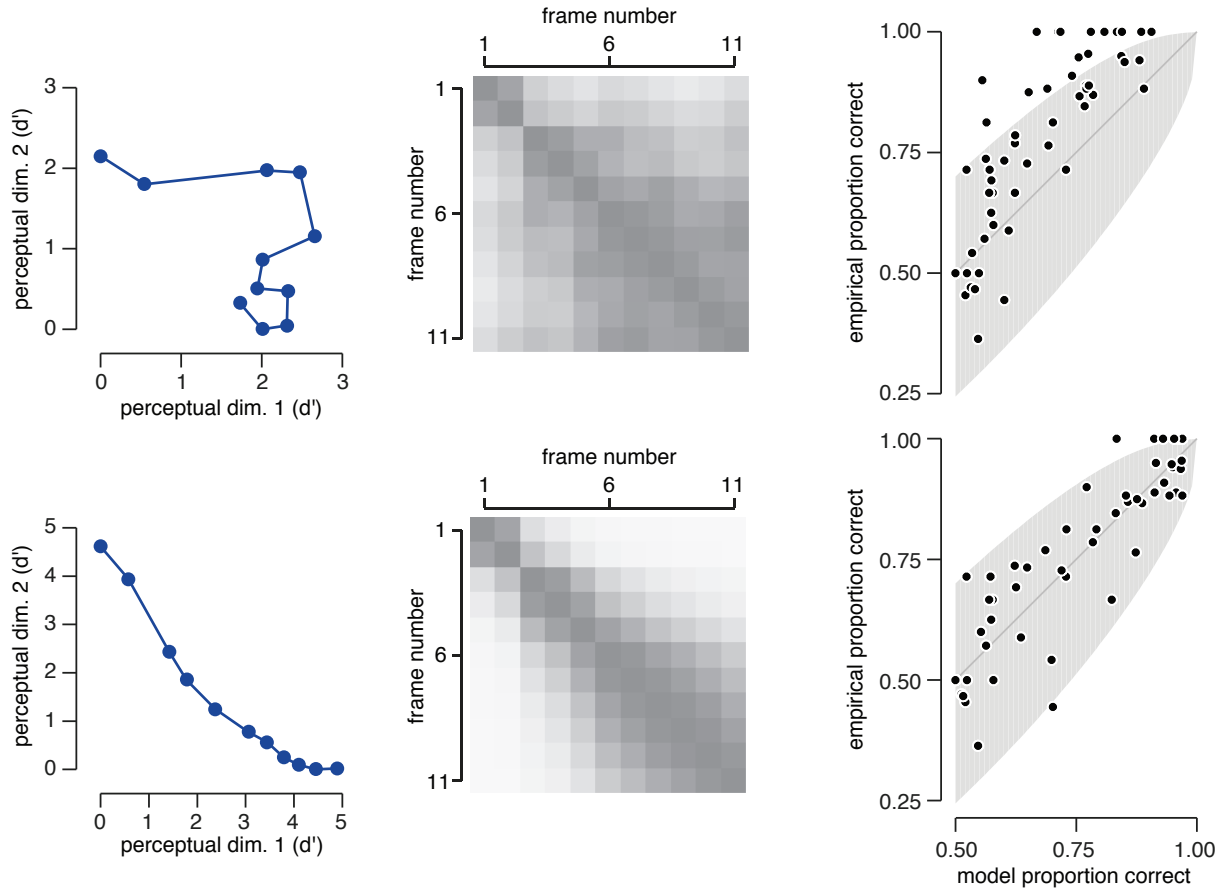


Figure 2.4 Inferring perceptual curvature from psychophysical data. Left column: 2-dimensional projections of 10-dimensional perceptual trajectories. Each point illustrates the centroid of a 2-D Gaussian distribution corresponding to the noisy perceptual representation of a frame in the sequence. Middle column: pattern of performance on the 'AXB' task predicted from the pairwise distances between points along the trajectory. Right column: match between empirical and model-predicted proportion correct (one point for each pair of frames), along with the 95% confidence interval expected from binomial variability (gray region). Top row: a curved perceptual trajectory predicts a slow increase in discriminability as frames are further separated in time, providing a poor match to the data. Bottom row: a straighter perceptual trajectory predicts a faster increase in discriminability and provides a good match to the data. Performing this analysis for many such perceptual trajectories, we can infer that these human judgments are consistent with low perceptual curvature.

left). The pairwise distances between points along the trajectory lead to a prediction about the discriminability we would have observed had this been the trajectory that the human observer used to inform their judgments (Noreen, 1981) (**Fig. 2.4,** top middle). In particular, this allows us to measure the large discrepancy between the predicted and observed patterns of discriminability (**Fig. 2.4,** 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. 2.4,** bottom row). Given this, it is tempting to iteratively adjust this trajectory until arriving at the most likely one (similarly to non-linear dimensionality reduction methods (Roweis & Saul, 2000; Tenenbaum et al., 2000)), and reporting its curvature as was done in the intensity domain. But this two-step method is plagued by significant estimation bias when used with the amounts of data available from our experiments (**Fig. 2.5a**). 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 (**Fig. 2.5b**, see section on inference). For visualization purposes, we display the perceptual trajectory whose length and curvature are equal to the average across plausible perceptual trajectories (**Fig. 2.4,** bottom row).

Estimating perceptual curvature: experimental paradigm

Eighteen observers (6 females, 12 males; ages 19-30) with normal or corrected-to-normal vision participated in the experiment. 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; all others were naive to the purposes of the experiment.

Each trial of the experiment followed an 'AXB' paradigm, in which three images were

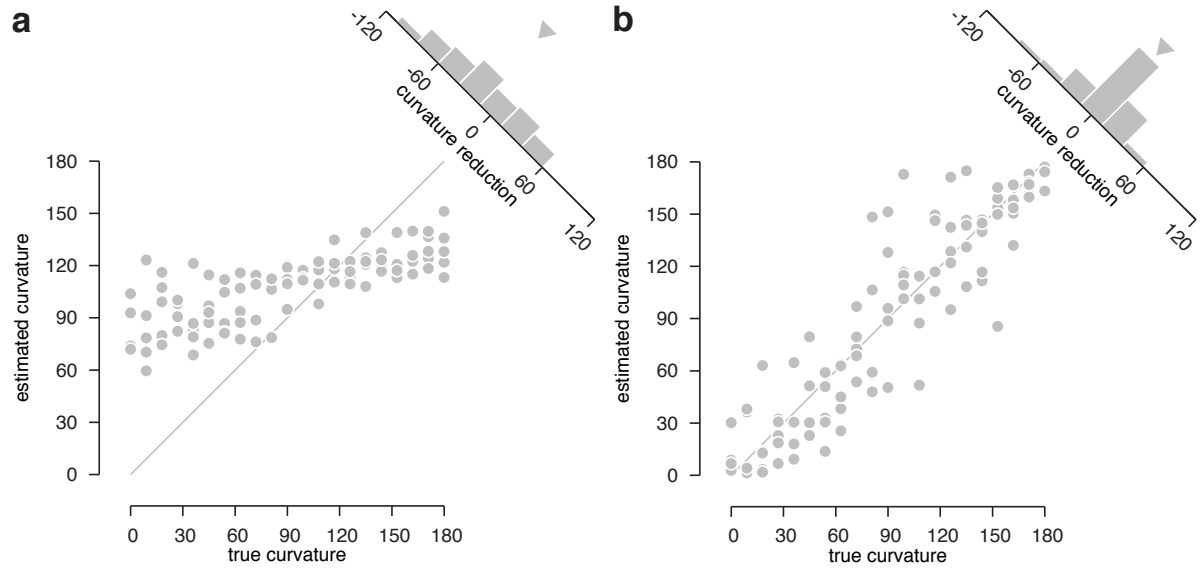


Figure 2.5 Recovery analysis for perceptual curvature estimation. We simulated 4 observers with different sensitivities viewing 21 different sequences with varying perceptual curvature and evaluated our ability to estimate the perceptual curvature from the same amount of data we use in our experiment. Simulated observers' sensitivities spans the range of human sensitivities, and perceptual curvatures vary from 0°to 180°. (**a**) Greedy, two-step estimation, that measures perceptual curvature from the most likely perceptual trajectory, is plagued by significant bias. (**b**) Our method, which estimates perceptual curvature by averaging over many plausible perceptual trajectories, is largely unbiased.

displayed, the first and the last being a random 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 which image was the unique one, and were given feedback after each trial. Sequences obtained from the Chicago Motion Database were luminance calibrated, and we displayed their intensities on a calibrated display. For the others, we transformed intensities with pixel-wise power (gamma) function, whose exponent we obtained from the meta-data. 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. Each observer performed 1,000 trials for each sequence on which they were tested, resulting in 18 trials for each pairwise comparison on average.

Estimating perceptual curvature: observer model

We wish to infer the perceptual curvature of sequences 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 between $T+1$ 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

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

The probability of a correct response p_{ij} is a linear combination of the probability of successfully performing the AXB task p_{ij}^{AXB}, and the probability of successfully guessing the correct answer $(\frac{1}{2})$ weighted by the lapse rate λ

$$p_{ij} = (1 - 2\lambda) p_{ij}^{\mathrm{AXB}} + \lambda$$

The discriminability of two frames does not determine their relative locations in perceptual space, or the shape of their associated noise distributions. However, in order 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, and the probability of successfully performing the task may then be expressed using standard methods from signal detection theory (Noreen, 1981):

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

where Φ is the cumulative distribution function of the standard normal.

Estimating perceptual curvature: inference

Intuitively, a natural procedure for estimating perceptual curvature consists of maximizing the likelihood of the perceptual locations $\{x_t\}_{t=0,\ldots,T}$ of each frame given the entire data set of an observer, and then computing the curvature of this trajectory. Although this is correct 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°(i.e. the most likely configuration of random vectors in a high-dimensional space; **Fig. 2.5a**). 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. While more complex, this procedure is substantially more robust than the greedy two-step process described above, and is nearly unbiased (**Fig. 2.5b**).

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 $\{v_t\}_{t=1,\ldots,T}$, 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 over 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:

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

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 (Wichmann & Hill, 2001)). Σ_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 $z = \{z_t^d, z_t^c, z_t^a, z^\lambda\}$. Direct curvature estimation amounts to maximizing the likelihood of these parameters in order to best account for the data, a form of empirical Bayes estimation. Computing the (log) likelihood of this prior requires marginalizing over local curvature variables, a high-dimensional integral that is intractable in practice:

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

Fortunately, variational methods provide a tractable lower bound on the likelihood which can be optimized with stochastic gradient methods (Jordan et al., 1999). By fitting an approximate Gaussian posterior $q_\phi(z|n, m)$ to the true posterior over local variables $p(z|n, m)$,

we replace the intractable integral with an analytical one:

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

We optimize this lower bound simultaneously with respect to the parameters of the prior and the approximate posterior, using a stochastic gradient descent algorithm (Kingma & Welling, 2013).

The example trajectories and curvature values in **Figures 3-5** are the result of this optimization procedure. When describing population data, we further reduce the variance of our curvature estimates by reporting the mean of 100 bootstrapped samples.

Evaluating curvature estimates with simulated observers

We compared the distribution of estimated perceptual curvatures to a null distribution, obtained by simulating observers whose internal curvature was matched to the intensity-domain curvature. This null distribution also provides a means of evaluating the variance and bias of our estimation procedure. 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 image sequence's pixel-intensities, we arrive at the perceptual 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 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 curvature estimate which reflects the null hypothesis for this set of sequences and observers. By comparing the distribution of perceptual curvature for human observers to that of their simulated counterparts, we can assess whether perceptual curvature differs significantly from intensity-domain curvature.

2.4 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- and perceptual-domain curvature of twelve natural image sequences which differed in content (experiment 1; see **Fig. 2.6a** for 3 frames from an example sequence, **Fig. 2.12,2.13** for all sequences). To visualize our analysis and gain an intuition for the results, we projected the intensity- and perceptual-domain representations for a single sequence and observer onto the first two principle components (**Fig. 2.6b**). 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 curvature = 8°). Moreover, this

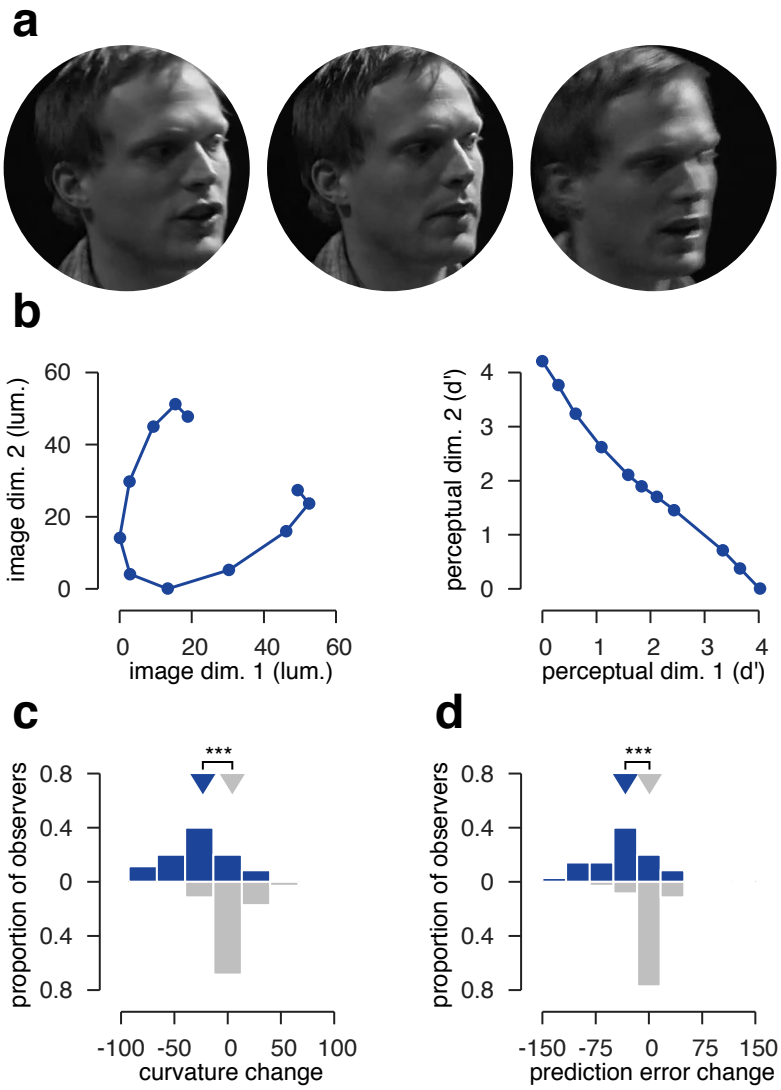


Figure 2.6 Curvature reduction for natural image sequences. (**a**) Initial, middle, and final frames of one such sequence (a man speaking). (**b**) Two-dimensional projections of an example sequence in the intensity domain (left) and in the inferred perceptual domain (right). (**c**) Change in curvature (left) and prediction error (right) from the intensity to the perceptual domain, for 12 natural image sequences and 18 observers (35 sequence/observer pairs total). Blue histogram: perceptual curvature estimated from human subject data. Gray histogram: perceptual curvature estimated from data simulated from model observers whose perceptual curvature is matched to their intensity-domain curvature, with all other parameters matched to those of the human observers. Triangles indicate the median of each distribution. *** $p < 0.001$. (**d**) Change in prediction error from the intensity to the perceptual domain. Same layout as (**c**).

curvature reduction is robust across sequences and observers (**Fig. 2.6c**, blue histogram; median curvature change = $-23°$; $p < 0.001$, 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 and rule out potential biases. 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 assumed these model observers based their responses on a perceptual representation whose curvature was matched to the pixel-domain curvature (Online Methods). When applied to these synthetic data, our analysis method found no reduction in curvature (**Fig. 2.6c**, gray histogram; median curvature change = $4°$; $p = 0.99$, 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 thus provide clear supporting evidence for the notion that the human visual system straightens curved natural videos.

If straightening the trajectory of natural videos is the mechanism by which the visual system supports temporal prediction, then these videos should be more predictable perceptually than they are in the intensity domain. We quantified the error committed by linear extrapolation in a given domain in the following manner. If we have observed the trajectory up to time t, linear extrapolation predicts the next frame to be $\hat{x}_{t+1} = x_t + v_t$,

resulting in the following error:

$$\begin{aligned} e_t &= \|x_{t+1} - \hat{x}_{t+1}\| \\ &= \|v_{t+1} - v_t\| \end{aligned}$$

We report the average prediction error $\hat{e}$, as a percentage of the average step size $\hat{d}$:

$$\begin{aligned} \hat{e} &= \frac{1}{T-1} \sum_{t=1}^{T-1} e_t \\ \hat{d} &= \frac{1}{T} \sum_{t=1}^{T} \|v_t\| \end{aligned}$$

Having computed the prediction error in the perceptual domain, we found that it was indeed significantly reduced both relatively to the intensity domain (**Fig. 2.6d**, blue histogram; median change in prediction error = -33%, $p < 0.001$), and relatively to the simulated controls that did not change the curvature of these videos (**Fig. 2.6d**, gray histogram; median difference in prediction error = -33%, $p < 0.001$). Hence, consistently with our original motivation, natural videos are made more predictable via perceptual straightening.

2.5 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 visual input. It does not by itself, however, indicate that this transformation is specifically targeted 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 hypothesize,

temporal straightening is targeted at sequences that occur naturally, then sequences that are unlikely to occur have no reason to be straight. On the contrary, these sequences will most likely be distorted by the nonlinear hierarchical transformations of the visual system (Poole et al., 2016).

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 have zero curvature in the intensity domain, but are highly unnatural in that the interpolated middle frames contain pixel-wise averages of two different images (see **Fig. 2.7a** for 3 frames from an example sequence, **Fig. 2.12,2.13** 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. 2.7b**). Consistent with our hypothesis, the perceptual trajectory of this artificial sequence is much more curved than the intensity trajectory (curvature change = 48°). This effect was just as prominent in the high-dimensional spaces (curvature change = 49°), and was consistent across all artificial sequences and observers (**Fig. 2.7c**, green histogram; median curvature change = 53°). Note that the 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. 2.7c**, gray histogram), significantly smaller than

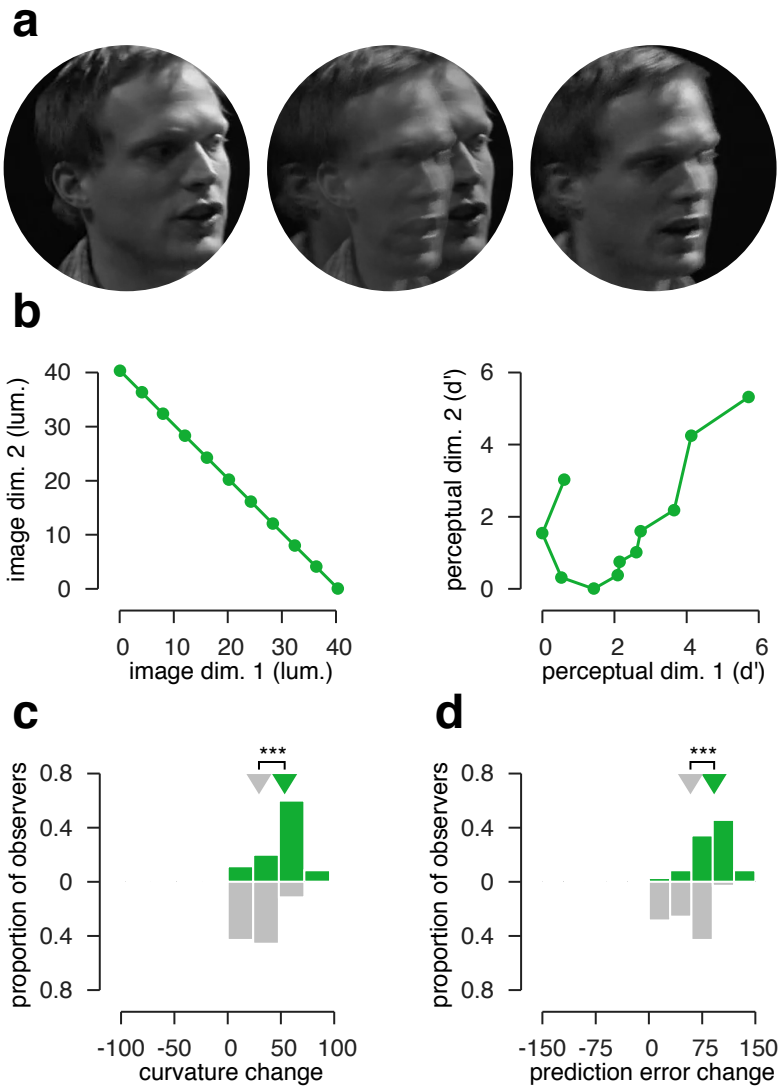


Figure 2.7 Curvature increase for artificial image sequences. (**a**) Initial, middle, and final frames of one such image sequence. Initial and final frames are identical to those of the corresponding natural sequence (**Fig. 2.6a**), whereas intermediate frames are generated by linearly interpolating (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**) Change in curvature from the intensity to the perceptual domain, for 12 artificial image sequences and 35 sequence/observer pairs. Green histogram: perceptual curvature estimated from human subject data. 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. (**d**) Change in prediction error from the intensity to the perceptual domain. Same layout as (**c**).

the 53° increase observed in our human subjects ($p < 0.001$).

Moreover, this increase in curvature was accompanied by a decrease in predictability. Indeed, the perceptual trajectories described by these unnatural sequences were less predictable compared to their intensity-domain trajectories (**Fig. 2.7d**, green histogram; median change in prediction error = 92%, p < 0.001) and to simulated controls who did not increase the curvature of these sequences (**Fig. 2.7d**, gray histogram; median difference in prediction error = 30%, p < 0.001).

