Natural scene statistics predict how humans pool information across space in the estimation of surface tilt

Seha Kim[1.] & Johannes Burge[1.,2.,3.]

[1.] Department of Psychology, University of Pennsylvania
[2.] Neuroscience Graduate Group, University of Pennsylvania
[3.] Bioengineering Graduate Group, University of Pennsylvania

ABSTRACT

Visual systems estimate the three-dimensional (3D) structure of scenes from information in two-dimensional (2D) retinal images. Visual systems use multiple sources of information to improve the accuracy of these estimates, including statistical knowledge of the probable spatial arrangements of natural scenes. Here, we examine how 3D surface tilts are spatially related in real-world scenes, and show that humans pool information across space when estimating surface tilt in accordance with these spatial relationships. We develop a hierarchical model of surface tilt estimation that is grounded in the statistics of tilt in natural scenes and images. The model computes a global tilt estimate by pooling local tilt estimates within an adaptive spatial neighborhood. The spatial neighborhood in which local estimates are pooled changes according to the value of the local estimate at a target location. The hierarchical model provides more accurate estimates of groundtruth tilt in natural scenes and provides a better account of human performance than the local model. Taken together, the results imply that the human visual system pools information about surface tilt across space in accordance with natural scene statistics.

KEYWORDS

natural scene statistics, vision, spatial pooling, 3D surface orientation, slant, tilt

INTRODUCTION

Estimating three-dimensional (3D) surface orientation from two-dimensional (2D) retinal images is one of the most critical functions of human vision [1]. To determine whether a surface can be walked on or used to hang a picture, its 3D orientation must be accurately estimated. Laboratory studies have typically examined the estimation of 3D orientation with isolated planar surfaces that are textured with simple patterns [2-13]. However, the estimation problem in the real world is often more complex than situations that are commonly studied in the lab. Real-world surfaces have varied textures and complicated geometries [14-16]. To understand how the estimation of 3D surface orientation works in the real world, it can be useful to study performance with stimuli that are as natural as possible. In images of natural scenes, statistical information about local image cues and the most probable three-dimensional spatial contexts can both provide useful information. In a variety of visual tasks with artificial stimuli, the spatial (i.e., global) context surrounding a given set of local cues can affect local percepts [17-21]; well known examples include the simultaneous contrast illusion, the simultaneous color contrast illusion, and the slant contrast illusion [22,23]. But it is not always clear how to account for these effects. The computer vision literature frequently models the use of spatial context, but infrequently provides insights into the computations that may underlie human performance [24]. In the vision literature, there have been attempts to develop quantitative models that capture the impact of spatial context on human perception, but the stimuli that these models apply to are often rather artificial [6,25]. There have also been multiple demonstrations that global context influences human perception of surface orientation in real-world 3D scenes [26], but these studies typically do not provide quantitative models that account for human performance.

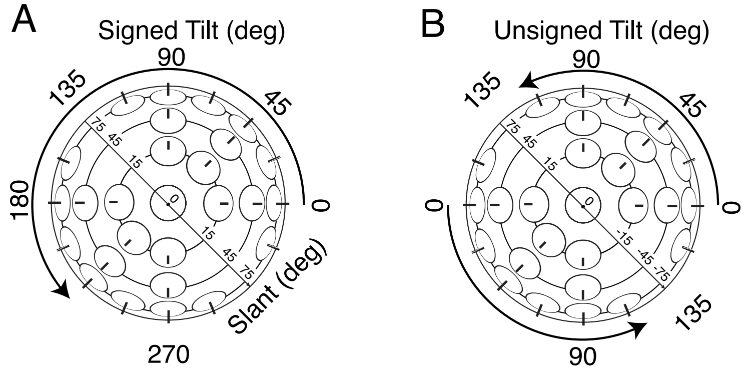

Figure 1. 3D surface orientation is fully described by slant and tilt. Slant is the angle indicating how much a surface is rotated out of the fronto-parallel plane. Tilt is the direction of slant, as quantified by the angle between the x-axis in the frontoparallel plane and the surface normal projected into the frontoparallel plane. **A** *Signed* tilt, defined on $[0^\circ, 360^\circ)$, and unsigned slant. **B** *Unsigned* tilt, defined on $[0^\circ, 180^\circ)$, and signed slant.

In this paper, we examine how humans incorporate spatial context to estimate 3D surface orientation in real-world scenes. Surface orientation is typically parameterized by slant and tilt [27]. Slant is the amount by which a surface is rotated away from an observer. Tilt is the direction in which the surface is rotated (Fig. 1). The current study focuses on the role of spatial context on tilt estimation. To use spatial context, visual systems must integrate (or pool) information across space. To model this pooling process, we propose a two-stage hierarchical model. In the first stage, 3D tilt is estimated at each of multiple spatial locations, using the joint statistics relating local cues in images to groundtruth tilts in natural scenes [15,16]. At each spatial location, the estimate is Bayes-optimal given measurements of three local image cues: local gradients of luminance, texture, and disparity. This estimate is referred to as a *local-model estimate* because it is based on local cues only. In the second stage, the local estimates are pooled across multiple spatial locations to obtain a *global-model estimate*. The global pooling rules are motivated by how tilts are spatially related in real-world scenes.

We examine the ability of the global model i) to estimate tilt in real-world scenes and ii) to predict human estimation of tilt in those same scenes. We find that the global pooling model provides more accurate estimates of tilt and better predictions of human performance than the local model. Additionally, we find that the size of pooling region that optimizes estimation performance is approximately the same as the size of the pooling region that optimizes the model predictions of human performance. The results suggest that the human visual system pools information over the spatial region that optimizes tilt estimation performance in natural scenes.

RESULTS

Natural scene statistics of tilt

In images of natural scenes, 3D surface orientations corresponding to nearby image locations are correlated. This is because natural scenes tend to be dominated by continuous surfaces, and surface discontinuities tend to be comparatively rare [28]. Visual systems that internalize and properly use the statistics governing these spatial relationships in natural scenes will outperform visual systems that do not. These scene statistics motivate the global pooling rules that are the primary focus of the paper.

To determine how surface tilts in real-world scenes are related across space, we analyzed a recently published database of natural stereo-images with precisely co-registered time-of-flight laser-based distance measurements at each pixel [15] (Fig. 2A). Groundtruth surface tilt was computed from the distance measurements at each pixel (Fig. 2B; see Methods).

The rules for pooling information across space that maximize estimation performance depend critically on how the variable to be estimated (e.g. tilt) is correlated across space. Unfortunately, the correlation of circular (i.e., angular) variables is notoriously unstable when the variables are highly dispersed, and it is known that local tilt in natural scenes is a highly dispersed circular variable [16]. This fact makes it difficult to precisely link measured scene statistics to optimal pooling rules. Thus, rather than to quantitatively specify the optimal pooling rules from first principles, we used the available natural scene statistics to motivate an exploration of plausible pooling rules.

As alternative to spatial correlation, we computed the *mean absolute tilt difference* $E\left[\left|\tau_i - \tau_j\right|\right]$ between tilts corresponding to image locations i and j. (Note that, in the main text, we use $\tau_i - \tau_j$ as notational shorthand for the circular distance between two angles; see Methods). Figure 2C shows the mean absolute tilt differences $E\left[\left|\tau_i - \tau_0\right|\right]$ in natural scenes for all possible spatial relationships in a spatial neighborhood surrounding the target tilt τ_0 at the center of an image region. Unsurprisingly, the tilt differences increase systematically as spatial distance increases; the iso-difference contours are approximately circular. This finding suggests that pooling local tilt estimates within a circular neighborhood centered on the target tilt will increase estimation accuracy (see below).

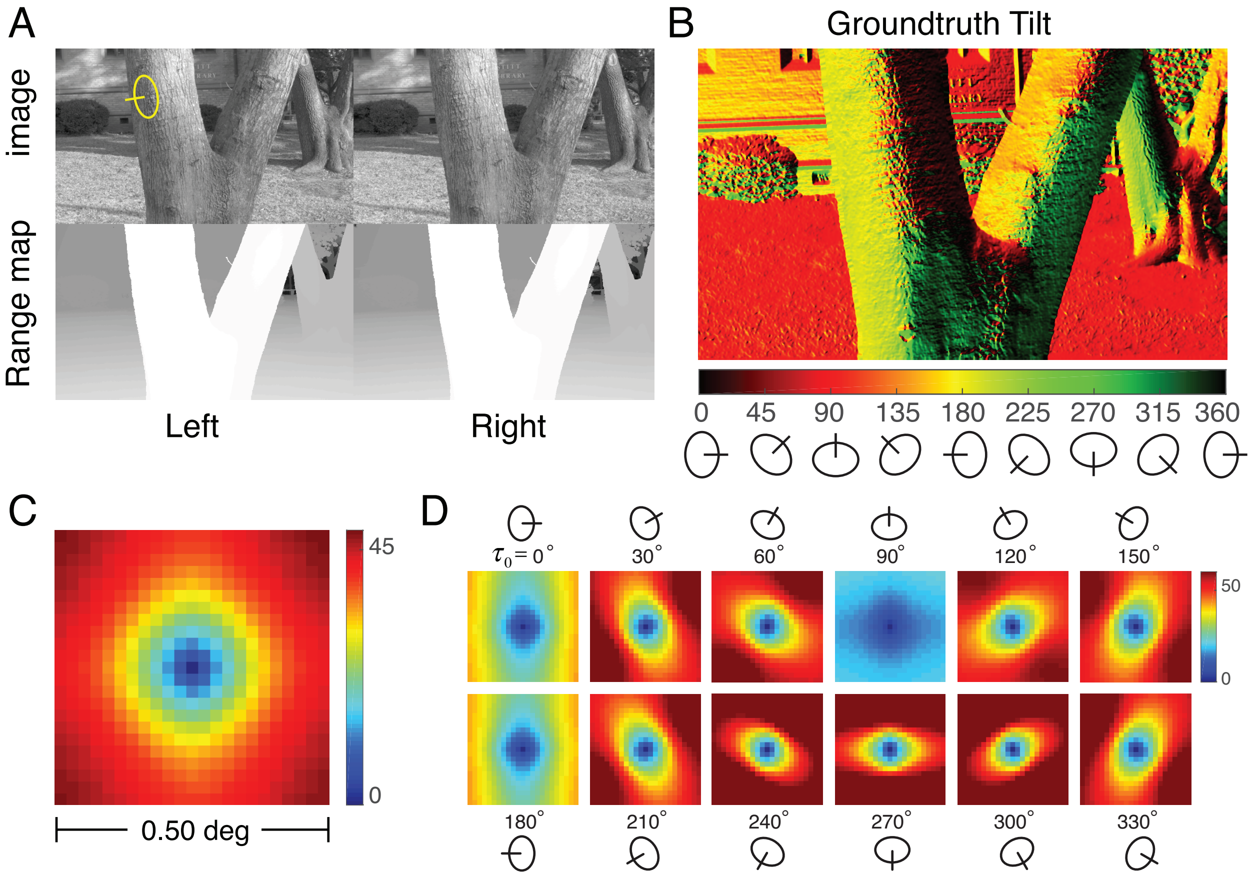

Figure 2. Spatial statistics of tilt in natural scenes. **A** Stereo images and stereo distance maps of real-world scenes. The distance data is co-registered to the image data at each pixel. **B** Groundtruth tilt corresponding to the image in A. Groundtruth tilt at each pixel is computed directly from the data in the distance maps. **C** Mean absolute tilt difference from the center target tilt as a function of spatial location. The color represents the tilt difference across all pixels in all images in the natural scene database. **D** Mean absolute tilt difference conditioned on the groundtruth tilt at the target location.

