Asymmetric ON-OFF processing of visual motion cancels variability induced by the structure of natural scenes

Juyue Chen[1], Holly B. Mandel[2], James E. Fitzgerald[3,#], Damon A. Clark[1,2,4,5#]

1 – Interdepartmental Neuroscience Program, Yale University
2 – Department of Molecular, Cellular, Developmental Biology, Yale University
3 – Janelia Research Campus, Howard Hughes Medical Institute
4 – Departments of Physics, Yale University
5 – Department of Neuroscience, Yale University

# - These authors contributed equally to this work. Correspondence:
fitzgeraldj@janelia.hhmi.org (J.E.F.), damon.clark@yale.edu (D.A.C.)

Abstract

Animals detect motion using a variety of visual cues that reflect regularities in the natural world. Experiments in animals across phyla have shown that motion percepts incorporate both pairwise and triplet spatiotemporal correlations that could theoretically benefit motion computation. However, it remains unclear how visual systems assemble these cues to build accurate motion estimates. Here we used systematic behavioral measurements of fruit fly motion perception to show how flies combine local pairwise and triplet correlations to reduce variability in motion estimates across natural scenes. By generating synthetic images with statistics controlled by maximum entropy distributions, we show that the triplet correlations are useful only when images have light-dark asymmetries that mimic natural ones. This suggests that asymmetric ON-OFF processing is tuned to the particular statistics of natural scenes. Since all animals encounter the world's light-dark asymmetries, many visual systems are likely to use asymmetric ON-OFF processing to improve motion estimation.

Introduction

For any visual system, motion estimation is an important but computationally challenging task. To accurately extract motion signals from complex natural inputs, visual systems should take advantage of all useful information. One source of information lies in the stable statistics of the visual input, that is, in the regularities of natural scenes (Geisler, 2008). A strong version of this hypothesis is that visual systems are tuned, through evolution and experience, to the statistics of natural environments (Chichilnisky and Kalmar, 2002; Olshausen and Field, 1996; Simoncelli and Olshausen, 2001; Srinivasan et al., 1982). However, it remains unclear how visual systems use the statistics of natural scenes and the motion signals in them to aid in motion estimation (Salisbury and Palmer, 2016; Sinha et al., 2018).

Motion computation has long been understood algorithmically as selective responses to specific spatiotemporal correlations (Fitzgerald et al., 2011; Poggio and Reichardt, 1973; Potters and Bialek, 1994). For example, canonical models propose that animals extract motion from visual signals by detecting *pairwise* spatiotemporal correlations (Adelson and Bergen, 1985; Hassenstein and Reichardt, 1956). Higher-order correlations could also contribute to motion

computation, and Bayes optimal visual motion estimators can be written as a sum of terms specialized for detecting different correlation types (Poggio and Reichardt, 1973; Potters and Bialek, 1994).This mathematical result follows from a Volterra series expansion, which provides a general and systematic way to represent nonlinear computational systems. Higher-order correlations are also empirically relevant. For example, triplet spatiotemporal correlations are the next lowest-order terms after pairwise correlations, and both humans and flies perceive motion in "glider" stimuli that isolate triplet spatiotemporal correlations (Clark et al., 2014; Hu and Victor, 2010). The sensitivity to triplet spatiotemporal correlations shows that motion perception incorporates cues neglected by canonical motion detectors.

Interestingly, perceptual sensitivities to triplet spatiotemporal correlations prove that visual systems must consider the polarity of contrast when computing motion. This is because triplet correlations flip sign when contrast polarities are inverted, which means that the perceptual contribution of triplet correlations to a motion estimator reverses when all input contrasts are inverted. This contrast-polarity dependent motion processing has been hypothesized to be an adaptation to natural scenes, especially to the light-dark asymmetry of the contrast distribution (Clark et al., 2014; Fitzgerald and Clark, 2015; Fitzgerald et al., 2011; Leonhardt et al., 2016; Nitzany and Victor, 2014). For example, simulated motion detectors that were optimized to estimate motion in natural scenes exhibited contrast-polarity dependent responses similar to flies (Fitzgerald and Clark, 2015; Leonhardt et al., 2016). These modeling studies suggest that contrast-polarity dependent responses emerge from performance optimization in natural scenes, but they do not show that real visual systems use their sensitivity to triplet spatiotemporal correlations to improve motion estimates. This limitation arises because previous experimental studies measured sensitivities to only a few triplet correlations (Clark et al., 2014; Leonhardt et al., 2016). However, one cannot assess the utility of individual correlations in isolation (Clark et al., 2014; Nitzany et al., 2016), and naturalistic visual signals contain many spatiotemporal correlations with diverse spatiotemporal structures. Moreover, although previous analyses recognized that some kind of light-dark asymmetry is required for triplet correlation sensitivity to emerge in optimized motion estimators (Clark et al., 2014; Fitzgerald and Clark, 2015; Fitzgerald et al., 2011; Leonhardt et al., 2016), they did not discover which statistical regularities within natural scenes were sufficient for the observed motion signals to improve accuracy.

Here we filled these gaps by systematically measuring the nonlinearities in *Drosophila* visual motion detection and relating them to light-dark asymmetries in natural scenes. We first systematically characterized low-order components of the fly's motion computation algorithm by modeling its visually-evoked turning behavior with a Volterra series expansion (Clark et al., 2011; Clark et al., 2014; Fitzgerald et al., 2011; Marmarelis and McCann, 1973; Poggio and Reichardt, 1973; Salazar-Gatzimas et al., 2016). Through this framework, we extended canonical pairwise (second-order) motion computation models by adding a triplet (third-order) component that accounts for contrast-polarity dependent motion computation. We evaluated the performance of the inferred algorithm across an ensemble of moving natural images and discovered that the third-order component improves velocity estimates by canceling image-induced variability in the second-order component. Finally, we leveraged maximum entropy distributions to develop a method for generating synthetic images with precisely controlled contrast statistics. This method revealed that the skewness of natural images allows the fly's sensitivity to triplet spatiotemporal correlations to improve its canonical motion estimates.

Results

The structure of natural scenes induces variability in second-order motion estimates

To evaluate how canonical motion detectors performed with natural scene inputs, we simulated responses of the Hassenstein-Reichardt Correlator (HRC) to rigidly translating natural scenes. The HRC exemplifies canonical motion detectors, which rely exclusively on pairwise spatiotemporal correlations to estimate motion (Adelson and Bergen, 1985; Hassenstein and Reichardt, 1956) (**Figure 1A**). It can be equivalently written as a motion energy model (Adelson and Bergen, 1985). We used a database of natural, panoramic photographs to create naturalistic motion stimuli (Meyer et al., 2014). In particular, we first converted the photographs' luminance signals into local contrast signals (**Figure 1B, Supplementary Figure 1-1**). We then rigidly translated these natural images at various horizontal velocities to simulate full-field motion signals (Badwan et al., 2019; Dror et al., 2001; Fitzgerald and Clark, 2015; Leonhardt et al., 2016). This rigid translation of images mimics the motion produced by an animal's pure rotation, during which visual objects all move at the same rotational velocity and occlusion does not change over time. Real motion through an environment generates more complex signals than this, but rigid translations are straightforward to compute and rotational visual stimuli are known to induce the rotational optomotor response that we focus on in this manuscript.

The spatiotemporal contrast signals from these (image, velocity) pairs were used as inputs to the HRC model, and we evaluated the model's output for fixed image velocities across different scenes (**Figure 1C**, **Methods**). The model generated a mean response that was linearly tuned for small velocities, peaked at around 130º/s, and then decayed to zero for fast speeds (**Figure 1C** *green line*). However, we observed substantial variance about the mean response, and this variance implies that different natural scenes generated different second-order motion estimates, even when moving at the same velocity (**Figure 1C** *green shading*). This is consistent with the finding that canonical second-order motion detectors generate variable responses with natural scene inputs (Dror et al., 2001; Fitzgerald and Clark, 2015; Sinha et al., 2018).

Next, we sought to investigate how the higher-order structure of natural scenes influences the performance of the second-order motion estimates. Though canonical motion detectors use only pairwise spatiotemporal correlations, higher-order statistics of static images, such as contrast kurtosis, influence the detector's variance (Clark et al., 2014; Fitzgerald and Clark, 2015). To demonstrate this, we generated a synthetic image set in which we preserved the second-order statistics of natural scenes, including their spatial correlation function and contrast variance, but eliminated all higher-order structure (**Figure 1D, Methods**). When the higher-order structure was eliminated, the HRC's average tuning was unchanged, but there was a marked decrease in the variance (**Figure 1C** *purple*). This demonstrates that higher-order structure in natural scenes induces variability in canonical motion estimates.

Modeling fly motion computation with second- and third-order Volterra kernels

To investigate how real visual systems compute motion, we wanted to systematically characterize an animal's motion computation at the algorithmic level (Marr and Poggio, 1976). Motion computation requires a nonlinear transformation to form a motion estimate from the visual stimulus (Borst and Egelhaaf, 1989; Fitzgerald et al., 2011; Poggio and Reichardt, 1973).

We approximated this nonlinear transformation using a Volterra series expansion (Marmarelis and McCann, 1973; Marmarelis, 2004; Schetzen, 1980; Wiener, 1966). Similar to the Taylor series from calculus, the Volterra series is a polynomial description of a nonlinearity, with a first-order kernel that describes linear transformations, a second-order kernel that captures quadratic terms, and higher-order kernels that combine to represent a wide variety of nonlinearities beyond the second-order. However, many polynomial terms can be needed to describe some nonlinearities. For instance, the polynomial description of a compressive, saturating nonlinearity is inefficient, and it can be easier to describe such transformations using alternative nonlinear model architectures, such as linear-nonlinear cascade models (Dayan et al., 2001). We emphasize that the Volterra kernel description is explicitly algorithmic, as it aims to summarize the overall system processing without considering the mechanisms leading to this processing.

Volterra kernels are useful for studying visual motion processing because they allow us to rigorously group response properties by their order (Fitzgerald et al., 2011; Potters and Bialek, 1994), thereby permitting us to clearly describe both canonical and contrast polarity-dependent components of the behavior. For example, the second-order kernel is equivalent to the canonical motion detecting algorithms, as it explains the sensitivity to pairwise spatiotemporal correlations (Fitzgerald et al., 2011; Salazar-Gatzimas et al., 2016). Second-order Volterra kernels, along with related spike-triggered covariance methods (Bialek and van Steveninck, 2005; Sandler and Marmarelis, 2015; Schwartz et al., 2006), have been used to model second-order behavior and neural processing in flies and primates (Clark et al., 2011; Marmarelis and McCann, 1973; Poggio and Reichardt, 1973; Rust et al., 2005; Salazar-Gatzimas et al., 2016). However, the second-order kernel cannot capture the system's sensitivity to triplet spatiotemporal correlations. We therefore minimally extended the depth of the Volterra series expansion to include the third-order kernel. The third-order kernel directly measures sensitivities to triplet spatiotemporal correlations and probes ON/OFF asymmetries in motion processing.

Experimental measurements of Volterra kernels in fly behavior

We focused on how the fly responds to correlations between nearest-neighbor pixels in the visual input, which corresponded roughly to a single ommatidium separation (Buchner, 1976) (**Figure 2A**). The second-order kernel describes how the behavioral response is influenced by the product of contrasts at each pair of spatiotemporal points in the visual input (**Figure 2B** *blue*). In comparison, the third-order kernel describes how the response is influenced by the product of contrasts at each triplet of spatiotemporal points in the visual input (**Figure 2B** *green*). Note that triplet spatiotemporal correlations could in principle be computed across three distinct spatial locations, but our analysis focused on triplet spatiotemporal correlations distributed across two nearest-neighbor pixels (**Figure 2, Figure 4**).

In order to extract Volterra kernels, especially higher-order ones, we needed a large amount of data. We thus developed a high-throughput setup to measure turning in walking flies in response to visual stimuli (**Figure 2C**) (Creamer et al., 2018; Creamer et al., 2019). In this setup, a fly's optomotor turning response serves as a readout of its motion perception (Götz and Wenking, 1973; Hassenstein and Reichardt, 1956), which allowed us to characterize the fly's motion computation algorithm by measuring its visually-evoked turning response. Flies spend a large portion of their lives standing and walking on surfaces, making walking optomotor responses ethologically critical (Carey et al., 2006).

We extracted the second- and third-order Volterra-kernels with reverse-correlation methods. To do this, we presented flies with spatiotemporally uncorrelated binary stimuli on a panoramic screen around the fly, measured their turning responses, and correlated the behavior at each time to the stimuli preceding it (**Methods**, **Appendix A, Supplementary Figure 2-1**) (Clark et al., 2011; Mano and Clark, 2017; Salazar-Gatzimas et al., 2016). The measured second-order kernel showed positive and negative lobes (**Figure 2D**). The positive lobe below the diagonal indicates that flies turned to the right when presented with positive correlations in the rightward direction. This second-order kernel is consistent with classical models of motion computation and with previous neural and behavioral measurements (Clark et al., 2011; Marmarelis and McCann, 1973; Salazar-Gatzimas et al., 2016). The measured third-order kernel also showed both positive and negative values (**Figure 2E**), and we will dissect its detailed structure later in this manuscript. However, we first set out to evaluate how the third-order kernel contributed to motion estimation across an ensemble of moving natural images.

The third-order kernel improves velocity estimation for moving natural scenes

The kernels were fit to turning behavior, so the output of the model to moving visual stimuli is the predicted optomotor turning response. Following previous work (Clark et al., 2014; Fitzgerald and Clark, 2015; Leonhardt et al., 2016; Poggio and Reichardt, 1973; Potters and Bialek, 1994; Sinha et al., 2018), we hypothesized that optomotor turning responses provide a proxy for the fly's velocity estimate. Using the fitted behavioral model, we could thus investigate how accurately the fly's velocity estimate tracks the true image velocity. We evaluated the fly's motion computation performance with a simple and specific metric: when an entire natural image translates rigidly with constant velocity, how accurately does the behavioral algorithm predict the image velocity (**Figure 3A**)? Specifically, does the fly use its sensitivity to triplet spatiotemporal correlations to improve velocity estimation?