2.6 Curvature preservation of naturalistic videos

We interpreted the outcome of experiment 2 as evidence that the nonlinear computations underlying perceptual straightening are targeted for 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 representations, regardless of whether they are natural or artificial. To resolve this ambiguity, we characterized the perceptual 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. 2.8a**). 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 (e.g. due to the onset of fog). The temporal straightening hypothesis predicts that these sequences should be much straighter than their unnatural counterparts. **Fig. 2.8b** shows the low-dimensional projection of the perceptual trajectory, for an example observer in experiment 3. In this case, the perceptual distortion of the contrast-varying sequence is minimal (low-dimensional curvature change = 9°; high-dimensional curvature

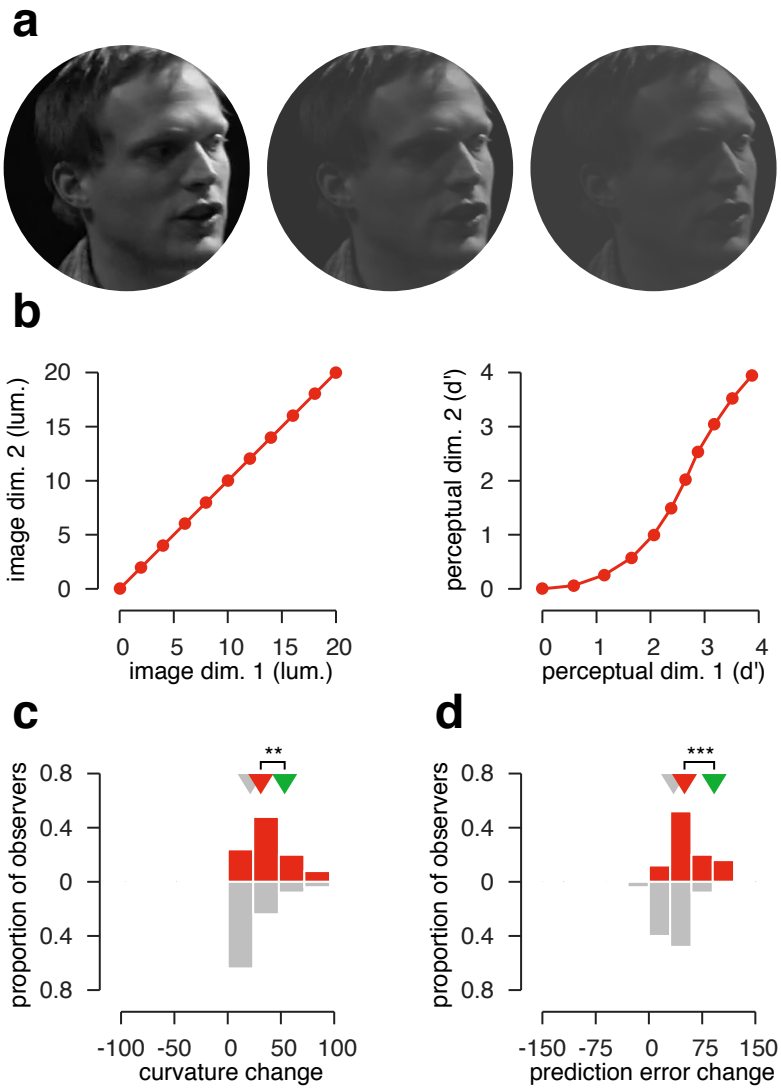


Figure 2.8 Curvature conservation for naturalistic, intensity-linear image sequences. (**a**) Initial, middle, and final frames of one such sequence. Initial frame is identical to those in figures 2.6 and 2.7, and the rest are generated by gradually reducing contrast. (**b**) Low-dimensional projections of an example sequence in the intensity domain (left) and in the inferred perceptual domain (right). (**c**) Change in curvature from the intensity to the perceptual domain, for 10 sequences and 25 sequence/observer pairs. Red histogram: perceptual curvature estimated from human subject data. 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). Green triangle is copied from **Fig. 2.7c**, showing that naturalistic sequences are significantly less curved than their artificial counterparts. ** $p < 0.01$. (**d**) Change in prediction error from the intensity to the perceptual domain. Same layout as (**c**).

change = 18°). We found this result to be consistent across all sequences and observers. Specifically, although these sequences experienced a significant increase in curvature (median difference between human observers and simulated controls = 12°, $p < 0.01$), their perceptual curvature was significantly smaller than that of the artificial sequences (difference in median curvature = 22°, $p < 0.01$). Similarly, these sequences were significantly more predictable than their unnatural counterparts (difference in median prediction error = 42%, $p < 0.001$). Hence, consistent with the temporal straightening hypothesis, the human visual system is able to preserve the linearity and predictability of straight, naturalistic videos.

2.7 Computational basis of perceptual straightening

Finally, we wondered 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 curvature. To probe this hypothesis, we examined responses of a two-stage model that mimics the nonlinear functional properties of the early visual system (**Fig. 2.9,** top). 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 (Mante et al., 2008; Berardino et al., 2017). The second stage further transforms this representation using a set of oriented filters whose responses are squared and combined over phase, capturing the nonlinear behaviors of complex cells in primary visual cortex (area V1) (Adelson & Bergen, 1985). We constructed response trajectories by applying the model to each video frame independently, and evaluated the curvature of the model

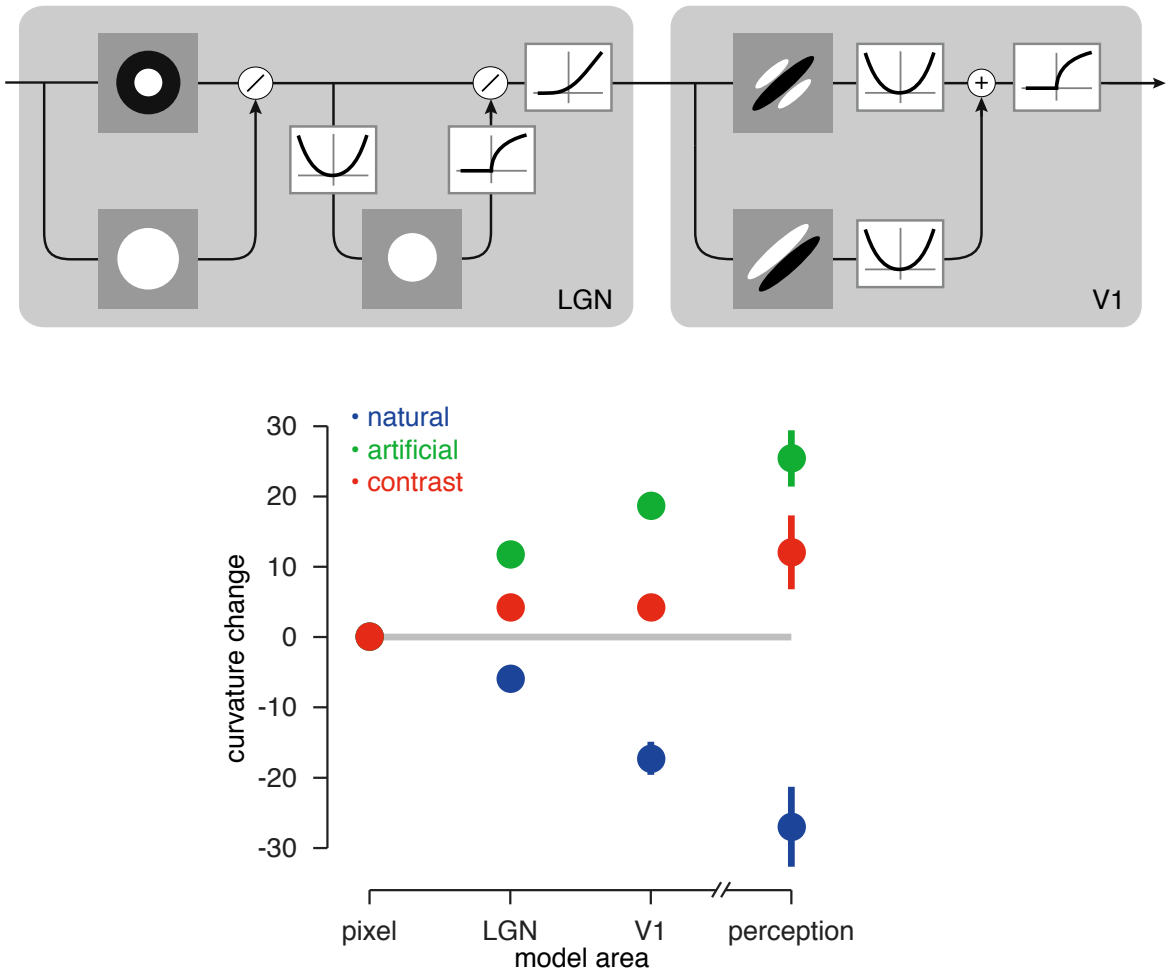


Figure 2.9 Changes in curvature for a hierarchical model of the early visual system. Top: Two-stage cascade architecture describing computations found in the retina, LGN, and V1. The first stage performs bandpass filtering, followed by luminance and contrast gain control. The second stage decomposes 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). Bottom: 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). We report the perceptual data as the median difference between human observers' and simulated controls' curvature, to correct for any estimation bias.

response vectors 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 the naturally straight sequences (**Fig. 2.9,** bottom). These effects were clearly evident in the first stage of the model, and became stronger in the second, although still falling short of the perceptual effects observed in our human subjects.

We also tested the straightening capabilities of artificial neural networks constructed from many stages of rectified linear filters. These models have shown impressive capabilities when optimized for object recognition (LeCun et al., 2015), and have been proposed as candidate models of biological vision (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Tacchetti et al., 2017). We wondered whether the ability of these networks to 'untangle' image manifolds associated with object categories (DiCarlo & Cox, 2007) might also extend to straightening of natural videos. To test this, we evaluated the changes in curvature in each stage of the AlexNet architecture (Krizhevsky et al., 2012). Unlike the simple biological models and our human subjects, we found that this model did not straighten the time-course of any of the natural videos (**Fig. 2.10)**. We tested several other current deep neural network architectures used for image classification (Ioffe & Szegedy, 2015; Simonyan & Zisserman, 2015; He et al., 2016; Huang et al., 2017), and found that all of them increased the curvature of natural videos (**Fig. 2.11**). 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 do exhibit straightening. We thus conclude that optimizing such networks for static object recognition fails to endow them with temporal straightening capabilities.

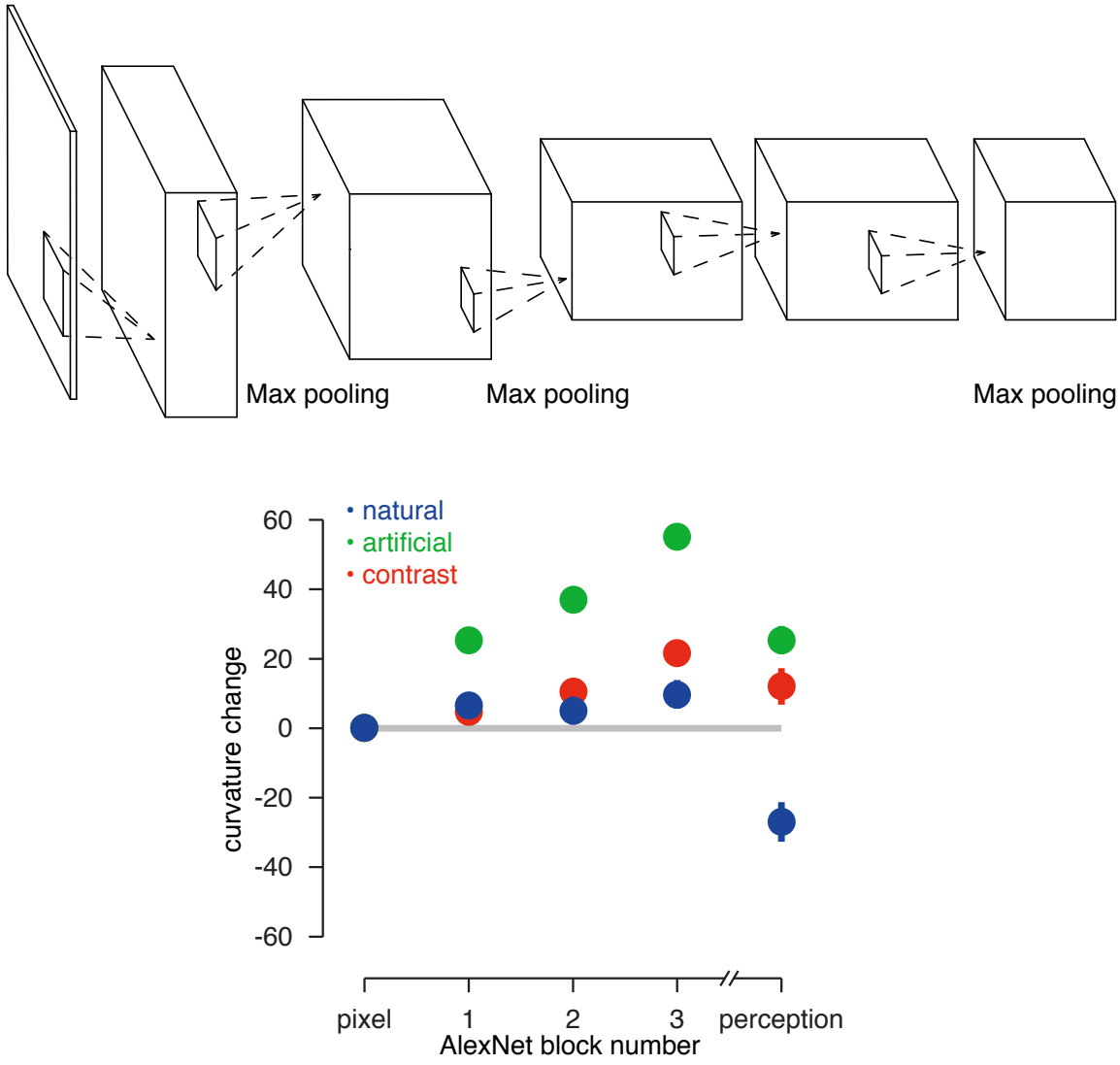


Figure 2.10 Changes in curvature for an artificial neural network trained for object recognition. Top: Model architecture, alternating convolution, rectification, and max-pooling. Bottom: Change in curvature induced by this network, for the same sequences as in **Fig. 2.9.** Despite strong performance on object classification, this model does not straighten natural sequences.

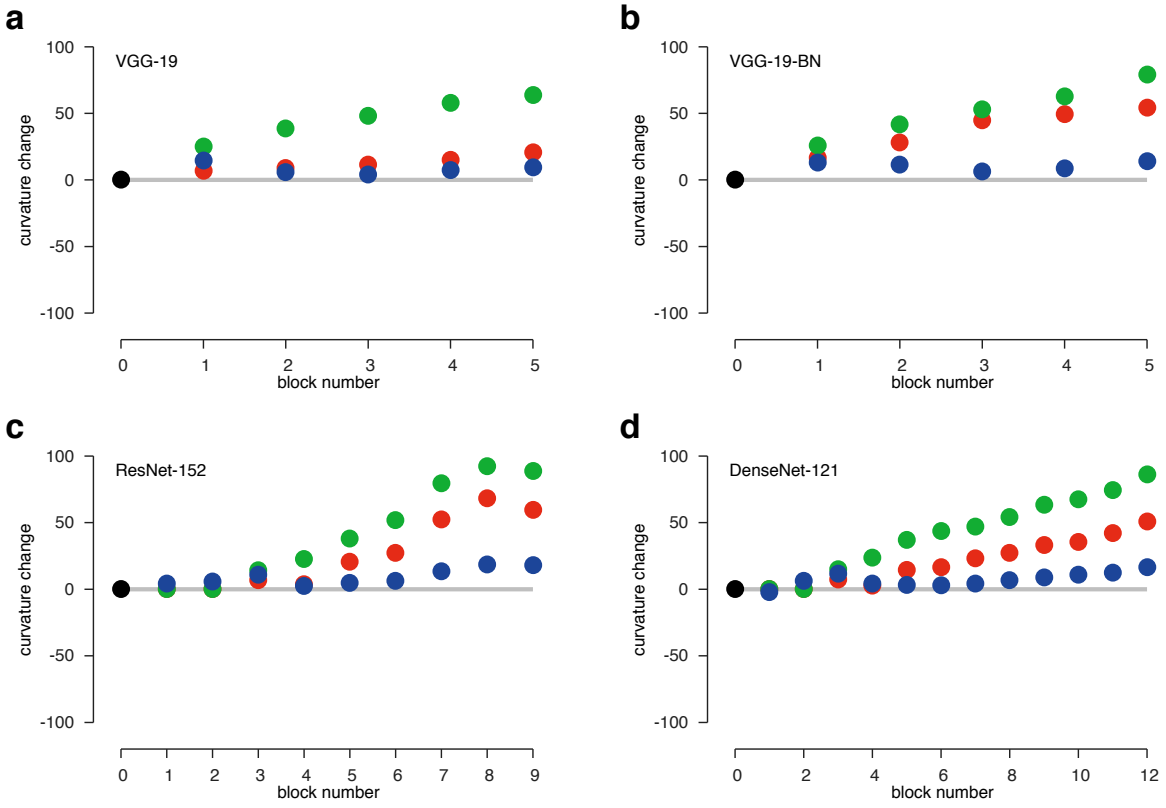


Figure 2.11 Changes in curvature in contemporary deep neural network architectures. Despite their strong performance in object recognition, none of these architectures straighten natural videos. (**a**) 19-layer VGG architecture(Simonyan & Zisserman, 2015). (**b**) 19-layer VGG architecture with batch normalization(Ioffe & Szegedy, 2015). (**c**) 152-layer Residual Network architecture(He et al., 2016). (**d**) 121-layer Dense Network architecture(Huang et al., 2017).

2.8 Discussion

We have introduced temporal straightening, a principle that 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 predictability to a specific timescale, corresponding to the interval between frames. Despite these restrictions, we have found human perceptual capabilities (and models thereof) to behave as predicted. How would these results generalize to the predictability of continuous streams of images, over different time scales? We used sequences with sampling rates roughly matched to the integration times of photoreceptors (30 frames/second or less), whose response would likely be similar to those under static presentation. However, downstream areas that process visual input over longer timescales would likely 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, therefore seems necessary to

test predictability or 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 (Barlow, 1961). Efficient coding posits that sensory representations are structured so as to preserve information in natural signals, while reducing redundancy and minimizing the use of neural resources (e.g., 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. Beyond this bottleneck, however, coding efficiency may no longer suffice to fully explain the form or specifics of sensory processes (Barlow, 2001; Simoncelli & Olshausen, 2001; Machens et al., 2005; Geisler, 2008). 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 likely 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 (Srinivasan et al., 1982; Rao & Ballard, 1999; Tishby et al., 1999; Bialek et al., 2006; Palmer et al., 2015). Temporal straightening offers a simple and readily testable instantiation of the temporal prediction hypothesis.

Temporal straightening also has some similarity with the "untangling hypothesis" that has been proposed as a normative explanation for the visual representations underlying object recognition capabilities (DiCarlo & Cox, 2007). 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 their visual appearance that arises from changes in viewing conditions and configuration. This hypothesis posits that the goal of the system is to produce a population representation that can be linearly decoded for object categorization, 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. As such, it could provide a practical means of learning such an untangled representation in an unsupervised manner, one of the most important open problems in machine learning (LeCun et al., 2015). 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 many 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 response 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 (Fukushima, 1980; Serre et al., 2007; Yamins et al., 2014). The curvature estimation methodology we have developed is agnostic to the partic-

ular form of experimental measurement, and we have begun to explore its application to physiological data (spiking responses recorded with multi-electrodes; Bai et. al. SfN 2018) to directly evaluate the curvature of representations in different visual areas (relatively to their respective noise distributions), and their contributions to perceptual straightening.

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 (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Berardino et al., 2017). By failing 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. On the other hand, 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 (Hénaff & Simoncelli, 2015), and we have used the perceptual straightness of these sequences as a measure for comparing candidate models of the human visual system (O.J. Hénaff, R.L.T. Goris, E.P. Simoncelli, *Cosyne Abstr.*, II-72, 2016).

While 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 force for 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 (Li & DiCarlo, 2008) or scales (Li & DiCarlo, 2010) can change the invariance properties of single cells in visual area IT, as well as the robustness of perceptual discriminability in human observers (Cox et al., 2005). 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 proven useful, both in testing for the efficiency of visual representations, or in showing that they can be learned by maximizing efficiency in the representation of naturally occurring stimuli (Barlow, 1961; Barlow, 2001; Simoncelli & Olshausen, 2001). Temporal straightening, as a task-relevant generalization of efficient coding, holds promise to fulfill an analogous role in higher-level visual areas.

2.9 Supplementary Methods

Stimuli

We measured the perceptual curvature of 12 natural image sequences that are representative of the diversity found in real videos (experiment 1; **Fig. 2.12,2.13**, blue path). To constrain our choice of sequences, we established a list of attributes that distinguish natural

	objects	textures	camera	rigid	flexible
water		✓			✓
prairie		✓			✓
egomotion		✓	✓	✓	
ice		✓		✓	✓
dam		✓			✓
leaves		✓		✓	✓
bees	✓			✓	✓
chironomus	✓				✓
Dogville	✓		✓	✓	✓
smile	✓			✓	✓
walking	✓			✓	✓
boats	✓		✓	✓	

Table 2.1 Diversity of the natural sequences used in our experiment.

videos. These pertain to the content of the videos (discrete, isolated objects vs. dense textures) as well as the types of motion and transformations over time (camera motion, rigid object motion and flexible/articulated object motion). **Table 2.1** indicates the diversity of the chosen set. We obtained 8 of these (*water*, *prairie*, *egomotion*, *ice*, *bees*, *carnegie-dam*, *leaves-wind*, *chironomus*) from the Chicago Motion Database, 1 from a feature film (*Dogville*; Lions Gate Entertainment, 2003), and 2 from the LIVE Video Quality Database (Seshadrinathan et al., 2010) (*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 f/s). However, because curvature can only be resolved when successive frames are sufficiently discriminable, we temporally downsampled 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, approximately 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 3–4 observers (41 sequence/observer pairs).