Richer statistical structure is revealed when the tilt differences are conditioned on the central target tilt (i.e., $E\left[\left|\tau_i - \tau_0\right| \mid \tau_0\right]$). The size and shape of the neighborhood within which tilts are most similar change dramatically with the target tilt; the iso-difference contours are approximately elliptical (Fig. 2D). For example, if the tilt at the center of a spatial area is equal to 0º (e.g., the side of a tree trunk), tilts at spatial locations above and below the center are more likely to be similar to the target tilt than tilts to the left and right. Thus, for a central target tilt of 0º, a vertically elongated pooling region may be appropriate. For a central target tilt of 90°, spatial locations to the left and right of the center are most likely to be similar to the target tilt, and neighboring tilts are likely to be similar over a larger area. More generally, the statistics suggest that the pooling region should be elongated in a direction that is orthogonal to the tilt direction. Visual systems that use local estimates, and pool them adaptively consistent with these statistical regularities, have the potential to outperform visual systems that pool local estimates with fixed (non-adaptive) neighborhoods. Before developing specific models that will investigate these ideas, we describe a psychophysical experiment that we performed to determine how humans estimate tilt in natural scenes.

Psychophysical experiment

To test how human observers estimate 3D tilt in natural scenes, we performed a psychophysical experiment. The scenes were obtained from the same database that we used to analyze the natural scene statistics [15] (see above). Two sets of 3600 scene locations were randomly sampled from the database under specified constraints. We only sampled regions with surfaces within a specified range of distances, slants, and image contrasts. The constraints ensured that the image cues were measureable by the human visual system, that the task was well defined, and that the stimuli could be presented without artifacts on our display system (see Methods).

Scenes were displayed on a large stereo-projection system at a 3m viewing distance (Fig. 3A). The display system creates retinal images that provide a close approximation to the retinal images and stereo-viewing geometry that viewing the original scene would have created [16]. On each trial, the observer viewed a scene location through a 3° diameter circular stereoscopic window. The task was to estimate the 3D tilt of the surface at the center of each stereoscopic window. Observers indicated their estimates with a mouse-controlled graphical probe (Fig. 3B). Human tilt estimation performance is presented alongside modeling results in subsequent sections.

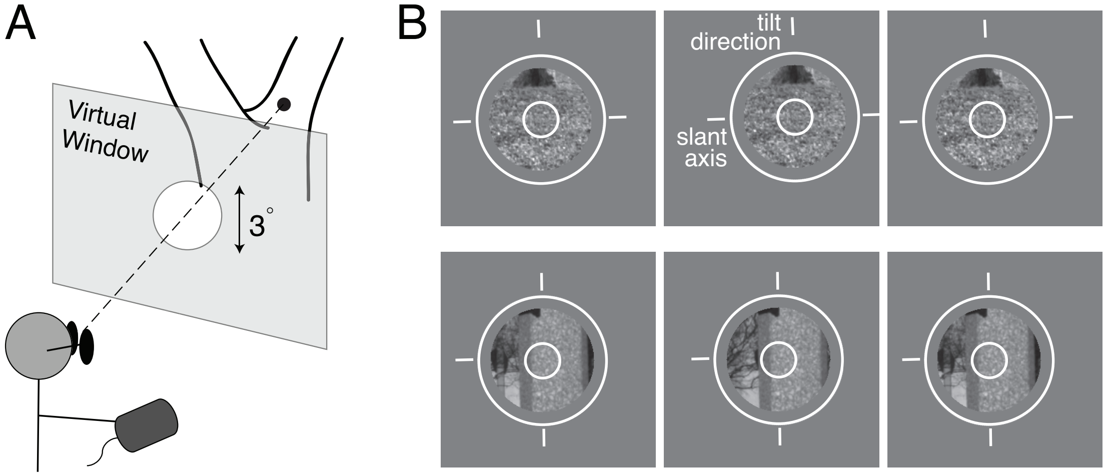

Figure 3. Human tilt estimation experiment. **A** Human observers binocularly viewed real-world scenes through a circular aperture with a 3° diameter that was positioned stereoscopically in front of the scene. **B** Example of stimuli. Left-eye, right-eye, and left-eye images (for both uncrossed and crossed fusion). The patches are displayed with a graphical probe (white circle and three tick marks). Observers rotated the probe to align the middle tick mark with the perceived tilt for the surface point in the center of the window.

Modeling

The proposed model of tilt estimation has two hierarchical processing stages. In the first stage, local estimates are computed from image cues that are extracted from natural images. In the second stage, global estimates are obtained by pooling the local estimates within a spatial neighborhood centered on a target location; the global pooling rules are motivated by our statistical analyses (cf. Fig. 2CD) of how tilts are spatially related in natural scenes. These two processing stages are described in order.

The modeling effort described here builds on previous work in two ways [16]. The primary development is in how the current model makes use of spatial context; local estimates are pooled based on the statistics of surface orientation in real-world scenes. A secondary development is that the local estimates are now of *signed* tilt (i.e., both tilt magnitude and sign) rather than of only *unsigned* tilt (i.e., tilt magnitude; Fig. 1). These developments allow us to investigate the manner in which humans pool estimates of signed tilt across space.

Local signed tilt estimation

The first stage of the hierarchical model estimates local tilt, relying on the statistics relating local image cues and surface tilt in natural scenes. The scene statistics are compiled from hundreds of millions of samples from the previously mentioned database of natural scenes[15]. The local estimation stage is based closely on a previously published local model that predicted many of the successes and failures of human tilt estimation in natural scenes [15,16]. However, the previous model had a shortcoming; it provided estimates only of unsigned tilt. The local model proposed here provides estimates of both tilt magnitude and sign (i.e., signed tilt).

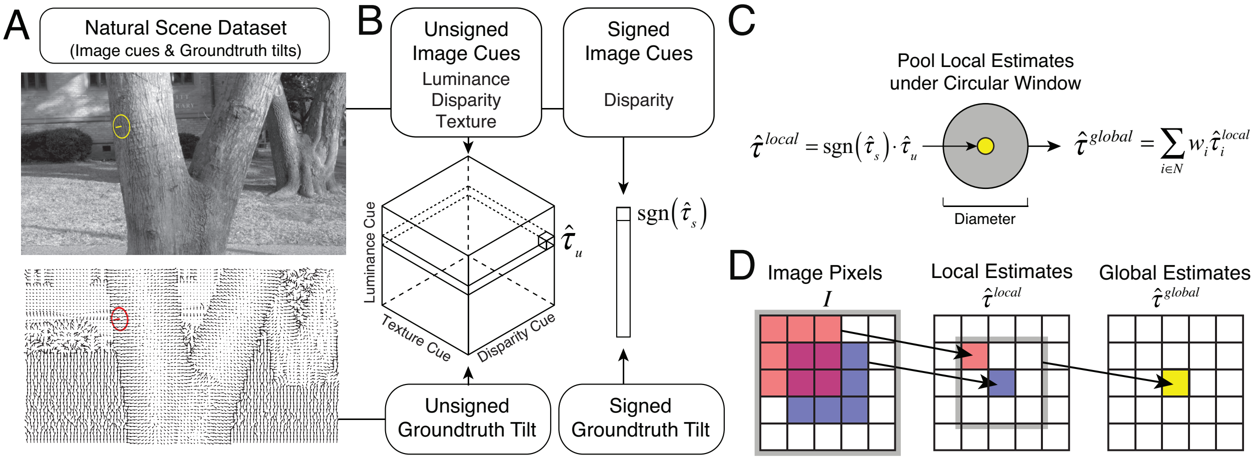

Figure 4. Constructing local and global models of tilt estimation. **A** Image cues and groundtruth tilt in natural scenes. Image cues are derived directly from photographic stereo images (top). Groundtruth tilt at each pixel is computed directly from the range data (cf. Fig. 2A). Here, groundtruth tilt is depicted with local surface normals instead of a colormap (cf. Fig. 2B). **B** The local model estimates tilt based on local image cues. Local estimates are obtained via lookup tables that store conditional means (i.e. posterior means) given all possible combinations of three quantized unsigned image cue values (i.e. 64^3 unique cue combinations), and one quantized signed image cue value (i.e. 64 unique cue values), as computed from the natural image database. We have previously verified that quantizing the cue values is not a primary limiting factor on the performance of the model [16]. **C** Pooling local estimates in a spatial pooling region centered on a target location. **D** Each global estimate is obtained by pooling local estimates over a spatial neighborhood. Each local estimate is obtained by combining cues that are computed from multiple pixels in the image. Thus, image information contributes to the global estimate from an area larger than the local estimate-pooling region.

The local model first computes the estimate of unsigned tilt given three unsigned image cues: local luminance, texture, and disparity gradients. The estimate of unsigned tilt $\hat{\tau}_u$ specifies the tilt *magnitude* and is equal to the mean of the posterior over unsigned tilt given the unsigned image cues,

$$\hat{\tau}_u = E\left[\tau_u \mid \mathbf{c}_u\right] = \sum\nolimits_u \tau_u p\left(\tau_u \mid \mathbf{c}_u\right), \qquad (1)$$

where τ_u is the unsigned groundtruth tilt, and $\mathbf{c}_u$ is a vector of three unsigned cue values (Fig. 4A cube). (Note the expression for the posterior mean in Eq. 1 is used as notational shorthand for the mean of a probability distribution over a circular variable; see Methods). The posterior mean is equivalent to the Bayes-optimal estimate assuming the circular analog to a squared-error cost function (see Methods).

The model then obtains the estimate of *tilt sign* $\mathrm{sgn}(\hat{\tau}_s)$ by computing the mean of the posterior over signed tilt given the signed disparity cue (Fig. 4A, bar), which is the only cue providing reliable information about sign. Specifically,

$$\mathrm{sgn}\left(\hat{\tau}_s\right) = \mathrm{sgn}\left(E\left[\tau_s \mid \mathbf{c}_s\right]\right) = \mathrm{sgn}\left(\sum\nolimits_s \tau_s p\left(\tau_s \mid \mathbf{c}_s\right)\right), \qquad (2)$$

where τ_s is the signed groundtruth tilt, and $\mathbf{c}_s$ is the signed disparity cue (Fig. 4A bar). The final local estimate of *signed* tilt is obtained by multiplying the estimate of tilt magnitude ($\hat{\tau}_u$ itself) by the estimate of tilt sign

$$\hat{\tau}^{local} = \hat{\tau}_u \cdot \operatorname{sgn}\left(\hat{\tau}_s\right). \tag{3}$$

(Note that an alternative approach would be to compute the posterior mean over signed tilt given all three image cues. Doing so, however, produces larger errors. We favor the more accurate method.) These local estimates are used as input to the second stage of the hierarchical model. Performance of the local model will be compared to the performance of models that use global pooling in subsequent sections.

Global tilt estimation

The second stage of the hierarchical model pools local estimates to improve performance. The pooling rules that we investigate are based on the statistical properties of tilt natural scenes. We have shown that groundtruth tilt signals exhibit statistical regularities across space (cf. Fig. 2C,D). Under these conditions, pooling local estimates has the potential to average out noise and improve performance [29]. But the benefit of averaging out noise by pooling must be balanced against the risk of averaging over groundtruth signals that are changing across space. Considering two extremes helps drive the point home. On one extreme, if local groundtruth signals are perfectly correlated across space, then the optimal pooling rule is to average all local estimates across the largest possible area. On the other extreme, if local groundtruth signals are perfectly uncorrelated across space (i.e., random), any spatial pooling at all degrades performance. Thus, to realize performance improvements, the pooling rules must be well matched to the governing statistics. If all local estimates are equally reliable, for example, the optimal pooling area should be determined by the spatial correlation of the groundtruth signals.