We sampled the velocities from a zero-mean Gaussian distribution with a standard deviation of 114 °/s: this distribution roughly matched turning distributions in walking flies (DeAngelis et al., 2019; Katsov and Clandinin, 2008). Crucially, because we measured the Volterra kernels, we could separate the fly's predicted output into two components: the canonical second-order response, $r^{(2)}$, and the non-canonical third-order response, $r^{(3)}$ (**Figure 3A**). The second-order response is the output from the second-order kernel, and it describes how the fly responded to naturalistic second-order spatiotemporal correlations in the stimulus. Similarly, $r^{(3)}$ is the output from the third-order kernel, and it describes how the fly responded to naturalistic triplet spatiotemporal correlations. This separation allowed us to ask how the pairwise and triplet correlations are individually and jointly used to estimate motion.

We quantified how well the model's responses predicted the image velocity using the Pearson correlation coefficient (Clark et al., 2014; Fitzgerald and Clark, 2015; Leonhardt et al., 2016). This metric supposes that the model response and image velocity are linearly related, and its value summarizes intuitively the mean-squared-error of the best linear fit between the model's output and the image velocity. When the correlation coefficient has an absolute value near 1, the model closely tracks image velocity, while a value near 0 indicates no linear relationship between model and image velocity. The responses derived from the second-order kernel, $r^{(2)}$, correlated positively with the true velocity (**Figure 3B** *blue*), indicating that the second-order

response matches the behavioral direction (Clark et al., 2011; Hassenstein and Reichardt, 1956; Salazar-Gatzimas et al., 2016). Interestingly, the isolated third-order response, $r^{(3)}$, anti-correlated with true image velocities (**Figure 3B** *green*). This means that the fly's third-order response on its own would predict that the fly turns in the direction opposite to the presented motion. However, when $r^{(3)}$ was added to $r^{(2)}$, the accuracy of the full motion estimator increased by ~25% compared to $r^{(2)}$ alone (**Figure 3B** *red*, **Figure 3C**). This important result shows that the third-order responses improve velocity estimates only in conjunction with second-order responses.

To understand this counterintuitive finding, it's useful to recognize that the second-order response is influenced by both the image velocity and the structure of the natural scene. For example, recall that the output of the HRC depended both on the velocity of motion and on the particular image that was moving (**Figure 1C**). Thus, one way to improve the accuracy of the response is to reduce scene-dependent variability in the second-order estimate. To investigate whether this interpretation explained the observed improvement, we calculated the residuals of the second-order responses by subtracting the best linear fits of the image velocity and plotted them against the third-order responses. We found that the third-order signal was strongly anticorrelated with this scene-induced residual in the second-order response (**Figure 3D**). This means that the fly's sensitivity to triplet spatiotemporal correlations indeed canceled scene-dependent variability in the second-order motion estimator to improve the accuracy of motion estimation across natural scenes.

Since the magnitude of the second-order kernel and third-order kernel were each measured experimentally, our model combined $r^{(2)}$ and $r^{(3)}$ with a 1:1 ratio. Nevertheless, we were interested in whether the fly could have done better with alternate weighting coefficients, so we fit a linear regression model to reweight $r^{(2)}$ and $r^{(3)}$ to best predict image velocity. Strikingly, we found the optimized relative weighting between $r^{(2)}$ and $r^{(3)}$ was near one, and the performance of the best weighted model was only marginally better than the empirical model (**Figure 3C** *gray***).** Thus, the measured second- and third-order kernels were weighted near optimally for performance in naturalistic motion estimation.

We also wanted to understand how the improvement added by $r^{(3)}$ depended on the parameters of our simulation. To see how it depended on the width of the image velocity distribution, we varied the standard deviation over an order of magnitude. The improvement did not depend strongly on the variance of the velocity (**Supplementary Figure 3-1**). We also asked how the contrast computation affected the performance of the measured algorithm. When we previously converted luminance into contrast signals, we computed local contrasts on a length scale of 25º (measured by full-width-at-half-maximum), because that is the approximate spatial scale of surround inhibition measured in flies (Arenz et al., 2017; Freifeld et al., 2013; Srinivasan et al., 1982). When we swept this spatial scale from 10° to 75°, the improvement added by the third-order kernel first increased, peaked at around 30º, and then decreased to negative values after 40º (**Supplementary Figure 3-2A-E**). When we computed the contrast over time, instead of space, we observed improvements on timescales less than 100ms, comparable to measured timescales involved in early visual neurons that compute temporal derivatives (Behnia et al., 2014; Srinivasan et al., 1982; Yang et al., 2016) (**Supplementary Figure 3-2FG**). However, the third-order term hurt performance when contrasts were computed on longer timescales. These results

show that contrast computations influence the utility of the measured third-order kernel, with maximal utility occurring in a regime that approximately matches the contrast computation of the fly eye.

Visualizing the measured third-order kernel with impulse responses

Since the measured third-order Volterra kernel improved motion estimates, we wanted to characterize it in more detail. To better visualize the third-order kernel, we rearranged its elements in an impulse response format (**Figure 4AB Methods**). The impulse response of a system is its output when presented with a small and brief input, called an impulse. This impulse may consist of a change in contrast at a single point, in which case the impulse response captures the linear response of the system. Analogously, if the impulse consists of a contrast triplet over three points in space and time, then the triplet impulse response captures the system's response to the interactions of those three points, after accounting for those responses already explained by linear or second-order impulse responses.

Triplet correlation impulse responses are useful because they allow one to rapidly digest how different triplet correlations will affect behavior. For example, in **Figure 4A,** we colored three occurrences of triplets that have the same spatiotemporal structure and their corresponding triplet impulse responses. We set the origin of the impulse responses to be the most recent point in the triplet, because the system could not respond to the interaction of three points before all three points were presented. Since the spatiotemporal structures of these three triplets are the same, the three impulse responses have the same shape. Note that a negative impulse, consisting of an odd number of dark elements within the triplet, would drive turning in the opposite direction. In **Figure 4B,** we represented impulse responses of different triplet with colormaps and used ball-stick cartoons to show the relative temporal distances between the points in each triplet. The predicted time course of the behavioral effect is easy to discern, and the kernel predicts that the behavioral consequences of triplet correlations will last almost a second. We can more compactly understand the relative magnitudes of the behavioral effects by summing the impulse responses over time (**Figure 4C**) (Salazar-Gatzimas et al., 2016). As expected, the impact of different triplet correlations varies significantly in both direction and magnitude.

Verification of the third-order kernel measurement

We verified the reliability of our third-order kernel measurement in two ways. First, we tested the significance of the measured kernel directly. We extracted an ensemble of null kernels by applying the reverse-correlation analysis to the measured behavioral responses and temporally-shifted visual stimuli (**Methods**). By comparing summed kernel elements in the empirical and null kernels, we found that many terms in the third-order kernel were statistically significant at the $p = 0.05$ level (**Figure 4C**). Significance was especially common when the temporal distance between the points in the triplet spatiotemporal correlation was less than 0.1 seconds.

Second, we measured the fly's sensitivity to triplet spatiotemporal correlations with third-order glider stimuli (**Figure 4D, Supplementary Figure 4-1**). Third-order glider stimuli are binary stimuli that lack pairwise correlations and are enriched in specific triplet spatiotemporal correlations (Clark et al., 2014; Hu and Victor, 2010). We used the measured third-order kernel to predict responses to the glider stimuli. Most of the measured responses were quantitatively predicted by the third-order kernel (**Figure 4D**). Several gliders elicited smaller behavioral

responses compared to the kernel prediction; such differences might be attributable to induced long-range spatial correlations in glider stimuli (Clark et al., 2014; Hu and Victor, 2010), which are not captured by our measured nearest-neighbor kernel. Nevertheless, the successes revealed by this independent experimental test strongly suggest that we had enough statistical power to reliably fit the third-order kernel to the behavioral data.

The second- and third-order kernels share temporal structure.

Multiple models propose that sensitivity to pairwise and triplet spatiotemporal correlations could emerge simultaneously from the same nonlinear step in the fly brain (Fitzgerald and Clark, 2015; Leong et al., 2016; Leonhardt et al., 2016). We were thus curious whether the measured second- and third-order kernels had a common temporal structure. To compare the second-order and third-order kernels, we simplified the third-order kernel to a two-dimensional approximation (**Supplementary 4-2ABC**), rearranged the second-order kernel into the impulse response format (**Supplementary 4-2D**), and computed summed kernel strengths to obtain one-dimensional representations for both kernels (**Figure 4E**). We compared the second- and third-order kernel elements at the same temporal offsets (**Figure 4E** *top***)**. In the case of pairwise correlations, the temporal offset was determined by the temporal distance between the left and the right points, and in the case of triplet correlations, the temporal offset was determined by the average temporal distance between the left and right points. The summed kernel strengths showed that the second-order and third-order kernels had similar sensitivities to temporal delays between the input pixels, with peak sensitivity at the shortest delays in our experiment (**Figure 4E** *bottom*). An analysis employing the singular value decomposition yielded similar results, and also showed comparable kinetics in the behavioral responses to pairwise and triplet correlations (**Supplementary 4-2EFG**). These similarities suggest that the second- and third-order responses originate in common physiological processes.

Comparing the measured third-order kernel to optimal motion estimators

A recent theoretical study proposed several motion detectors whose parameters were optimized for velocity estimation in natural scenes (Fitzgerald and Clark, 2015) (**Supplementary Figure 4-3**). In order to compare our measured third-order kernel to those of these optimized motion detectors, we presented stochastic binary stimuli to these detectors and extracted their third-order kernels using reverse-correlation. We found that the third-order kernels of the optimized models were usually similar to each other (**Figure 4F**), which is consistent with prior analyses (Fitzgerald and Clark, 2015). The measured third-order kernel consistently agreed with the optimized kernels in its signs, and in some cases the kernels were also similar in magnitude. However, certain kernel elements differed markedly between the optimized models and the behaviorally measured kernel. Perhaps most noticeably, the behavioral kernel was much smaller than the optimized kernels for correlations whose spatiotemporal structure involved large delays between the points (**Figure 4F**, third and fifth kernel elements). Such differences between the optimized models and the measured behavior could indicate suboptimalities in the fly brain. However, they could also result from unrealistic constraints imposed on the model optimization, such as fixed temporal processing and restricted model structures (Fitzgerald and Clark, 2015; Leonhardt et al., 2016). The measured kernel therefore provides valuable new data to inform theoretical work assessing the optimality of biological motion estimators.

Positive skewness is sufficient for the third-order kernel to improve motion estimates.
Which features of natural images allow the measured third-order kernel to improve motion estimates? The natural scene dataset is comprised of heterogeneous individual images (**Figure 5A**), so we calculated the contrast mean, variance, skewness, and kurtosis of each image individually. The variance describes the scale of the contrast variation; the skewness quantifies imbalance between contrasts above and below the mean; and the kurtosis roughly characterizes the frequency of extreme bright and dark points. Each of these statistics showed a wide distribution over the image ensemble **(Supplementary Figure 5-1ABCD)**. These statistics were also highly dependent on each other: a positively skewed image often had high variance and was highly kurtotic (**Figure 5B, Supplementary Figure 5-1E**). These strong relationships make it difficult to isolate the effects of individual statistics within the image ensemble.

To isolate the effects of individual statistics, we therefore generated several different synthetic image datasets that have alternate contrast statistics (**Methods**). To generate each image, we constructed a synthetic contrast distribution and sampled pixel contrasts from this distribution (**Appendix B, Supplementary Figure 5-2A**). In this way, we could manipulate the statistics of the image by constraining various statistics of the distribution. In particular, we constrained the distribution to have specific lower-order moments, such as mean, variance, and skewness. A distribution is not solely determined by its lower-order moments, so there can be many distributions sharing the same lower-order moments (**Supplementary Figure 5-2A**). Among all such distributions, we chose the most random one, known as the maximum entropy distribution (MED) (Berger et al., 1996; Jaynes, 1957; Schneidman et al., 2006; Victor and Conte, 2012). Because we can specify these lower-order moments independently, we can ask whether specific statistics are *sufficient* to generate the improvement added by the third-order signal.

We began by generating two synthetic image datasets **(Methods)**. In the first dataset, we generated a synthetic image for each natural image that had the same contrast mean and variance. To do this, we first found a MED whose contrast mean and variance matched those of the natural image (**Figure 5C,** *green*). We then generated a single synthetic image by sampling from this MED (**Figure 5D** *green*). In the second dataset, we required the synthetic image to have the same contrast mean, variance, and skewness as the original image (**Figure 5CD,** *brown*). By retaining the skewness, these synthetic images retained naturalistic light-dark asymmetries. We then asked how the third-order response affected velocity estimates across these two synthetic image datasets. When only the mean and variance of natural scenes were retained, the third-order response was near zero (**Supplementary Figure 5-2A** *green*) and did not improve motion estimation (**Figure 5E**). However, when the synthetic scenes were constrained to be naturalistically light-dark asymmetric, the improvement added by the third-order kernel was recovered (**Figure 5E**), with magnitude comparable to what was observed for the natural scene dataset (**Figure 3C**).

Finally, we wanted to see whether the degree of skewness controlled the magnitude of the improvement. We therefore generated synthetic image datasets in which we systematically varied the image skewness (**Figure 5FG**). In these synthetic images, the degree of skewness determined how much the third-order response could improve the full motion estimate (**Figure 5H**). When the images were negatively skewed, the third-order response correlated with image velocity (**Supplementary Figure 5-3C**). However, since it also positively correlated with the

residual in the second-order response, adding it to the second-order response decreased the model's overall performance (**Figure 5H, Supplementary Figure 5-3CD**). When the images were positively skewed, the third-order response became anticorrelated with the residual in the second-order response and thus improved the overall motion estimates (**Figure 5H, Supplementary Figure 5-3CD**). These synthetic image sets show that positive image skewness is sufficient for the third-order signal to improve motion estimates. Thus, the measured algorithm for motion estimation leverages light-dark asymmetries found in natural scenes to improve motion estimates.