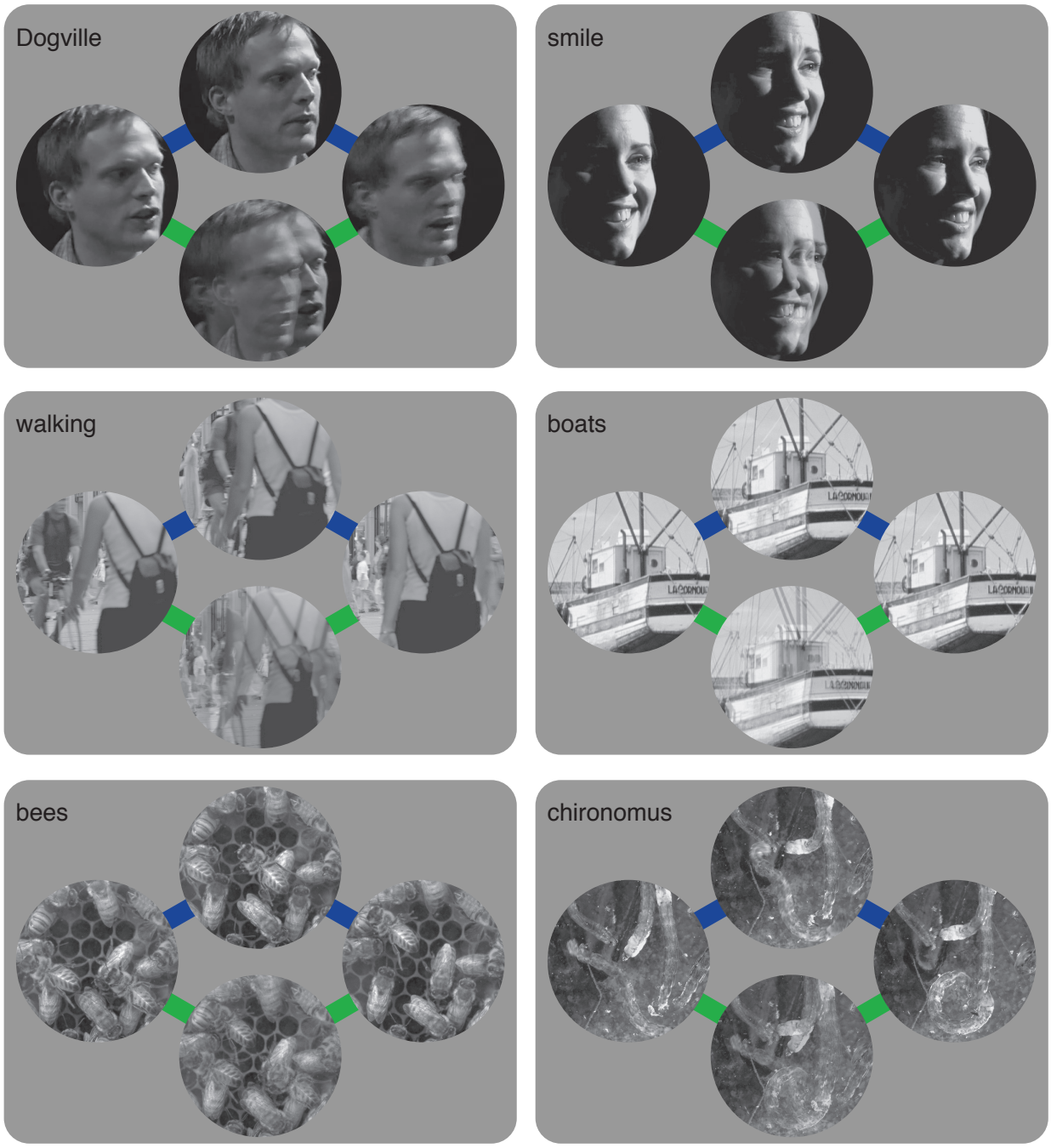


Figure 2.12 Initial, middle, and final frames from the first 6 natural and artificial sequences used in our experiment. Natural image sequences follow the top (blue) path, whereas artificial sequences follow the bottom (green) path between the same end-points.

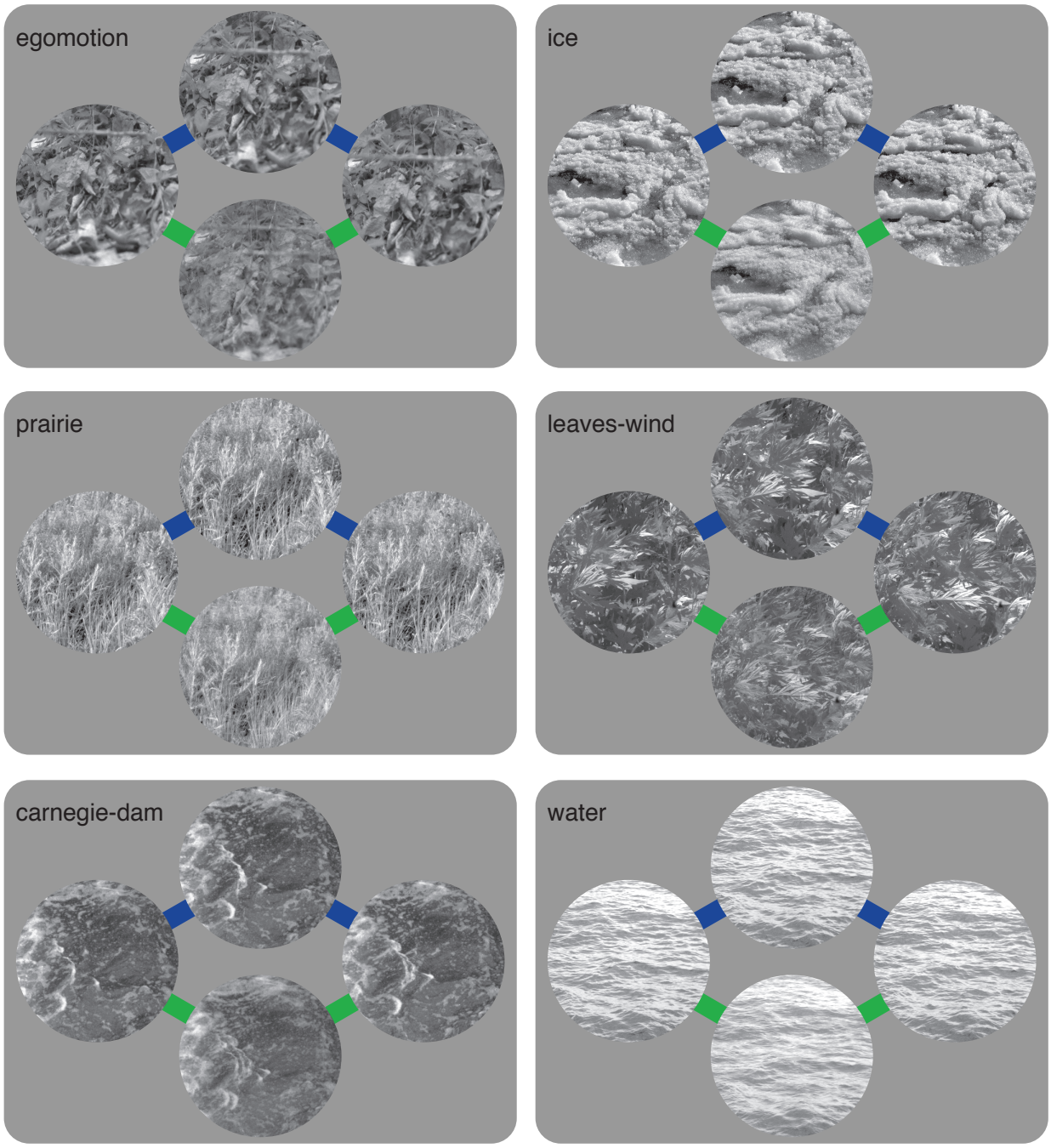


Figure 2.13 Initial, middle, and final frames from the last 6 natural and artificial sequences used in our experiment. Natural image sequences follow the top (blue) path, whereas artificial sequences follow the bottom (green) path between the same end-points.

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 2–4 observers per sequences and 35 sequence/observer pairs.

Each observer also was shown an artificial sequence that faded linearly between the first and the last frame of the corresponding natural sequence (experiment 2; **Fig. 2.12,2.13**, green path). Finally, 9 of these observers also viewed 10 natural, intensity-linear sequences that were generated by manipulating the contrast of a single frame (25 sequence/observer pairs; experiment 3). Five such sequences varied the contrast of the first frame of the *water*, *walking* (twice), *bees* and *boats* sequences, respectively, from 50% to 100% in equal 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.

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 humans 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 populations, which yielded similar results (**Fig. 2.14**, top row). Specifically, the simulated populations we present in **Figures 2.6-2.8** display a median change in

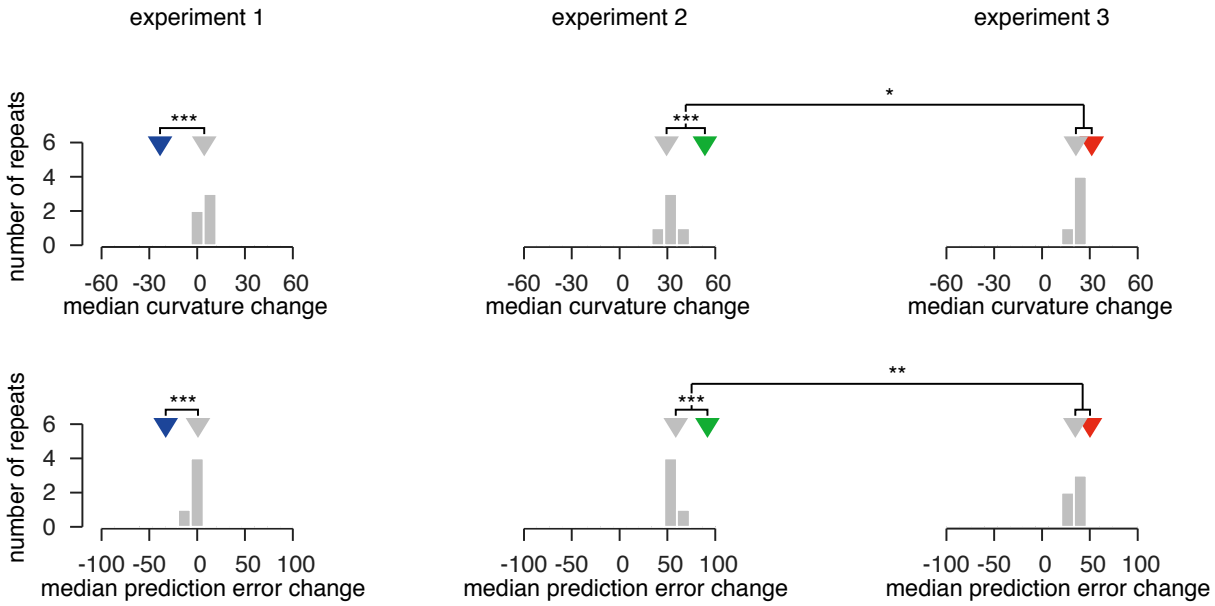


Figure 2.14 Recovery analysis across simulated populations. Top row: distribution of curvature change across simulated controls. In **Fig. 2.6-2.8** we show a single, typical, population of simulated controls (whose median change in curvature is depicted here by a gray arrow). The simulation process is inherently variable, as is the subsequent recovery due to finite numbers of subjects and trials. Here we evaluate the dispersion of the median curvature change across repetitions of the simulation and recovery procedure (gray histogram). Experiments 1 and 2: the curvature change for human observers is much larger than for any of the simulated controls ($p < 0.001$, two-tailed Z-test). Experiment 3: human observers increase curvature relatively to controls much less than in experiment 2 ($p = 0.02$, two-tailed Z-test). Bottom row: distribution of prediction error change across simulated controls. Similar results as for curvature change (experiment 1: $p < 0.001$, experiment 2: $p < 0.001$, experiment 3: $p = 0.004$). Together, these experiments show that the typical simulated populations shown in **Fig. 2.6-2.8** are representative of the distribution across simulated populations.

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 average change in curvature observed in human observers is significantly greater than for all simulated populations ($p < 0.001$, two-tailed Z-test; **Fig. 2.14**, left and middle). For experiment 3, the median change in curvature displayed by human observers relatively to simulated controls is significantly smaller than in experiment 2 ($p = 0.02$; **Fig. 2.14**, right). The same was true for changes in prediction error (**Fig. 2.14**, bottom row; experiment 1: $p < 0.001$, experiment 2: $p < 0.001$, experiment 3: $p = 0.004$). As such, the simulated populations presented in **Figures 2.6-2.8** 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 a one of primary visual cortex. The retinal model composes spatial center-surround filtering, luminance and contrast gain control and a rectifying nonlinearity (Berardino et al., 2017). The model of primary visual cortex uses a set of multiscale, oriented and band-pass filters (a "steerable pyramid" (Simoncelli & Freeman, 1995)), followed by squaring, summing over quadrature pairs, and a square-root nonlinearity to mimic the action of complex cells (Adelson & Bergen, 1985). We used 6 scales, and 4 orientations, excluding the high- and low-pass residual bands. The retinal model was optimized to match foveal perceptual discriminability judgments of human observers (Berardino et al., 2017), but our images were presented at 2 degree eccentricity. To approximate the loss of visual acuity in the parafovea (Green, 1970), we spatially downsampled 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 (downsampling 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.

As a control, we also evaluated the curvature of the same sequences as represented in each layer of a convolutional neural network (known as "AlexNet") trained for object recognition (Krizhevsky et al., 2012). The network contains a sequence of 5 rectified convolutional layers, with max pooling after the 1^{st}, 2^{nd} and 5^{th} 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 (Simonyan & Zisserman, 2015) with and without batch normalization (Ioffe & Szegedy, 2015), a 152-layer Residual Network (He et al., 2016), and a 121-layer Dense Network (Huang et al., 2017), 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.

Chapter 3

Neural straightening of natural videos

3.1 Introduction

In the previous chapter, we found behavioral evidence for the hypothesis that human observers straighten the time-course of natural videos, thereby making them more predictable. These measurements likely reflect the structure of the neural representations we use to make those perceptual judgments, but do not speak to where in the brain these neurons are nor how their representations are formed. In particular, it could be that all of the perceptual effects we uncovered are the result of computations performed by higher visual areas that encode the properties of objects and scenes. On the other hand, if the entire visual hierarchy has learned to straighten natural videos, we might expect individual areas to straighten the features they are selective for. Our model of early visual areas suggests that their computations could be used to support perceptual straightening, so we set out to perform a direct physiological test of neural straightening in primary visual cortex. This chapter, which describes the results of a collaboration with Yoon Bai, Julie Charlton, Ian

Nauhaus and Robbe Goris, provides evidence for neural straightening of natural videos in macaque primary visual cortex.

Specifically, we designed a physiological experiment based on the same natural and artificial sequences used in our perceptual experiments. While recording the activity of populations of cells in primary visual cortex, we presented individual frames from these sequences. By organizing the responses to individual frames according to their temporal ordering, we have direct access to the trajectory described by an image sequence in the recorded neural space. We then evaluate the curvature of this trajectory and compare it to that of the intensity-domain trajectory. These measurements show that, consistent with human discriminability judgments, primary visual cortex straightens the time-course of natural videos while distorting artificial ones. Moreover, a significant portion of perceptual straightening can be attributed to the neural straightening found in V1.

These measurements also allow us to address one of the limitations of our perceptual measurements, which only evaluate the predictability of natural videos at a particular time-scale. In contrast to discriminability judgments, which constrain the relative locations of pairs of frames, neural recordings provide us with the absolute locations of individual frames in neural space. As a result, these measurements allow us to assess the global structure of these sequences, in addition to the local curvature we measured in the previous chapter. These results show that the straightening performed by the visual system could exist at multiple timescales, enabling a range of different predictions.

3.2 Methods

Stimuli

The stimuli used in these experiments are similar to those used in the previous chapter, and we only recall them only briefly here. We used the *water*, *prairie*, *egomotion*, *bees*, *carnegie-dam*, *leaves-wind*, *chironomus*, *Dogville*, *smile* and *walking* sequences from our perceptual experiments (**Fig. 2.12,2.13**). We normalized each sequence to have a mean luminance of 0.25 (on a range of [0,1]) and standard deviation of 0.15. Frames were vignetted with a flat-topped raised cosine whose diameter was 15°. Frames from all sequences were presented in a random order, for 200 ms each with a 100 ms inter-stimulus interval. We presented these images at their original resolution (512×512 pixels) and at half the original resolution (upsampling the 256×256 pixel central crop to 512×512 pixels). As in the previous experiment, we presented the original, natural sequences as well as artificial sequences created by linearly interpolating between the end-frames of the natural sequence. In all, 2 classes (natural and artificial) of 10 sequences each were presented at 2 different resolutions. Each sequence having 11 frames, 440 unique images were presented over the course of an experiment, each one being repeated 50 times.

Physiology

The physiological data reported in this thesis were collected from three anesthetized, paralyzed macaque monkeys (Macaca fascicularis). Animals were initially anesthetized with ketamine (10 mg/kg, i.m.) and pretreated with atropine (0.04 mg/kg, i.m.). Anesthesia was maintained throughout the experiment with sufentanil citrate (4-20 µg/kg/h, i.v.). Animals were paralyzed using pancuronium bromide (0.1-0.2 mg/kg/h, i.v.) and artificially

ventilated using a small animal respirator (Harvard Apparatus or Ugo Basile). The EKG, EEG, SpO2, heart rate and body temperature were monitored continuously to judge the animal's health and maintain proper anesthesia. Eyes were dilated with 1% atropine and corneas protected with contact lenses. Refraction of the eyes were determined for a monitor distance at 65 cm. We used neural responses recorded from electrodes to determine ophthalmic lenses yielding the best spatial frequency tuning curves. Retinoscopy was also performed to determine ophthalmic lenses.

The skull over the occipital lobe was thinned and a custom-made imaging well was attached over regions of V1 and V2 to collect intrinsic imaging data. Within the well perimeter, we made craniotomies and durotomies (approximately 20 mm $\times$ 20 mm) to expose the brain. Periodic gratings (at various orientations and spatial frequencies) were presented to reveal functional maps such as orientation columns and ocular dominance columns in V1. Functional maps derived from intrinsic imaging determined the V1/V2 border. We made extracellular recordings from areas V1 and V2 using multi-channel electrodes (NeuroNexus). Spike waveforms were isolated from extracellular voltage traces collected from each channel using Kilosort (Pachitariu et al., 2016). For every isolated unit, spiking activity was collected within a time window that was synchronized to stimulus presentation. Response latencies of 40 ms were assumed for V1 recordings and 50 ms for V2 recordings.

3.3 Estimating neural curvature

Each frame in a sequence will cause the cells we record from to fire at a different rate. We are interested in the trajectory of the vector of rates over time (**Fig. 3.1,** black trajectory), and its curvature in particular. On a given trial, we observe a noisy version of this trajectory, due to the stochastic nature of neural responses. Having measured these

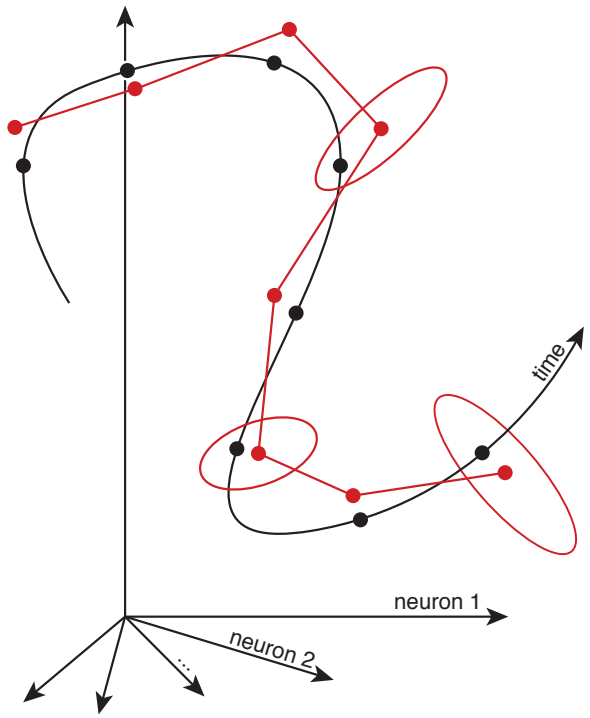


Figure 3.1 Inferring neural curvature. Each frame in a sequence induces a vector of responses, whose average is depicted in black. Estimating this trajectory from a finite number of trials (using a trial average) appears much more curved (red trajectory; **Fig. 3.2,** left). However, each set of observations for a given frame is consistent not just with the trial-average, but with a family of vectors determined by variance of each neuron, and the noise correlations between neurons (red ellipses). Searching for a value of the curvature that is consistent with all plausible trajectories, we arrive at a robust estimate of neural curvature (**Fig. 3.2,** right).

noisy trajectories over many trials, it is tempting to estimate the curvature of the true, underlying trajectory by computing a trial-averaged trajectory and its curvature (**Fig. 3.1,** red trajectory). This procedure, however, is plagued by the same biases we encountered in the previous chapter (**Fig. 3.2,** left). Indeed, because we must estimate the trajectory from a finite number of trials, some amount of noise will always contaminate our trial-averaged trajectory, making it appear more curved than it really is (**Fig. 3.1,** difference between red and black trajectories). We have developed a nearly unbiased procedure for estimating the curvature of a neural trajectory by considering not just the single most plausible rate-trajectory (i.e. the trial average), but all trajectories that are reasonably consistent with the data (**Fig. 3.1,** red ellipses). We evaluate the likelihood of any neural trajectory by modeling the variance-to-mean relationship of each neuron (Goris et al., 2014) and the correlations between neurons (Rabinowitz et al., 2015). By reporting the value of curvature

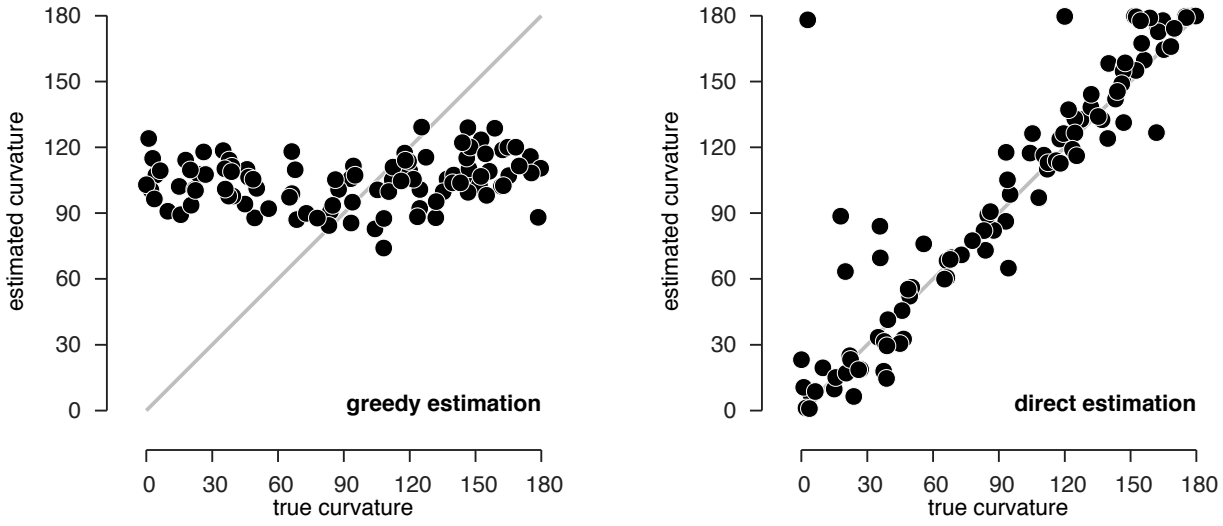


Figure 3.2 Recovery analysis for neural curvature estimation. We simulated 100 neural trajectories whose curvature was randomly sampled from from 0° to 180°. All other parameters (number of cells, overall firing rates, super-Poisson variability, noise correlations, dimensionality, path length) were chosen to match those of a neural trajectory in our V1 dataset. We then simulated an experiment with the same number of trials as in the actual experiment, and attempted to recover the known curvature from these data. (**a**) Greedy estimation, in which we first estimate the most likely trajectory (by averaging over trials), and then calculate its curvature, is plagued by significant bias. (**b**) Direct curvature estimation, which reports the value of curvature most consistent with *all* plausible trajectories, has minimal bias and relatively small variance.

that is most consistent with all plausible trajectories, we arrive at an estimate of curvature which is largely unbiased (**Fig. 3.2,** right).