The global pooling models proposed here compute a global tilt estimate at a given target location from a weighted sum of the local estimates in a spatial neighborhood centered on the target location (Fig. 4B). The specific weights and the neighborhood together represent the pooling rule. Specifically, the global estimate is given by

$$\hat{\tau}^{global} = \sum_{i \in N} w_i \hat{\tau}_i^{local} \quad , \tag{4}$$

where N is the spatial neighborhood, and w_i is the weight for each local tilt estimate within the neighborhood [30-32]. (Note that Eq. 4 is notational shorthand for the weighted circular mean; see Methods). Interestingly, pooling local estimates causes image information from an area larger than the pooling region to contribute to each global estimate (Fig. 4C).

We examine the performance of global pooling relative to the local model (Fig. 5A) in the context of two global pooling strategies: *fixed circular pooling* and *adaptive elliptical pooling* (Fig. 5BC). The fixed circular pooling model uses the same pooling area regardless of the tilt at the target location in the center of the pooling region (Fig. 5B). The adaptive elliptical pooling model changes the pooling region with the target tilt (Fig. 5C). Each of these pooling strategies is motivated by the natural scene statistics shown in Fig. 2C and Fig. 2D, respectively, and is discussed in more detail below.

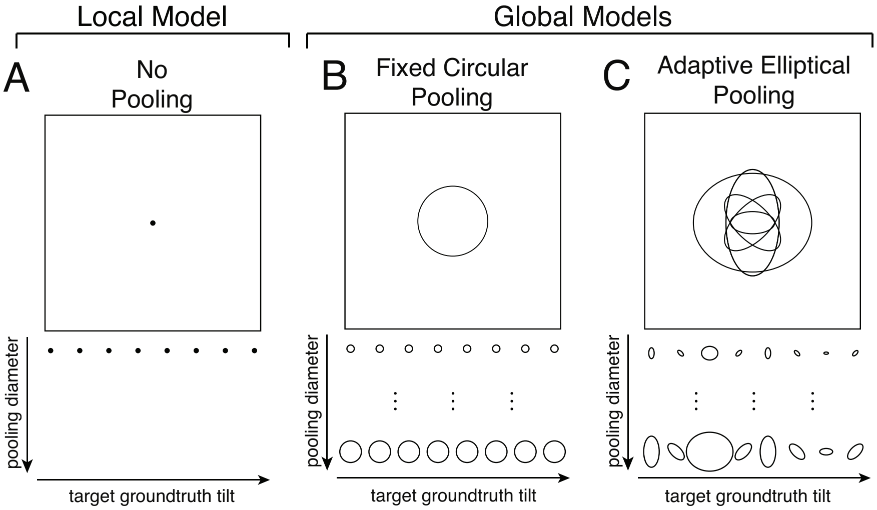

Figure 5. Local and global models for tilt estimation. **A** The local model obtains a local tilt estimate given three local image cues. **B** The fixed circular pooling model uses a circular pooling region with the same size for all target groundtruth tilts (cf. Fig. 2C). **C** The adaptive elliptical pooling model uses an adaptive pooling region with a different size, aspect ratio, and orientation for each groundtruth tilt (cf. Fig. 2D). As the average area of the adaptive elliptical pooling region changes, the relative area, orientation, and aspect ratio of the pooling regions are held fixed,

We quantify performance of each model in two ways. First, for a given pooling strategy, we analyze how pooling changes the accuracy of groundtruth tilt estimation. Second, we analyze how well a given pooling strategy accounts for human performance. If the human visual system uses global context to estimate tilt, then human responses should be better predicted by a global model that uses spatial context than by a local model that uses only local image cues. By comparing the neighborhood sizes that optimize groundtruth tilt estimation and that maximize the prediction of human performance, we gain insight into the pooling strategy that humans use when estimating tilt in natural scenes.

Fixed Circular Pooling: Modeling Results and Human Performance

We start by considering a model with a fixed circular pooling. A fixed circular pooling region is centered on the target, and it equally weights each local estimate in the pooling region (i.e., $w_i = w_j$ for all *i* and *j*; Fig. 4B). The circular shape of the pooling region is motivated by the circular shape of the iso-similarity contours in natural tilt statistics (see Fig. 2C). For this model to be optimal, two conditions must be satisfied, assuming zero noise correlations. The first condition is that all groundtruth tilts within the pooling region must have the same value. The second condition is that local estimates, regardless of their value, must provide equally reliable information about the groundtruth tilt that gave rise to the estimate. Although neither condition can be strictly true, there is some empirical justification for each. First, groundtruth tilts within a sufficiently small circular area tend to be quite similar (Fig. 2C). Assuming that all tilts are equal within the pooling region is therefore not an unreasonable approximation, provided the pooling region is not too large. Second, probability distributions over groundtruth tilt given a local estimate with a particular value (obtained from the local model) $p\left(\tau \mid \hat{\tau}^{local}\right)$ are approximately shift-invariant (Fig. 6); each local estimate is thus an equally reliable predictor of groundtruth tilt regardless of its value [16]. Pooling local tilt estimates with equal weights in a circular region is thus a reasonable starting point for investigating the degree to which spatial pooling can improve performance.

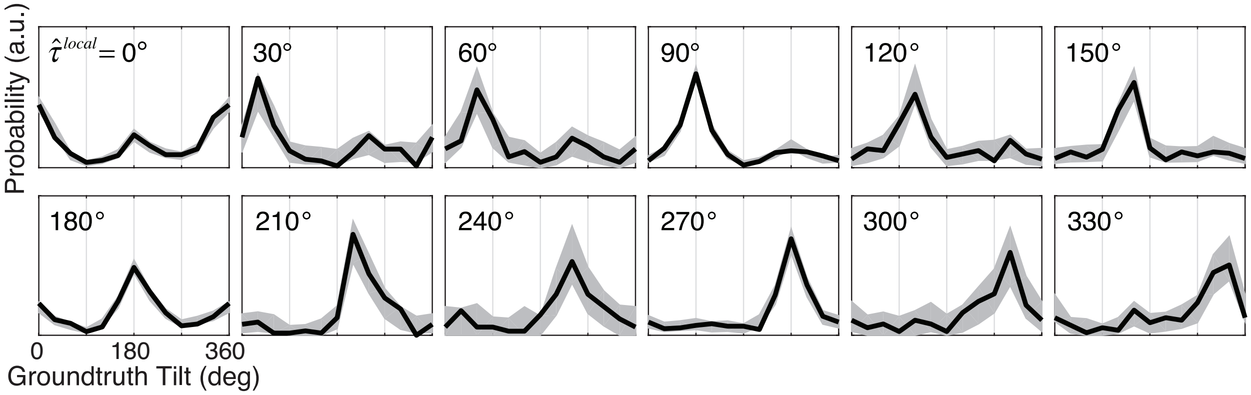

Figure 6. Conditional distributions of groundtruth tilt $p\left(\tau \mid \hat{\tau}^{local}\right)$ given the value of the *local* tilt estimate. For example, the fifth subplot in the first row shows the distribution of groundtruth tilts given that the local tilt estimate had a value of 120° $p\left(\tau \mid \hat{\tau}^{local} = 120°\right)$. The fact that the conditional distributions of groundtruth tilts are approximately shift-invariant indicates that each local tilt estimate, regardless of its value, provides approximately equally reliable information about the groundtruth tilt. Gray regions represent 95% confidence intervals from Monte Carlo simulations of 1000 experimental datasets. Confidence intervals at non-cardinal tilts (e.g. $\hat{\tau}^{local}$ = 30º, 60º, 120º, 150º, etc.) are larger in part because the local model produces fewer non-cardinal tilt estimates, in keeping with the prior probability distribution over tilt, which has peaks at the cardinal tilts (e.g. τ =0º, 90º, etc.).

To analyze the fixed circular pooling model, we examined how performance changes as a function of the pooling region diameter. First, we determined the size of the pooling region that produces the best estimates of groundtruth tilt in natural scenes (Fig. 7). Second, we determined the size of the pooling region that maximizes the model prediction of human estimates (Fig. 8).

To evaluate the model's groundtruth tilt estimation performance, we computed the *estimation error* between model estimates and groundtruth tilts across the entire stimulus ensemble used in the psychophysical experiment. The estimation error is the circular distance between model estimate and groundtruth tilt. We express neighborhood size by the diameter of the pooling region. Mean estimation error across all stimuli is plotted as a function of the pooling diameter of the circular region. With fixed circular pooling, estimation error decreases as the pooling diameter increases until it reaches a critical pooling diameter that optimizes performance (Fig. 7). The critical pooling diameter is approximately 1.0º. As pooling diameter increases further, estimation error begins to increase. Once pooling diameters exceed 3.5º, the global model fails to outperform the local model (Fig. 7; dashed line). These results show that, for a range of pooling diameters, the global model with fixed circular pooling provides more accurate estimates of tilt than does the local model. The overall benefit of global pooling is small (~10º), but it is robust. To ensure that this result is not an artifact of the experimental stimulus set, we analyzed estimation errors with a Monte Carlo simulation on 1000 sets of randomly sampled stimuli. The results show that global pooling consistently reduces estimation error; the performance improvements are not due to the particular sample of stimuli used in the psychophysical experiment.

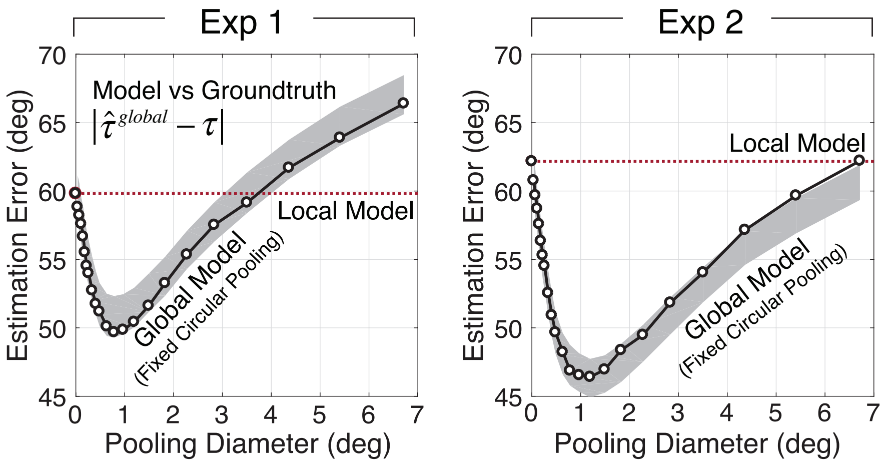

Figure 7. Groundtruth tilt estimation errors from the global model with fixed circular pooling. Mean estimation error is plotted as a function of the diameter of pooling region. Mean estimation errors are computed across all tilts. The dashed line shows mean estimation error for the local model; the local model does not pool local estimates and thus has a pooling diameter of 0º. Monte Carlo simulations on 1000 randomly sampled stimulus sets were used to obtain 95% confidence intervals on estimation error (gray area). Data from Exp. 1 and Exp. 2 are shown in the left and right columns, respectively.