Discussion

In this study, we first fit a Volterra series expansion to model the fly's turning behavior in response to binary stochastic stimuli, and both second- and third-order terms in the Volterra series contributed to the turning behavior. We then evaluated the model's output when it was presented with an ensemble of rigidly translating natural scenes. There, the second and third-order terms of the model combined to produce outputs that better correlated with image velocities. There is no *a priori* reason to assume that a model fit to explain turning behavior would necessarily predict the image velocity. Therefore, these results can be taken together to motivate the hypothesis that the magnitude of the fly's turning response is determined by an internal estimate of velocity. Furthermore, this estimate is specifically tailored for natural environments, since we found that the third-order kernel relies on light-dark asymmetries that are present in natural scenes but not in arbitrary images. Since skewed scenes are prevalent across natural environments, and many visual systems exhibit ON-OFF asymmetric visual processing (Chichilnisky and Kalmar, 2002; Jin et al., 2011; Mazade et al., 2019; Pandarinath et al., 2010; Ratliff et al., 2010; Zaghloul et al., 2003), many animals are likely to use similar strategies for motion perception.

Direct demonstration that triplet correlations improve velocity estimation for natural images

The idea that features of the visual motion computation serve to improve performance in natural environments is conceptually appealing and theoretically powerful. For example, prior studies have found that optimized motion detectors had triplet correlation sensitivities similar to those measured from the fly visual system (Fitzgerald and Clark, 2015; Leonhardt et al., 2016). Although it is intriguing that biologically-relevant response properties emerged in optimized motion detectors, the link from contrast-polarity dependent motion computation to naturalistic motion estimation remained indirect. Here we have provided the first direct demonstration that third-order components of the fly's motion computation algorithm improve velocity estimation for moving natural scenes. This direct demonstration had not been possible before because prior measurements were limited to a narrow range of correlations, and it was unclear how the measured cues interacted with unmeasured components of the motion estimation algorithm. For example, here we found that the third-order kernel appeared counterproductive when viewed in isolation, but the way flies incorporated triplet correlations was easily interpretable via the deficits of the second-order motion estimator. Remarkably, this problem could have persisted if motion computation needed to be understood in the context of additional visual motion cues that involve longer spatial-scales or higher-order nonlinearities, which have been neglected in this study. This suggests that it might be sufficient to comprehensively characterize motion computation with a few local and low-order visual cues, which is encouraging for the approach

outlined here. On the other hand, a mechanistically accurate model of visual motion processing could eventually summarize the relevant cues in a more succinct and less abstract way.

Flies use triplet correlations to cancel scene-dependent variability in second-order cues

We found that the third-order response in flies was anti-correlated with the natural image velocity (**Figure 3BC**). Nevertheless, they improved velocity estimation when added to the second-order responses (**Figure 3D**). This result appears counter-intuitive at first but can be understood. Spatiotemporal correlations are influenced by both the motion and local structure of the scene, but motion-driven behaviors should ignore fluctuations stemming from the scene's structure as much as possible. Pairwise and triplet spatiotemporal correlations are related to contrast variance and skewness, respectively, which are correlated across the ensemble of natural scenes (**Figure 5B**). This means that fluctuations in second-order signals tend to be accompanied by fluctuations in third-order signals. Therefore, with the right weighting, second-order and third-order signals may collaborate to reduce the image-induced signal fluctuations in the motion estimate (Clark et al., 2014). Indeed, we found that the third-order responses improved motion estimates because they helped to cancel variability in the second-order responses induced by the structure of natural scenes. This finding highlights a generally important but underappreciated point about cue combination and population coding in neural systems. Although a neuron's tuning is often used as a proxy for its involvement in stimulus processing, even untuned neurons can contribute productively to downstream decoding if their responses are correlated with noise in the tuned neuronal population (Zylberberg, 2017).

Large computational benefits can underlie small behavioral effects

The HRC has explained a large number of behavioral phenomena and neural responses, and it is reasonable to ask how much we've gained by extending its second-order algorithmic description to a third-order one. The magnitude of behavior elicited by the third-order kernel is small compared to the second-order kernel's contribution. However, the magnitude of the behavioral effect and the magnitude of the underlying performance gain can differ significantly. For example, the largest reported turning response elicited by a third-order glider stimulus is less than 20% of that elicited by second-order glider stimuli (Clark et al., 2014), yet the performance gain afforded by a motion detector designed to detect its defining third-order correlation exceeded 30% (Fitzgerald and Clark, 2015). Similarly, here we predict that many natural images would elicit little output from the third-order kernel (**Figure 3B**), yet third-order responses improved the correlation between the model output and velocity by ~ 25% (**Figure 3C**). These performance gains are modest, but they are comparable to the inarguable benefits provided by spatial averaging and could be ecologically relevant to the fly (Dror et al., 2001; Salazar-Gatzimas et al., 2018). Since we've only approximated the system with a spatially-localized low-order polynomial, we also expect that the improvements we observed here represent a lower bound on the total effects provided by the full mechanism underlying light-dark asymmetric motion processing. Indeed, longer-range and higher-order motion detection models can nearly double motion estimation accuracy while predicting realistic glider response magnitudes (Fitzgerald and Clark, 2015). It will be interesting to investigate whether mechanistically accurate models that explain the origin of the third-order kernel also reveal larger performance improvements.

Asymmetric ON-OFF processing could affect motion processing across the animal kingdom. Here we showed that flies systematically exploit contrast asymmetries in natural scenes to improve their visual motion estimates. This resonates with previous work showing that many visual systems process ON and OFF signals asymmetrically to improve other aspects of visual processing (Kremkow et al., 2014; Mazade et al., 2019; Pandarinath et al., 2010; Ratliff et al., 2010). Moreover, flies and vertebrates share striking anatomical and functional properties in their motion detection circuits (Borst and Helmstaedter, 2015; Clark and Demb, 2016; Sanes and Zipursky, 2010), likely because they are solving similar problems with similar constraints. We thus expect that many visual systems use ON-OFF asymmetric processing to improve visual motion perception. Nevertheless, it remains unclear how similar or different the details of such strategies will be across the animal kingdom. Indeed, although both primates and insects respond to third-order glider stimuli, their patterns of response differ (Clark et al., 2014; Hu and Victor, 2010; Nitzany et al., 2017). ON-OFF asymmetric visual processing also varies in other ways, and there is evidence that contrast adaptation in ON and OFF pathways is different between primate and salamander retinas (Chander and Chichilnisky, 2001).

Adaptations to differences in both habitat and early sensory processing could potentially explain these divergences. Here we found positive and negative contrast skewness in different terrestrial scenes (**Supplementary Figure 5-1 C**), but positive skewness was most prevalent across the scenes. If certain habitats feature scenes that are predominantly negatively skewed, our work predicts that animals living in these habitats should have opposite third-order responses to flies (**Figure 5H**). More generally, natural scenes in different habitats are known to be similar in some statistics and different in others (Balboa and Grzywacz, 2003; Burkhardt et al., 2006). Interestingly, positive skewness might be particularly common in natural luminance distributions, because luminance signals are the product of many independent factors that generically combine to produce log-normal distributions (Richards, 1982). Nevertheless, the skewness level that matters for motion detecting circuits also depends on earlier processing operations in the eye (**Supplementary Figure 3-2C)**. In our numerical experiments, we computed local contrast signals. The spatial scales of this preprocessing influenced the statistics of the resulting contrast, which in turn influenced the performance of the motion computation. Biologically, this suggests that signal processing in early visual circuits can strongly influence how downstream circuits organize their computations (Dror et al., 2001; Fitzgerald and Clark, 2015). Alternatively, early sensory processing might be tailored to accommodate the computational requirements of downstream processing. These possibilities are not mutually exclusive, and in both cases, the early visual processing must work in concert with the downstream motion detectors to form robust and consistent perceptions.

Volterra kernels systematically characterize nonlinear motion computations

The algorithm used by the visual system to extract motion signals is a nonlinear transformation from light detection to motion estimates. There are numerous ways to characterize a nonlinear system. In many cases, visual neuroscientists have purposefully designed stimuli, such as sinusoidal gratings, plaids, and gliders, to probe specific nonlinearities in the system (Clark et al., 2014; Creamer et al., 2018; Euler et al., 2002; Fisher et al., 2015; Haag et al., 2016; Hu and Victor, 2010; Movshon et al., 1985; Rust et al., 2005; Salazar-Gatzimas et al., 2018; Salazar-Gatzimas et al., 2016). In other cases, they have used stochastic stimuli to fit simple predictive

models, such as linear-nonlinear models, generalized linear models, cascade models, and normalization models to capture a restricted but relatively broad set of biologically-plausible nonlinearities (Dayan et al., 2001; Leong et al., 2016; Maheswaranathan et al., 2018; McIntosh et al., 2016; Salazar-Gatzimas et al., 2016; Simoncelli and Heeger, 1998).

Here we approximated the nonlinear system with Volterra kernels (Marmarelis and Naka, 1972; Wiener, 1966). This represents a general and systematic approach to nonlinear system identification, since (1) one need not make strong assumptions about the system to measure its kernels, (2) higher-order kernels can in principle be added to characterize the system arbitrarily well, and (3) a complete set of kernels predicts the system's output for arbitrary input signals. One major limitation of this approach is that higher-order kernels become progressively more difficult to fit as the number of kernel elements increases. This makes the approach most practical when a few low-order terms already capture conceptually important variables. Here we leveraged the fact that second-order kernel capture the canonical models for visual motion estimation while third-order kernel probes ON/OFF asymmetries in motion processing. These two kernels can be related to distinct statistics of natural scenes.

Polynomial approximations to complex nonlinear systems have also been useful in other domains of neuroscience. For example, the experimental phenomenon of frequency-dependent long-term potentiation can be explained by extending canonical pairwise spike-timing dependent plasticity models to include the relative timing of three spikes (Pfister and Gerstner, 2006; Sjöström et al., 2001).This makes learning sensitive to third-order correlations (Gjorgjieva et al., 2011). In the field of texture perception, researchers have long sought low-order statistics that explain whether two patterns are texturally discriminable (Julesz, 1962; Julész et al., 1973; Julész et al., 1978). Similar to our findings for motion perception, both natural scene statistics and upstream visual processing play important roles (Hermundstad et al., 2014; Portilla and Simoncelli, 2000; Tkačik et al., 2010). As a final example, understanding how neural network structure impacts dynamics was aided by formally expanding the network's connectivity matrix into low-order connectivity motifs (Hu et al., 2014; Trousdale et al., 2012). These motifs might relate to measurable properties of the neocortex (Song et al., 2005).

Velocity estimation is a useful approximation to motion computation

In this paper, we evaluated the fly's motion computation algorithm by measuring the accuracy of velocity estimation. Prior studies have often hypothesized that velocity estimation is a key requirement of motion processing and optomotor circuitry (Clark et al., 2014; Dror et al., 2001; Fitzgerald and Clark, 2015; Fitzgerald et al., 2011; Poggio and Reichardt, 1973; Potters and Bialek, 1994). It is thus reassuring that a model that better fit optomotor behavior also predicted image velocity more accurately. Nevertheless, motion computation is involved in perceptual tasks beyond the optomotor response, such as detecting looming stimuli (Card and Dickinson, 2008; Zacarias et al., 2018). In such tasks, the goal may not be to estimate the velocity of the visual object, but spatiotemporal correlations might nevertheless be useful (Nitzany and Victor, 2014). In addition, fly motion detecting neurons respond to static sinusoids or local luminance changes without obvious relevance for motion processing (Fisher et al., 2015; Gruntman et al., 2018; Salazar-Gatzimas et al., 2018), which suggests that motion computation algorithms might be jointly optimized alongside the detection of other visual features. Finally, visual systems have to coordinate with motor systems to achieve accurate sensorimotor transformations, so one

should take the properties of the motor system into consideration when evaluating the performance of a motion computation algorithm (Dickinson et al., 2000). Future work could consider more sophisticated evaluation metrics that better reflect the total ethological relevance of the visual environment to the fly. As we continue to distill the factors that combine to set the ultimate performance criteria, the use of velocity estimation is likely to remain a simple, useful, and insightful approximation.

Potential mechanisms underlying the measured light-dark asymmetries

Visual systems in both vertebrates and invertebrates split into ON and OFF pathways that process light and dark signals separately and asymmetrically (Balasubramanian and Sterling, 2009; Chichilnisky and Kalmar, 2002; Clark and Demb, 2016; Leonhardt et al., 2016; Ratliff et al., 2010; Ravi et al., 2018; Sagdullaev and Mccall, 2005; Salazar-Gatzimas et al., 2018). Some of these differences could result from biological constraints. Others could be an ethologically relevant adaptation to light-dark asymmetries found in the natural world. Either way, it is difficult to extrapolate from asymmetric neuronal processing of light and dark signals to functional asymmetries in downstream processing, including behavior. In this study, we used the behavioral turning responses to measure asymmetries in the flies' motion computation algorithm, instead of examining ON and OFF processing channels at the neuronal level. Since optomotor sensitivity to triplet spatiotemporal correlations is necessarily a functional consequence of underlying asymmetric visual signal processing, we could thus directly link light-dark asymmetries in natural scenes to the functional impact of ON-OFF asymmetric neural circuitry. It is similarly important to identify additional light-dark asymmetric behaviors that can clarify the functional role of other light-dark asymmetries in visual processing.

Having established the functional relevance of ON-OFF asymmetric visual processing, it is next important to find its neural implementation. Previous work has suggested that front-end nonlinearities could account for certain optomotor illusions in flies (Bülthoff and Götz, 1979), and it is conceivable that such nonlinearities could generate contrast asymmetric motion responses (Clark and Demb, 2016; Fitzgerald et al., 2011). However, several simple front-end nonlinearities can improve motion estimation without inducing the observed triplet correlation responses (Fitzgerald and Clark, 2015). Alternatively, nonlinear processing at the level of direction-selective T4 and T5 neurons could also generate the asymmetries we observed here. Indeed, differentially affecting T4 and T5 activity, either through direct silencing or by manipulating upstream neurons, alters the behavioral responses of flies to triplet correlations (Clark et al., 2014; Leonhardt et al., 2016), and parallel experiments in humans similarly find that contrast-asymmetric responses are mediated by neurons separately modulated by moving ON and OFF edges (Clark et al., 2014). Yet it remains unclear whether asymmetric responses of T4 and T5 are inherited from upstream neurons. For instance, contrast adaptation could differ between the two pathways (Chichilnisky and Kalmar, 2002), and incompletely rectified inputs to T4 and T5 could generate asymmetrical responses to light and dark (Salazar-Gatzimas et al., 2018). The weightings of T4 and T5 signals in downstream circuits could also result in contrast asymmetric phenomena. This rich landscape of possibilities motivates us to think that multiple mechanisms are likely to be involved. By measuring behavior and distilling the abstract algorithmic properties of the system, we will be able to constrain the contributions of individual circuit components without confining ourselves to an overly narrow class of mechanistic models.