Problem statement

For a given sequence, we observe the responses of D cells to T images, repeated K times. Let $n^k_{t,d}$ be these spike count of cell d in response to image t on trial k, $\boldsymbol{n}^k_t$ the vector of such responses across cells, and $\boldsymbol{N}$ the dataset of such vectors. On the surface, estimating the curvature of this trajectory seems much more straightforward than in the previous chapter. We simply need to estimate the underlying rate of each neuron in response to each image

by computing a trial average $\hat{\boldsymbol{\lambda}}_t$, and measure the curvature $\hat{c}$ of its trajectory over time:

$$\begin{aligned}
\hat{\boldsymbol{\lambda}}_t &= \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{n}_t^k \\
v_t &= \hat{\boldsymbol{\lambda}}_t - \hat{\boldsymbol{\lambda}}_{t-1} \\
\hat{v}_t &= \frac{v_t}{\|v_t\|} \\
c_t &= \arccos(\hat{v}_t \cdot \hat{v}_{t+1}) \\
\hat{c} &= \frac{1}{T-2}\sum_{t=2}^{T-1} c_t
\end{aligned}$$

However simple, this estimation procedure suffers from a considerable bias (**Fig. 3.2,** left). Rather than committing to a single estimate of the trajectory and report its curvature, we seek the value of the curvature that is most consistent with all plausible trajectories. More formally, we wish to infer an estimate c^* of the curvature that is most likely given the data $\boldsymbol{N}$, averaging over all plausible trajectories $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda}_t\}_t$:

$$\begin{aligned}
c^* &= \arg\max_c \mathbb{P}(\boldsymbol{N}|c) \\
&= \arg\max_c \int \mathbb{P}(\boldsymbol{N}|\boldsymbol{\Lambda})\mathbb{P}(\boldsymbol{\Lambda}|c)d\boldsymbol{\Lambda}
\end{aligned}$$

This requires us to define a likelihood function $\mathbb{P}(\boldsymbol{N}|\boldsymbol{\Lambda})$ to measure how plausible a given trajectory $\boldsymbol{\Lambda}$ is, given the observed spike counts. In particular, we will need to model the variance-to-mean relationship of individual neurons, as well as their correlations. We also need to define the prior $\mathbb{P}(\boldsymbol{\Lambda}|c)$ which expresses how compatible these trajectories are with a particular value of the curvature. Finally, we will need to (approximately) perform the high-dimensional integral over plausible trajectories. Given the numbers of frames (11 in

our experiments) and cells (26–61 in our experiments), a trajectory lives in a continuous space with hundreds of dimensions and exact computation of this integral is intractable. Variational tools allow us to derive a lower bound on the objective we wish to maximize, which we optimize instead, leading to largely unbiased curvature estimates (**Fig. 3.2,** right).

Likelihood

Let $\boldsymbol{n}_t^k$ be the vector of spike counts on a given trial, and $\boldsymbol{\lambda}_t$ the associated rate vector. Given that images are presented over the course of the experiment in a random order, we can assume that the spike counts in response to different images and on different trials are independent from one another:

$$\mathbb{P}(\boldsymbol{N}|\boldsymbol{\Lambda}) = \prod_{k=1}^{K}\prod_{t=1}^{T}\mathbb{P}(\boldsymbol{n}_t^k|\boldsymbol{\lambda}_t)$$

Our problem therefore reduces to modeling a vector of spike counts on a given trial $\boldsymbol{n}$ with a vector of spike rates $\boldsymbol{\lambda}$.

Poisson

This simplest likelihood function describes the spike counts from every neuron as independent from one another, and distributed according to a Poisson given the underlying rate:

$$\begin{aligned}\mathbb{P}(\boldsymbol{n}|\boldsymbol{\lambda}) &= \prod_{d=1}^{D}\mathbb{P}_{\text{Poisson}}(n_d|\lambda_d)\\ &= \prod_{d=1}^{D}\frac{\lambda_d^{n_d}\exp(-\lambda_d)}{n_d!}\end{aligned}$$

Modulated Poisson

The previous model assumes neurons spike independently from one another, and yet neurons are known to exhibit shared variability. Following the work of Goris et al., 2014 and Rabinowitz et al., 2015, we model the 'noise correlations' between neurons by assuming their rates are modulated by a common gain factor $\boldsymbol{g}$. Specifically, we assume that they remain independent and Poisson-distributed *conditioned* on a multiplicative gain factor $\boldsymbol{g}$:

$$\begin{aligned}\mathbb{P}(\boldsymbol{n}|\boldsymbol{\lambda}) &= \int d\boldsymbol{g}\ \mathbb{P}(\boldsymbol{g})\ \mathbb{P}(\boldsymbol{n}|\boldsymbol{\lambda}, \boldsymbol{g}) \\ &= \int d\boldsymbol{g}\ \mathbb{P}(\boldsymbol{g})\ \mathbb{P}_{\text{Poisson}}(\boldsymbol{n}|\boldsymbol{\lambda} \odot \boldsymbol{g})\end{aligned}$$

where $\odot$ denotes the element-wise product between two vectors. If $\boldsymbol{g}$ has independent components which are Gamma-distributed, this integral has an analytical solution (a product of negative binomial distributions) and we recover the 'Modulated Poisson' model from (Goris et al., 2014). Allowing for the components of $\boldsymbol{g}$ to be correlated however enables us to model dependencies between spike counts. Specifically, we choose the gain to be a multivariate, log-normal random variable, whose components are correlated:

$$\boldsymbol{g} = \exp[\boldsymbol{\epsilon}_p + \boldsymbol{L}\boldsymbol{\epsilon}_s]$$

where $\boldsymbol{\epsilon}_p$ and $\boldsymbol{\epsilon}_s$ are private and shared sources of noise respectively (which we concatenate into a single vector $\boldsymbol{\epsilon}$). The coupling matrix $\boldsymbol{L}$ captures the dependencies between neurons, and is shared across images and trials. In the case of shared gain fluctuations, evaluating the likelihood involves an integral over many dimensions (i.e. the number of cells) which is not practical. Nevertheless, variational inference provides a tractable lower bound on this

likelihood which we can optimize instead (Jordan et al., 1999):

$$\begin{aligned}\log \mathbb{P}(\boldsymbol{n}|\boldsymbol{\lambda}) &= \log \int d\boldsymbol{g}\ \mathbb{P}(\boldsymbol{g})\ \mathbb{P}(\boldsymbol{n}|\boldsymbol{\lambda}, \boldsymbol{g}) \\ &= \log \int d\boldsymbol{\epsilon}\ \mathbb{P}(\boldsymbol{n}|\boldsymbol{\lambda}, \boldsymbol{\epsilon})\ \mathbb{P}(\boldsymbol{\epsilon}) \\ &\geq \mathbb{E}_{\mathbb{P}_\phi(\boldsymbol{\epsilon}|\boldsymbol{n})}\left[\log \mathbb{P}(\boldsymbol{n}|\boldsymbol{\lambda}, \boldsymbol{\epsilon})\right] - D_{KL}[\mathbb{P}_\phi(\boldsymbol{\epsilon}|\boldsymbol{n})||\mathbb{P}_\theta(\boldsymbol{\epsilon})]\end{aligned}$$

where the likelihood $\mathbb{P}(\boldsymbol{n}|\boldsymbol{\lambda}, \boldsymbol{\epsilon})$ is again Poisson conditioned on the value of $\boldsymbol{\epsilon}$. We choose the prior $\mathbb{P}_\theta(\boldsymbol{\epsilon})$ and approximate posterior $\mathbb{P}_\phi(\boldsymbol{\epsilon}|\boldsymbol{n})$ to be Gaussian with a diagonal covariance, and their KL divergence is thus available in closed form. The mean and variance of the prior over private noise $\boldsymbol{\epsilon}_p$ are chosen such that the mean of the corresponding gain is 1, and its variance accounts for the amount of super-Poisson variability in each neuron. The prior over shared noise $\boldsymbol{\epsilon}_s$ is a standard normal. The trajectory of the rate $\boldsymbol{\Lambda}$, as well as the parameters of the prior and approximate posterior are all learned simultaneously, using stochastic gradient descent (Kingma & Welling, 2013; Kingma & Ba, 2014).

Inference

Having connected the observed data to the trajectory of the vector of rates, we now re-parameterize these rates in terms of the quantities of interest, namely, the length and curvature of the trajectory. Specifically, we start out by re-parameterizing each rate vector $\boldsymbol{\lambda}_t$ as a point-wise function of a real-valued 'pre-rate' vector $\boldsymbol{y}_t$:

$$\boldsymbol{\lambda}_t = f_\lambda(\boldsymbol{y}_t)$$

where f_λ is a elementwise, rectifying function such as a soft-plus. This ensures the rate is positive, but we can also choose f_λ to reflect the geometry of the likelihood function. Finally, since the T 'pre-rate' vectors $\boldsymbol{y}_t$ live in a $T-1$ dimensional subspace, we need only model their relative locations in this low-dimensional subspace. Specifically, we re-parameterize these high-dimensional vectors $\boldsymbol{y}_t$ as a functional of low-dimensional vectors $\boldsymbol{x}_t$ and an embedding matrix $\boldsymbol{E}$:

$$\boldsymbol{y}_t = \boldsymbol{E}\ \boldsymbol{x}_t$$

Without knowing this subspace in advance, we learn (and marginalize over) the embedding matrix $\boldsymbol{E}$. As in the previous chapter, we now re-parameterize the trajectory x as a function of distances d_t, curvatures c_t, and directions of curvature $\hat{a}_t$:

$$\begin{aligned}
\boldsymbol{x}_t &= \boldsymbol{x}_{t-1} + \boldsymbol{v}_t \\
\boldsymbol{v}_t &= d_t \hat{\boldsymbol{v}}_t \\
\hat{\boldsymbol{v}}_t &= \cos(c_t)\hat{\boldsymbol{v}}_{t-1} + \sin(c_t)\hat{\boldsymbol{a}}_t
\end{aligned}$$

These are themselves re-parameterized in terms of real-valued variables, on which we place the following priors:

$$\begin{aligned}
d_t &= f_d(\ z_t^d) & z_t^d &\sim \mathcal{N}(f_d^{-1}(d^*),\ \sigma_d^2\ ) \\
c_t &= \quad z_t^c & z_t^c &\sim \mathcal{N}(\quad c^*\ ,\ \sigma_c^2\ ) \\
\hat{\boldsymbol{a}}_t &= f_{\hat{a}_t}(\boldsymbol{z}_t^{\boldsymbol{a}}) & \boldsymbol{z}_t^{\boldsymbol{a}} &\sim \mathcal{N}(\quad \boldsymbol{0}\ ,\ \boldsymbol{\Sigma_a}) \\
\boldsymbol{E} &= f_{\boldsymbol{E}}(\ \boldsymbol{Z}^{\boldsymbol{E}}) & \boldsymbol{Z}^{\boldsymbol{E}} &\sim \mathcal{N}(\quad \boldsymbol{0}\ ,\ \boldsymbol{I}\ )
\end{aligned}$$

where f_d is a smooth rectifying function, $f_{\hat{\boldsymbol{a}}_t}$ ensures that $\hat{\boldsymbol{a}}_t$ is of unit length and orthogonal to $\hat{\boldsymbol{v}}_{t-1}$ and $\boldsymbol{\Sigma_a}$ controls the effective dimensionality and aspect-ratio of the trajectory. The function $f_{\boldsymbol{E}}$ implements the Graham-Schmidt algorithm, which ensures that the columns of $\boldsymbol{E}$ are an orthonormal family of vectors in the high-dimensional neural space (i.e. $\boldsymbol{E}$ is a tight-frame). This guarantees that the metric properties of the trajectory $\{\boldsymbol{y}_t\}_t$ (i.e. all pairwise distances, path length, and curvature) are identical to those of the trajectory $\{\boldsymbol{x}_t\}_t$.

Finally, let $\boldsymbol{\theta} = \{d^*, c^*, \sigma_d, \sigma_c, \boldsymbol{\Sigma_a}\}$ be the parameters governing priors of the latent variables $\boldsymbol{z} = \{z_t^d, z_t^c, \boldsymbol{z}_t^{\boldsymbol{a}}, \boldsymbol{Z}^{\boldsymbol{E}}\}$. Let $\mathbb{P}_\phi(\boldsymbol{z}|\boldsymbol{N})$ be an approximate posterior over the latents. By combining the variational lower bound on these latents with the one obtained for the likelihood in the previous section, we derive our final objective:

$$
\begin{aligned}
\log \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{N}) &\geq -D_{KL}[\mathbb{P}_\phi(\boldsymbol{z}|\boldsymbol{N})||\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{z})] + \mathbb{E}_{\mathbb{P}_\phi(\boldsymbol{z}|\boldsymbol{N})}\left[\log \mathbb{P}(\boldsymbol{N}|\boldsymbol{z})\right] \\
&\geq -D_{KL}[\mathbb{P}_\phi(\boldsymbol{z}|\boldsymbol{N})||\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{z})] + \sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}_{\mathbb{P}_\phi(\boldsymbol{z}|\boldsymbol{N})}\left[\log \mathbb{P}(\boldsymbol{n}_t^k|\boldsymbol{z})\right] \\
&\geq -D_{KL}[\mathbb{P}_\phi(\boldsymbol{z}|\boldsymbol{N})||\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{z})] \\
&\quad + \sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}_{\mathbb{P}_\phi(\boldsymbol{z}|\boldsymbol{N})}\mathbb{E}_{\mathbb{P}_\phi(\boldsymbol{\epsilon}_t^k|\boldsymbol{n}_t^k)}\left[\log \mathbb{P}(\boldsymbol{n}_t^k|\boldsymbol{z}, \boldsymbol{\epsilon}_t^k)\right] - D_{KL}[\mathbb{P}_\phi(\boldsymbol{\epsilon}_t^k|\boldsymbol{n}_t^k)||\mathbb{P}_\theta(\boldsymbol{\epsilon}_t^k)]
\end{aligned}
$$

3.4 Neural straightening of natural videos

The perceptual results from the previous chapter indicate that natural image sequences are significantly straightened by the human visual system. Our modeling results suggest that a significant portion of this straightening could be accomplished by computations in the early stages of the visual system. We tested this hypothesis by presenting the same

stimuli used in our psychophysical experiments to macaque monkeys, and recording the activity of populations of cells in primary visual cortex. To visualize the trajectories of an example natural video in the domain of pixel-intensities and neural activity, we projected them onto the first principle components. In the intensity domain, this trajectory is highly curved (**Fig. 3.3a**, left; curvature = 54°). In the domain of neural responses, it is slightly straighter (**Fig. 3.3a**, middle; curvature = 52°). In the high-dimensional pixel-intensity and neural domains, this difference was much more prominent (pixel-domain curvature = 105°, neural curvature = 67°). Moreover, this change in curvature was robust across sequences and neural populations (**Fig. 3.3a**, right; median change in curvature = $-11°$, $p < 0.001$, two-tailed Wilcoxon signed-rank test).

We ensured that this change in curvature was not an artifact of our estimation method by simulating a family of neural trajectories that were matched to the ones we measured in terms of the firing rates they induced and their overall discriminability, but that displayed the same pairwise distance structure as the pixel-intensity trajectories. In particular, their curvature was identical to that in the intensity domain. When evaluated on these simulated trajectories, our curvature estimation method finds no change relatively to the intensity domain (change in curvature = $-1°$, $p = 0.42$). Moreover, this analysis shows that the reduction in curvature measured in the recorded neural trajectories is significantly greater than the variability in our estimates ($p < 0.001$, test on the difference between neural and simulated trajectories). Hence, consistently with our perceptual data and hierarchical modeling, primary visual cortex significantly contributes to the straightening of natural videos.

Contrarily to our behavioral results, in which trajectories are inferred from the relative locations (or discriminability) of nearby frames, these data provide the absolute positions

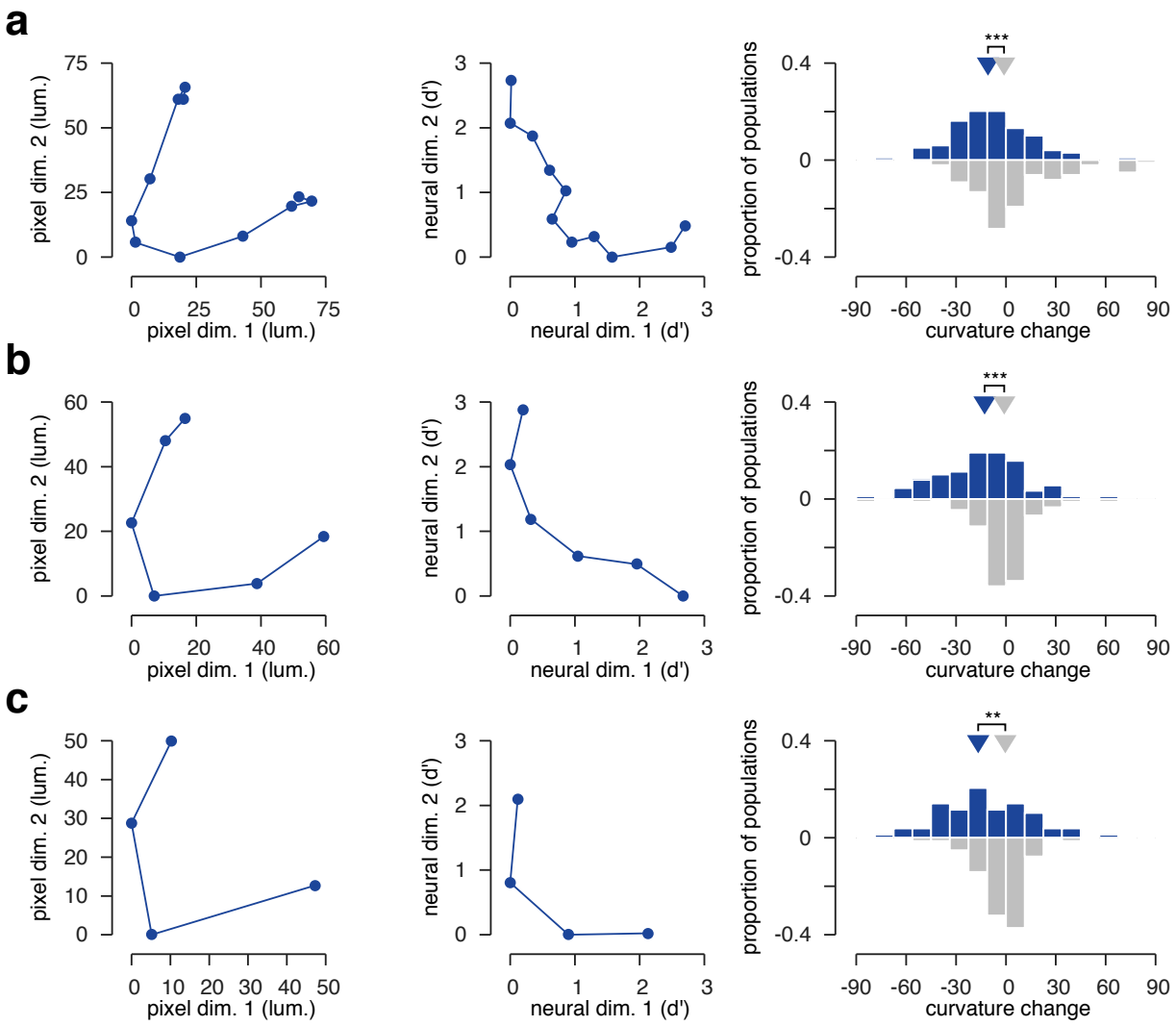


Figure 3.3 Neural straightening of natural image sequences at multiple timescales. (**a**) Neural straightening of natural image sequences at the finest timescale. Left, middle: Two-dimensional projections of an example sequence in the pixel-intensity (left) and neural (middle) domains. Right: Change in curvature from the intensity to the neural domain, for 20 natural image sequences and 5 neural datasets (100 neural trajectories in total). Blue histogram: neural curvature estimated from original recordings. Gray histogram: curvature estimated from simulated trajectories whose neural curvature is matched to intensity-domain curvature, with all other parameters matched to those of the neural trajectories. Triangles indicate the median of each distribution. *** $p < 0.001$. (**b**) Neural straightening of natural image sequences at a coarser timescale (subsampled by a factor of 2). (**c**) Neural straightening of natural image sequences at the coarsest timescale (subsampled by a factor of 3). ** $p < 0.01$.

of each frame in neural space. As a result, we can infer from them the global structure of each neural trajectory, in addition to the local curvature. In particular, we asked how curved these sequences are at multiple timescales. Indeed, measuring curvature between adjacent frames in a discretely sampled sequence indicates how predictable the trajectory is *at the timescale at which the trajectory was sampled*. Yet real world tasks require making predictions at many different timescales. How, then, does the property of neural straightening hold up at other timescales? We answered this question by constructing new sequences that were subsampled versions of the previous set. In **Fig. 3.3b** we subsample the neural data (and the associated video frames) by a factor of 2, and re-estimate the corresponding trajectories. At this coarser timescale, the difference between the low-dimensional pixel-intensity and neural trajectories is clear (**Fig. 3.3b**, left, middle; pixel-domain curvature = 42°, neural curvature = 30°), as it is in the high-dimensional ambient spaces (pixel-domain curvature = 111°, neural curvature = 32°). Moreover, this reduction is curvature was consistent across all sequences and neural populations (**Fig. 3.3b**, right; difference between neural and simulated trajectories = −9°, $p < 0.001$). Measuring curvature at an even coarser timescale (subsampling by a factor of 3) led to similar, although more variable (due to smaller amounts of data) results (**Fig. 3.3c**; difference between neural and simulated trajectories = −11°, $p = 0.025$). Together, these results show that the neural representation of natural videos in primary visual cortex facilitates their predictability at multiple timescales.