To examine whether global pooling predicts human performance better than the local model, we computed the *prediction error* between model estimates and human estimates across the stimuli used in the psychophysical experiment. The prediction error is the circular distance between the model estimate and the human estimate. Mean prediction error across all stimuli is plotted as a function of the diameter of the circular pooling region. Just as with estimation error, prediction error decreases as pooling diameter increases, until a critical diameter is reached (Fig. 8). The pooling diameter that minimizes prediction error is between 1.0º and 2.0º. This diameter is similar to the diameter that minimizes estimation error (Fig. 8). The same result holds for individual human observers; the pooling diameter that minimizes prediction error is between 1.0-1.5º for four of five human observers (Fig. S1). This result suggests that the human visual system pools local estimates over an area that is sized to balance the benefits (i.e., averaging out measurement noise) and the costs (i.e., pooling over irrelevant different groundtruth tilts) to maximize accuracy.

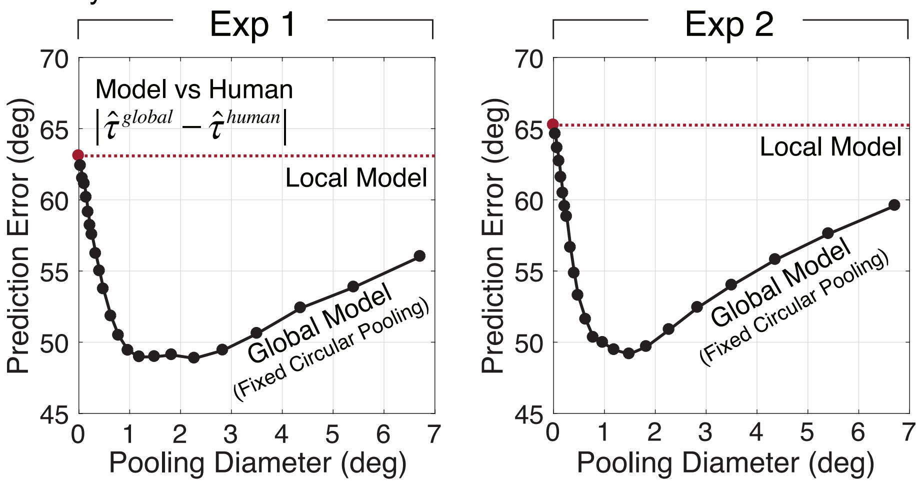

Figure 8. Human prediction errors from the global model with fixed circular pooling. Mean prediction error is plotted as a function of the diameter of pooling region. Mean prediction errors are computed across all tilts and human observers. Data from Exp. 1 and Exp. 2 are shown in the left and right columns, respectively.

Adaptive Elliptical Pooling: Modeling Results and Human Performance

Pooling local tilt estimates within a fixed circular neighborhood confers a performance benefit compared to no pooling at all. Our analyses of natural scene statistics show that the spatial neighborhood in which nearby tilts are most similar to the target tilt depends on the target tilt itself (see Fig. 2D). These elliptical regions of similarity suggest that a strategy more sophisticated than fixed circular pooling may yield additional performance improvements. The adaptive elliptical pooling model pools local estimates within an elliptical neighborhood that changes adaptively with the target tilt. The orientation, aspect ratio, and relative size of the elliptical pooling regions were fit to and fixed by the scene statistics in Fig. 2D (see Methods). The results of these fits are shown in Fig. 9A.

To determine the performance of the adaptive elliptical pooling model, we varied the average size of the elliptical pooling regions while keeping the pattern of relative size, orientation, and aspect ratio fixed, and then computed groundtruth tilt estimation errors and human prediction errors. These errors were compared to the errors obtained with fixed circular pooling and the local model. However, direct comparison with fixed circular pooling is complicated by the relative size and shape changes associated with adaptive elliptical pooling. To address this problem, we defined the *equivalent diameter* of a given ellipse as the diameter of the circle $D = 2\sqrt{A/\pi}$ that has the same area A as the ellipse. The *average equivalent diameter* $\bar{D} = 2\sqrt{\bar{A}/\pi}$ corresponds to the average ellipse area $\bar{A} = \sum_{\tau} A_i$ across target tilts where A_i is the elliptical area associated with each groundtruth tilt τ_i. For a given average equivalent diameter, the areas of the adaptive elliptical pooling regions across different target tilts are proportional to the areas of the ellipses fit to the natural scene statistics (see Figs. 5C, 9AB).

To enable direct comparison of the two global pooling models, the average equivalent diameter of the adaptive pooling model is matched with the diameter of a fixed circular pooling region. Then, estimation and prediction errors from the two models are plotted against each other as a function of average equivalent diameter (Fig 9CD). Adaptive elliptical pooling causes a small but robust improvement in estimation performance (Fig. 9C; blue curve); the minimum estimation errors from adaptive elliptical pooling were lower than those errors from fixed circular pooling on 1000 randomly sampled sets of stimuli (inset in Fig 9C).

In the analysis just presented, however, the elliptical pooling area (e.g., relative size, orientation, aspect ratio) was chosen based on the groundtruth tilt at the target location in the center of each pooling region. The groundtruth tilt at the target location (or any other location) is, of course, not directly available to the visual system. Therefore, it is important to ask whether adaptive pooling provides a benefit when the local estimate from the local model, instead of the groundtruth tilt, is used to determine the adaptive pooling region. We found that results are essentially unchanged (Fig. 9C; orange curve). When the local estimate is used, adaptive pooling improves global tilt estimation performance compared to fixed circular pooling, although the benefit in error is smaller than when groundtruth tilt is used.

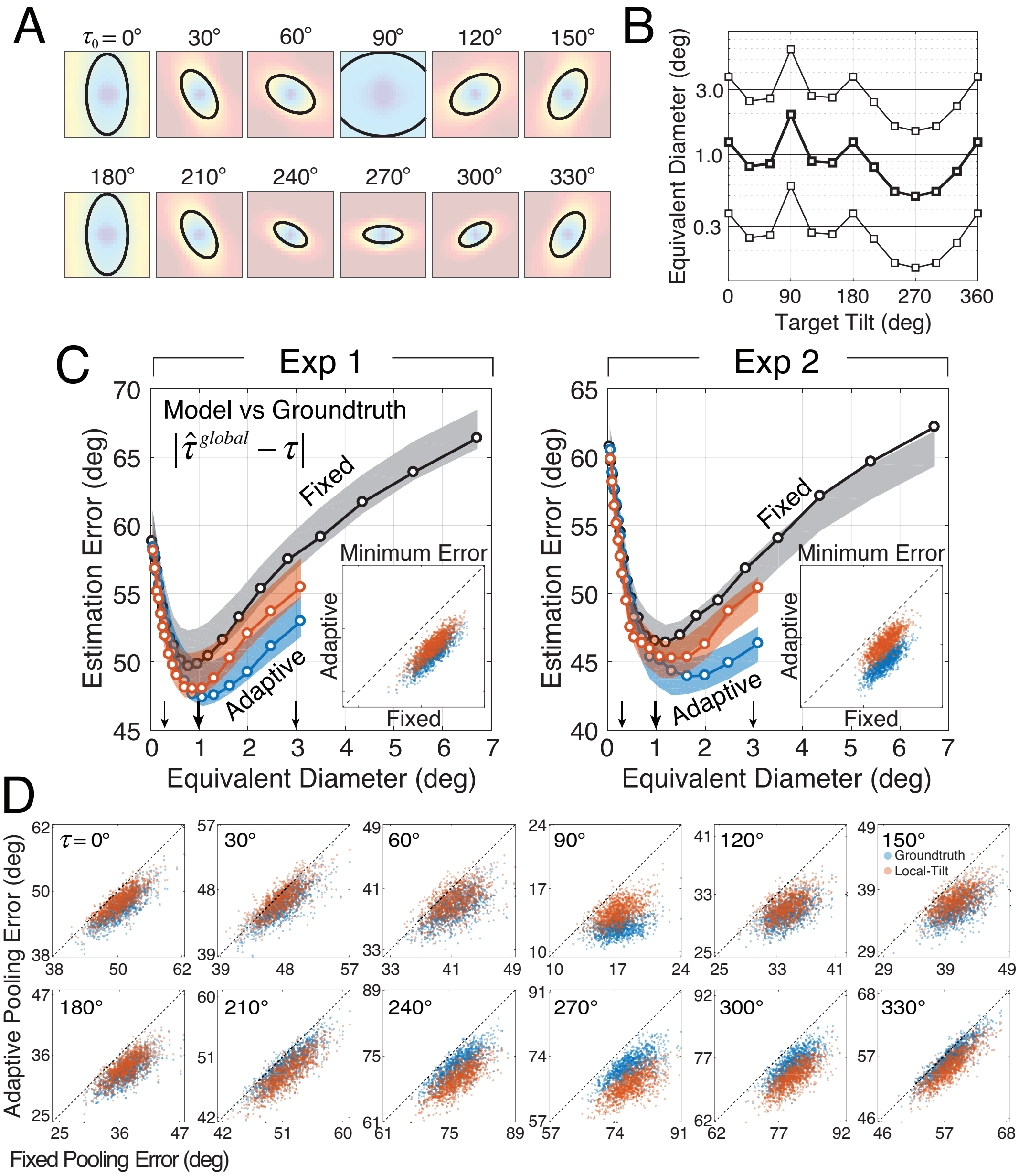

Figure 9 Estimation errors of adaptive elliptical pooling model. **A** The adaptive elliptical pooling area dictated by the target tilt. **B** The relative elliptical pooling area for different target tilts. As the average *equivalent diameter* increases or decreases, the relative size (i.e., *equivalent diameter*) remains in a fixed proportion. **C** Estimation error (model estimate vs. groundtruth tilt) as a function of pooling area. Performance is plotted for two adaptive pooling strategies: the groundtruth-based strategy chooses the elliptical pooling region based on the groundtruth tilt at the target location (blue); the local-estimate-based strategy chooses the elliptical pooling region based on the local tilt estimate as the target location (orange). The insets show simulation results that compare performance of the adaptive elliptical pooling model vs. the fixed circular pooling model on 1000 matched randomly sampled stimulus sets. Computing the prediction errors on matched stimulus sets isolates the impact of the model, and prevents stimulus variability from unduly affecting the results. Both adaptive pooling models (groundtruth-tilt-based, blue; local-estimate-based, orange) outperform the fixed circular pooling model on nearly all stimulus sets (i.e., data is below positive diagonal). **D** Simulation results, just as in C insets, except that estimation error is shown as a function of groundtruth tilt (subpanels). The fact that the majority of points lie below the dashed unity line, indicating that adaptive elliptical pooling outperforms fixed circular pooling in tilt estimation at all groundtruth tilts for the task of estimating groundtruth tilt in natural scenes.

The improvement of overall estimation performance by adaptive elliptical pooling leaves open the possibility that adaptive pooling produces a large benefit at one or only a small number of target tilts while hurting performance at other target tilts. If the natural scene statistics indeed govern the pooling rules that optimize performance, a performance improvement should be observed at each target groundtruth tilt. To check, we examined the performance of the fixed circular pooling vs. adaptive elliptical pooling at each target tilt (Fig. 9D; Fig. S2). Adaptive elliptical pooling outperforms fixed circular pooling at all target tilts regardless of whether the groundtruth tilt or the local tilt estimate at the target location is used to determine the pooling region. The results indicate that adaptive elliptical pooling improves performance compared to fixed circular pooling at each individual target tilt and provides further evidence that pooling rules governed by natural scene statistics improve estimation performance.