Relating algorithm and implementation in fly visual motion estimation

David Marr famously asserted that neural computation needs to be understood at both the algorithmic and implementational levels (Marr and Poggio, 1976). The benefits of this dual understanding go both ways. On one hand, the brain is immensely complicated, and an algorithmic theory can provide an invaluable lens for making sense of its details. On the other hand, the nuances of neuronal implementation can lead to new algorithmic questions and mechanistically satisfying answers. Marr used the optomotor response of flies to articulate his philosophy over forty years ago, and the community is still leveraging this problem to unravel the subtle relationships between algorithm and mechanism in the brain. The HRC model provided the first algorithmic theory of fly visual motion estimation, and this model's insights into the roles of spatial separation, differential time delays, and nonlinear signal integration have now been verified mechanistically (Arenz et al., 2017; Fisher et al., 2015; Haag et al., 2016; Leong et al., 2016; Salazar-Gatzimas et al., 2016; Takemura et al., 2017). They still provide the bedrock of our understanding. Yet the precise mathematical form and mechanistic origin of the nonlinearity remain controversial, with different papers pointing out compelling roles for membrane voltages, intracellular calcium signals, and ON-OFF pathways (Badwan et al., 2019; Gruntman et al., 2018; Haag et al., 2016; Leong et al., 2016; Leonhardt et al., 2016; Salazar-Gatzimas et al., 2018; Wienecke et al., 2018). None of this complexity invalidates the core insights of the HRC, nor does the HRC's domain of success warrant apathy towards the fundamental importance of these unexpected findings. Instead of algorithm and mechanism providing parallel or hierarchical goals, they should be treated as parts of one integrated understanding of the circuit.

Figures and Supplementary Figures.

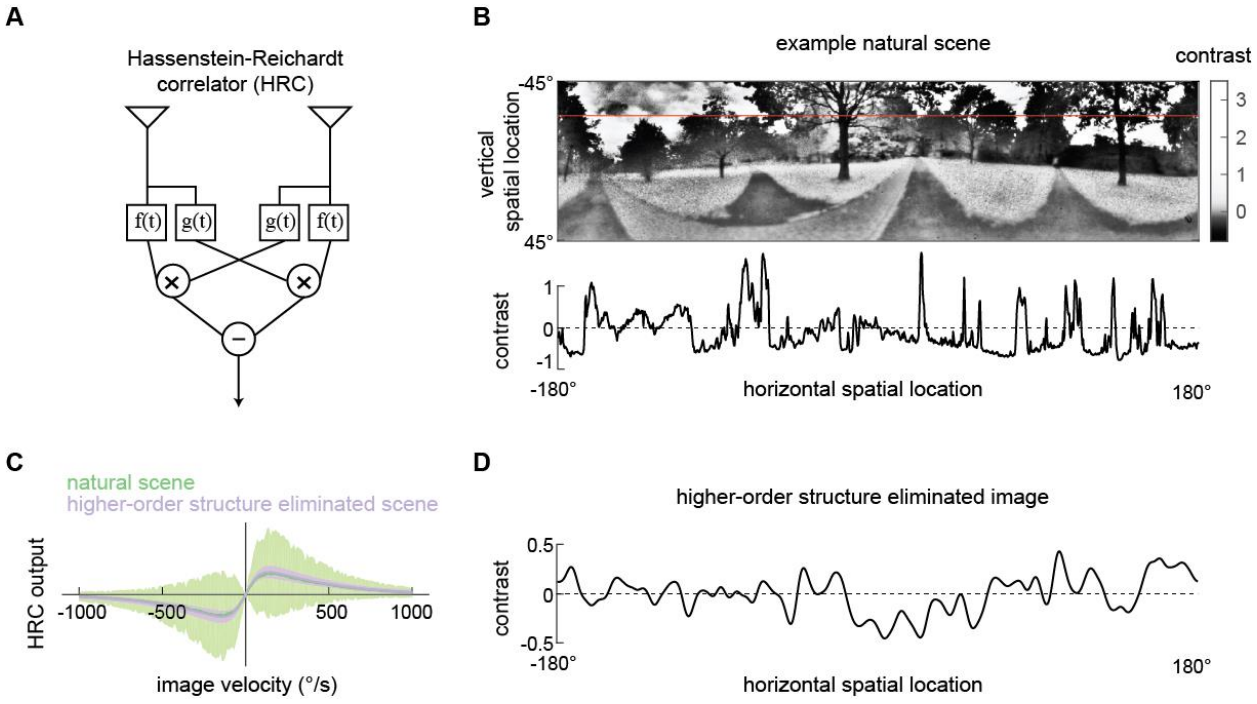


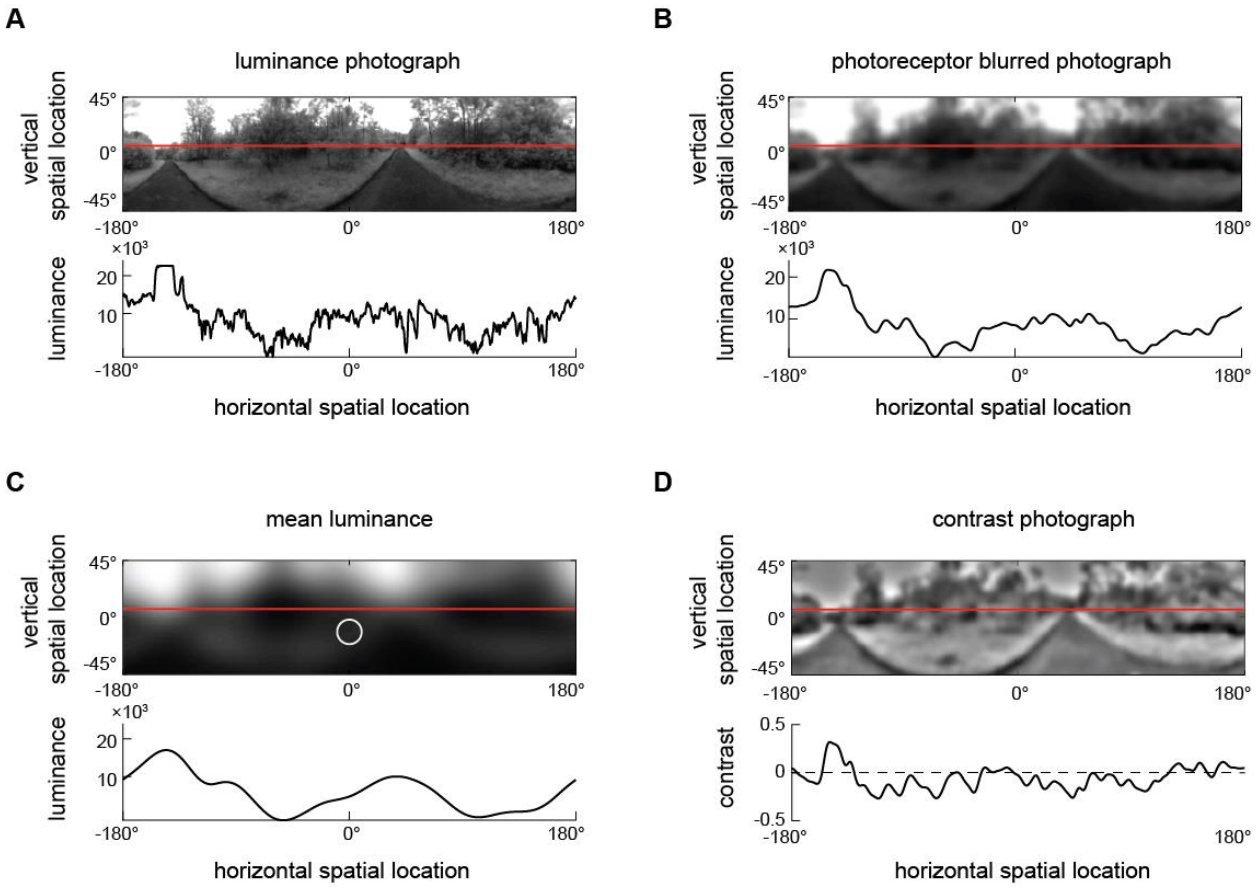
Figure 1. Second-order motion detector performs poorly with natural scene inputs.

A. Schematics of Hassenstein-Reichardt correlator (HRC). One half of the HRC receives inputs from two nearby points in space, which are filtered in time with filters $f(t)$ and $g(t)$, and then multiplied together. The full HRC receives outputs from two symmetric halves with opposite direction tuning and subtracts two outputs.

B. An example 2-dimensional photograph from a natural scene dataset (*top*), including a 1-dimensional section (image) through the photograph (*bottom*), indicated by the red line. So that the image can be viewed clearly, the contrasts in the photograph were mapped onto gray levels so that an equal number of pixels were represented by each gray level.

C. Average response (*line*) and variance (*shaded*) of the outputs of an HRC (equivalent to a motion energy model (Adelson and Bergen, 1985)) when presented with naturalistic motion at various velocities. Images were sampled from natural scenes (*green*) or from a synthetic image dataset in which all higher-order structure was eliminated (*purple*, see **Methods**).

D. Example synthetic image in which all higher-order structure was eliminated.



Supplementary Figure 1-1. Converting luminance signals into contrast signals (see **Methods**).

A. An example 2-dimensional luminance photograph (*upper*) from the dataset, including a 1-dimensional section through the photograph (*bottom*), indicated by the red line. The luminance is shown in log scale in the 1-dimensional image.
B. The 2-dimensional photograph from (A) filtered with a 2-dimensional Gaussian filter (FWHM = 5.3°) to represent the spatial filtering of the fly's ommatidia. This yielded the photograph I_{blur}.
C. The 2-dimensional photograph from (B) filtered with a 2-dimensional Gaussian filter (FWHM = 25°) to represent local mean luminance. The white circle represents the FWHM of the filter. This yielded the photograph I_{mean}.
D. The 2-dimensional photograph from (A) in units of contrast. We calculated the contrast by computing $c = \frac{I_{blur} - I_{mean}}{I_{mean}}$.

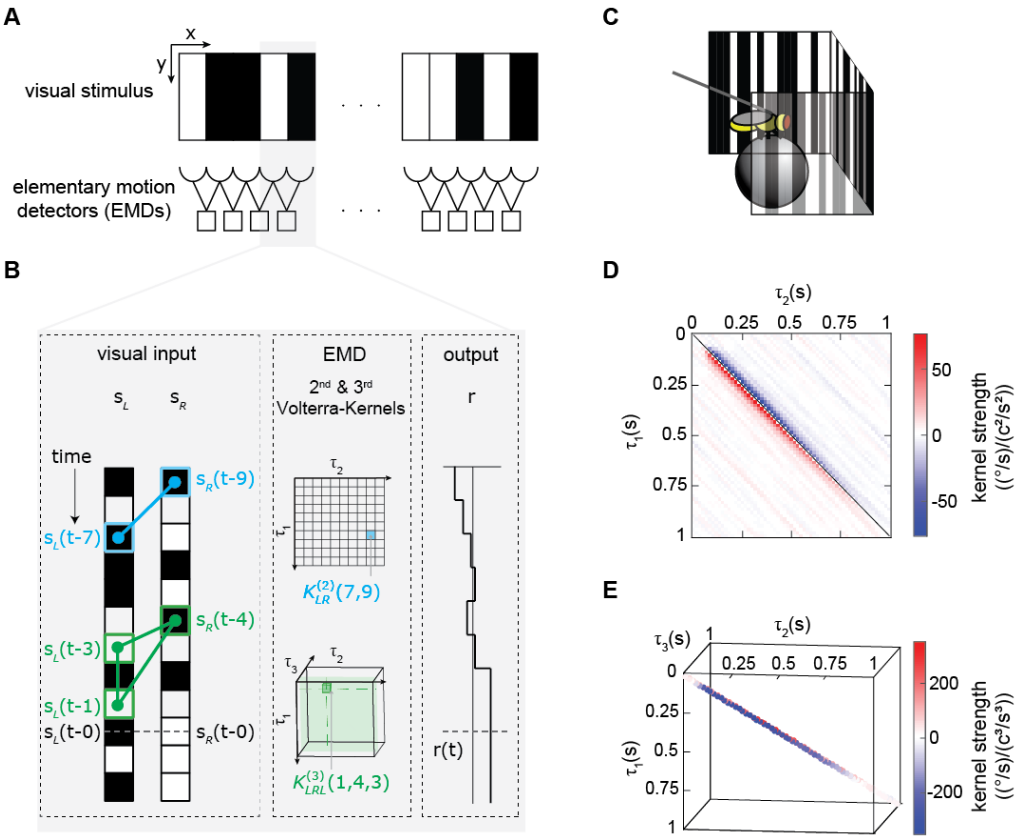
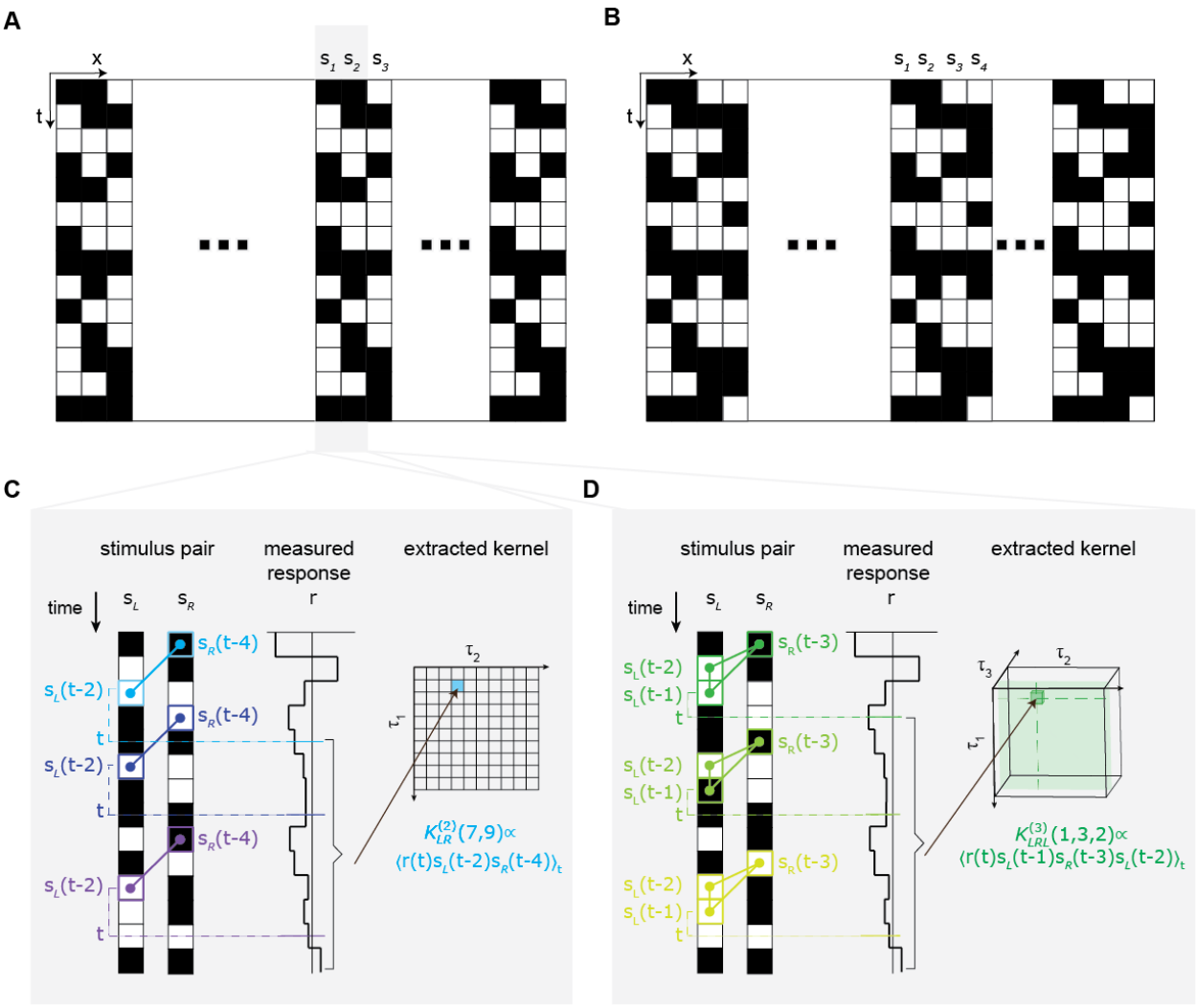


