Global estimation of signed 3D surface tilt from natural images

Seha Kim and Johannes Burge

The ability of human visual systems to estimate 3D surface orientation from 2D retinal images is critical. But the computation to calculate 3D orientation in real-world scenes is not fully understood. A Bayes optimal model grounded in natural statistics has explained 3D surface tilt estimation of human observers in natural scenes (Kim and Burge, 2018). However, the model is limited because it estimates only *unsigned* tilt (tilt modulo 180°). We extend the model to predict *signed* tilt estimates and compared with human signed estimates. The model takes image pixels as input and produces optimal estimates of tilt as output, using the joint statistics of tilt and image cues in natural scenes (Figure 1). The image cues to tilt are the directions of luminance, texture, and disparity gradients in a local area on the image. To estimate signed tilt, the disparity cue is used as a *signed* tilt cue, and the luminance and texture cues are used as *unsigned* tilt cues. Given a particular set of local image cues, the model computes the minimum mean squared error (MMSE) estimate, which is equal to the posterior mean over signed tilt. We found that the signed MMSE estimates were well aligned with human signed tilt estimates on the identical set of stimuli (Figure 2). Next, we pooled the local MMSE estimates across the space to obtain a *global* tilt estimate. Given that local MMSE estimates are unbiased predictor of groundtruth tilt with nearly equal reliability, the global pooled estimates are also near-optimal. The global estimates even better explained human tilt estimation (Figure 3). We conclude that this computational model provides a tool to understand how human visual systems make the best use of 2D image information to compute local estimates and integrate a global estimate of 3D surface tilt in complex natural scenes using the local estimates.

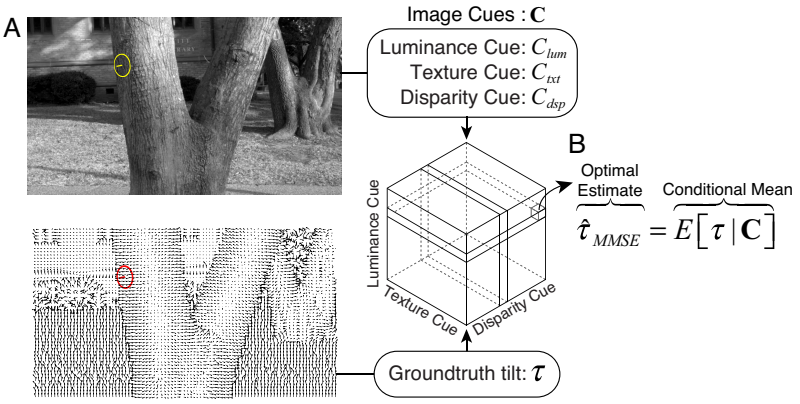

Figure 1. Minimizing mean squared error (MMSE) model. (A) The model uses a large set of data (around 10^8 samples) with natural scene images, which provide image cues, and the co-registered range maps, which provide groundtruth tilts. (B) For given image cues, the mean of the posterior of tilt (formed from the samples of groundtruth tilt in the database) is the *optimal* tilt that minimizes mean squared errors of estimation. In practice, the mean is computed as the running sum of the samples. The optimal tilts are stored in an 'estimate cube.'

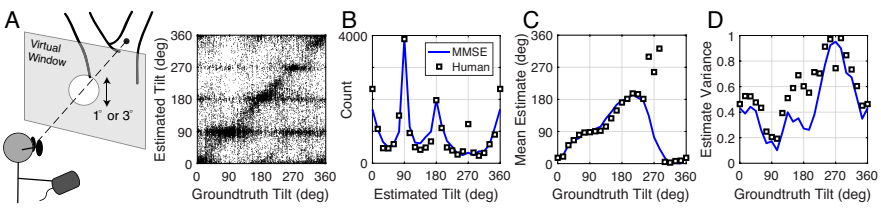

Figure 2. Signed human and local MMSE estimates tested with the same set of natural image patches. (A) Experiment setup and human tilt estimates. Model and human observers have a similar pattern of (B) the distribution of estimates and (C) bias and (D) variance of estimates as a function of tilt.

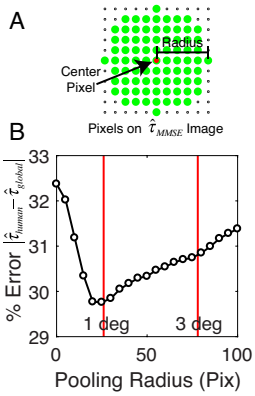

Figure 3. Difference between human estimates and global estimates by the model. (A) Global estimate are computed as the mean of MMSE tilts under a circular pooling area. (B) The percentage of mean difference between human and global estimates to the maximum difference (i.e. 180°) is plotted as a function of the size of pooling area.

Reference: Kim, S. & Burge. J. (2018). The lawful imprecision of human surface tilt estimation in natural scenes. *eLife* (7), e31448. http://doi.org/10.7554/eLife.31448