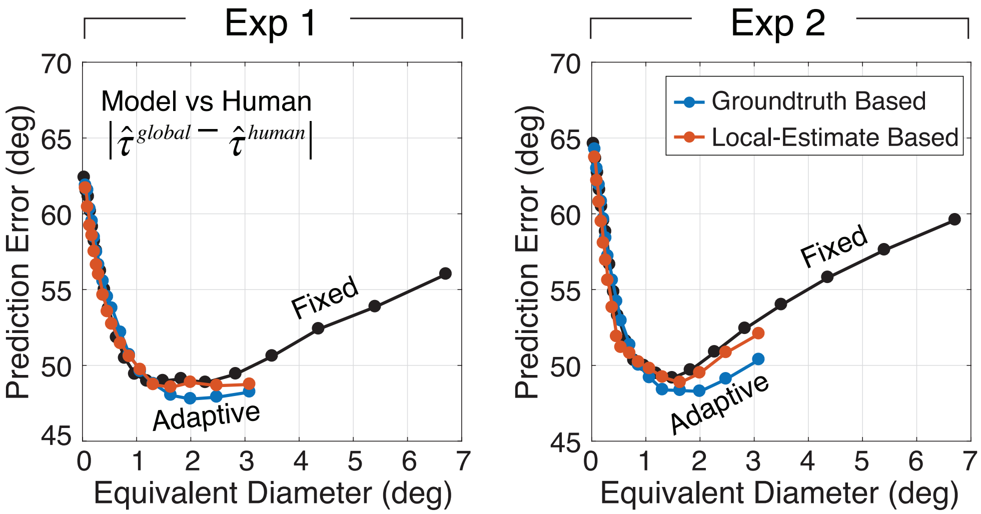

Figure 10. Prediction error of adaptive elliptical pooling model. Human prediction error (model estimate vs. human estimate) is plotted as a function of pooling area (i.e., equivalent diameter). Performance is plotted for two adaptive elliptical pooling strategies: the groundtruth-based strategy chooses the elliptical pooling region based on the groundtruth tilt at the target location (blue); the local-estimate-based strategy chooses the elliptical pooling region based on the local tilt estimate as the target location (orange). For comparative purposes, performance is also plotted for the fixed circular pooling model (black; same data as Fig. 9C).

.

The story is a bit different when it comes to the prediction of human prediction error. Adaptive elliptical pooling and fixed circular pooling provide equivalently good predictions of human performance regardless of whether the pooling region is determined by the groundtruth tilt or the local estimate at the target location (Fig. 10); similar patterns of performance are obtained for individual human observers (Fig. S1). Human prediction error therefore does not allow us to discriminate between fixed circular and adaptive elliptical pooling in so far as their ability to predict human performance. To determine which of the two models provides a better account of human performance, additional analyses are necessary.

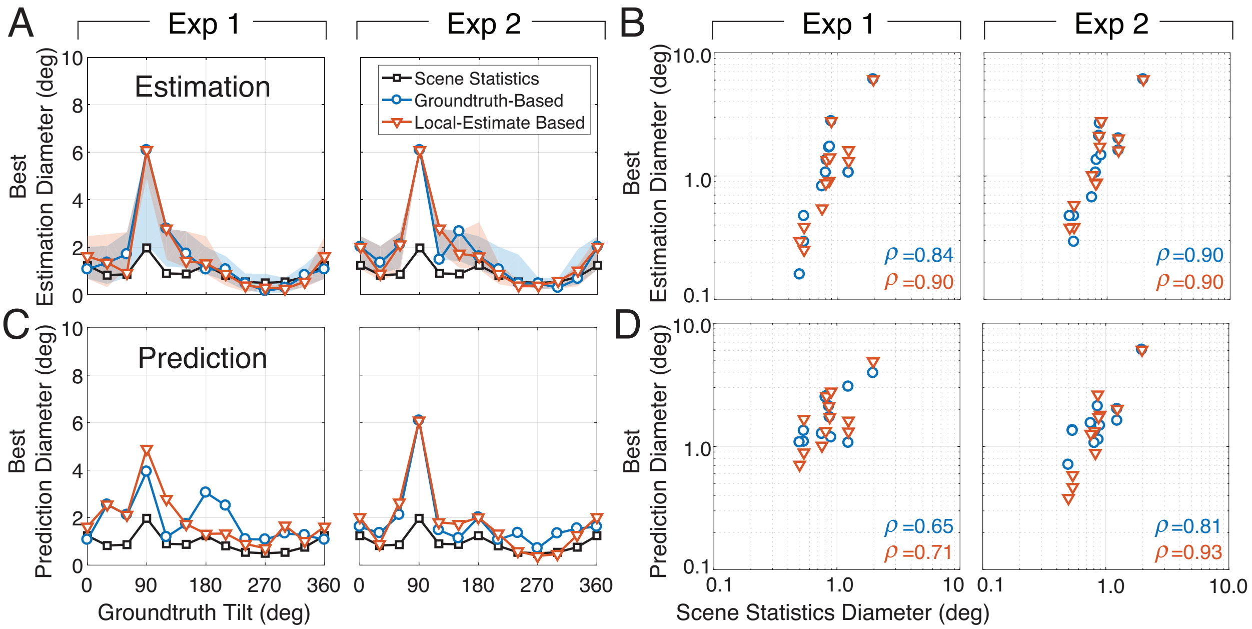

Figure 11. Adaptive pooling regions predicted by natural scene statistics predict the pooling regions that maximize performance at each groundtruth tilt. **A** Equivalent pooling diameters fit to the natural scene statistics (black; same data as Fig. 9B) and equivalent pooling diameters that minimize estimation error (groundtruth-based, blue; local-estimate-based, orange), plotted as a function of groundtruth tilt. The left and right columns represent data from Exp. 1 and Exp. 2, respectively. **B** Best estimation diameters are correlated with the diameters fit to the natural scene statistics. **C** Equivalent pooling diameter fit to the natural scene statistics and equivalent pooling diameters that minimize prediction error, plotted as a function of groundtruth tilt. **D** Best prediction diameters are correlated with the diameters fit to the natural scene statistics. All correlations were significant at the level of p<0.05; all but one were significant at the level of p<0.001.

To further examine whether adaptive or fixed pooling provides a better account of human visual processing, we determined the pooling region size that best accounts for human performance at each target tilt. If the fixed pooling model is the best account of human performance, human pooling at all target tilts should be best accounted for similarly sized pooling regions. Otherwise, the size of the best pooling region should vary systematically with the tilt at the target location. In the analyses presented thus far, the areas used in the adaptive elliptical pooling model to estimate groundtruth tilt and predict human performance were fixed by fits to the natural scene statistics (Figs. 2D,5C,9AB). But the areas determined by these fits do not necessarily match the areas that maximize the accuracy of groundtruth tilt estimation or the prediction of human performance. Thus, we independently determined the size of the elliptical pooling region that maximizes performance (i.e., minimizes error) at each groundtruth tilt. Fig. 11A plots the best equivalent pooling diameter for estimation at each groundtruth tilt against the equivalent pooling diameters that were fit to the natural scene statistics. The correlation is strong, both for the groundtruth-based and for the local-estimate-based adaptive pooling strategies. Fig. 11B plots the pooling diameters that best predict human performance at each groundtruth tilt. The same conclusions hold. Similar results are obtained if the natural scene statistics are fit over a larger spatial area (Fig. S3). Thus, for both groundtruth- and local-estimate-based adaptive strategies, the best pooling diameters for estimating groundtruth tilt and predicting human performance are tightly correlated with those obtained by fits to the natural scene statistics. These results favor the adaptive elliptical pooling model over the fixed circular pooling model as the best account of human visual processing, given that the fixed circular pooling strategy predicts no change in pooling diameter with groundtruth tilt. Natural scene statistics therefore provide a solid prediction for how signals should be pooled across space to maximize the estimation of groundtruth tilt and the prediction of human performance.

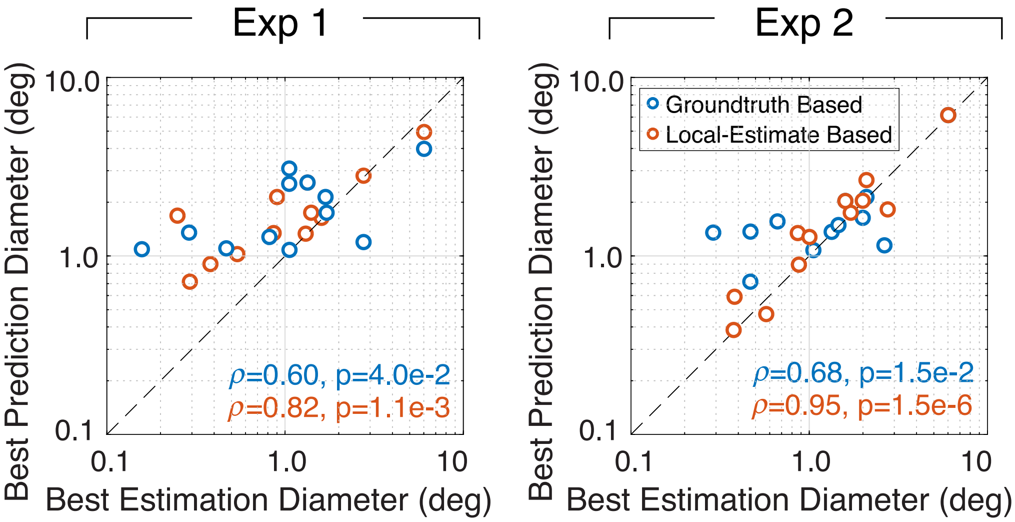

Figure 12. Pooling diameters that maximize estimation performance predict those that maximize the prediction of human performance. Each data point represents the diameter that maximizes performance for a different groundtruth tilt at the target location (cf. Fig. 11). The actual sizes of the pooling regions that maximize estimation performance in are similar to the sizes that maximize the prediction of human performance for both groundtruth-based (blue) and local-estimate-based (orange) strategies.

Finally, we compared the absolute sizes of the best pooling diameters (cf. Fig. 11) for groundtruth tilt estimation against those that are best for predicting human performance. In some sense, this is the most direct test of the hypothesis that natural scene statistics guide how humans pool information across space in surface tilt estimation. If humans use the pooling regions that yield the most accurate performance, humans are doing the right thing. The absolute sizes of the adaptive pooling regions that maximize estimation and prediction performance are strongly related to one another (Fig. 12AB). In fact, when the size of the adaptive pooling region is based on the local tilt estimate at the target location, the best estimation diameters (i.e., the equivalent pooling diameters that maximize estimation performance) are nearly the same as the best prediction diameters (i.e., the pooling diameters that maximize the model ability to predict human performance; Fig. 12B).

DISCUSSION

In this paper, we have analyzed the spatial statistics of surface tilt in natural scenes, and found that a hierarchical model of surface tilt estimation that pools local tilt estimates according to these statistics provides more accurate estimates of groundtruth tilt and better predictions of human tilt estimation than a principled model that bases performance only on local image cues [15,16]. Additionally, we have shown that the spatial scale of the pooling region that maximizes groundtruth tilt estimation performance is similar to the spatial scale that optimizes the model ability to predict human performance. Together, these findings show that natural scene statistics predict how humans pool information across space in surface tilt estimation.

The evidence for pooling

We have shown that a two-stage hierarchical model of visual information processing provides a good account of tilt estimation in natural scenes. To be confident that the second (i.e., pooling) stage of the model is necessary, the performance improvements of the global model should be due at least in part to pooling local estimates, and not due simply to the fact that the global model uses image information from a larger spatial area than the local model (cf. Fig. 4C). One approach to demonstrating this is to examine whether the performance of a local (i.e., one-stage) model, that uses information from the same area of the image as the global model, can equal the ability of the global model to estimate groundtruth tilt and predict human performance. A local model that computes cue values from an area of 1.25deg closely approximates the image area from which the best global model implicitly uses image information. (The image area that contributes to each global estimate is the sum of the areas of the pooling region and the Gaussian derivative operator used to compute the local image cues; see Methods.) This area-matched local model underperforms the best global model. Note that we cannot exhaustively examine all possible local models. We therefore cannot rule out the possibility that there exists some local model—that, for example, uses a different set of local cues—that can achieve performance equivalent to the global model. However, at least in the space of models that we have considered, the demonstrated benefit of global pooling cannot be trivially explained by the fact that the global model implicitly uses image information over a larger area than the local model. The demonstrated links between the scene statistics and the best performing adaptive global model suggests that pooling according to the natural scene statistics benefits performance.