Figure 2. We modeled the fly's motion computation algorithm with second- and third-order Volterra kernels and extracted the kernels using reverse-correlation (**Appendix A, Methods**).

A. Visual depiction of our model, where the fly's motion computation system consists of a spatial array of elementary motion detectors (EMDs), and each EMD receives inputs from two neighboring spatial locations. We presented flies with vertically uniform stimulus with 5°-wide pixels, roughly matching ommatidium spacing.

B. Diagram showing how the output of one EMD at time t, indicated by gray dashed line, is influenced by the second- and third-order products in the stimulus. *Left*: visual inputs of one EMD, s_L and s_R. The visual stimulus contained products of pairwise and triplet points with various spatiotemporal structure. One specific pairwise product is highlighted (*blue barbell*) and one specific triplet product is highlighted (*green triangle*). *Middle*: The motion computation of the EMD is approximated by the second-order kernel (*blue*) and the third-order kernel (*green*). The second-order kernel (*blue*) $K_{LR}^{(2)}(\tau_1, \tau_2)$ is a two-dimensional matrix. For example, the response at time t is influenced by the products of $s_L(t-7)$ and $s_R(t-9)$ with weighting $K_{LR}^{(2)}(7,9)$. The third-order kernel (*green*) $K_{LRL}^{(3)}(\tau_1, \tau_2, \tau_3)$ is a three-dimensional tensor. The response at time t is influenced by $s_L(t-1)s_R(t-4)s_L(t-3)$ with weighting $K_{LRL}^{(3)}(1,4,3)$. *Right*: turning response at time t is influenced by all pairwise and triplet products in the visual stimulus, with weightings given by the second- and third-order kernel elements.

C. Diagram of the fly-on-a-ball rig. We tethered a fly above a freely-rotating ball, which acted as a two-dimensional treadmill. We presented stochastic binary stimuli, and measured fly turning responses.
D. The extracted second-order kernel. The color represents the magnitude of the kernel, with red indicating rightward turning and blue indicating leftward turning to positive pairwise spatiotemporal correlations. Above the diagonal line, the matrix represented left-tilted pairwise products (example in B) and below the diagonal line represents right tilted pairwise products.
E. The extracted third-order kernel. For visualization purposes, we show only the two diagonals with the largest magnitude.



732

733 **Supplementary Figure 2-1**. Using reverse-correlation to extract second- and third-order kernels
734 from the measured turning response to stochastic binary stimulus (**Methods, Appendix A**).

735 A. The 3-bar-block stochastic binary stimulus (**Methods**). The screen is 270° wide, and
736 discretized into 5° bars. Within each 3-bar-block, contrasts of 3 bars were randomly and
737 independently sampled to be black or white over space and time. The block repeats itself
738 across space 18 times.
739 B. The 4-bar-block stochastic binary stimulus (**Methods**), and block repeated itself across
740 space 14.5 times.
741 C. Extracting the second-order kernel using reverse correlation. *Left*: visual stimulus at two
742 neighboring spatial locations s_L and s_R, and examples of a second-order feature $s_L(t-2)s_R(t-4)$ at three different times t. *Middle*: Measured optomotor turning response of a
743 fly. *Right*: To calculate a given element $(\tau_1 = 2, \tau_2 = 4)$ in the second-order kernel, we
744 calculated the product between the turning response $r(t)$ at time t and the corresponding
745 second-order feature $s_L(t-\tau_1)s_R(t-\tau_2)$, and averaged the products over all times t.
746

D. Extracting the third-order kernel with reverse correlation. *Left*: visual stimulus at two neighboring spatial locations, s_L and s_R, and examples of the third-order feature $s_L(t-1)s_R(t-3)s_L(t-2)$ at three different times t. *Middle*: Measured optomotor turning response of a fly. *Right*: To measure a given element ($\tau_1 = 1, \tau_2 = 3, \tau_3 = 2$) in the third-order kernel, we calculated the product between the turning response $r(t)$ at time t and the corresponding third-order feature $s_L(t-\tau_1)s_R(t-\tau_2)s_L(t-\tau_3)$, and averaged the products over all times t.

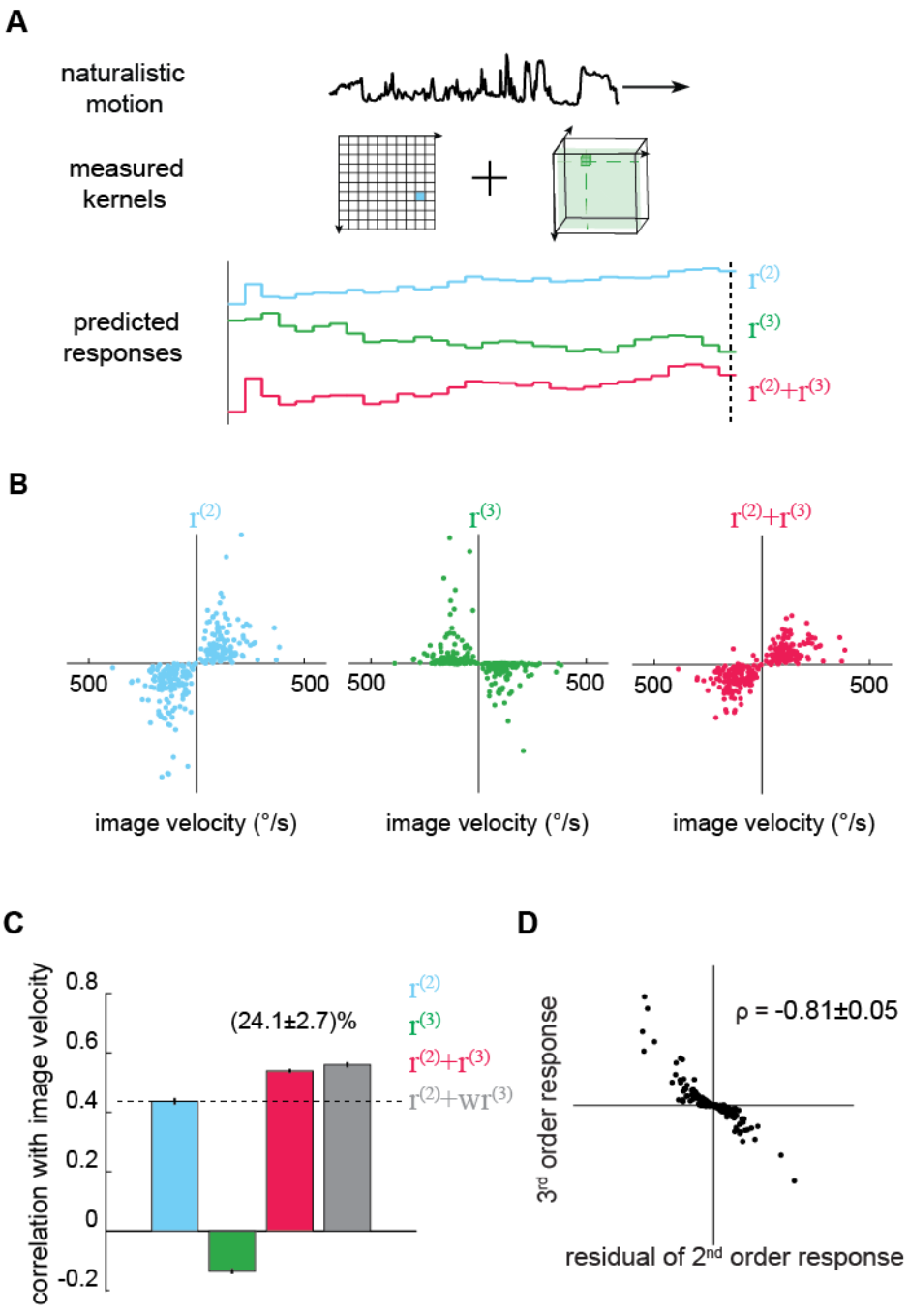


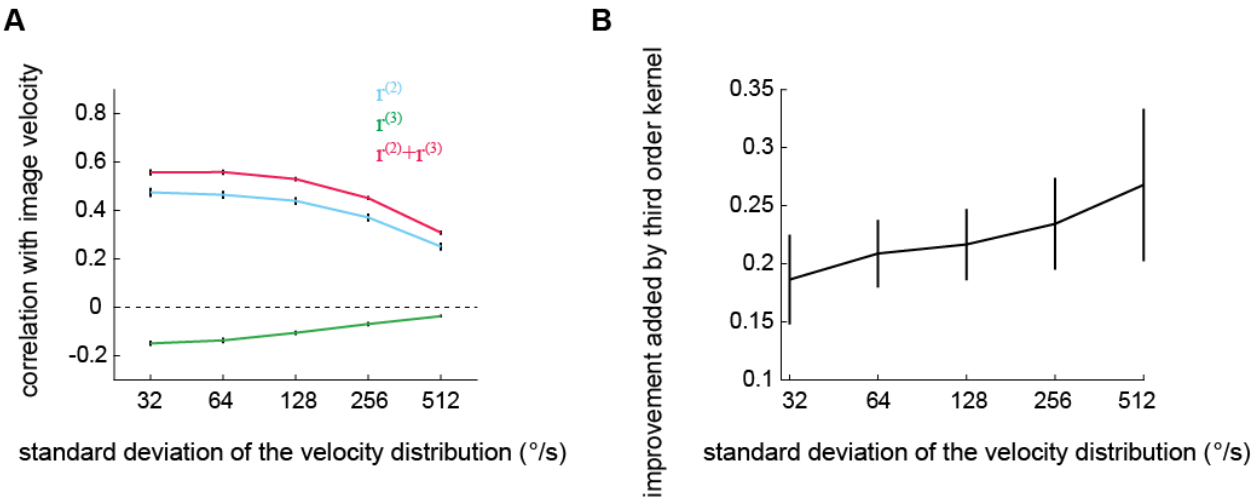
Figure 3. The third-order kernel improved motion estimation in natural scenes.

A. Predicting responses of the second- and third-order kernels to rigidly moving scenes. *Top*: natural scenes rigidly translating with constant velocities. *Middle*: cartoon of the second- and third-order kernels. *Bottom*: second-order response (*blue*), third-order response (*green*), the predicted motion estimate (*red*) is the summation of $r^{(2)}$ and $r^{(3)}$.

B. Scatter plot of $r^{(2)}, r^{(3)}$ and $r^{(2)} + r^{(3)}$ against image velocity over the ensemble of images. 10000 independent trials were simulated, and 1000 trials were plotted here.