3.5 Neural distortion of artificial videos

As in the previous chapter, we wanted to verify that the straightening we have uncovered in primary visual cortex is targeted to natural videos. It could be that all videos, whether

or not they are natural, see their curvature reduced by the computations found in V1. The straightening hypothesis predicts that artificial sequences that are behaviorally irrelevant have to reason to be straight. Moreover, if we consider how rare straight sequences in high-dimensional spaces are, we would predict that unnatural sequence should instead be made more curved.

We tested this hypothesis by presenting the same pixel-linear sequences used in our perceptual experiments. Specifically, these sequences fade linearly from one frame in a sequence to another (distant) frame. The resulting intermediate frame are highly unnatural in that they contain a superimposition of two different images. Although an example sequence is perfectly straight in the intensity domain, it is curved in the domain of neural firing rates (**Fig. 3.4a**, left, middle; low-dimensional curvature = 42°, high-dimensional curvature = 82 °). This increase in curvature is robust across sequences and neural populations (**Fig. 3.4a**, right; median curvature increase = 64°) and significantly greater than the increase in curvature found for simulated trajectories whose curvature was equal to that of intensity-domain sequences (median increase in curvature = 33°; $p < 0.001$, difference between neural and simulated trajectories).

It could be that this increase in curvature is limited to fine timescales. Indeed, were the directions in which these sequences angle to cancel each other out over time, these artificial sequences could remain straight at coarser timescales. Hence, we measured the neural curvature of these sequences at multiple timescales and found a significant increase in curvature for these sequences at a coarser timescale (**Fig. 3.4b**, subsampling by a factor of 2; difference between neural and simulated trajectories = 31°, $p < 0.001$), and at the coarsest timescale (**Fig. 3.4c**, sampling by a factor of 3; difference between neural and simulated trajectories = 28°).

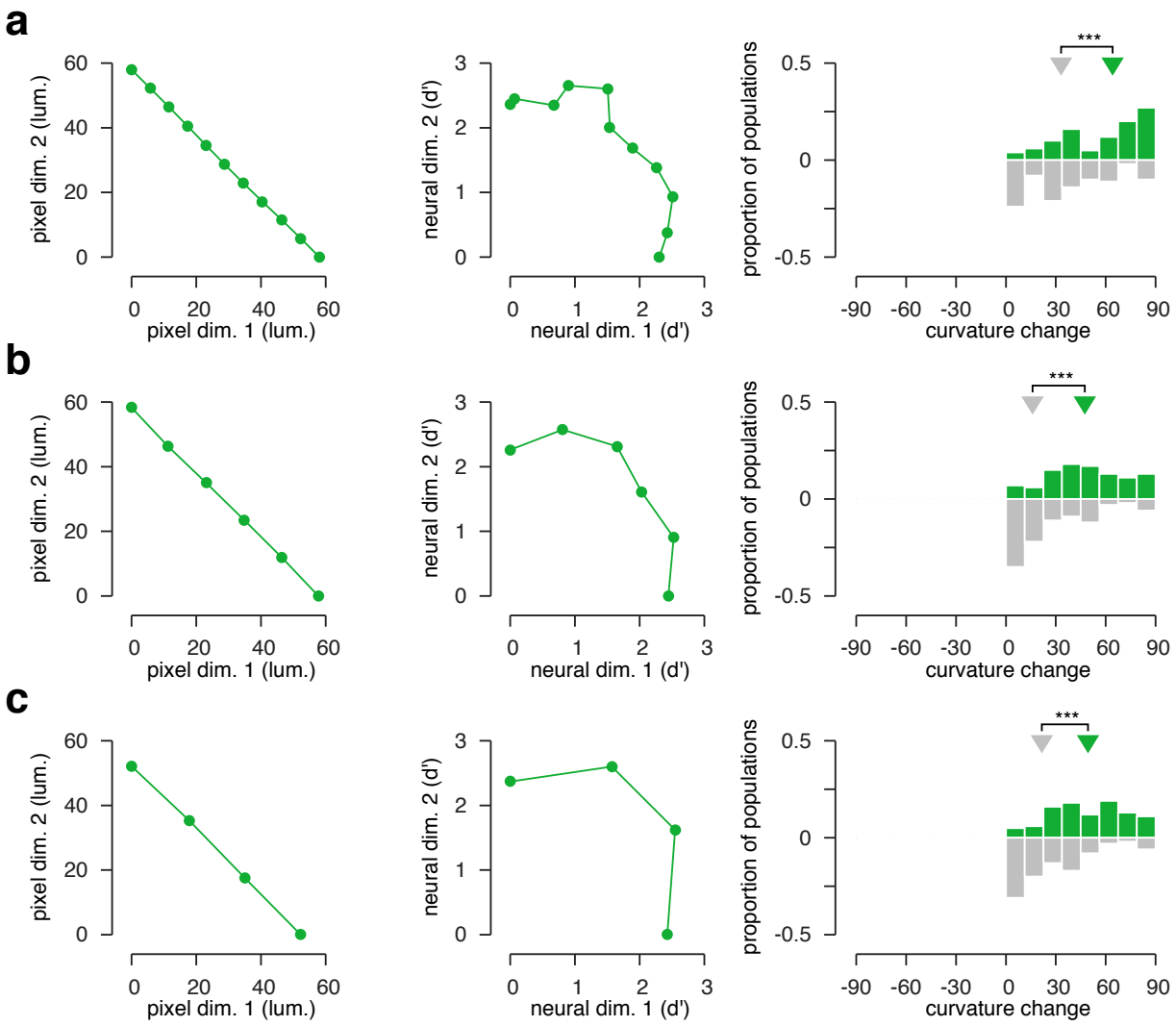


Figure 3.4 Neural distortion of artificial image sequences at multiple timescales. (**a**) Neural distortion of artificial image sequences at the finest timescale. Left, middle: Two-dimensional projections of an example sequence in the intensity (left) and neural (middle) domains. Right: Change in curvature from the intensity to the neural domain, for 20 natural image sequences and 5 neural datasets (100 neural trajectories total). Green histogram: neural curvature estimated from original recordings. Gray histogram: curvature estimated from simulated trajectories whose neural curvature is matched to intensity-domain curvature, with all other parameters matched to those of the neural trajectories. Triangles indicate the median of each distribution. *** $p < 0.001$. (**b**) Neural distortion of artificial image sequences at a coarser timescale (subsampled by a factor of 2). (**b**) Neural distortion of artificial image sequences at the coarsest timescale (subsampled by a factor of 3).

3.6 Neural straightening predicts perceptual straightening

Having confirmed that primary visual cortex contributes to the selective straightening of natural videos, we asked whether the magnitudes of these changes in curvature were consistent with those found perceptually. **Fig. 3.5a** shows that the curvature reduction for natural sequences found in primary visual cortex accounts for a significant fraction of that found perceptually (average curvature change (neural) = $-11°$, average curvature change (perceptual) = $-33°$). Moreover, we find this change in curvature to be well predicted by the two-stage model of visual cortex we used in the previous chapter. As a reminder, the first stage of this model captures the functional properties of the retina and lateral geniculate nucleus with center-surround filtering followed by luminance and contrast gain control. The second stage decomposes the output of the first into a set oriented sub-bands, and measures the local energy in each one. If the curvature estimated from our neural recordings reflects the activity of a combination of simple and complex cells, and simple cells represent an intermediate stage of computation (between the LGN and V1 complex cells), we would expect the neural curvature reduction to be contained within that found in each of the model stages. This is indeed what we find, as the curvature reduction in the model LGN is $-7°$ whereas the curvature reduction in the model V1 is $-23°$. Similarly, the average curvature increase in curvature was well predicted by our model and perceptual data (curvature increase for model LGN = $12°$, curvature increase for macaque V1 = $26°$, curvature increase for model V1 = $18°$, curvature increase for perception = $24°$).

Having verified the average curvature found in primary visual cortex, we asked whether the pattern of change in curvature across sequences was consistent with the one found perceptually. For each sequence, we computed the average change in curvature, across neural

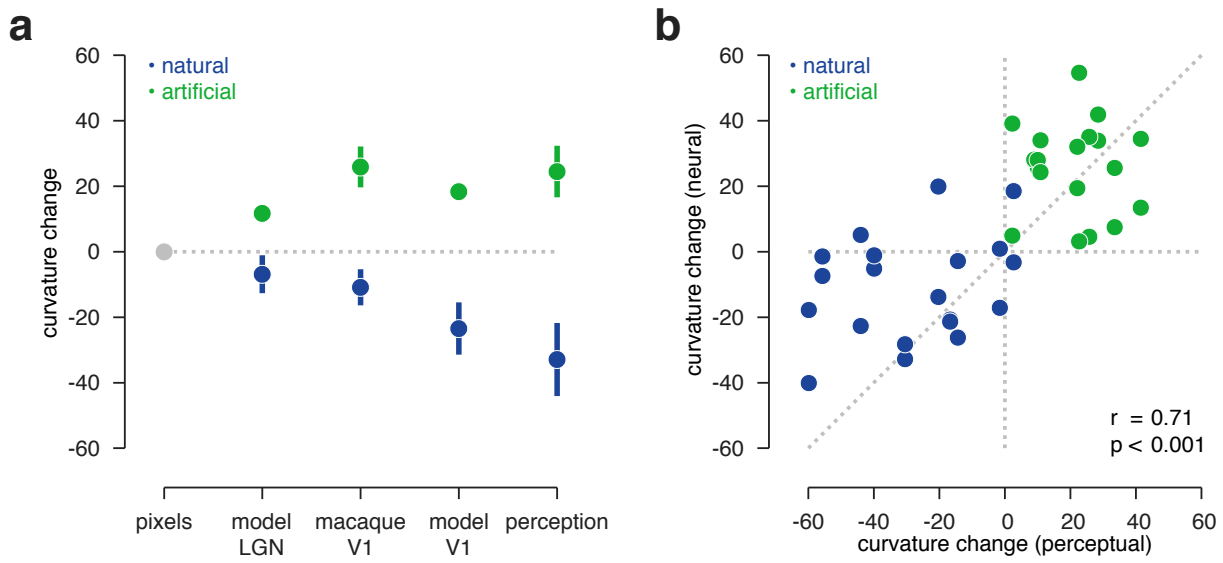


Figure 3.5 Neural changes in curvature are predictive of perceptual changes in curvature (**a**) Average curvature increase found in primary visual cortex accounts for a significant fraction of the perceptual change. The neural change in curvature is also well predicted by a hierarchical model of visual cortex, which cascades divisive normalization and non-linear pooling. (**b**) The pattern of change in curvature across sequences in primary cortex is significantly correlated with that found perceptually.

recordings for V1, and across human observers for perception. Although the reduction in curvature found perceptually is systematically greater than that found in primary visual cortex, the two patterns are highly correlated (**Fig. 3.5b**; $r = 0.71$, $p < 0.001$). From this analysis we conclude that although downstream computations will be necessary to fully explain our perceptual results, early visual areas including primary visual cortex provide a substantial contribution to the changes in curvature found perceptually.

3.7 Neural straightening beyond V1

Given the role of early visual areas in perceptual straightening, we asked whether downstream areas consolidate the neural straightening found in primary visual cortex. To that

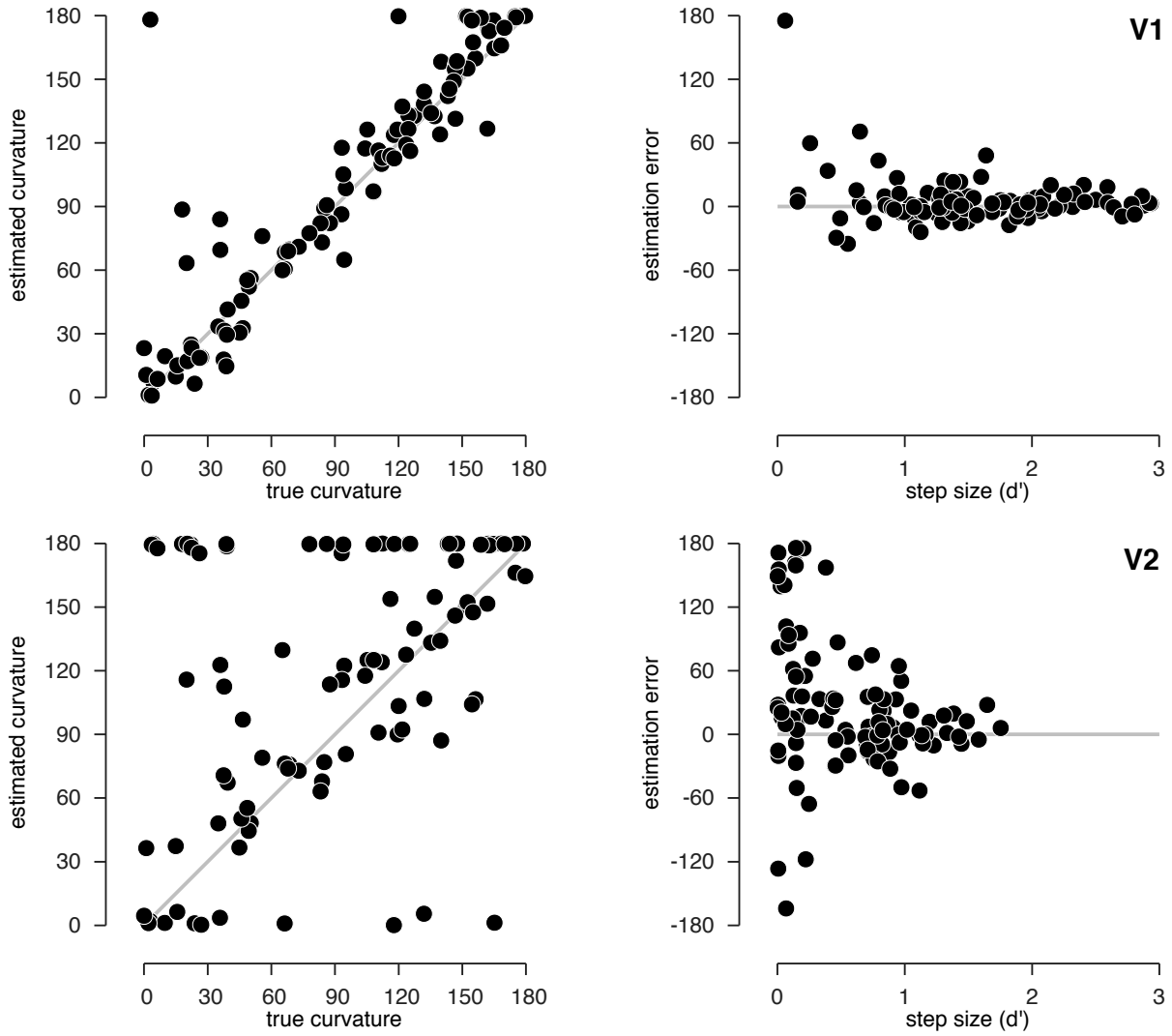


Figure 3.6 Recovery analysis for V1 and V2 populations. Top row: when simulated trajectories are matched to the parameters of V1 (e.g. in terms of their path length), we successfully recover known curvature values (left). Nonetheless, the precision with which we resolve neural curvature depends on the discriminability of successive frames. As the signal-to-noise ratio increases, so does the precision of our estimation (right). Bottom row: when simulated trajectories are matched to the parameters of V2, the recovery analysis fails (left). This can be largely explained by the fact that neural trajectories in V2 are significantly shorter than in V1 (right). In particular, very few V2 trajectories are in the regime in which the recovery is reliable (d' greater than 1).

effect, we gathered a dataset identical in size to the one we have discussed, from populations of neurons in the second visual area (V2). These data did not, however, enable us to answer the question we set out to test. We found that our ability to resolve neural curvature crucially depended on the overall path length of neural trajectories (**Fig. 3.6,** right column). Indeed, if successive frames are too close to each other, estimated curvature becomes overly sensitive to measurement noise. Firing rates were significantly lower in V2 than in V1 (presumably due to anesthesia), causing neural trajectories to be much shorter (**Fig. 3.6,** top right and bottom right). Consequently, our measurements in V2 did not allow us to reliably estimate the curvature of these trajectories (**Fig. 3.6,** bottom left). Several factors can be leveraged to address this issue:

- **More data.** A small number of neural trajectories in V2 are long enough for us to be able to reliably estimate their curvature. There are currently not enough of them for us to be able to make any statements about their curvature as a whole, but additional datasets might give us the statistical power necessary to answer this question.

- **More discriminable frames.** A new experimental design, in which successive frames are sampled more coarsely in time (and are therefore more discriminable), or in an awake animal, could place neural trajectories in V2 in the admissible regime for neural curvature estimation.

- **Better inference.** Our direct curvature inference method has freed us of much the the biases found in two-step, greedy estimation (**Fig. 3.2)**. Yet the variational objective we optimize only approximates the likelihood of neural curvature, and more flexible approximate posteriors have been shown to tighten this lower-bound, improving model fits (Rezende et al., 2014; Salimans et al., 2015; Rezende & Mohamed, 2016).

We are currently looking into each of these directions.

3.8 Discussion

In this chapter, we have validated the contribution of primary visual cortex to perceptual straightening. Indeed, we found that populations of neurons in V1 significantly straighten the time-course of natural videos, while distorting artificial ones. Moreover, these changes in curvature are predictive of the changes in curvature found perceptually, indicating that the straightening found in V1 could contribute to perceptual straightening.

In addition, we have leveraged our neural measurements to assess the straightening of natural videos at multiple timescales. These analyses revealed that changes in curvature do not only occur at the timescale at which we sampled our videos (and at which we had measured perceptual straightening) but at several coarser timescales. This is of paramount behavioral importance, as different tasks require making predictions at different time horizons. For example, kicking a football might require making predictions hundreds of milliseconds into the future, whereas deciding where to run or who to pass to require longer-term predictions. We have discovered that V1 facilitates predictions at several timescales (tens of milliseconds to hundreds of milliseconds), but we do not know whether this will hold at longer timescales. Two factors seems to indicate that this straightening cannot hold at arbitrarily long ones. If populations of neurons must learn to straighten the time-course of natural videos from experience, they will be limited by the coherence (or temporal consistency) of the signal they have access to. In particular, a V1 cortical column seeing the world through a 2° aperture may not be able to straighten natural videos over as long a timescale as higher visual areas (Hasson et al., 2008). Secondly, different types of behavior might place further constraints on the predictability of natural signals. Eye-movements, for

example, introduce discontinuities in our stream of visual input several times per second. How we are able to reconstruct a coherent percept across saccades and form predictions over longer timescales is therefore a promising direction for future investigation.

Perhaps the most appealing aspect of our approach is its generality. In this chapter, we were able to cast three very different representations (human psychophysics, computational models, and primate physiology) into a single format in which they behave lawfully (**Fig. 3.5)**. This paves the way for the deployment of this analysis in downstream areas, for which functional models are still lacking. At best, visual neuroscience has managed to differentiate visual areas from their afferents using highly specialized stimuli: oriented gratings differentiate V1 from the retina (Hubel & Wiesel, 1962), contours and textures differentiate V2 from V1 (Peterhans & Heydt, 1989; Freeman et al., 2013), objects and scenes differentiate IT from V4 (Rust & DiCarlo, 2010). Were downstream visual areas to further straighten the trajectories of their afferents, we would be able to cast all of visual computation in terms of a single, behaviorally relevant objective.

Chapter 4

Geodesics of machine representations

4.1 Introduction

We have found the straightening of natural videos to be an effective statistic in distinguishing models of the visual system: in the second chapter, we were able to identify a biologically plausible model relatively to a generic one solely based on this criterion (**Fig. 2.9**). Again using this single parameter-free statistic, we were able to place the primary visual cortex between the domain of pixel-intensities and perception, and between different stages of a computational model (**Fig. 3.5**). This may be contrasted with current approaches in biological model comparison, in which artificial representations are regressed onto a biological representation using hundreds or thousands of parameters (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014).

However appealing, we do not know how effective this approach will be in shedding light on higher-level computations of the visual hierarchy. In particular, if computational models of V1—or V1 itself—already account for a substantial portion of perceptual straightening, will we be able to distinguish the functional properties of V4 and IT using this method? Although we may be able to make this statistic more discriminative by measuring curvature

over longer timescales (Hasson et al., 2008), testing the straightening properties of these representations for a fixed set of sequences might still prove ineffective in distinguishing higher-level representations. Furthermore, in addition to distinguishing the computations of successive visual areas, we would like to know *why* sequences are straightened by some areas and not others. Can we gain an intuition for what features in a sequence a particular representation is straightening, and which ones it is not? We hope to address all of these issues by no longer asking a particular representation how straight natural sequences are, but rather which sequences are straightest, and how natural they are.

A related method for probing the selectivities and invariances of a representation is image synthesis. In this procedure, given a representation and reference image, one samples from the set of images that are mapped to the same point. These images are highly informative in that their differences reveal the invariances of that representation. This has enabled the iterative refinement of a model of visual texture (Portilla & Simoncelli, 2000), testing a theory of hierarchical visual processing (Freeman & Simoncelli, 2011), and the differentiation of successive visual areas (Freeman et al., 2013). In a different context, synthesis has been applied to artificial neural networks to reveal their sensitivity to 'adversarial perturbations', in which two images that appear entirely different to humans are identified by the network as belonging to the same category (Szegedy et al., 2013). In all of these cases, synthetic images provide a means of verifying or falsifying the hypothesis that the invariances of an artificial representation are also invariances for a biological one.