Visual systems and the internalization of natural scene statistics

In recent years, a series of papers have provided evidence linking certain statistical aspects of natural images and scenes [15,28,33-38] to the design of the human visual system [39-41], and to the performance of ideal and human observers in perceptual tasks [14,16,37,42-51]. This broad program of research has, with varying degrees of rigor, invoked natural scene statistics to account for a strikingly diverse set of topics: how the shape of pupils changes across species in different ecological niches [41], where corresponding points are located in the two retinas [39,40], how biases in binocular eye movements manifest [48], how targets are detected in natural images [47], how image contours are perceptually grouped [37,42], how image orientation is estimated [45], how focus error is estimated[50,51], how binocular disparity is estimated [44,52,53], how image motion is estimated[46,49,54], how 3D tilt is estimated [16], and now, how cues to 3D tilt are pooled across space. Over this same period, numerous modeling frameworks have emerged that provide theoretical and computational methods for predicting and accounting for these links [52,55-58]. The coming years are likely to demonstrate more links between properties of natural scenes and functional properties of sensory-perceptual systems.

Adaptive spatial spooling

The current work indicates that the human visual system adaptively pools information across a spatial neighborhood that is closely related to the spatial neighborhood that maximizes the ability to estimate groundtruth tilt in natural scenes. A number of investigations have found evidence for adaptive spatial pooling. Local image properties (e.g., contrast) at a target location influence the spatial region over which information is integrated, both at the level of individual neurons and at the level of perception [59,60]. The current work shows that, in the domain of surface tilt estimation, the rules governing adaptive pooling are linked to the statistics of natural scenes. However, local estimates should only be pooled if they carry information about the same physical source; local estimates should not typically be pooled across depth boundaries [61-63]. The current work is limited in that it does not explicitly address how the visual system should avoid pooling across depth boundaries. This is left as a project for the future.

Spatial pooling and cue combination

The logic underlying the spatial pooling rules investigated here is closely related to the logic underlying standard theories of cue combination. Spatial pooling and cue combination both rely on the simple fact that multiple sources of information are better than one, provided that the sources are properly combined. In the current analysis of spatial pooling, the individual local tilt estimates play a similar role that individual cues play in cue combination. The difference between cue combination and spatial pooling is in the point of emphasis. Cue combination has most often been studied for cases in which multiple measurements of a single groundtruth value are available at the same spatial location (but see [31,64,65]). Spatial pooling, by contrast, focuses on the integration of pieces of information at multiple different spatial locations, which often correspond to multiple groundtruth values. Developing computational and experimental paradigms to rigorously explore these distinctions is an important goal for future work.

METHODS

Human observers

Four human observers participated in the two experiments; two authors and one naïve subject participated in Experiment 1 and one author and a different naïve subject participated in Experiment 2. Informed consent was obtained from participants before the experiment. The research protocol was approved by the Institutional Review Board (IRB) at the University of Pennsylvania (protocol number 824435) and is in accordance with the Declaration of Helsinki.

Natural scene database

Natural stimuli were sampled from a recently published natural scene database containing stereo-images with precisely co-registered distance data of natural scenes [15]. The images for left and right eyes were taken at two positions separated by a typical human inter-pupillary distance (6.5cm). Scenes were photographed such that no objects were nearer than 3m and such that all images were sharp. The left- and right-eye images associated with each of the 95 stereo pairs had a resolution of 1080x1920 pixels. The natural scenes depicted in the database contain buildings, streets, shrubs, trees, and open green areas.

Apparatus

Stereo-image patches were presented with a ViewPixx Technologies ProPixx projector fitted with dynamic polarization filter. Left and right eye images were temporally interleaved at a refresh rate of 120 Hz (60 Hz per eye). Projected images were displayed on a 2.0x1.2m Harkness Clarus 140XC polarization-maintaining screen and viewed through passive polarization maintaining goggles. At the 3m viewing distance, the screen subtended 36ºx21º of visual angle. The 3m screen distance minimized screen cues to flatness depth because the blur detection threshold is approximately 1/3D [66]. Head position was stabilized by a chinrest and a headrest. The display system nearly recreates the retinal images that would have been formed by the original scenes. The primary difference is that the overall intensity of the light reaching the eyes is lower because sunlight is more intense than the max intensity produced by the projector (84cd/m^2).

Experimental stimuli

Each natural scene was viewed binocularly in gray scale through a virtual stereoscopic aperture. The aperture had a diameter of 3° of visual angle and was positioned 5 arcmin of disparity in front of the surface that appeared at its center (Fig. 2A). Scene locations (i.e., patches) were sampled with a number of constraints. In Experiment 1, patches were excluded i) if the center pixel was associated with a surface slant of less than 30º or more than 75º, (ii) if the center pixel was associated with a surface distance that was less than 5m or larger than 50m, (iii) if the center pixel was in a half-occluded region, and (iv) if the root-mean-squared contrast of the patch was less than 5% or greater than 40%. In Experiment 2, all constraints were the same except the acceptable surface slants were between 30º and 60º instead of between 30º and 75º. Stimuli were selected so that all tilts were evenly represented in the experiments. For each of 24 bins that were 15º wide (24 bins x 15º = 360º), 150 stereo-image patches were selected for a total of 3600 unique patches (3600 = 24 x 150). In both experiments, the patches were displayed at the image location from which they were sampled.

Procedure

Data was collected in 24 blocks of randomly permutated trials. Each block lasted approximately 10 min. On each trial, observers estimated the tilt at the center of each patch. The task was to match the surface tilt angle with the orientation of a mouse-controlled graphical probe. The initial probe orientation was randomly selected. There was no time limit for the response. No feedback was provided.

Groundtruth tilt

Groundtruth tilt τ is computed from the distance data contained in the range map $\mathbf{r}$ that is co-registered to each natural image in the database. We defined groundtruth tilt $\tan^{-1}\left(\nabla_y \mathbf{r} / \nabla_x \mathbf{r}\right)$ as the orientation of the range gradient[1], where $\tan^{-1}(\cdot)$ is the four quadrant inverse tangent function. The range gradient was computed by first convolving the groundtruth distance data with a 2D Gaussian with space constant σ_{tilt} and then taking the partial derivatives in the x and y image directions[15,16]. Groundtruth tilt was computed using a space constant σ_{tilt} of 3arcmin, corresponding to analysis kernels of approximately 0.25°x0.25°.

Image cues to tilt

Image cue gradients were computed directly from the cue images. The disparity and luminance gradients were defined as the orientation of the local disparity and luminance gradients, $\tan^{-1}\left(\nabla_y cue / \nabla_x cue\right)$ and were computed by convolving the cue image with a 2D Gaussian having space constant σ_{cue} of 6arcmin and then taking the partial derivatives in the x and y image directions. The disparity image was computed from the left- and right-eye luminance images via standard local windowed cross-correlation [15,29,67]. The cross-correlation window had the same space constant as the derivative operator that computed the gradient. The texture cue was defined as the dominant orientation of the major axis of the local amplitude spectrum[10,15]. The unsigned cue values were obtained by taking the 180° modulus of the signed gradients.

Fitting elliptical pooling regions to scene statistics

To determine how 3D surface tilt is spatially related in natural scenes we computed the mean tilt difference as a function of spatial offset in the image (Fig. 2CD). Then, we fit a two-dimensional Gaussian to the map of tilt differences after scaling the map so the volume equaled 1.0. The aspect ratio and relative size of the elliptical pooling regions were determined from the covariance matrix of the best-fit Gaussian, the iso-level curves of which are ellipses. The orientation of the elliptical pooling regions was aligned with the orientation of the target tilt.

Mean absolute tilt difference

Tilt is a circular (i.e., angular) variable. The difference between two tilts is determined via the circular distance, the standard method of computing the difference between circular variables. The absolute circular distance between a pair of tilts at two different locations is given by

$$\left|\tau_i - \tau_0\right| = \left|\arg\left(\exp\left[j\left(\tau_i - \tau_0\right)\right]\right)\right|, \qquad (5)$$

where τ_0 is the tilt at the target location, τ_i is the tilt at a neighboring location, and $j = \sqrt{-1}$ is the imaginary unit number. The mean circular distance across pairs of tilts in a given spatial relationship is given by

$$E\left[\left|\tau_i - \tau_0\right|\right] = \arg\left(\sum_{k=1}^{N} \exp\left(j\left|\tau_{i_k} - \tau_{0_k}\right|\right)\right), \tag{6}$$

where N is the number of tilt pairs contributing to the mean. Tilt differences (i.e., circular means) are plotted as a function of spatial location (relative to a target location) in Fig. 2CD.

Local Model: Estimating tilt magnitude

Tilt magnitude (i.e., unsigned tilt) is estimated from three *unsigned* tilt cues, $\mathbf{c}_u = \left\{l_u, t_u, d_u\right\}$, where l_u is the unsigned luminance cue, t_u is the unsigned texture cue, and d_u is the unsigned disparity cue at the target location. The tilt estimate is the conditional mean $\hat{\tau}_u = E\left[\tau_u \middle| \mathbf{c}_u\right]$ given a triplet of image cue measurements $\mathbf{c}_u$ (Fig. 4). The conditional mean is identical to the mean of the posterior over unsigned tilt assuming a minimum circular distance cost function (i.e., analogous to the mean squared error cost function for linear variables). The posterior mean equals the sample mean from a large number of samples of τ_u in the natural scene database, assuming the samples are representative. Tilt is a circular variable. The conditional mean is thus given by

$$\hat{\tau}_u = E\left[\tau_u \middle| \mathbf{c}_u\right] = \arg\left(\frac{1}{N}\sum_{\tau_u \in \Omega} e^{i\tau_u}\right), \tag{7}$$

where Ω_u is the set of unsigned groundtruth tilts τ_u co-occurring with the triplet of cue values $\mathbf{c}_u$, and N is the number of tilt samples. On test images, the cue triplet is computed from the images and the optimal tilt is obtained from a lookup table (cf. estimate cube in Fig. 4A).

Local Model: Estimating tilt sign

Tilt sign is determined from the *signed* disparity cue only $\mathbf{c}_s = \left\{d_s\right\}$, where d_s is the signed disparity cue at the target location. The

$$\operatorname{sgn}\left(\hat{\tau}_s\right) = \operatorname{sgn}\left(E\left[\tau_s \middle| \mathbf{c}_s\right]\right) = \operatorname{sgn}\left(\arg\left(\frac{1}{N}\sum_{\tau_s \in \Omega} e^{i\tau_s}\right)\right), \tag{8}$$

where Ω_s is the set of signed groundtruth tilts co-occurring with the signed disparity values $\mathbf{c}_s$ and N is the number of tilt samples.

ACKNOWLEDGMENTS

This work was supported by NIH grant R01-EY028571 from the National Eye Institute and the Office of Behavioral and Social Science Research (JB), and NIH grant R01-EY011747 from the National Eye Institute (JB).