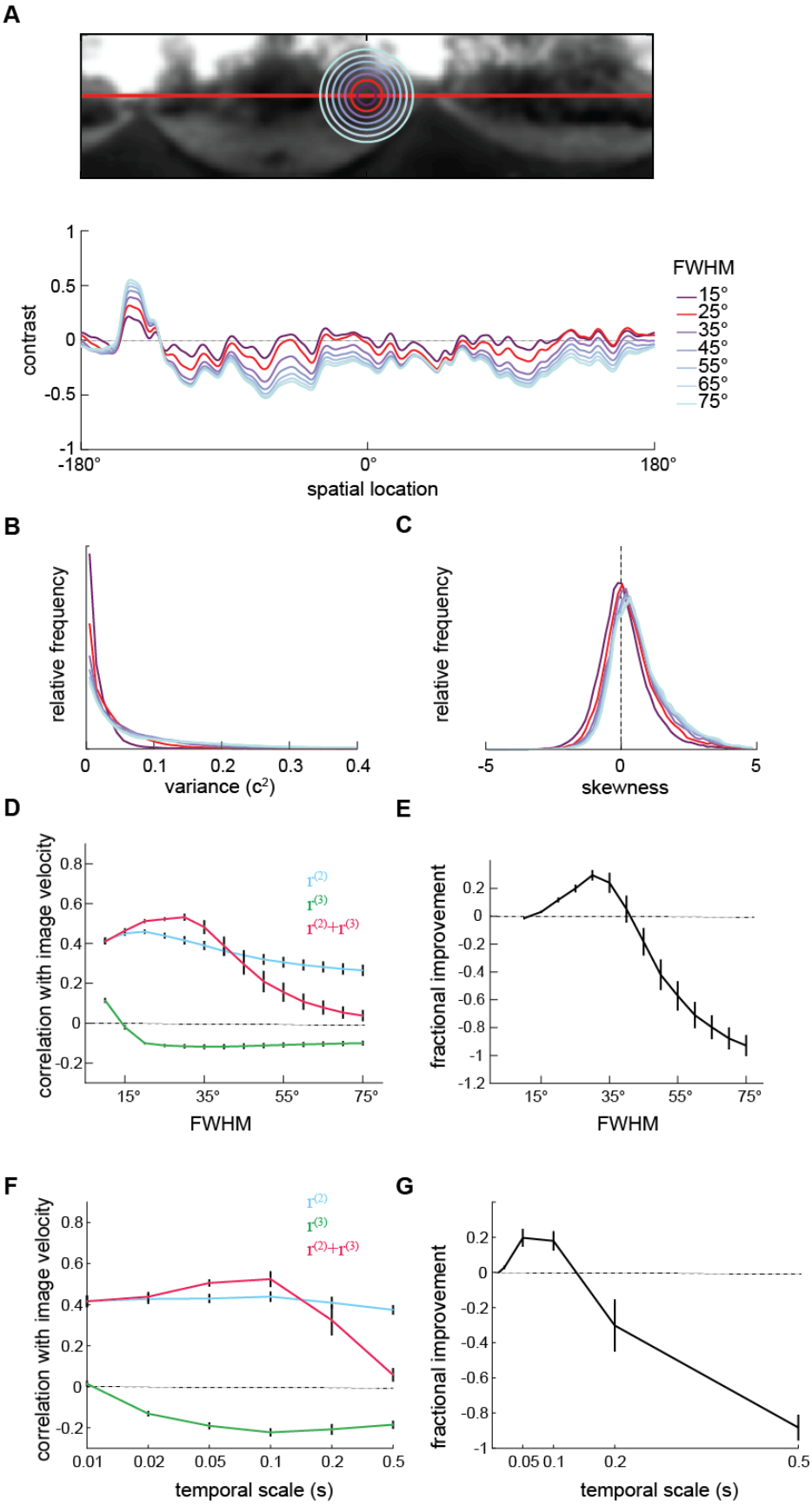
C. Pearson correlation coefficients between responses of each kernel and the true image velocities ($\rho = 0.44 \pm 0.01, -0.14 \pm 0.01, 0.54 \pm 0.01, 0.56 \pm 0.01$, from *left* to *right*; $w = 1.39 \pm 0.10$; mean $\pm$ SEM across 10 groups of 1000 trials).

D. Scatter plot between $r^{(3)}$ and the residual in $r^{(2)}$, computed by subtracting a scaled image velocity from $r^{(2)}$ (**Methods**). ρ represents the Pearson correlation coefficient mean $\pm$ SEM across 10 groups (**Methods**).



Supplementary Figure 3-1. The improvement added by the third-order kernel persists across a wide range of velocities.

A. Pearson correlation coefficient between the true image velocity and the second-order, third-order, and full outputs as a function of the standard deviation of the velocity distribution.
B. The improvement added by the third-order kernel as a function of the velocity standard deviation.


A
contrast
spatial location
FWHM
15°
25°
35°
45°
55°
65°
75°
B
relative frequency
variance (c²)
C
relative frequency
skewness
D
correlation with image velocity
FWHM
E
fractional improvement
FWHM
F
correlation with image velocity
temporal scale (s)
G
fractional improvement
temporal scale (s)

777

Supplementary Figure 3-2. The length scale of local mean luminance computation affected the performance of the measured kernels.

A. Local mean luminance was computed by filtering the luminance picture with a Gaussian filter (**Methods**). We varied the full-width-half-maximum (FWHM) of this Gaussian filter from 15º to 75º. *Top*: Example 2-dimensional blurred luminance photographs I_{blur}. The radius of the circle represents the length scale of the filter. *Bottom*: Example 1-dimensional contrast images computed with corresponding filters.
B. As the length scale of the Gaussian filter increased, the distribution of the variances of individual images shifted to higher values.
C. As the length scale of the Gaussian filter increased, the distribution of the skewness of individual images shifted to more positive values.
D. The performance of measured kernels changed as the length scale of the Gaussian filter increased.
E. The improvement added by the third-order response to the second-order response existed when FWHM varied from 15º to 40º, and peaked around 30º.
F. The performance of second- and third-order responses as a function of the timescale of the contrast computation.
G. The degree of improvement added by the third-order responses changed with the timescale of the contrast computation.

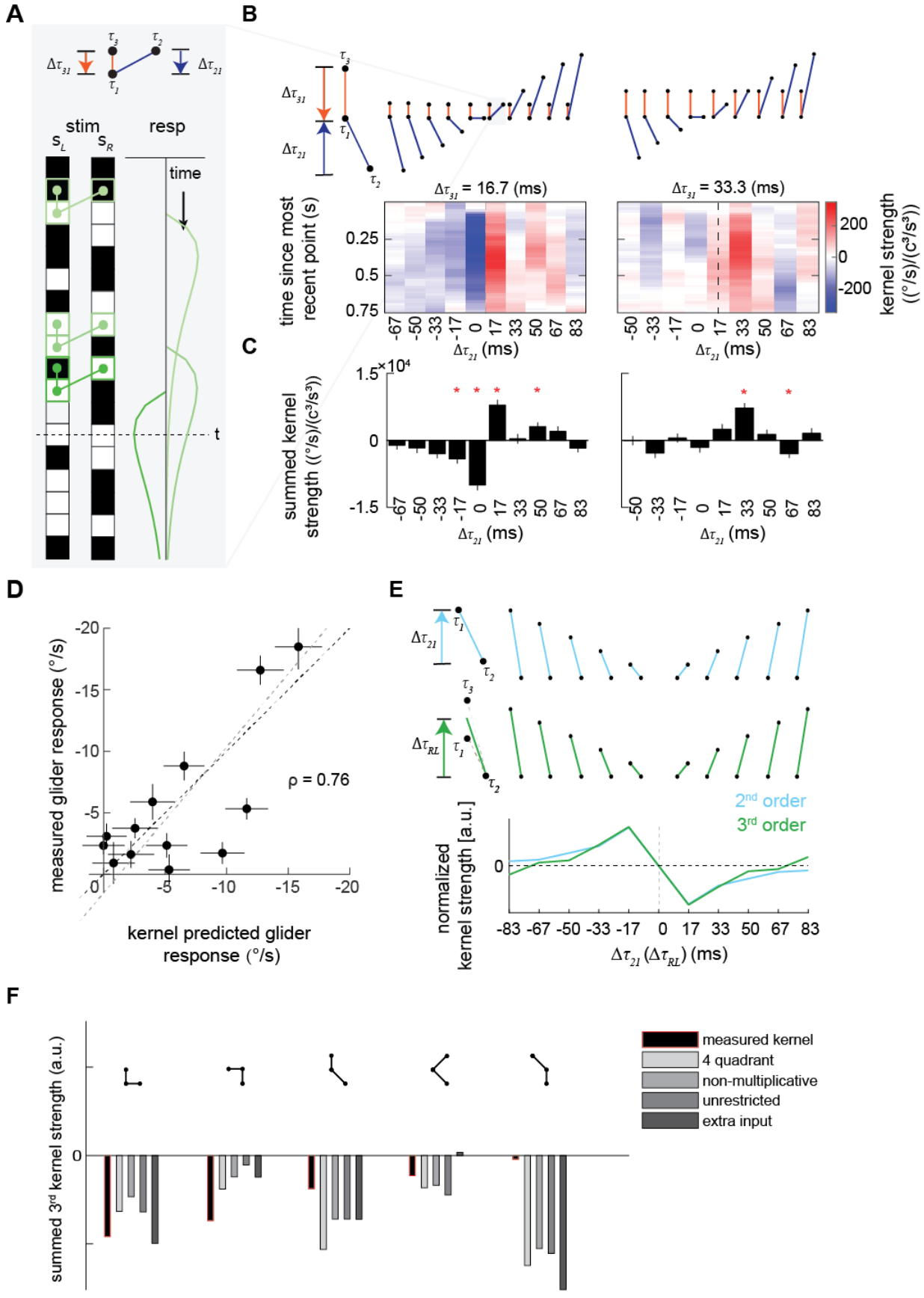
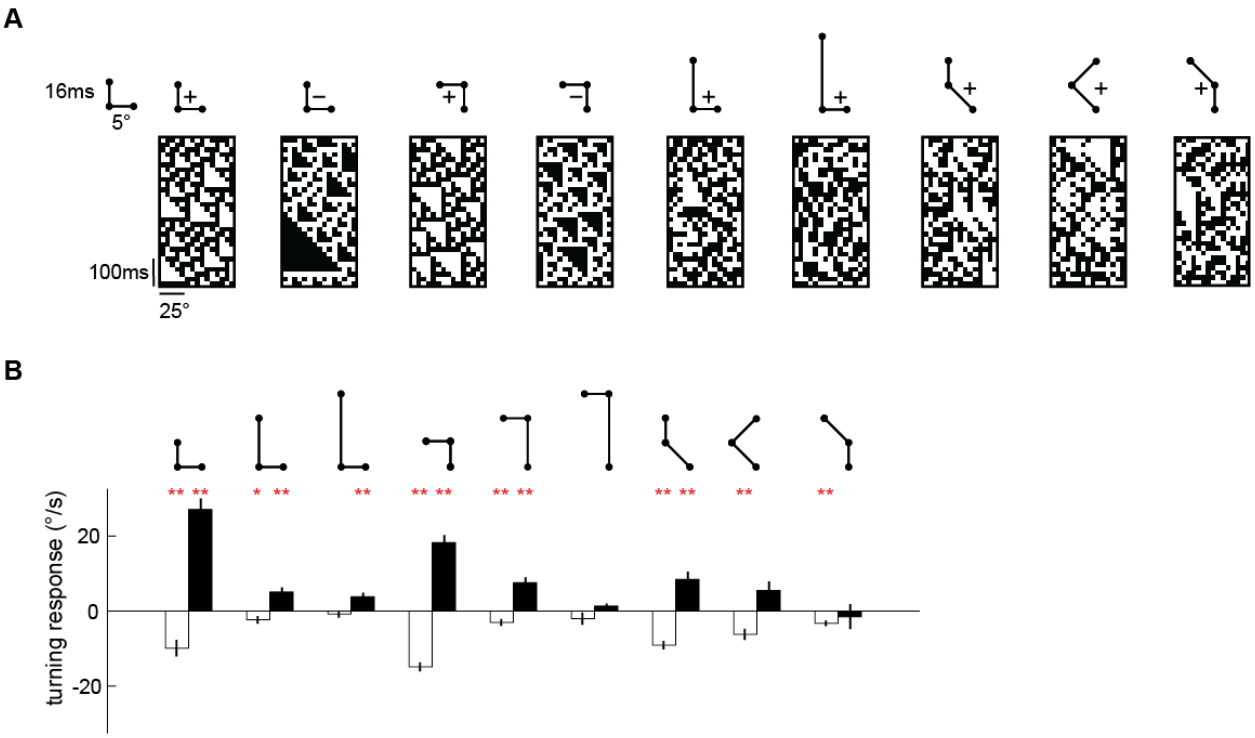

A
B
C
D
E
F
stim
resp
time
t
s_L
s_R
$\Delta\tau_{31}$
$\Delta\tau_{21}$
τ_1
τ_2
τ_3
$\Delta\tau_{31}$ = 16.7 (ms)
$\Delta\tau_{31}$ = 33.3 (ms)
time since most recent point (s)
0.25
0.5
0.75
$\Delta\tau_{21}$ (ms)
kernel strength ((°/s)/(c³/s³))
200
0
-200
summed kernel strength ((°/s)/(c³/s³))
1.5×10⁴
0
-1.5
measured glider response (°/s)
kernel predicted glider response (°/s)
ρ = 0.76
$\Delta\tau_{RL}$
2nd order
3rd order
normalized kernel strength [a.u.]
$\Delta\tau_{21}$ ($\Delta\tau_{RL}$) (ms)
summed 3rd kernel strength (a.u.)
measured kernel
4 quadrant
non-multiplicative
unrestricted
extra input

Figure 4. Characterization and validation of the measured third-order kernel.