But the synthesis test, in which human observers try to discriminate synthesized images, is one-sided: failures (i.e. visually distinct images) can reveal *inappropriate* invariances of a representation, but successes can mask a lack of *desired* invariances. Consider the standard case of translation-invariance. The Fourier amplitude spectrum (i.e., the set of magnitudes

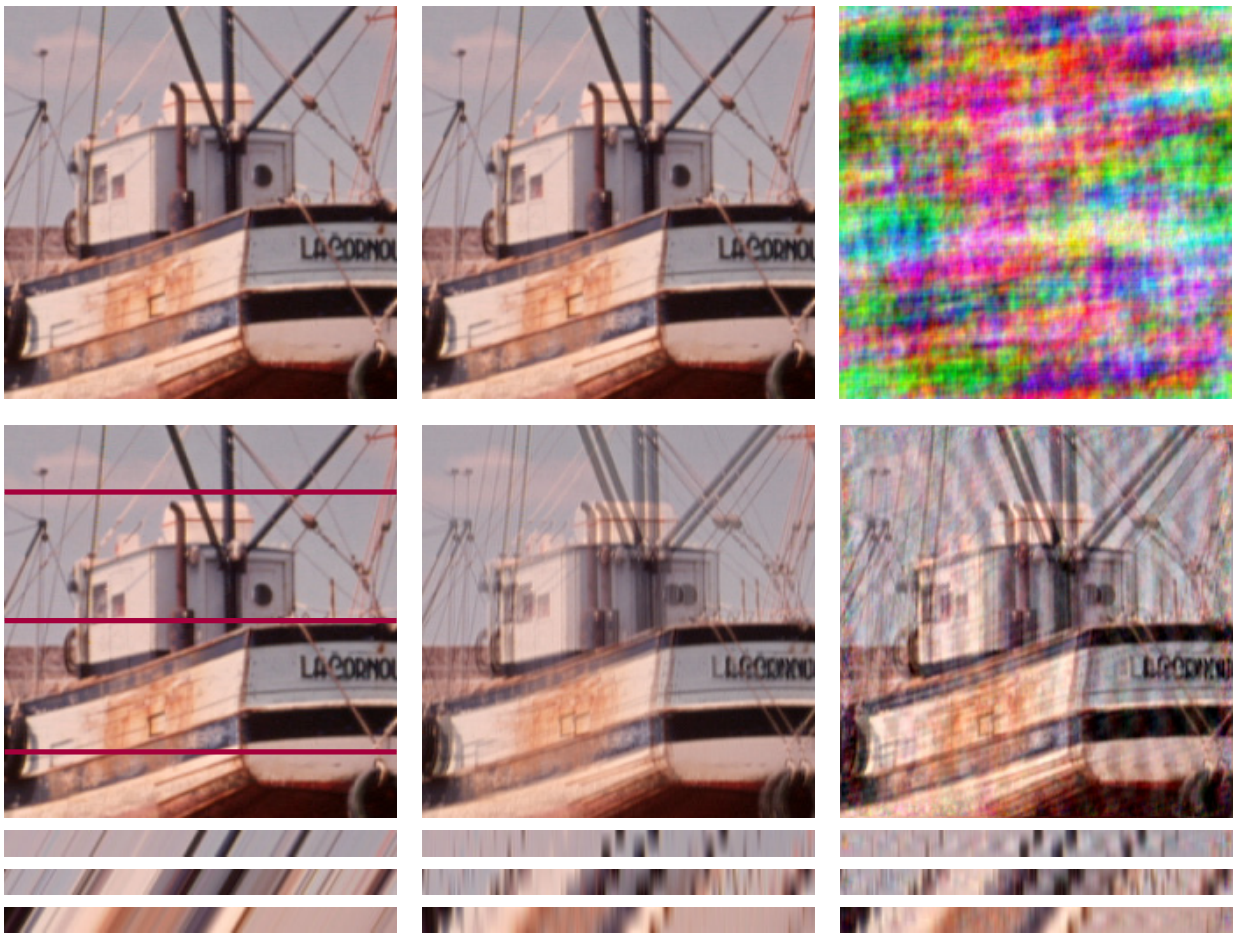


Figure 4.1 Geodesics can reveal either insufficient or excessive invariance, whereas synthesis reveals only the latter. **Top**: images synthesized so that their representation is matched to that of the ground truth image (left). Middle image has matching pixel intensities (i.e., it is identical to the ground truth image) and right image has matching Fourier magnitudes. **Bottom**: synthetic geodesic sequences connecting two translated copies of the same image, via different representations. Shown are the middle frame of each sequence, and below it, the temporal evolution of each row of pixels indicated by a horizontal red line. The ground truth transformation (left) is a translation (as can be seen from the diagonal lines in the temporal slice), but both geodesics deviate from the true transformation. The pixel representation fully constrains the image, and has no invariances, and thus the synthesized geodesic images are simply linearly interpolated between the initial and final images. The Fourier magnitudes, while translation-invariant, are also invariant to arbitrary phase perturbations, and the synthesized geodesic image contains Fourier components whose phases are shifted inconsistently.

of Fourier transform coefficients) provides a well-known example of a translation-invariant representation, but it is invariant to far more than translations, and this is immediately revealed by a synthesis test (figure **Fig. 4.1,** top right). On the other hand, simply representing an image with its raw pixel values (the identity representation) will trivially produce visually perfect synthetic examples (figure **Fig. 4.1,** top center) despite the fact that it has no invariance properties at all.

We seek a more general method of evaluation that penalizes a model for discarding too much information (as with synthesis) but also for discarding too little information. Each of these failures can be seen as an inadequacy of the image *metric* induced by the representation. Specifically, an image representation deforms the input space, bringing some images closer to each other while spreading others out, and thus inducing a new metric in image space. We can expose properties of this image metric by generating a geodesic sequence of images. Given an initial and final image, we synthesize a sequence of images that follow a minimal-length path in the response space of the representation. In the absence of any other constraints, this path will be a straight line connecting the representations of the two images; more generally, it will be the *straightest* realizable path connecting the two points. In the case where the two images differ by a simple transformation (e.g. a translation, figure **Fig. 4.1,** left column) that is not mapped to the straightest path connecting the two representations, the geodesic will differ from the original transformation connecting the images (figure **Fig. 4.1,** middle column). Similarly, if the representation is invariant to many transformations, the geodesic may correspond to a path that uses a mixture of transformations, and thus differ from the ground truth path (figure **Fig. 4.1,** right column). As a result, by visualizing whether a representation has straightened various deformations, representational geodesics can reveal both excessive and insufficient

invariance in an image model.

We develop an algorithm for synthesizing geodesic sequences for a representation, and use it to examine whether learned representations straighten various real-world transformations such as translation, rotation, and dilation. We find that a current state-of-the-art object recognition network fails to straighten these basic transformations. However, these failures point to a deficiency in the representation, leading to a simple way of improving it. We show that the improved representation is able to straighten a range of parametric transformations as well as generic distortions found in natural image sequences. By measuring the perceptual length of geodesics synthesized from different models, we show that the improved representation provides a better match to human perception. Finally, by generating geodesics from the hierarchical model of early visual areas used in previous chapters, we arrive at a set of stimuli that interpolate between the straightening of natural videos and distortion of unnatural ones.

4.2 Synthesizing geodesic sequences

Suppose we have an image representation, $y = f(x)$, where x is the vector of image pixel intensities and $f(\cdot)$ a continuous function that maps it to an abstract vector-valued representation y (e.g. the responses of an intermediate stage of a hierarchical neural network). Given initial and final images, we wish to synthesize a sequence of images that lies along the path of minimal length in the representation space (a *representational geodesic*). If the mapping is many-to-one (as is usually the case), this sequence of images is not unique. We resolve this ambiguity by selecting the representational geodesic that is also of minimal length in the space of images (i.e., a *conditional geodesic* in image space).

Objective function

In order to generate such a sequence, we optimize an objective function that expresses a discrete approximation of the problem, directly in terms of images sampled along the path. Given a desired sequence length N and initial and final images, $\{x_0, x_N\}$, we wish to synthesize a sequence of images, $\gamma = \{x_n; n = 0 \dots N\}$, lying along a geodesic in representation space. The representational path length is

$$L[f(\gamma)] = \sum_{n=1}^{N} \|f(x_n) - f(x_{n-1})\|_2$$

which is bounded by the representational energy

$$E[f(\gamma)] = \sum_{n=1}^{N} \|f(x_n) - f(x_{n-1})\|_2^2$$

thanks to the Cauchy-Schwartz inequality

$$L[f(\gamma)]^2 \leq NE[f(\gamma)]$$

with equality if and only if the representations are equispaced, which is encouraged by minimizing the representational energy. As a result, a path that meets this condition (e.g. the red curve in figure **Fig. 4.2)** while minimizing the representational energy $E[f(\gamma)]$ is a representational geodesic.

When the mapping to representation space is many-to-one, there are many possible solutions to this problem. To uniquely constrain the solution, we define an analogous

energy term that ensures that this path is also of minimal length in the image domain

$$E[\gamma] = \sum_{n=1}^{N} \|x_n - x_{n-1}\|_2^2$$

Since we are looking for the shortest path in image space that is also a geodesic in representation space, we minimize $E[\gamma]$ conditioned on the path also minimizing $E[f(\gamma)]$. Furthermore, during the optimization we constrain image pixel intensities to the [0, 1] range.

Optimization

We optimize this objective in three steps. First, we initialize the path with the minimum of $E[\gamma]$, which is simply a sequence of images that are linearly interpolated between the initial and final images. Next we minimize the representational geodesic objective $E[f(\gamma)]$. Finally, we minimize the image-domain geodesic objective, conditioned on staying in the set of representational geodesics.

Minimizing the representational geodesic objective in the second step requires optimizing an image for its representation via a non-linear function, and thus shares much of the non-convexity found in training deep neural networks. In particular, the curvature of the energy surface can vary widely over the course of the optimization. For this reason, we used the Adam optimization method (Kingma & Ba, 2014), which scales gradients by a running estimate of their variance, providing robustness to these changes in the energy landscape. We run Adam, using the default parameters, for 10^4 iterations to ensure that we reach the minimum of the representational geodesic cost.

To optimize the image-domain geodesic objective while constraining the solution to remain in the set of representational geodesics, we start by computing a descent direction for

the image-domain geodesic objective. We then project out the component of this direction that lies along the gradient of the representational geodesic objective. We take a step in that direction, then project back onto the set of representational geodesics by re-minimizing the representational geodesic cost (again using Adam), and repeat until convergence. We summarize our method with the following algorithm.

Conditional geodesic computation

Require: f: continuous mapping
Require: x_0, x_N: initial and final images
Require: N: number of steps along geodesic path ($N = 10$ in all our experiments)
Require: λ: gradient descent step size
Ensure: $\gamma = \{x_n; n = 0 \dots N\}$ minimizes $E[\gamma]$ conditioned on minimizing $E[f(\gamma)]$

$x_n \leftarrow \frac{N-n}{N} x_0 + \frac{n}{N} x_N \quad n \in \{0, 1, \dots N\}$ *initialize with pixel-based interpolation*
minimize $E[f(\gamma)]$ *project onto set of representational geodesics*
while γ has not converged **do**
  $d_r \leftarrow \nabla_\gamma E[f(\gamma)]$
  $d_p \leftarrow \nabla_\gamma E[\gamma]$
  $\widehat{d}_p \leftarrow d_p - \frac{<d_r, d_p>}{\|d_r\|_2^2} d_r$ *project out representational gradient*
  $\gamma \leftarrow \gamma - \lambda \widehat{d}_p$
  minimize $E[f(\gamma)]$ *re-project onto set of representational geodesics*
end while
return γ

Despite the non-convexity of the problem, we have good reason to believe that solving this optimization problem should be feasible for trained neural networks. Since the output of the first layer is equal to the convolution of the input image with a filter bank, our problem is similar in complexity to optimizing the weights of the first layer of a network, for the same objective. Recent theoretical work shows that optimizing all layers of a network jointly makes the problem significantly more difficult than optimizing a single layer in isolation (Saxe et al., 2013). Hence optimizing $E[f(\gamma)]$ should be easier than training the

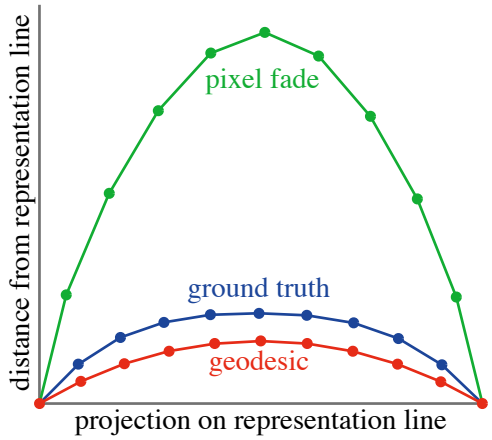


Figure 4.2 Deviation from the straight line connecting the representations of a pair of images, for different paths in representation space. Due to the nonlinearity of the representation (the third stage of L_2 pooling of a deep neural network, see section 4.3) the geodesic deviates slightly from the straight line. The ground truth transformation (here, a translation) deviates similarly, indicating that the representation has straightened the transformation to a large extent. For reference, a pixel-based interpolation deviates significantly more from a straight line. Axes are in the same units, normalized by the distance separating the end point representations. Knots along each curve indicate samples used to compute the path.

full network for recognition. In practice we were able to solve the optimization problem for a variety of deep networks.

It should be noted that if the mapping $f(\cdot)$ is not surjective, not all vectors in the representation space are attainable from an input image. Specifically, if the mapping is nonlinear (as for most representations of interest) the set of attainable vectors is non-convex, and vectors lying along the straight line connecting two representations are not necessarily attainable. As such, we can only expect to find a geodesic path whose representation is *as close as possible* to this straight line by minimizing the representational geodesic cost $E[f(\gamma)]$. **Figure 4.2** shows an example of this, for the case of image translation. By construction, the geodesic is closer to a straight line in representation space than either the ground truth transformation or a pixel interpolation. The ground truth transformation lies close to the geodesic, indicating that this representation has almost (but not completely) straightened this transformation. The differences between these two paths can be made explicit by visualizing the geodesic sequence, as detailed in the following section.

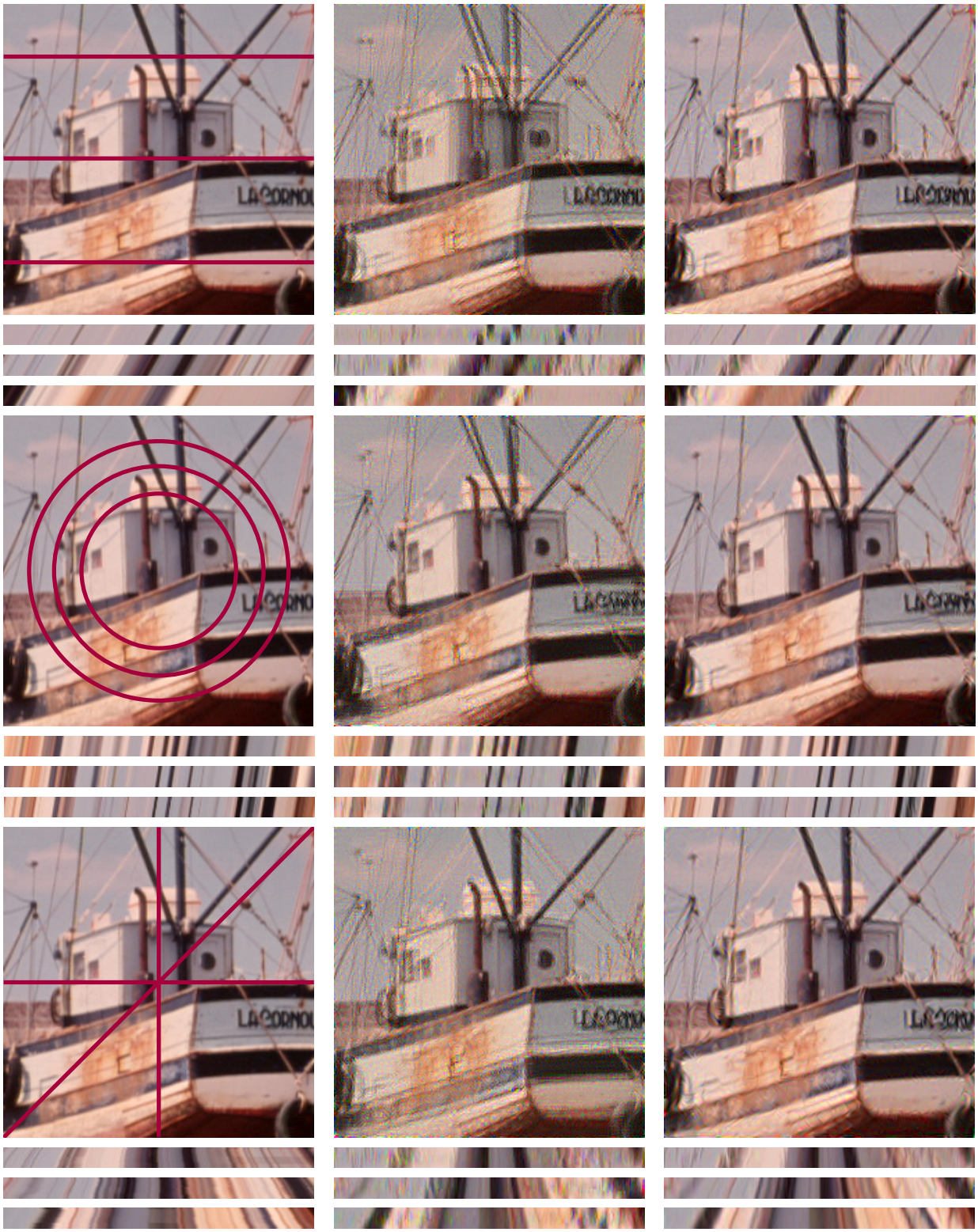
4.3 Visualizing geodesic sequences

We used our geodesic framework to examine the invariance properties of the 16-layer VGG network (Simonyan & Zisserman, 2015), which we chose for its conceptual simplicity and strong performance on object recognition benchmarks. As a "representation" for our tests, we used the output of the third stage of pooling. Each stage of this continuous non-linear mapping is constructed as a composition of three elementary operations: linear filtering, half-wave rectification, and max pooling (which summarizes a local region with its maximum). We followed the preprocessing steps described in the original work: images are rescaled to the $[0, 255]$ range, color channels are permuted from RGB to BGR, and the mean BGR pixel value, $[104, 117, 124]$, is subtracted. We verified that our implementation could replicate the published object recognition results.

Geodesics as a diagnostic tool

We first examined whether this representation linearizes basic geometric transformations: translation, rotation and dilation. To do so, we compute the geodesic sequence between two images that differ by one of these transformations, and compare it to the ground truth se-

Figure 4.3 ***(following page)*** Comparison of geodesic sequences for VGG network representation with max pooling (middle column) and VGG network with L_2 pooling (right column) with ground truth sequence (left column). Three different types of geometric transformation are tested: horizontal translation (top), rotation around the center (middle), dilation about the center (bottom). As in **Fig. 4.1,** square images are the middle frame from the corresponding sequence, and underneath is the temporal evolution of three image slices, taken along the red lines shown in the left column. The original VGG network is unable to linearize these transformations (as indicated by the 'double exposure' in the middle frame, and the discontinuous temporal slices), whereas the same VGG network with L_2 pooling (right column) induces a geodesic that is close to ground truth.



ground truth | VGG network, max pooling | VGG network, L_2 pooling

quence obtained by incremental application of the same transformation. The extent of the overall transformation determines the difficulty of this task: all representations (even trivial ones) will produce geodesics that are close to the ground truth for very small transformations, whereas all are likely to fail for very large transformations. For our discriminative test we chose intermediate values: an 8 pixel translation, a 4° rotation, and a 10% dilation.

We found that the VGG network, despite its impressive classification performance, failed to linearize these simple geometric deformations and produced geodesics with salient aliasing artifacts (**Fig. 4.3,** middle column). Given that no subsampling is used in the convolutional layers, we attributed this failure to the max pooling layers, which subsample the representation by a factor of 2 in each direction, despite their small spatial extent (a 2×2 pooling region). To avoid aliasing artifacts when subsampling by a factor of 2, the Nyquist theorem requires blurring with a filter whose cutoff frequency is below $\frac{\pi}{2}$. Following this indication, we replaced the max pooling layers with L_2 pooling:

$$L_2(x) = \sqrt{g * x^2}$$

where the squaring and square-root operations are point-wise, and the blurring kernel $g(\cdot)$ is chosen as a 6×6 pixel Hanning window that approximately enforces the Nyquist criterion. This type of pooling is often used to describe the behavior of neurons in primary visual cortex (Vintch et al., 2015), and also bears resemblance to the complex modulus used in the "scattering transform" (Mallat, 2011) which has been shown to be robust to smooth deformations.

We found that this modified VGG network not only produced geodesic sequences that were free of most aliasing artifacts, but also linearized these geometric transformations convincingly, as can be seen in the temporal slices of the geodesic (**Fig. 4.3,** right column).

This confirms that, as with the Fourier magnitude and the scattering transform, smooth, quadratic pooling operators are able to linearize local deformations. Unlike the Fourier magnitude however, the locality and hierarchical nature of these representations tailors their invariances to a much more limited set of transformations. Furthermore, this demonstrates the power of geodesics as a visualization tool for understanding learned representations. Not only does this diagnostic report a deficiency of a representation (**Fig. 4.3,** middle column), it also points to the mechanism of this failure, suggesting a simple way to improve the model.

This suggests that the VGG network's performance on object recognition tasks could be improved by substituting max pooling with L_2 pooling, and retraining the network to decode this new representation. Indeed, the added invariance of this representation could enable the network to generalize to new viewing conditions more robustly.

Disambiguating spatial scale and nonlinear complexity with geodesics

Thus far we have found that a deep representation is able to linearize a range of real-world transformations (**Fig. 4.3,** right column) whereas a shallow one (e.g., the pixel intensities) is not (**Fig. 4.1,** middle column). It is unclear, however, whether the improved invariance of the deep representation is due to the spatial extent over which it computes its responses, or its nonlinear complexity. Indeed, as we progress up the hierarchy of a neural network, the effective input region for each unit (the "receptive field") increases in size, simply due to cascaded convolution and subsampled pooling. At the same time, the complexity of the representation increases as a longer sequence of non-linear operations are composed.