REFERENCES

1. Marr D. Vision. New York: W H Freeman & Company; 1982.
2. Knill DC. Ideal observer perturbation analysis reveals human strategies for inferring surface orientation from texture. Vision Research. 1998;38: 2635–2656.
3. Knill DC. Surface orientation from texture: ideal observers, generic observers and the information content of texture cues. Vision Research. 1998;38: 1655–1682.
4. Hillis JM, Watt SJ, Landy MS, Banks MS. Slant from texture and disparity cues: optimal cue combination. Journal of Vision. 2004;4: 967–992. doi:10.1167/4.12.1

5. Todd JT. The visual perception of 3D shape. Trends Cogn Sci. 2004;8: 115–121. doi:10.1016/j.tics.2004.01.006
6. Li A, Zaidi Q. Three-dimensional shape from non-homogeneous textures: carved and stretched surfaces. Journal of Vision. 2004;4: 860–878. doi:10.1167/4.10.3
7. Norman JF, Todd JT, Norman HF, Clayton AM, McBride TR. Visual discrimination of local surface structure: slant, tilt, and curvedness. Vision Research. 2006;46: 1057–1069. doi:10.1016/j.visres.2005.09.034
8. Malik J, Rosenholtz R. Computing local surface orientation and shape from texture for curved surfaces. International Journal of Computer Vision. 1997.
9. Burge J, Girshick AR, Banks MS. Visual-haptic adaptation is determined by relative reliability. J Neurosci. 2010;30: 7714–7721. doi:10.1523/JNEUROSCI.6427-09.2010
10. Fleming RW, Holtmann-Rice D, Bulthoff HH. Estimation of 3D shape from image orientations. Proc Natl Acad Sci. 2011;108: 20438–20443. doi:10.1073/pnas.1114619109/-/DCSupplemental
11. Marlow PJ, Todorović D, Anderson BL. Coupled computations of three-dimensional shape and material. Curr Biol. 2015;25: R221–2. doi:10.1016/j.cub.2015.01.062
12. Girshick AR, Banks MS. Probabilistic combination of slant information: Weighted averaging and robustness as optimal percepts. Journal of Vision. 2009;9: 8–8. doi:10.1167/9.9.8
13. Rosenberg A, Cowan NJ, Angelaki DE. The visual representation of 3D object orientation in parietal cortex. J Neurosci. 2013;33: 19352–19361. doi:10.1523/JNEUROSCI.3174-13.2013
14. Hansen BC, Essock EA. A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes. Journal of Vision. 2004;4: 1044–1060. doi:10.1167/4.12.5
15. Burge J, McCann BC, Geisler WS. Estimating 3D tilt from local image cues in natural scenes. Journal of Vision. 2016;16: 2. doi:10.1167/16.13.2
16. Kim S, Burge J. The lawful imprecision of human surface tilt estimation in natural scenes. eLife. 2018;7. doi:10.7554/eLife.31448
17. Yang Y, Blake R. On the accuracy of surface reconstruction from disparity interpolation. Vision Research. 1995;35: 949–960.
18. Parkes L, Lund J, Angelucci A, Solomon JA, Morgan M. Compulsory averaging of crowded orientation signals in human vision. Nat Neurosci. 2001;4: 739–744. doi:10.1038/89532
19. McDermott J. Psychophysics with junctions in real images. Perception. 2004;33: 1101–1127. doi:10.1068/p5265
20. Oliva A, Torralba A. The role of context in object recognition. Trends Cogn Sci. 2007;11: 520–527. doi:10.1016/j.tics.2007.09.009
21. Galleguillos C, Belongie S. Context based object categorization: A critical survey. Computer Vision and Image Understanding. Elsevier Inc; 2010;114: 712–722. doi:10.1016/j.cviu.2010.02.004
22. Adelson EH. Lightness Perception and Lightness Illusions. In: Gazzaniga M, editor. The New Cognitive Sciences. 2nd ed. Cambridge, MA: MIT Press; 2000. pp. 339–351. doi:10.3758/bf03195016
23. van Ee R, Banks MS, Backus BT. An analysis of binocular slant contrast. Perception. 1999;28: 1121–1145. doi:10.1068/p2961
24. Saxena A, Chung SH, Ng AY. 3-d depth reconstruction from a single still image. International Journal of Computer Vision. Springer; 2008;76: 53–69. doi:10.1007/s11263-007-0071-y
25. Mamassian P, Landy MS. Observer biases in the 3D interpretation of line drawings. Vision Research. 1998;38: 2817–2832. doi:10.1016/s0042-6989(97)00438-0

26. Todd JT, Koenderink JJ, van Doorn AJ, Kappers AM. Effects of changing viewing conditions on the perceived structure of smoothly curved surfaces. J Exp Psychol Hum Percept Perform. 1996;22: 695–706.
27. Stevens KA. Slant-tilt: The visual encoding of surface orientation. Biol Cybern. 1983;46: 183–195. doi:10.1007/BF00336800
28. Iyer AV, Burge J. Depth variation and stereo processing tasks in natural scenes. Journal of Vision. 4 ed. 2018;18: 1–22. doi:10.1167/18.6.4
29. Banks MS, Gepshtein S, Landy MS. Why is spatial stereoresolution so low? J Neurosci. 2004;24: 2077–2089. doi:10.1523/JNEUROSCI.3852-02.2004
30. Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. Nature. 2002;415: 429–433. doi:10.1038/415429a
31. Gepshtein S, Burge J, Ernst MO, Banks MS. The combination of vision and touch depends on spatial proximity. Journal of Vision. 2005;5: 1013–1023. doi:10.1167/5.11.7
32. Murray RF, Morgenstern Y. Cue combination on the circle and the sphere. Journal of Vision. 2010;10: 15. doi:10.1167/10.11.15
33. Jones JP, Palmer LA. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J Neurophysiol. 1987;58: 1233–1258.
34. Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. 1996;381: 607–609. doi:10.1038/381607a0
35. Yang Z, Purves D. Image/source statistics of surfaces in natural scenes. Network. 2003;14: 371–390.
36. Adams WJ, Elder JH, Graf EW, Leyland J, Lugtigheid AJ, Muryy A. The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. Sci Rep. 2016;6: 35805. doi:10.1038/srep35805
37. Geisler WS, Perry JS. Contour statistics in natural images: grouping across occlusions. Vis Neurosci. 2009;26: 109–121. doi:10.1017/S0952523808080875
38. Iyer A, Burge J. The statistics of how natural images drive the responses of neurons. Journal of Vision. 2019;19: 4. doi:10.1167/19.13.4
39. Cooper EA, Burge J, Banks MS. The vertical horopter is not adaptable, but it may be adaptive. Journal of Vision. 2011;11. doi:10.1167/11.3.20
40. Sprague WW, Cooper EA, Tosic I, Banks MS. Stereopsis is adaptive for the natural environment. Science Advances. American Association for the Advancement of Science; 2015;1: e1400254–e1400254. doi:10.1126/sciadv.1400254
41. Banks MS, Sprague WW, Schmoll J, Parnell JAQ, Love GD. Why do animal eyes have pupils of different shapes? Science Advances. 2015;1: e1500391. doi:10.1126/sciadv.1500391
42. Elder JH, Goldberg RM. Ecological statistics of Gestalt laws for the perceptual organization of contours. Journal of Vision. 2002;2. doi:10.1167/2.4.5
43. Geisler WS, Perry JS, Super BJ, Gallogly DP. Edge co-occurrence in natural images predicts contour grouping performance. Vision Research. 2001;41: 711–724.
44. Burge J, Fowlkes CC, Banks MS. Natural-scene statistics predict how the figure-ground cue of convexity affects human depth perception. J Neurosci. 2010;30: 7269–7280. doi:10.1523/JNEUROSCI.5551-09.2010
45. Girshick AR, Landy MS, Simoncelli EP. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. Nat Neurosci. 2011;14: 926–932. doi:10.1038/nn.2831
46. Burge J, Geisler WS. Optimal speed estimation in natural image movies predicts human performance. Nat Commun. 2015;6: 7900. doi:10.1038/ncomms8900
47. Sebastian S, Abrams J, Geisler WS. Constrained sampling experiments reveal principles of detection in natural scenes. Proc Natl Acad Sci. 2017. doi:10.1073/pnas.1619487114
48. Gibaldi A, Banks MS. Binocular Eye Movements are Adapted to the Natural Environment.

J Neurosci. 2019. doi:10.1523/JNEUROSCI.2591-18.2018
49. Chin BM, Burge J. Predicting the partition of behavioral variability in speed perception with naturalistic stimuli. bioRxiv. 2019;: 1–37. doi:10.1101/601161
50. Burge J, Geisler WS. Optimal defocus estimation in individual natural images. Proc Natl Acad Sci. 2011;108: 16849–16854. doi:10.1073/pnas.1108491108
51. Burge J, Geisler WS. Optimal defocus estimates from individual images for autofocusing a digital camera. Proceedings of SPIE; 2012. p. 82990E. doi:10.1117/12.912066
52. Burge J, Geisler WS. Optimal disparity estimation in natural stereo images. Journal of Vision. 2014;14. doi:10.1167/14.2.1
53. Goncalves NR, Welchman AE. “What Not” Detectors Help the Brain See in Depth. Curr Biol. 2017;27: 1403–1412.e8. doi:10.1016/j.cub.2017.03.074
54. Kane D, Bex P, Dakin S. Quantifying “the aperture problem” for judgments of motion direction in natural scenes. Journal of Vision. 2011;11. doi:10.1167/11.3.25
55. Weiss Y, Simoncelli EP, Adelson EH. Motion illusions as optimal percepts. Nat Neurosci. 2002;5: 598–604. doi:10.1038/nn858
56. Wei X-X, Stocker AA. A Bayesian observer model constrained by efficient coding can explain “anti-Bayesian” percepts. Nature Publishing Group. 2015;18: 1509–1517. doi:10.1038/nn.4105
57. Burge J, Jaini P. Accuracy Maximization Analysis for Sensory-Perceptual Tasks: Computational Improvements, Filter Robustness, and Coding Advantages for Scaled Additive Noise. PLoS Comput Biol. 2017;13: e1005281. doi:10.1371/journal.pcbi.1005281
58. Jaini P, Burge J. Linking normative models of natural tasks to descriptive models of neural response. Journal of Vision. 2017;17: 16. doi:10.1167/17.12.16
59. Sceniak MP, Ringach DL, Hawken MJ, Shapley R. Contrast's effect on spatial summation by macaque V1 neurons. Nat Neurosci. 1999;2: 733–739. doi:10.1038/11197
60. Tadin D, Lappin JS, Gilroy LA, Blake R. Perceptual consequences of centre-surround antagonism in visual motion processing. Nature. 2003;424: 312–315. doi:10.1038/nature01800
61. Ing AD, Wilson JA, Geisler WS. Region grouping in natural foliage scenes: image statistics and human performance. Journal of Vision. 2010;10: 10.1–19. doi:10.1167/10.4.10
62. Vilankar KP, Golden JR, Chandler DM, Field DJ. Local edge statistics provide information regarding occlusion and nonocclusion edges in natural scenes. Journal of Vision. 2014;14. doi:10.1167/14.9.13
63. Ehinger KA, Adams WJ, Graf EW, Elder JH. Local depth edge detection in humans and deep neural networks. 2017. pp. 2681–2689.
64. Takahashi T. A novel view of hearing in reverberation. Neuron. 2009;62: 6–7. doi:10.1016/j.neuron.2009.04.004
65. Ernst MO. Optimal Multisensensory Integration: Assumptions and Limits. In: Stein B, editor. The New Handbook of Multisensory Processes. MIT Press Cambridge, MA; 2012. pp. 527–543.
66. Sebastian S, Burge J, Geisler WS. Defocus blur discrimination in natural images with natural optics. Journal of Vision. 2015;15: 16. doi:10.1167/15.5.16
67. Tyler CW, Julesz B. Binocular cross-correlation in time and space. Vision Research. Elsevier; 1978;18: 101–105.