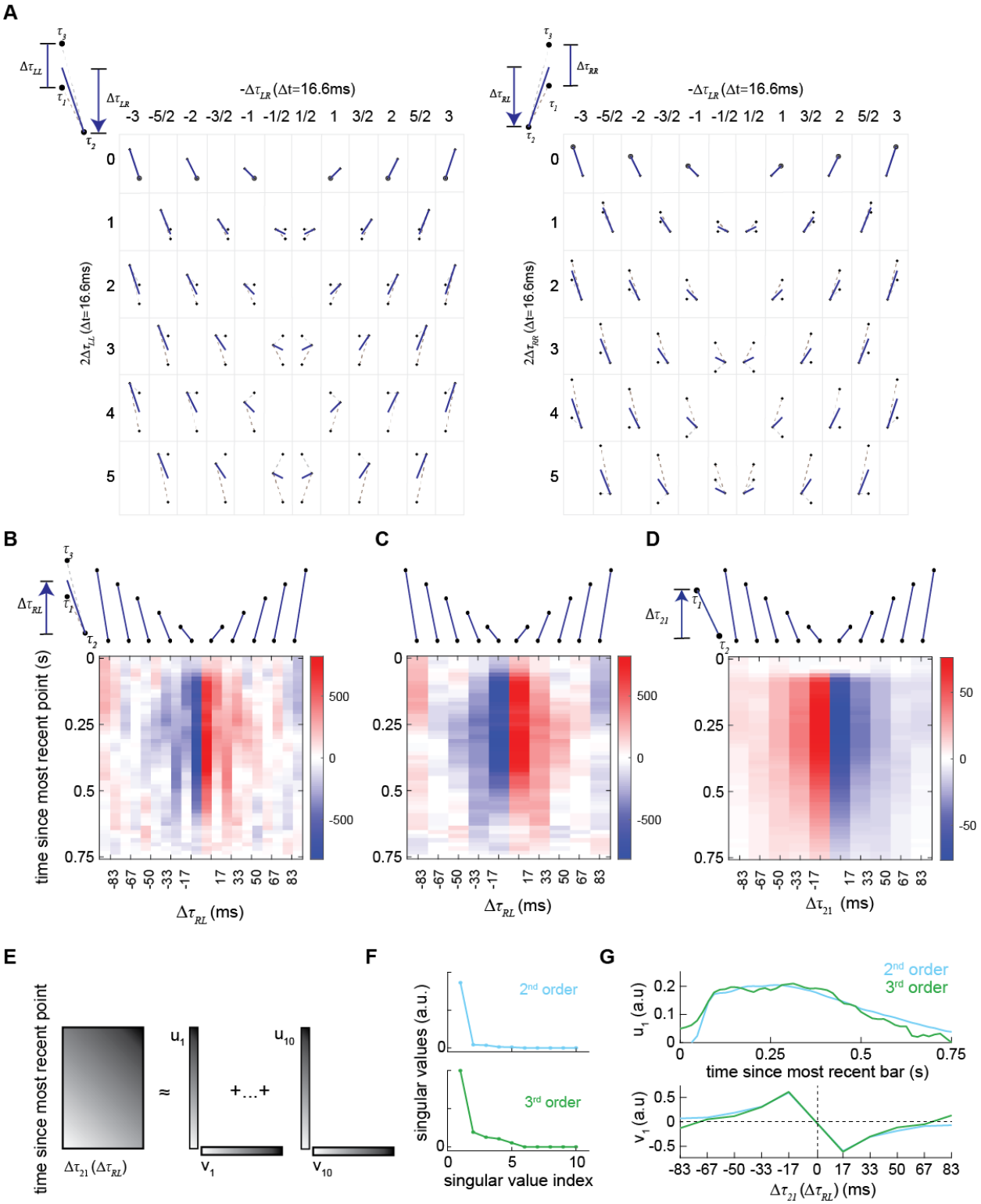
A. Triplet impulse response description. *Top*: the ball-stick diagram represents the relative spatiotemporal position of three points in a triplet. The red line denotes the temporal distance between the two left points, $\Delta\tau_{31}$, and the blue line denotes the temporal distance between the more recent point on the left and the sole right point, $\Delta\tau_{21}$. *Bottom*: Three specific example occurrences of the triplet elicit three impulse responses. The response at time t is the sum of the impulse responses to all previous occurrences of the triplet. The first triplet (*lightest green*) involves two black points and one white point, so their product is positive, and it elicits an impulse response with positive sign. The triplet occurs far from current time t, so its influence on the current response is small. The last triplet (*darkest green*) involves two white points and one black point, so the product is negative and it elicits an impulse response with flipped sign. It is close to current time t, and has a large influence on the current response.
B. Third-order kernel visualized using an impulse-response format (**Methods**). *Top*: the ball-stick diagrams as in (A). *Bottom*: the color map plots the "impulse response" to the corresponding triplets, and color represents the strength of the kernel. Different panels represent different $\Delta\tau_{31}$. In each color map, $\Delta\tau_{31}$ is fixed, the columns represent $\Delta\tau_{21}$, and the rows represent the time since the most recent point in each triplet. The dashed lines indicate the place where the right point is in the middle of the two left points in time.
C. The summed strength of the third-order kernel along each column in A. Error bars represent SEM calculated across flies (n = 72), and significance was tested against the null kernel distribution (*p < 0.05, two-tailed z-test (**Methods**)).
D. The scatter plot between the measured responses to third-order glider stimuli (**Methods**, **Supplemental Figure 4-1**) against responses predicted by the third-order kernel (**Methods**). The correlation between the predicted and measured responses is 0.76. Black dashed line is unity; gray dashed line is the linear best fit.
E. The measured second- and third-order kernel share temporal structures. *Top:* the ball-stick diagrams represent the relative spatiotemporal positions of the two points in each pair (*blue*), and three points in each triplet (*green*). *Bottom*: The kernel strength of the second-order kernel (*blue*) and third-order kernel (*green*) summed across all elements sharing the same spatiotemporal structures, i.e. summed over rows in **Supplementary Figure 4-2 CD** (**Methods**).
F. The extracted third-order kernels from four optimized motion detectors (**Supplementary Figure 4-3,** (Fitzgerald and Clark, 2015)) compared to the measured kernel from the fly. The summed kernel strength is summed across all elements which shared the same spatiotemporal structures diagramed above.



Supplementary Figure 4-1. Flies turned in response to third-order spatiotemporal correlations presented in binary glider stimuli.

A. Example space-time plots of third-order glider stimuli, in which the third-order spatiotemporal correlations were imposed. The ball-stick diagram above each space-time plot represents the relative spatiotemporal structure of the imposed triplet correlations. With the four-parameter scheme (**Methods**), they are (from left to right): $(1,0,L,+1)$, $(1,0,L,-1)$, $(1,1,R,+1)$, $(1,1,R,-1)$, $(2,0,L,+1)$, $(3,0,L,+1)$, $(1,-1,L,+1)$, $(2,1,R,+1)$, $(1,2,R,+1)$.

B. The white (black) bar represents the measured responses to the positive (negative) glider averaged across "left" and "right" gliders (**Methods**). These correspond to gliders (from left to right): $(1,0,L,\pm1)-(1,0,R,\pm1)$, $(2,0,L,\pm1)-(2,0,R,\pm1)$, $(3,0,L,\pm1)$-$(3,0,R,\pm1)$, $(1,1,R,\pm1)-(1,1,L,\pm1)$, $(2,2,R,\pm1)-(2,2,L,\pm1)$, $(3,3,R,\pm1)-(3,3,L,\pm1)$, $(1,-1,L,\pm1)-(1,-1,R,\pm1)$, $(2,1,R,\pm1)-(2,1,L,\pm1)$, $(1,2,R,\pm1)-(1,2,L,\pm1)$. The mean and SEM were calculated across flies. The p-values (**p<0.01, *p<0.05, Student t-test against null hypothesis of no response) and number of flies tested for each glider type is reported in **Table 1**.



Supplementary Figure 4-2. Rearranging the second- and third-order kernels and computing singular value decompositions on the rearranged kernels.

A. Diagrams to demonstrate how the third-order kernel was rearranged. The "ball-stick" diagram represents the spatiotemporal structure of the triplet correlations. *Left:* $\Delta\tau_{LR}$ is the average time difference between the left points and the right point, and $2\Delta\tau_{LL}$ is the temporal separation between two left points. In the grid, the rows represent $2\Delta\tau_{LL}$ in the unit of frame $\Delta t = 16.7ms$, and the columns represent $-\Delta\tau_{LR}$. In the first row, the temporal distance between two left points is zero, so two left points overlap with each other, which is represented by a solid dot surrounded by an empty circle. The z-axis, which is not shown here, represents times since the most recent point. *Right:* Similar to *left* panel, but in this case, there are two points on the right and one point on the left. In the grid, the rows represent $2\Delta\tau_{RR}$, and the columns represent $\Delta\tau_{RL} = -\Delta\tau_{LR}$. Positive $\Delta\tau_{RL}$ means the left point is more recent than the right points (**Methods**).

B. 2-dimensional representation of the third-order kernel by summing the rearranged third-order kernel elements in (A) and (B) over $\Delta\tau_{RR}$ and $\Delta\tau_{LL}$ (**Methods**). In this new format, rows represent time since the most recent point, and columns describe the temporal distance between right and left points, with negative intervals meaning that the right point is more recent than the left point.

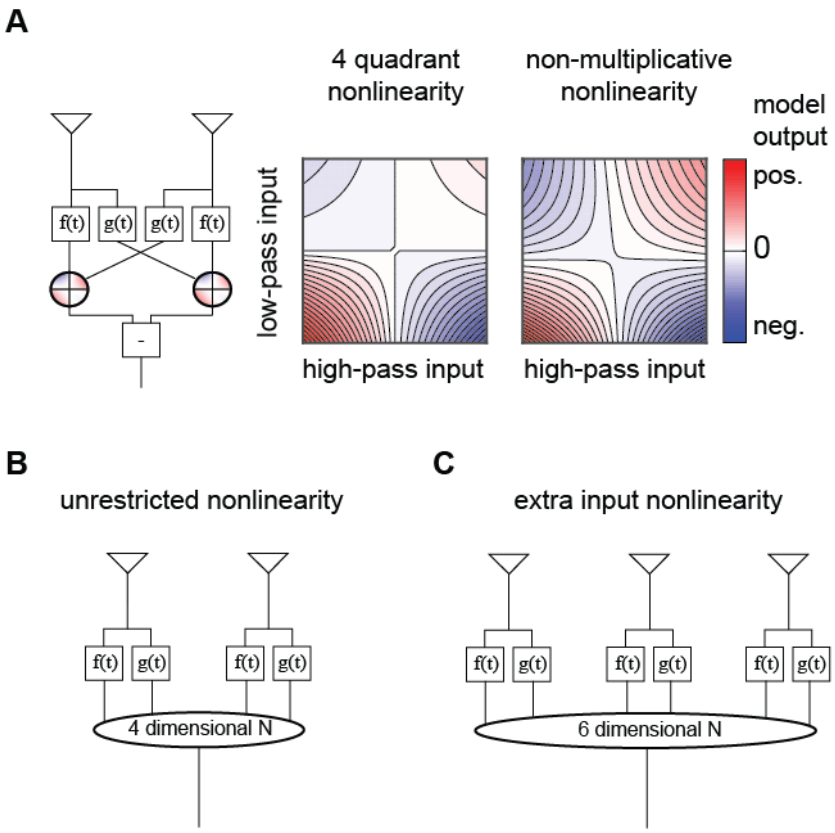
C. 2-dimensional representation of the third-order kernel created by summing neighboring columns in (B).

D. Second-order kernel visualized using an impulse-response format (**Methods**). *Top*: the ball-stick diagrams represent the relative spatiotemporal structure of the pairwise spatiotemporal correlations. *Bottom*: the color map denotes the "impulse response" to the corresponding pairs, and the color represents the kernel strength, the columns represent $\Delta\tau_{21}$, and the rows represent the time since the most recent point in the pair.

E. Diagram of the singular value decomposition of a matrix. The matrix can be approximated by a series of outer products of vectors.

F. The singular values of the second-order kernel (*blue*) and the third-order kernel (*green*).

G. The optomotor turning response of the fly can be modeled as a correlation-sensitive module that represents motion computation and a low-pass filter module that represents the dynamical change of the turning behavior (Clark et al., 2011). The left vector (*top*) could be interpreted as the dynamics in a behavioral module and the right vector (*bottom*) as a correlation-sensitivity module.



Supplementary Figure 4-3. Four models optimized to estimate image velocities in natural scenes, adapted from (Fitzgerald and Clark, 2015)

A. 4-quadrant and non-multiplicative model, these two models differ in the nonlinear interaction of the high- and low-pass inputs.
B. Unrestricted nonlinearity model, in which all interactions up to fourth-order were permitted between the four inputs from 2 points in space.
C. Extra input model, in which all interactions up to fourth-order were permitted between the 6 inputs from 3 points in space.

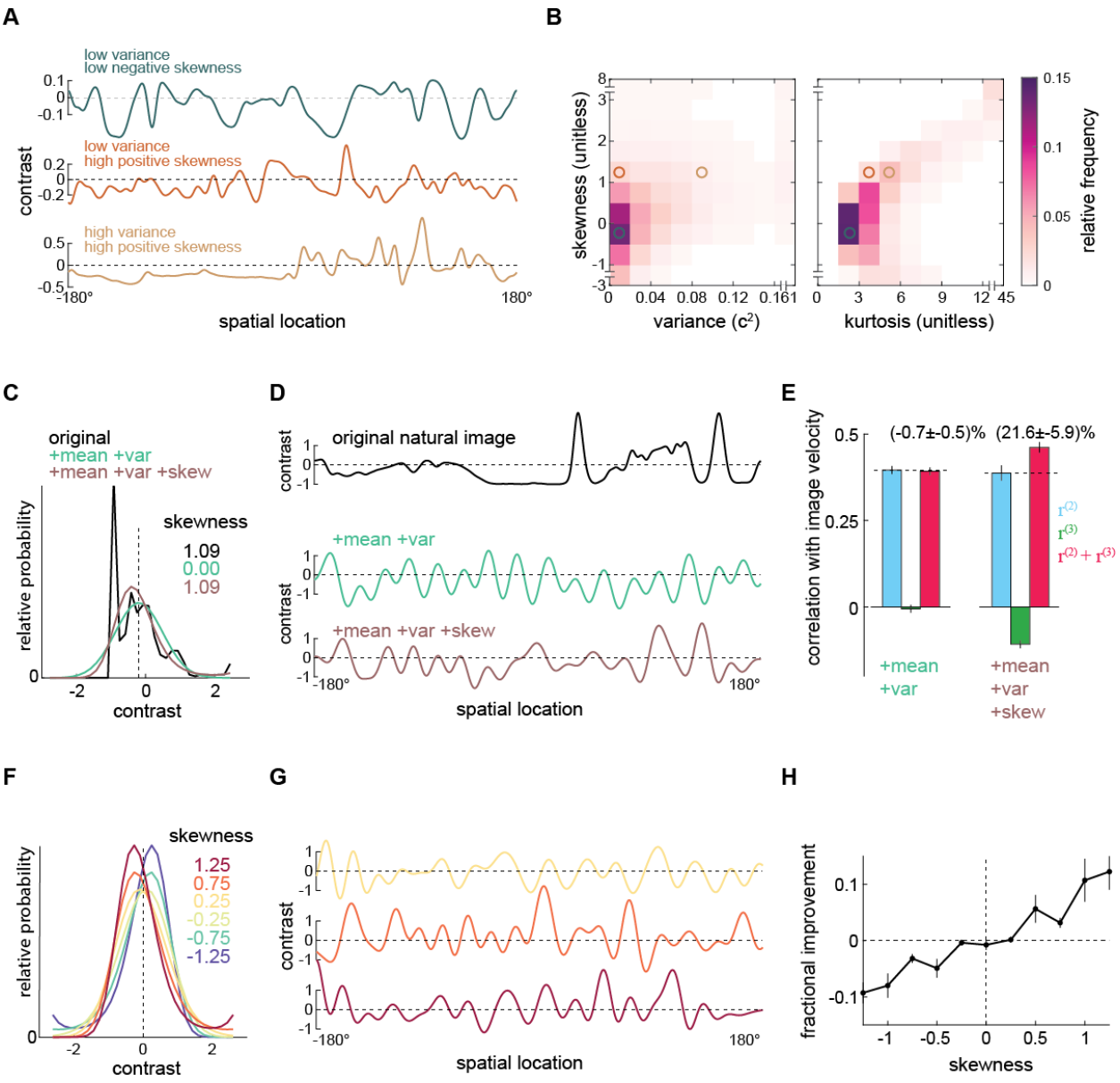


Figure 5. Positive skewness is sufficient for the third-order kernel to improve motion estimation.