In order to separate these two effects, we varied the complexity of the representation while keeping the size of the receptive field constant. For an artificial neuron, the receptive

Figure 4.4 Even when matched for "receptive field size", shallow representations cannot linearize translations as well as deep ones. From left to right: geodesics generated from 1st, 2nd and 3rd pooling layers, with receptive field sizes approximately matched by altering the spatial extent of the L_2 pooling (to 36×36, 18×18 and 6×6 pixels, respectively). As the complexity of the representation increases, so does the quality of the corresponding geodesic.

field quantifies the strength of the connection between a location in the image and that neuron's activity, and can be measured by computing the magnitude of the gradient of the neuron's activity with respect to the image. Hence, the receptive field of a non-linear neuron changes as a function of the input image. In order to measure the extent of a neuron's receptive field across all images, we averaged the magnitude of the gradient of its activity over a large set of white noise images. We generalized this method to measuring the receptive field of an entire population by computing the average magnitude of the gradient of an entire 'cortical column', or set of hidden units at a given location.

Using this method, we measured the receptive field size of the representation used in our previous experiments (third pooling layer of the VGG network with L_2 pooling). We then computed geodesics from shallower representations (first and second pooling layers of the

VGG network) for which we increased the pooling extent (from 6×6 to 36×36 and 18×18 respectively) in order to match the receptive field size of the deep representation. These experiments show that shallower layers, despite being matched for receptive field size, are unable to linearize translations as well as deeper ones (**Fig. 4.4**). Interestingly, we find a gradual increase in the quality of the geodesics as the complexity of the representation increases. Hence the curvature of representational geodesics, more than their dimensionality, is essential for capturing these non-linear deformations of the image.

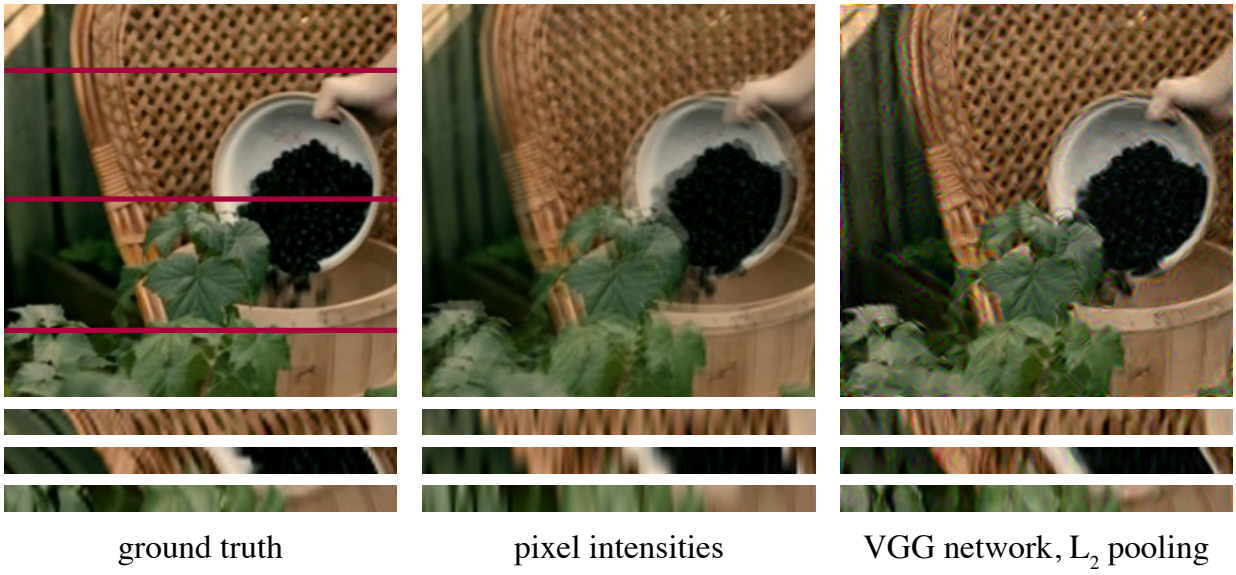


Figure 4.5 Comparison of geodesic sequences for a pixel-based representation (middle column) and the VGG network with L_2 pooling (right column) for a natural movie (left column). Geodesic sequences are generated between the first and last frame of the original movie. The pixel intensity representation fails to linearize the sequence, while the VGG network with L_2 pooling induces a geodesic that is close to the original movie. The main deficiency in this geodesic is due to temporal aliasing, where periodic structure in the image is shifted backwards relatively to the rest of the image (see first and second temporal slices).

Linearizing natural image sequences

Having tested for the modified VGG network's ability to linearize simple parametric transformations, we asked whether it can linearize compositions of these transformations that arise in natural image sequences. To explore this, we extracted 5 frames from the movie *Melancholia* and generated a geodesic from the first to the last of these. We find that this geodesic smoothly transitions between the two images, and captures much of the true temporal evolution of the video (**Fig. 4.5,** left and right panels). Relative to the original sequence, the only errors it produces are due to well known problems in motion estimation. A large component of the transformation in the video is an out-of-plane rotation due to the camera panning, creating a composition of translations and dilations throughout the image. In a region of the image with periodic structure (e.g. the woven cane texture of the chair), the motion between the two end frames is ambiguous, because the translation between them exceeds one half of this period. This problem, known as *temporal aliasing*, can be seen in the temporal slices, which reveal that the back of the chair is smoothly shifted in the opposite direction of the rest of the image (**Fig. 4.5,** right panel). In motion estimation, this problem is usually solved using a coarse-to-fine approach, in which the motion of the low frequencies is estimated first, and used to condition (or initialize) motion estimates derived from higher frequencies. This method can be naturally embedded in our framework by generalizing the nested conditionalization of geodesic objective functions (section 2.1). That is, each layer of the network can impose its own geodesic constraints, conditioned on those imposed by deeper layers. This hierarchical construction provides a means of solving the problem of temporal aliasing, and more generally should allow the network to linearize a broader class of transformations.

4.4 Perceptual model comparison using geodesics

Visual inspection of the geodesics synthesized by different models allows us to intuit that the metric induced by some of them is much better matched to human perception than others. In this section, we would like to quantify that match, as we did in chapter 2 using the straightness of the representation of natural sequences. Specifically, given a pair of reference images and a set of sequences connecting them (**Fig. 4.6)**, each model orders these sequences according their path length in their own representational space. If one of these sequences happens to be its own geodesic connecting those two images, we are guaranteed that it will be the shortest of that set (**Fig. 4.2)**. If we are then able to measure the perceptual length of these sequences according to human observers, we will arrive at a new ranking of these sequences, and the match between human and model rankings will allow us to evaluate the perceptual relevance of each model. Specifically, we are interested in finding which model produces a geodesic with the shortest perceptual length. Since the geodesic objective also minimizes the variability in step-sizes along the path, we would also expect a perceptually accurate model to produce a geodesic with equally discriminable increments.

We construct this psychophysical paradigm as a more constrained version of the one used in the second chapter, given that we seek only to measure the length of various perceptual trajectories, not their curvature. We presented human observers with the task

Figure 4.6 ***(following page)*** Perceptual model comparison using geodesics. Four sequences connect an image and its translated copy. Left/green: an incremental translation, representing the 'ground truth' interpolation. Middle-left/red: a pixel-wise interpolation (fade). Middle-right/yellow: a geodesic synthesized from the 3rd pooling layer of the VGG network. Right/blue: a geodesic synthesized from the 3rd pooling layer of a modified VGG network, in which the max-pooling operation is replaced by an L_2 pooling operation.

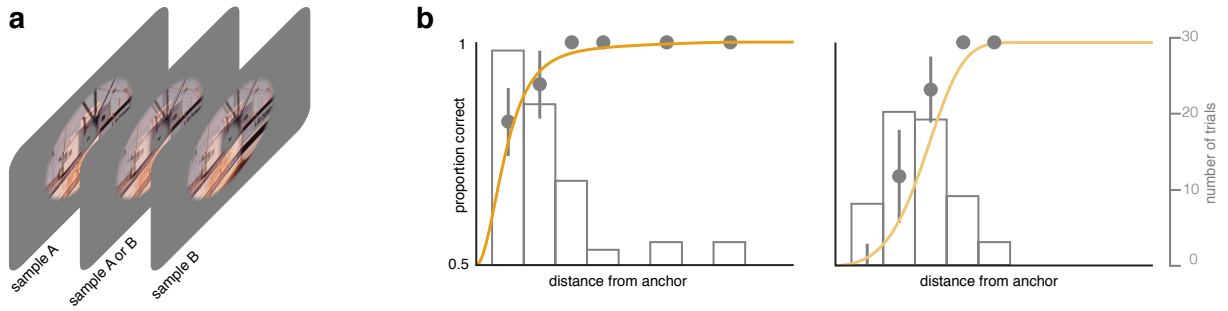


Figure 4.7 Psychophysical paradigm for perceptual model comparison. (**a**) A single trial in the experiment, in which the two images being compared are the first and last of a sequence of 3 images, with the second being equal to one of the other two. Observers are tasked with reporting which image is unique. Each frame is presented for 200 ms in an annulus whose inner and outer radii are equal to 2° and 10° respectively. (**b**) Proportion correct in this task, for a given observer and two 'reference' (or 'anchor') images, over many trials. Trials are preferentially sampled around the steepest part of the (inferred) psychometric curve.

of discriminating pairs of images in an 'ABX' paradigm. On a given trial, two images were presented at either end of a sequence of 3 images, with the middle frame being equal to one of the other two. Observers were asked to indicate which, of the first or the last frame, was the unique one (**Fig. 4.7a**). Rather than sampling these pairs of images uniformly, we used a staircase procedure. Having selected one ('reference') image in the pair randomly, we chose the second image such that it would lie in the steepest part of the psychometric curve (in the case of an 'AXB' task, situated at a performance level of 68%; **Fig. 4.7b**), using a '2 up, 1 down' scheme.

In order to infer the perceptual path length from these measurements, we estimated the locations (in a 1-dimensional space) of each image that best account for the data. Specifically, the perceptual representation of each image was modeled as a 1-dimensional Gaussian distribution with unit variance (**Fig. 4.8,** top row), whose overlap with neighboring distributions determined their discriminability (**Fig. 4.8,** middle row). This predicted pattern of discriminability could then be compared to the observed discriminability of pairs

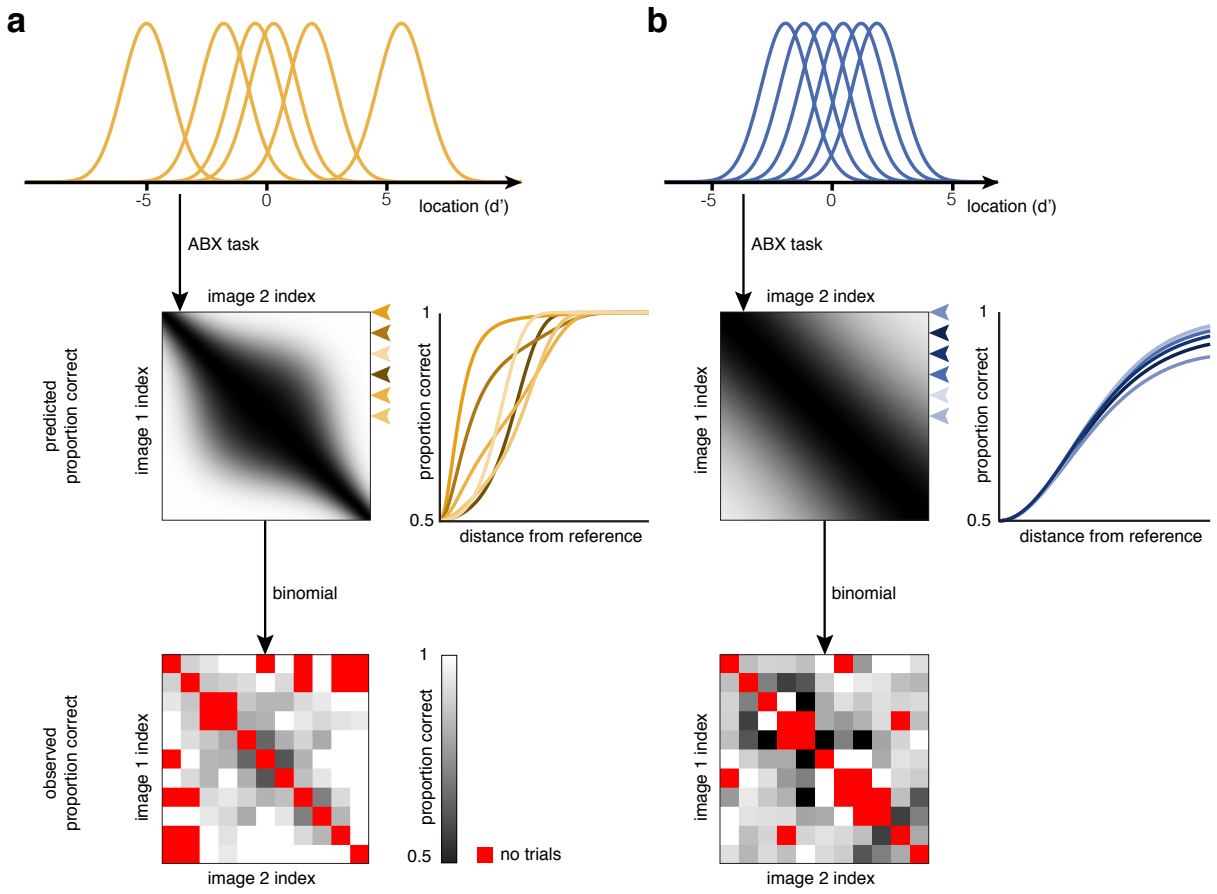


Figure 4.8 Inferring perceptual path length from pairwise discriminability. We associate each frame from a geodesic sequence to a 1-dimensional, unit-variance Gaussian (top, showing 6 out of 11 such distributions). The overlap between these distributions determines a pattern of discriminability between pairs of frames in a sequence (middle, each psychometric curve is a horizontal slice through the discriminability matrix, at the level of the corresponding arrow). This predicted pattern of discriminability can be compared to the measured pattern of discriminability (bottom). (**a**) Perceptual locations inferred from the discriminability of frames synthesized from the VGG network are far apart and irregularly spaced. (**b**) Perceptual locations inferred from the discriminability of frames synthesized from the modified VGG network are close together and evenly spaced.

of frames measured in our experiment (**Fig. 4.8,** bottom row). For one particular observer, this analysis shows that the frames of the geodesic synthesized from the VGG network are far apart and unevenly spaced, as evidenced by steep and irregular psychometric curves (**Fig. 4.8a**). In contrast, the frames synthesized from a the modified VGG network are close to each other and regularly spaced, as evidenced by consistently shallow psychometric curves (**Fig. 4.8b**).

We performed this analysis on three observers viewing four sequences connecting a pair of translated copies of the same image. The first of these sequences (**Fig. 4.6,** left/green) smoothly translates from one image to the next. The second (**Fig. 4.6,** middle-left/red) fades linearly between the two, which is also a geodesic for any linear representation of the image domain. The third is a geodesic from the original VGG network (**Fig. 4.6,** middle-right/green). The fourth is a geodesic from a VGG network whose max-pooling layers have been replaced with L_2 pooling. Consistent with the intuition gained from visual inspection of these sequences, the 'ground truth' translation consistently has the shortest perceptual length and the most regular steps along the path (**Fig. 4.9,** left/green). The the pixel-based interpolation and the geodesic synthesized from the original VGG network, in contrast, are much longer perceptually and have irregular step sizes (**Fig. 4.9,** middle). A simple modification to the VGG network allows it to generate a geodesic sequence whose discriminability is much lesser and more regular (**Fig. 4.9,** right/blue). However simple, this example shows that measuring the perceptual properties of geodesic sequences can provide a powerful test of a representation's match to biological representations.

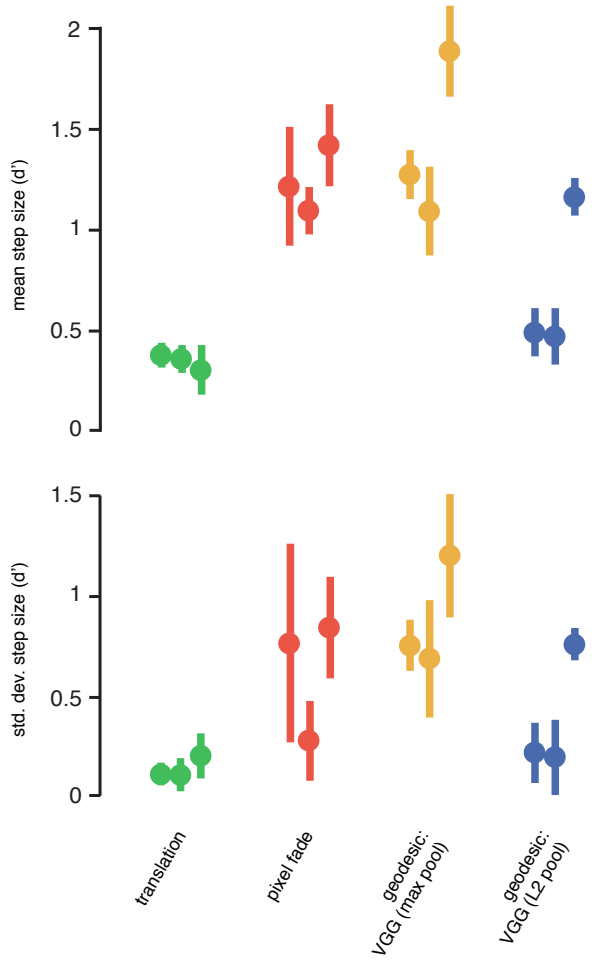


Figure 4.9 Perceptual model comparison across observers. Each point denotes the mean step size (i.e. path length divided by the number of steps, in this case, ten; top) or the variability in step sizes (bottom) for one observer and sequence.

4.5 Dissecting perceptual straightening with geodesics

Having showcased the power of geodesic synthesis as a diagnostic tool, we asked whether geodesic sequences could shed light on the straightening properties we uncovered in the previous chapters. In particular, the third chapter showed that perceptual straightening could arise through the incremental transformations occurring in successive visual areas. Moreover, we found these same visual areas to distort the trajectories of unnatural sequences. Synthesizing geodesic sequences that are straight according to an intermediate stage of our hierarchical model of visual processing suggests a way of interpolating between these two results. Indeed, a sequence that is straight according to the LGN stage (**Fig. 4.10** top, red) is most likely curved in the image domain due to the non-linear operations of that stage. This implies that the LGN stage will straighten LGN-geodesics. However, these sequences being otherwise unnatural, downstream stages have no reason to straighten them and will likely distort them. Similarly, if our hierarchical model of early visual processing provides a good account of those computations, a V1-geodesic (**Fig. 4.10** top, yellow) should be straightened by the human LGN and V1, but tangled afterwards.

We verified these predictions perceptually. Using the same psychophysical protocol and analysis methodology as in chapter 2, we inferred the perceptual curvature of a set of natural sequences (*egomotion*, *leaves-wind*, *carnegie-dam* and *walking*), as well as pixel-domain-, LGN-, and V1-geodesics interpolating between their end-frames. Comparing the perceptual curvature of these sequences to that in the pixel-domain again allows us to assess the amount of straightening for each of these sequences. Pixel-domain-geodesics (the 'unnatural' sequences used in previous chapters), which should be tangled by all stages of the visual system indeed experience maximal perceptual distortion (average change in

curvature = 41°; **Fig. 4.10** bottom, green). LGN-geodesics, which are straightened by the first stage of the model and tangled by subsequent ones experience less perceptual distortion (average change in curvature = 27°; $p < 0.05$, Wilcoxon signed-rank test; **Fig. 4.10** bottom, red). Similarly, V1-geodesics, which are straightened by the first two stages of the model and tangled by subsequent ones experience a change in curvature that is lower still (average change in curvature = $-7°$, $p < 0.01$; **Fig. 4.10** bottom, yellow). Finally, natural sequences, which we assume to be straightened by the entire visual system, experience a maximal reduction in curvature (average change in curvature = $-28°$, $p < 0.05$; **Fig. 4.10** bottom, blue).