Supplement

Natural scene statistics predict how humans pool information across space in the estimation of surface tilt

Seha Kim[1.] & Johannes Burge[1.,2.,3.]

[1.] Department of Psychology, University of Pennsylvania
[2.] Neuroscience Graduate Group, University of Pennsylvania
[3.] Bioengineering Graduate Group, University of Pennsylvania

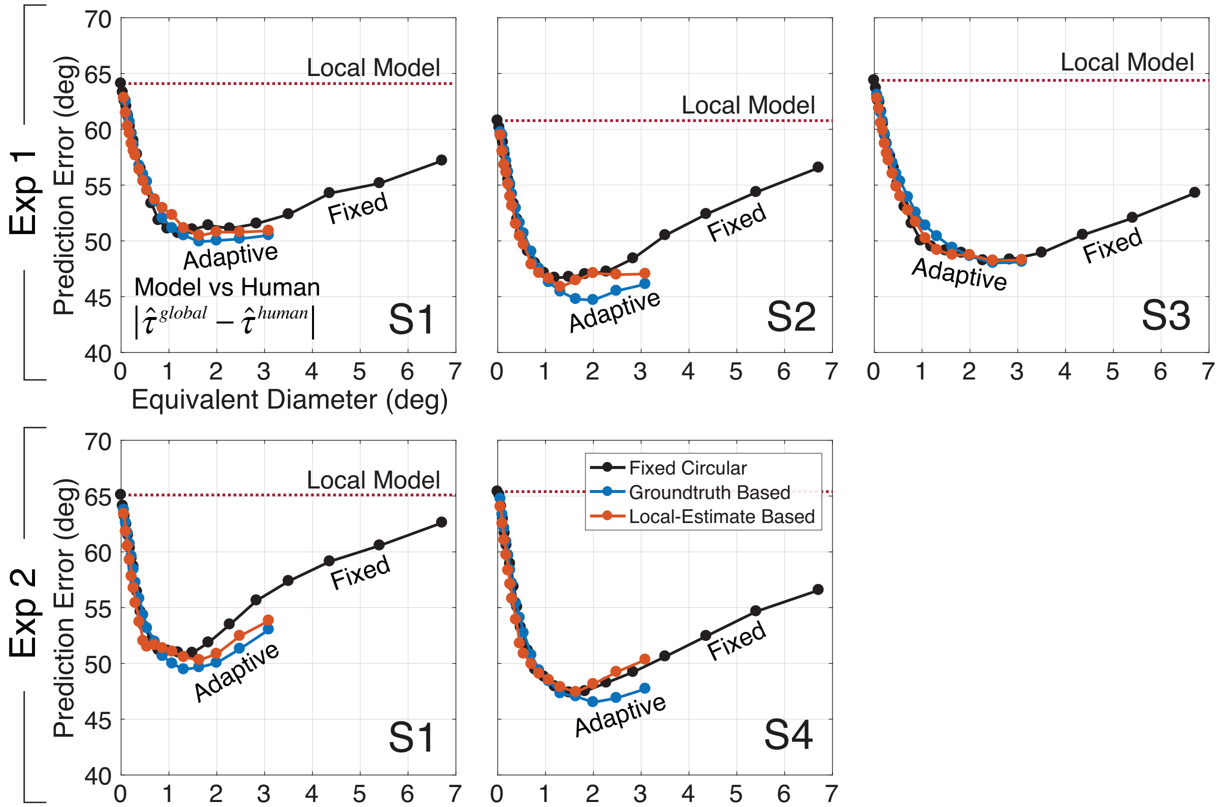

Figure S1. Prediction error for individual human observers. Prediction error is shown for the fixed circular pooling model (black), the groundtruth-based adaptive elliptical pooling model (blue), and the local-estimate-based adaptive elliptical pooling model (red). The top and bottom rows indicate results from Exp. 1 and Exp. 2, respectively. The pooling region that minimizes prediction error for all models and all human observers (except observer S3) corresponds to an equivalent pooling diameter between 1º and 2º.

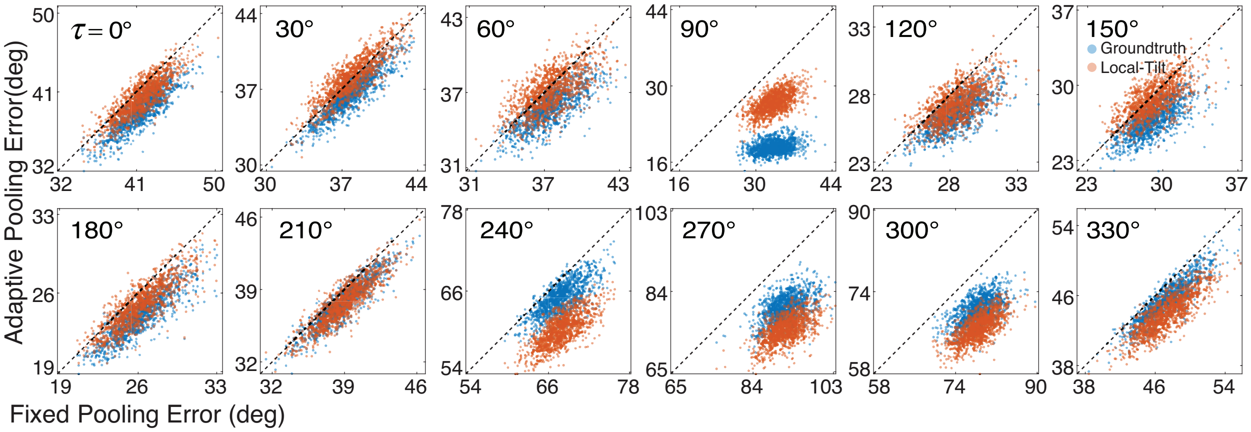

Figure S2. Estimation error with fixed circular vs. adaptive elliptical pooling for different groundtruth tilts in Exp. 2. Each point represents the mean estimation error in a randomly sampled stimulus set across stimuli at a given groundtruth tilt. Estimation error with fixed circular pooling (black) is plotted against estimation error with adaptive elliptical pooling based on the groundtruth tilt (blue) and the tilt estimate (red) at the target location. Computing the prediction errors on matched stimulus sets isolates the impact of the model, and prevents stimulus variability from unduly affecting the results. The fact that the majority of points lie below the dashed unity line, indicating that adaptive elliptical pooling outperforms fixed circular pooling for the task of estimating groundtruth tilt in natural scenes.

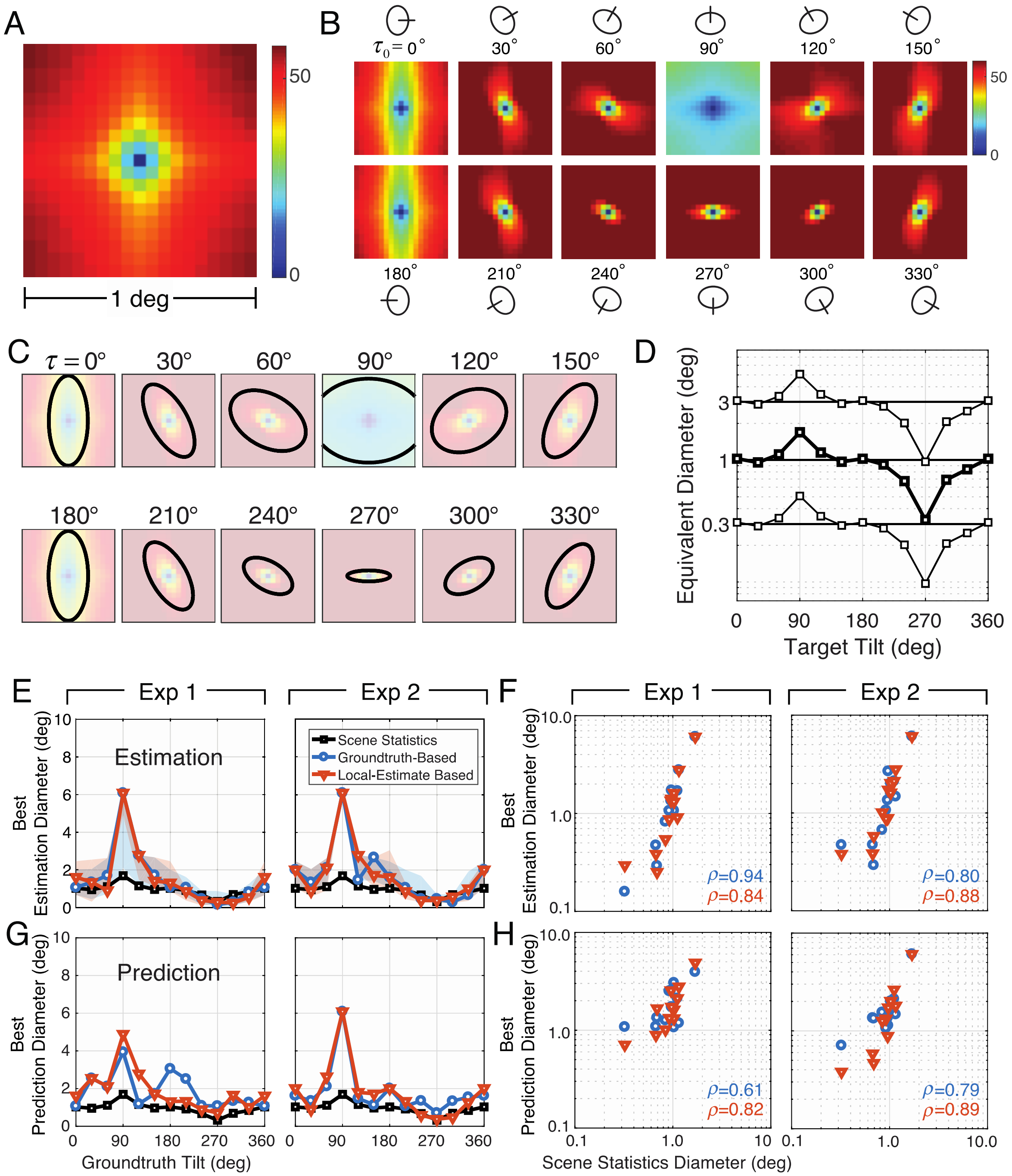

Figure S3. Robustness of natural scene statistics predictions. **A** Spatial statistics of tilt in natural scenes over a 1deg area. Mean absolute tilt difference as a function of spatial location relative to a target location. **B** Mean absolute tilt difference conditioned on the groundtruth tilt at the target location. **C** Fits to the scene statistics in B. **D** Equivalent diameters of the fits to the scene statistics in C. **E** Adaptive pooling regions predicted by natural scene statistics predict the pooling regions that maximize performance at each groundtruth tilt. Equivalent pooling diameters fit to the natural scene statistics and equivalent pooling diameters that minimize estimation error (groundtruth-based, blue; local-estimate-based, orange), plotted as a function of groundtruth tilt. The left and right columns represents data from Exp. 1 and Exp. 2, respectively. **F** Best estimation diameters are correlated with the diameters fit to the natural scene statistics. **G** Equivalent pooling diameter fit to the natural scene statistics and equivalent pooling diameters that minimize prediction error, plotted as a function of groundtruth tilt. **H** Best prediction diameters are correlated with the diameters fit to the natural scene statistics.