Synthesizing geodesics from hierarchical models of the visual system therefore produces sequences that can engage straightening or distortion in different visual areas, in a way that can be experimentally controlled. This suggests that the visual system does not partition the set of all image sequences simply into 'natural' and 'unnatural' sequences, but rather in varying degrees of naturalness, as evidenced by intermediate levels of straightening. Moreover, it could be that different visual areas have different criteria for naturalness, with deeper stages in the hierarchy having increasingly stringent criteria. Geodesic sequences could provide a way of discerning these criteria, as stimuli used in physiological experiments similar to those of chapter 3. In particular, if the perceptual definition of 'naturalness' and its corresponding straightening properties are built through successive refinement by individual visual areas, we should be able to dissociate these areas based on their straightening of different classes of stimuli. For example, we would expect primary visual cortex to straighten its LGN-afferents when presented with a V1-geodesic, but to tangle them when presented with a LGN-geodesic. If the LGN on the other hand straightens both of these sequences, this would indicate that primary visual cortex indeed has a more conservative

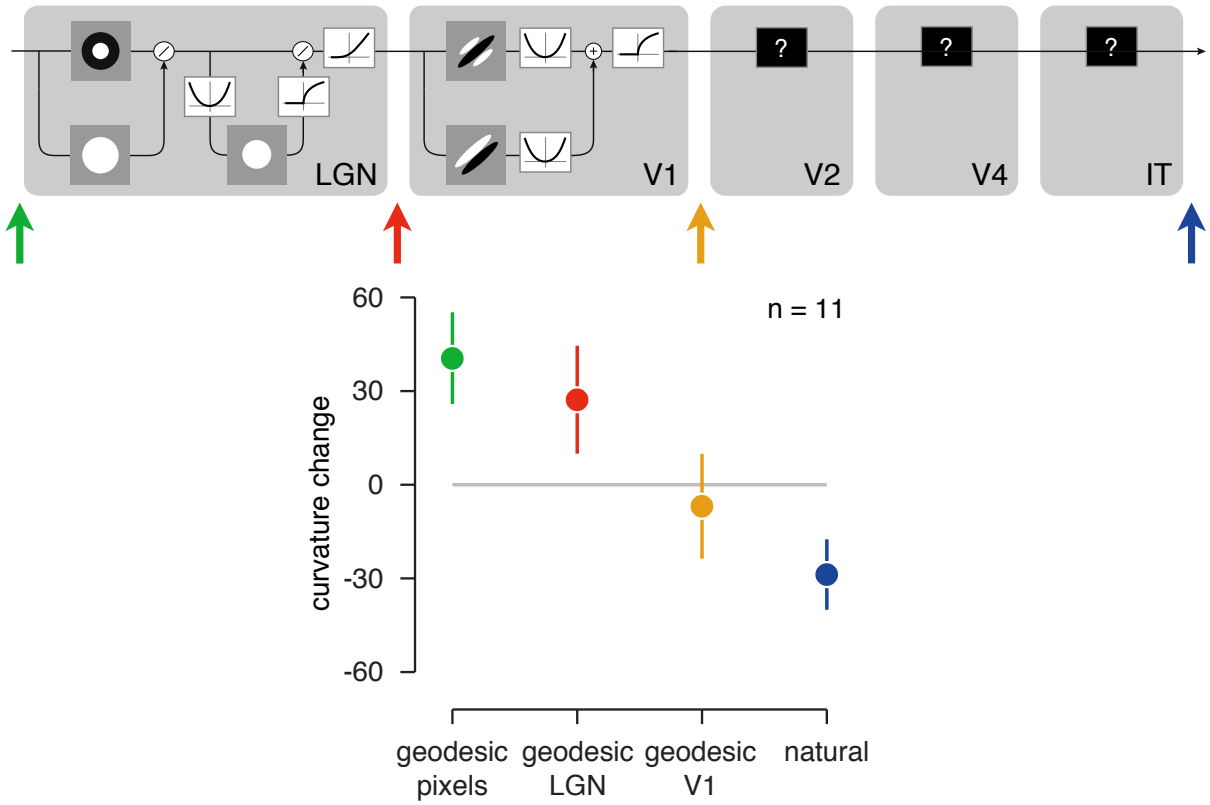


Figure 4.10 Dissecting perceptual straightening with geodesics. Top: a two-stage hierarchical model of early stages of the visual system, as part of a larger hierarchy of unknown modules. Green, red and yellow arrows represent geodesic sequences that are constructed to be straight according to the corresponding stage of the model (pixels, LGN, and V1, respectively). The blue arrow represents natural sequences, which are assumed to be pixel to perceptual domain) experienced by each class of sequences. Bottom: change in curvature (from the pixel- to the perceptual domain) for each class of sequences and 11 observers.

definition of naturalness.

4.6 Discussion

In previous chapters, we found that measuring the straightness of natural images sequences in different representations to be a discriminative test of their biological plausibility (**Fig. 2.9)**. In this chapter, we have arrived at an even more powerful test by instead asking which sequences are straightest according to a representation. This has not only enabled us to distinguish a generic representation from a more biologically plausible one (**Fig. 4.9)**, but also offered insight into the discrepancy between a generic representation and human perception (**Fig. 4.3)**. This understanding has allowed us to design a representation that is better matched to biological vision.

We have applied our analysis to pairs of images that differ according to simple transformations such as translation, rotation, and dilation. Asking whether human and machine representations straighten the trajectories of parametrically transformed images amounts to a direct test of the "untangling hypothesis" (DiCarlo & Cox, 2007). We have found that human observers indeed straighten these trajectories, but that generic neural networks trained for object recognition do not (**Fig. 4.3, Fig. 4.9)**. This suggests that these artificial representations are solving tasks such as object recognition using a very different strategy from humans, consistently with the observation that deep networks are vulnerable to adversarial attacks (Szegedy et al., 2013). Our visualization method has pointed to architectural modifications that have enabled these networks to straighten parametric deformations, and investigating whether similar modifications could lend robustness to adversarial attacks therefore appears to be a fruitful future direction.

The synthesis of geodesic sequences can be applied to arbitrary image pairs, including

but not limited to parametrically transformed images (**Fig. 4.3)** and frames from natural videos (**Fig. 4.5)**. For example, generating geodesics between two arbitrary images from the same object category can reveal whether object identity is an invariant of a representation. An affirmative answer implies that, back in the representation space, all of the images along the geodesic could be correctly identified using a linear decoder. Furthermore, geodesics synthesized within and between object classes could be an effective way of probing the metric properties of higher-level visual areas.

Finally, our method suggests a natural extension to hierarchical representations. Our geodesic sequences were computed by minimizing path length in the pixel domain, conditioned on minimizing path length in a network representation. This process could be applied recursively in a hierarchical representation, minimizing path length at each stage conditioned on minimal path length at higher stages. The resulting image sequences, in turn, could be used to characterize the changes in metric properties across visual areas. Conditional image synthesis has been used effectively to isolate novel selectivity in particular visual areas (Rust & DiCarlo, 2010; Freeman et al., 2013). Conditional geodesic synthesis, by generalizing this method to sequences of images, holds promise to isolate novel straightening properties along the ventral hierarchy.

Chapter 5

Conclusion

We have proposed the temporal straightening hypothesis, a mechanism for how biological organisms could solve tasks that require making predictions. Specifically, we conjectured that the visual system might straighten the temporal trajectory of internal representations of natural videos in order to facilitate their predictability. The first two chapters of this thesis are direct tests of this hypothesis: in the first we found that human perceptual judgments are consistent with this hypothesis, in the second that the activity of populations of neurons in primary visual cortex is likewise. Moreover, we found that the changes in curvature in each modality (perceptual or neural) were highly predictive of one another, suggesting that the straightening found in primary visual cortex could support perceptual straightening. Together, these results suggest that the straightening of natural videos could be used as a statistic for unifying the computations found at different stages in the ventral stream. Indeed, this framework enabled us to to cast perceptual judgments, the activity of cortical neurons, and computational models into a single format in which they behave lawfully. Future work will investigate whether we can frame the properties of downstream visual areas in terms of their straightening properties. Doing so would allow us to not only evaluate the nonlinear transformations of the visual hierarchy, but also understand their

relevance for a range of behavioral needs.

To design our experimental tests of the straightening hypothesis, we have made two simplifying assumptions. First, we have limited our analyses to the non-linear spatial processing of the visual system. This has allowed us to use standard psychophysical methods and signal detection theory to infer perceptual curvature, greatly simplifying the interpretation of those results. It has also allowed us to simplify the physiological protocol and analysis by assuming independent, random presentation of stimuli and a stationary response distribution. And yet real-world predictions must certainly use the full spatio-temporal context of our environment. For example, motion and other long-term dependencies provide strong cues regarding future outcomes. Measuring the perceptual and neural trajectories of contiguous videos is therefore a necessary step towards linking our results to a more natural behavioral context.

Our second simplifying design choice was to limit our measurements to short sequences sampled at their original frame rate (usually 30 frames/s). This made our experiments significantly more tractable, but limits the conclusions we can make about the predictability of these signals over longer timescales. Indeed, a reduction in curvature at a particular timescale (or sampling rate) indicates an increase in the predictability of those signals at that timescale, but does not necessarily generalize to others. We have found empirical evidence that straightening happens at coarser timescales (up to hundreds of milliseconds), but we do not know whether this will hold at longer timescales. Repeating our physiological experiments with longer sequences or coarser rates would allow us to address this question for neural representations. Measuring perceptual curvature at multiple timescales, however, might require a new psychophysical protocol that measures super-threshold perceptual distances (Maloney & Yang, 2003). Connecting the straightening properties of perceptual

and neural representations at different timescales to the statistics of natural videos therefore appears to be a highly promising future direction of investigation. In particular, it seems impossible for a given representation of natural scenes to be predictable at arbitrarily long timescales. In contrast, downstream areas that encode more abstract statistics in the signal could be better equipped for making longer-term predictions (Hasson et al., 2008). Understanding how the functional properties of the visual hierarchy enable predictions at multiple timescales could as a result be of central importance for building intelligent systems.

The final chapter of this thesis used the straightening of natural videos as a framework for investigating the selectivities and invariances of artificial representations, and their match to human vision. Specifically, we developed a methodology for generating the straightest sequence connecting a pair of reference images, according to a particular representation. This analysis revealed a failure of current artificial neural networks to straighten parametrically transformed image sequences (such as translations, rotations, and dilations) despite humans finding them perfectly natural. In contrast, the 'geodesics' of these representations contained severe artifacts, appearing highly unnatural. Importantly, these sequences proved to be highly interpretable and pointed to a simple architectural modification which significantly improved their perceptual relevance.

Geodesic sequences provide a means of bridging the gap between the straightening hypothesis and the functional properties of neural populations and perception. Indeed, the straightening hypothesis only states that internal representations should be structured to facilitate *their own* predictability in time. It does not state what features of the environment these areas should or should not represent. On one hand, this has been an useful asset that has allowed us to apply our theory to representations whose functional properties are

still unknown (human perception and deep neural networks being prime examples). But a more complete description of the role of straightening in the visual hierarchy must connect it to the functional properties of individual visual areas. For example, not all natural sequences are straightened by the same amount. What features in a sequence determine its straightening or distortion? Synthesizing geodesics from a computational model that captures those features provides a means of testing candidate answers to this question. Indeed, geodesics of our hierarchical model of early vision enabled us to interpolate between the straightening of natural sequences and the distortion of unnatural ones. This suggests that the features driving perceptual straightening are those that emerge from the increasingly complex properties of the visual hierarchy. Further confirmation would require evaluating the neural curvature of these sequences, and showing that it also interpolates between that found for natural and unnatural sequences. In particular, synthesizing geodesics from hierarchical models of the ventral stream could prove to be an effective means of dissecting the contributions of individual areas to the straightening of natural videos.

Finally, however elegant casting known computations into a single framework may be, the most useful application of the straightening hypothesis would be in deriving the functional properties of mid- to high-level visual areas that so far have eluded mathematical formalism. Such an endeavor would require optimizing a representation to straighten the time-course of natural videos. Further terms in the objective would also be necessary, to determine which aspects of the environment the representation should attempt to predict. There is a general consensus that the low-level detail of future states is not behaviorally relevant, and several attempts have been made to establish this distinction automatically (Goroshin et al., 2015; Lotter et al., 2016; Mathieu et al., 2016). More formally, limiting oneself to predicting the statistics of the future, rather than its actual outcome, appears

to be a very promising direction (Babaeizadeh et al., 2018; Buesing et al., 2018; Denton & Fergus, 2018). Despite these recent advances in machine learning, how biological systems decide what to represent and predict about their environment remains an open question. More fundamentally, how are groups of neurons able to evaluate the predictions they make about the world, and calibrate their internal model? Given the central importance of unsupervised learning in artificial intelligence, answering this question may be a crucial step in building machines that understand the world like we do.

Bibliography

Adelson, E. H. & Bergen, J. R. (1985). "Spatiotemporal energy models for the perception of motion". In: *Journal of the Optical Society of America A* 2(2), 284.

Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R. & Levine, S. (2018). "Stochastic Variational Video Prediction". In:

Barlow, H. (1961). *Possible principles underlying the transformations of sensory messages.*

Barlow, H. (2001). "Redundancy reduction revisited". In: *Network: Computation in Neural Systems* 12(3), 241–253.

Battaglia, P. W., Hamrick, J. B. & Tenenbaum, J. B. (2013). "Simulation as an engine of physical scene understanding". In: *Proceedings of the National Academy of Sciences.*

Berardino, A., Ballé, J., Laparra, V. & Simoncelli, E. P. (2017). "Eigen-Distortions of Hierarchical Representations". In: (Nips).

Bialek, W., De Ruyter Van Steveninck, R. R. & Tishby, N. (2006). "Efficient representation as a design principle for neural coding and computation". In: *IEEE International Symposium on Information Theory - Proceedings*, 659–663.

Buesing, L., Weber, T., Ere, S., Ali Eslami, S. M., Rezende, D., Reichert, D. P., Viola, F., Besse, F., Gregor, K., Hassabis, D. & Wierstra, D. (2018). "Learning and Querying Fast Generative Models for Reinforcement Learning". In:

Chechik, G., Globerson, a., Tishby, N. & Weiss, Y. (2005). "Information bottleneck for Gaussian variables". In: *J Mach Learn Res.*

Cloney, R. A. (1978). "Ascidian metamorphosis: review and analysis." In: *Settlement and Metamorphosis of Marine Invertebrate Larvae*.

Cox, D. D., Meier, P., Oertelt, N. & DiCarlo, J. J. (2005). "'Breaking' position-invariant object recognition." In: *Nature neuroscience* 8(9), 1145–7.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*.

Denton, E. & Fergus, R. (2018). "Stochastic Video Generation with a Learned Prior". In:

DiCarlo, J. J. & Cox, D. D. (2007). "Untangling invariant object recognition". In: *Trends in Cognitive Sciences* 11(8), 333–341.

Freeman, J. & Simoncelli, E. P. (2011). "Metamers of the ventral stream". In: *Nature Neuroscience*.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P. & Movshon, J. A. (2013). "A functional and perceptual signature of the second visual area in primates". In: *Nature Neuroscience*.

Fukushima, K. (1980). "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: *Biological Cybernetics* 36(4), 193–202.

Geisler, W. S. (2008). "Visual Perception and the Statistical Properties of Natural Scenes". In: *Annual Review of Psychology* 59(1), 167–192.

Ghahramani, Z. & Roweis, S. T. (1999). "Learning Nonlinear Dynamical Systems using an EM Algorithm". In: *Advances in neural information processing systems*.

Goris, R. L. T., Movshon, J. A. & Simoncelli, E. P. (2014). "Partitioning neuronal variability". In: *Nature Neuroscience* 17(6), 858–865.

Goroshin, R., Mathieu, M. & LeCun, Y. (2015). "Learning to Linearize Under Uncertainty". In: *NIPS*.

Green, D. G. (1970). “Regional variations in the visual acuity for interference fringes on the retina”. In: *The Journal of Physiology*.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J. & Rubin, N. (2008). “A hierarchy of temporal receptive windows in human cortex.” In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 28(10), 2539–50.

Haykin, S. (2001). *Kalman Filtering and Neural Networks*.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Hénaff, O. J. & Simoncelli, E. P. (2015). “Geodesics of learned representations”. In: *arXiv* 1511.06394, 1–9.

Hong, H., Yamins, D. L., Majaj, N. J. & Dicarlo, J. J. (2016). “Explicit information for category-orthogonal object properties increases along the ventral stream”. In: *Nature Neuroscience*.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. (2017). “Densely connected convolutional networks”. In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. Vol. 2017-Janua, pp. 2261–2269.

Hubel, D. H. & Wiesel, T. N. (1962). “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”. In: *Journal of Physiology* 160(1), 106–154.2.

Hung, C. P., Kreiman, G., Poggio, T. & DiCarlo, J. J. (2005). “Fast readout of object identity from macaque inferior temporal cortex”. In: *Science*.

Ioffe, S. & Szegedy, C. (2015). “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *ICML* 7(6), 1–9.

Jazwinski, A. H. (1970). “Stochastic processes and filtering theory”. In: *IEEE Transactions on Automatic Control*.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. (1999). "Introduction to variational methods for graphical models". In: *Machine Learning* 37(2), 183–233.

Julier, S. J. & Uhlmann, J. K. (1997). "New extension of the Kalman filter to nonlinear systems". In: *Int Symp AerospaceDefense Sensing Simul and Controls*.

Kalman, R. E. (1960). "A New Approach to Linear Filtering and Prediction Problems". In: *Journal of Basic Engineering*.

Khaligh-Razavi, S. M. & Kriegeskorte, N. (2014). "Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation". In: *PLoS Computational Biology* 10(11).

Kingma, D. P. & Ba, J. L. (2014). "Adam: A Method for Stochastic Optimization". In: *arXiv preprint arXiv:1412.6980*, 1–15.

Kingma, D. P. & Welling, M. (2013). "Auto-Encoding Variational Bayes". In:

Krishnan, R. G., Shalit, U. & Sontag, D. (2015). "Deep Kalman Filters". In: *arXiv*.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances In Neural Information Processing Systems*, 1–9.

Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. (2015). "Human-level concept learning through probabilistic program induction". In: *Science*.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). "Deep learning". In: *Nature* 521(7553), 436–444.

Li, N. & DiCarlo, J. J. (2008). "Unsupervised Natural Experience Rapidly Alters Invariant Object Representation in Visual Cortex". In: *Science* 321(5895), 1502–1507.

Li, N. & DiCarlo, J. J. (2010). "Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex". In: *Neuron* 67(6), 1062–1075.

Lotter, W., Kreiman, G. & Cox, D. (2016). “Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning”. In:

Machens, C. K., Gollisch, T., Kolesnikova, O. & Herz, A. V. (2005). “Testing the efficiency of sensory coding with optimal stimulus ensembles”. In: *Neuron* 47(3), 447–456.

Mallat, S. (2011). “Group Invariant Scattering”. In: *arXiv.org* math.FA.

Maloney, L. T. & Yang, J. N. (2003). “Maximum likelihood difference scaling.” In: *Journal of vision* 3(8), 573–85.

Mante, V., Bonin, V. & Carandini, M. (2008). “Functional Mechanisms Shaping Lateral Geniculate Responses to Artificial and Natural Stimuli”. In: *Neuron* 58(4), 625–638.

Mathieu, M., Couprie, C. & Lecun, Y. (2016). “Deep Multi-Scale Video Prediction Beyond Mean Square Error”. In:

Noreen, D. L. (1981). “Optimal decision rules for some common psychophysical paradigms”. In: *Proceedings of the Symposium in Applied Mathematics of the American Mathematical Society and the Society for Industrial and Applied Mathematics* 13(Mathematical psychology and psychophysiology), 237–279.

Pachitariu, M., Steinmetz, N., Kadir, S., Carandini, M. & Harris, K. D. (2016). *Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels*. Tech. rep.

Palmer, S. E., Marre, O., Berry, M. J. & Bialek, W. (2015). “Predictive information in a sensory population”. In: *Proceedings of the National Academy of Sciences of the United States of America*.

Peterhans, E. & Heydt, R. von der (1989). “Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity”. In: *The Journal of Neuroscience*.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J. & Ganguli, S. (2016). “Exponential expressivity in deep neural networks through transient chaos”. In:

Portilla, J. & Simoncelli, E. P. (2000). “Parametric texture model based on joint statistics of complex wavelet coefficients”. In: *International Journal of Computer Vision*.

Rabinowitz, N. C., Goris, R. L., Cohen, M. & Simoncelli, E. P. (2015). “Attention stabilizes the shared gain of V4 populations”. In: *eLife* 4(NOVEMBER2015).

Rao, R. P. N. & Ballard, D. H. (1999). “Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects”. In: *Nature Neuroscience* 2(1), 79–87.

Rezende, D. J. & Mohamed, S. (2016). “Variational Inference with Normalizing Flows”. In:

Rezende, D. J., Mohamed, S. & Wierstra, D. (2014). “Stochastic Back-propagation and Variational Inference in Deep Latent Gaussian Models”. In: *Proceedings of The 31st . . .*

Roweis, S. T. & Saul, L. K. (2000). “Nonlinear dimensionality reduction by locally linear embedding.” In: *Science (New York, N.Y.)*

Rust, N. C. & DiCarlo, J. J. (2010). “Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT.” In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 30(39), 12978–95.

Salimans, T., Kingma, D. P. & Welling, M. (2015). “Markov Chain Monte Carlo and Variational Inference: Bridging the Gap”. In: *Proceedings of the 32nd International Conference on Machine Learning*.

Saxe, A. M., McClelland, J. L. & Ganguli, S. (2013). “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks”. In: *arXiv.org* cs.NE.

Serre, T., Oliva, A. & Poggio, T. (2007). “A feedforward architecture accounts for rapid categorization”. In: *Proceedings of the National Academy of Sciences* 104(15), 6424–6429.

Seshadrinathan, K., Soundararajan, R., Bovik, a.C. & Cormack, L. (2010). “Study of Subjective and Objective Quality Assessment of Video”. In: *IEEE Transactions on Image Processing* 19(6), 1427–1441.

Shannon, C. E. (1948). “A Mathematical Theory of Communication”. In: *Bell System Technical Journal*.

Simoncelli, E. P. & Olshausen, B. A. (2001). “Natural image statistics and neural representation”. In: *Annu Rev Neurosci* 24, 1193–1216.

Simoncelli, E. & Freeman, W. (1995). “The steerable pyramid: a flexible architecture for multi-scale\nderivative computation”. In: *Proceedings., International Conference on Image Processing* 3(NOVEMBER 1995), 444–447.

Simonyan, K. & Zisserman, A. (2015). “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *International Conference on Learning Representations (ICLR)*, 1–14.

Spelke, E. S. (1990). “Principles of object perception”. In: *Cognitive Science*.

Srinivasan, M. V., Laughlin, S. B. & Dubs, A. (1982). *Predictive coding: a fresh view of inhibition in the retina*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. & Fergus, R. (2013). “Intriguing properties of neural networks”. In: *arXiv.org* cs.CV.

Tacchetti, A., Isik, L. & Poggio, T. (2017). “Invariant recognition drives neural representations of action sequences”. In: *PLoS Computational Biology* 13(12).

Tenenbaum, J. B., De Silva, V. & Langford, J. C. (2000). “A global geometric framework for nonlinear dimensionality reduction”. In: *Science*.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. (2011). *How to grow a mind: Statistics, structure, and abstraction*.

Tishby, N., Pereira, F. C. & Bialek, W. (1999). “The information bottleneck method”. In: *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing (University of Illinois, Urbana, IL), Vol 37, pp 368–377*. 1–16.

Vintch, B., Movshon, J. A. & Simoncelli, E. P. (2015). “A Convolutional Subunit Model for Neuronal Responses in Macaque V1”. In: *Journal of Neuroscience* 35(44), 14829–14841.

Wichmann, F. A. & Hill, N. J. (2001). "The psychometric function: I. Fitting, sampling, and goodness of fit". In: *Perception & Psychophysics* 63(8), 1293–1313.

Yamins, D. L. K., Hong, H. H., Cadieu, C. F., Solomon, E. A., Seibert, D. & DiCarlo, J. J. (2014). "Performance-optimized hierarchical models predict neural responses in higher visual cortex". In: *Proceedings of the National Academy of Sciences of the United States of America* 111(23), 8619–8624.