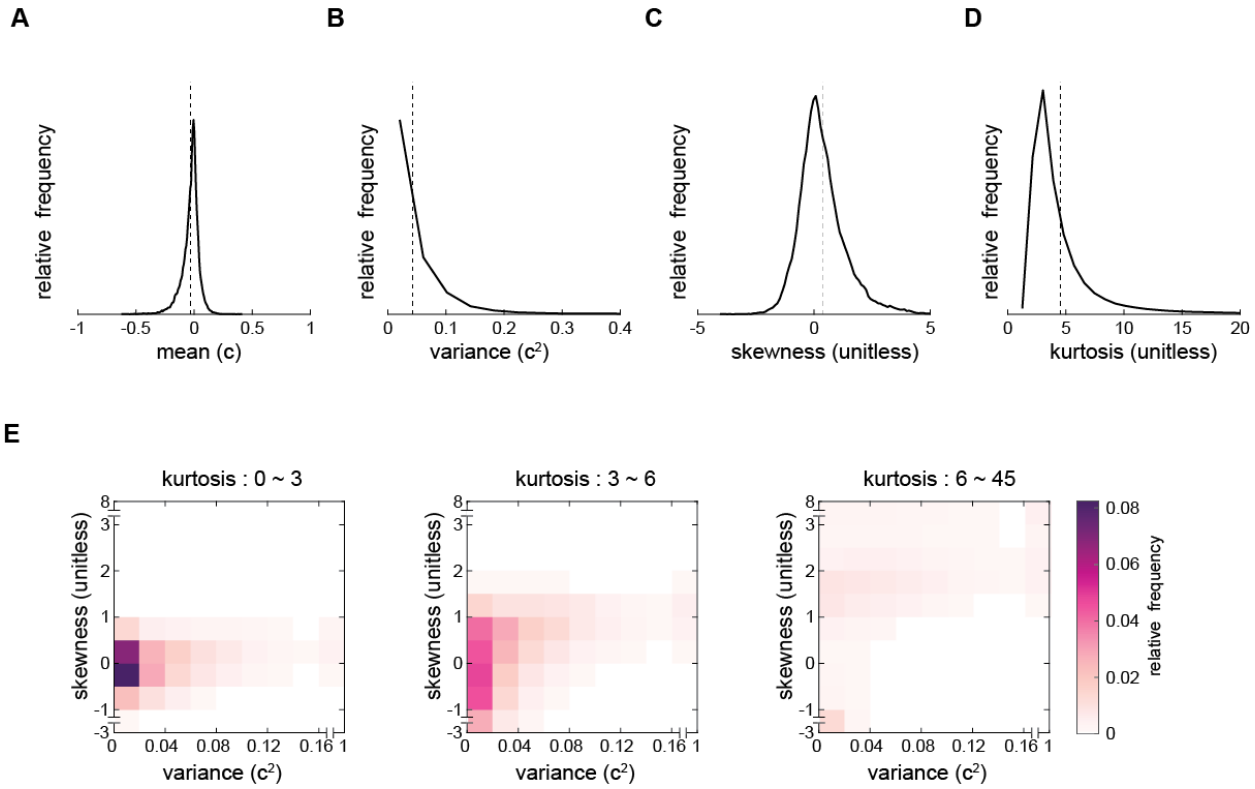
A. Three example natural scenes with different degrees of variance and skewness.

B. Joint-density maps of individual image statistics over the ensemble of natural images, showing the relationship between skewness and variance (*left*) and skewness and kurtosis (*right*).

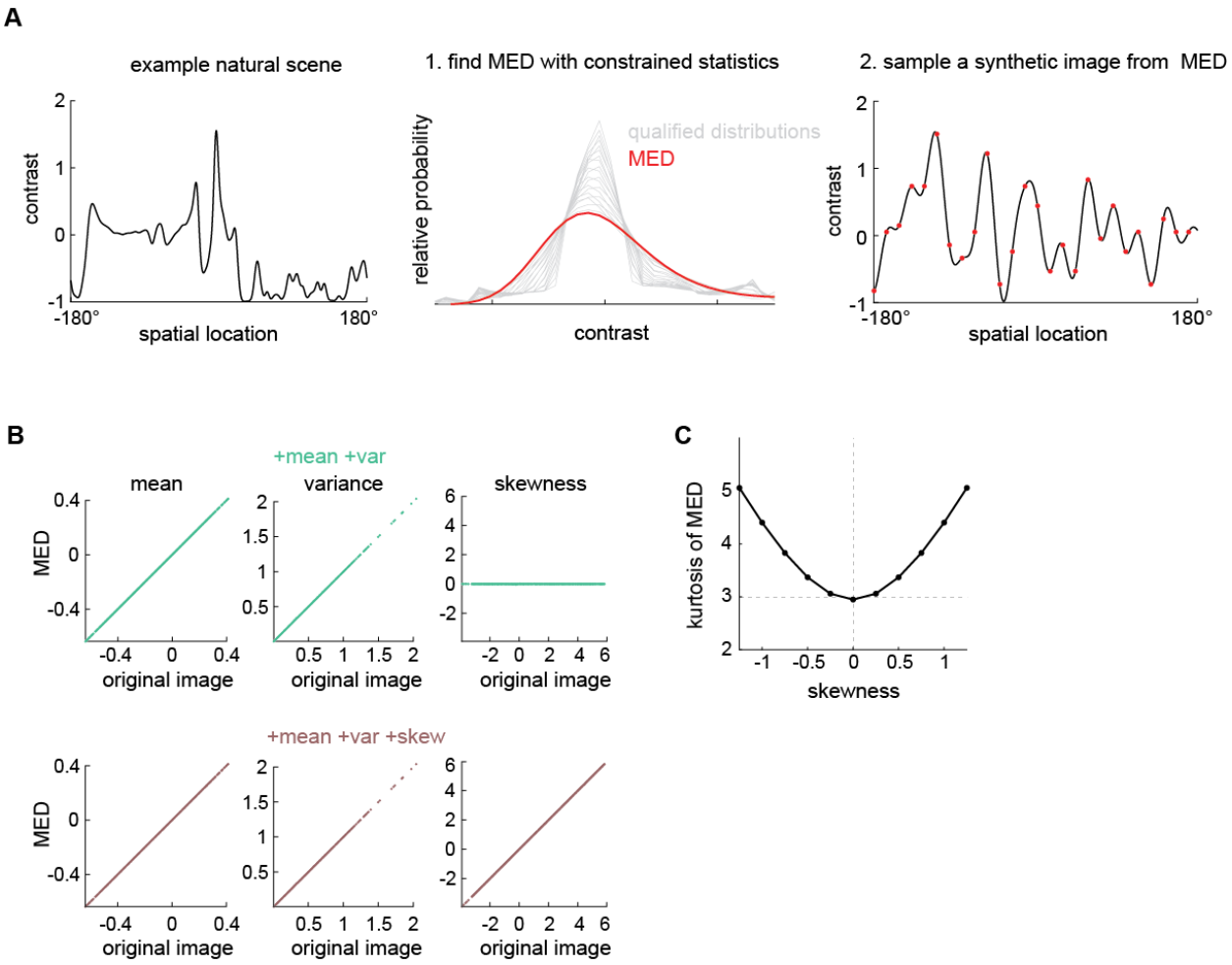
C. Contrast distributions of example images from the natural scene dataset and two synthetic image datasets. The natural image is shown (*black*), along with a maximum entropy distribution (MED) with matched mean and variance, denoted by +mean +var (*green*) and a MED with matched mean, variance, and skewness, denoted by +mean +var +skew (*brown*).

D. Example images from the natural scene dataset and two synthetic image datasets corresponding to three contrast distributions in (C).

E. The Pearson correlation coefficient between true image velocities and each kernel's responses in the two synthetic datasets +mean +var (*green*) and +mean +var +skew (*brown*)
F. Example of MEDs in six synthetic datasets, in which the image skewness ranged from -1.25 to 1.25.
G. Example of synthetic images in three synthetic datasets, corresponding to MEDs in (F) with constrained skewness of 0.25 (*top*), 0.75 (*middle*), and 1.25 (*bottom*).
H. Improvement added by the third-order response as a function of synthetic image skewness.



Supplementary Figure 5-1. Natural scenes have heterogenous contrast statistics.
A-D. Distribution of the contrast mean, variance, skewness and kurtosis of individual images within the entire natural scene ensemble. Mean is denoted by the vertical dashed line.
B. Image statistics are correlated with each other. Joint-density maps of variance and skewness with different ranges of kurtosis.

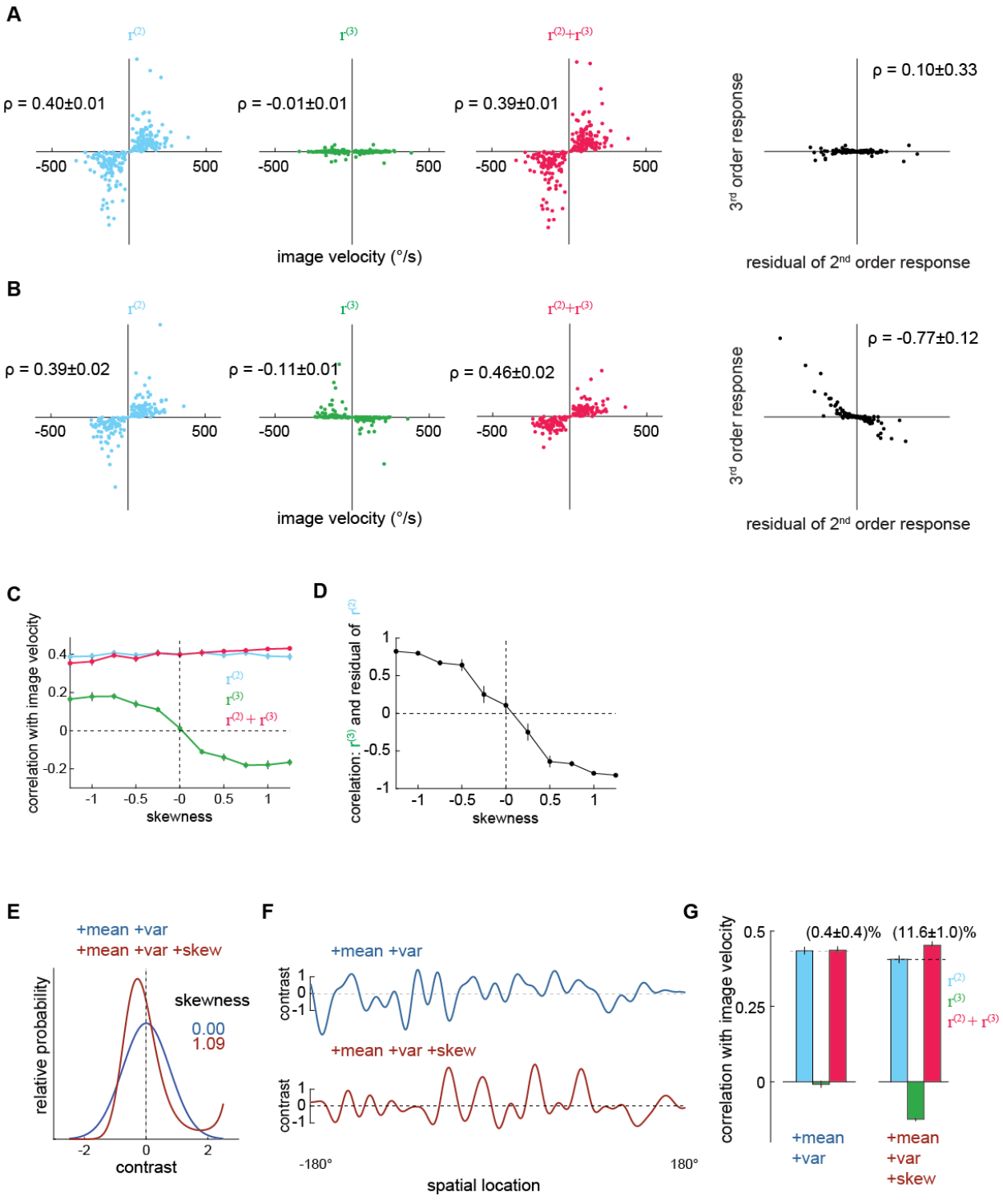


Supplementary Figure 5-2. Using maximum entropy distributions (MEDs) to generate synthetic images with controlled image statistics.

A. Two steps to generate one synthetic image with desired contrast statistics (**Methods**, **Appendix B**). *Left*: Select a natural image and calculate the mean, variance, and skewness of its contrast. *Middle*: There is a family of contrast distributions (*gray*) that have the same mean, variance, and skewness as the original natural image. One may use an optimization algorithm to find the distribution with the maximum entropy (*red*). *Right*: Sample independent contrast values from the solved MED and interpolate between them to induce local spatial correlations.

B. Validating that the MEDs possessed the constrained statistics. *Top row*: When mean and variance were constrained, the mean and variance of the MEDs matched those of the original images. The skewness of the MEDs was zero. *Bottom row*: When mean, variance, and skewness were constrained, the mean, variance, and skewness of the MEDs matched those of the original images.

C. The kurtosis of the MEDs (**Figure 5F**) with skewness constrained to be different values.



Supplementary Figure 5-3. The performance of the measured kernels in various synthetic image datasets.

A. Scatter plots between the model responses and true image velocities when we constrained the synthetic images to match only the mean and variance of the natural images.

B. As in (A), but in this case, we also constrained the synthetic images to match the skewness of the natural image.
C. The performance of the measured kernels as skewness varied in synthetic datasets.
D. The third-order responses as a function of the skewness in synthetic datasets.
E. In comparison to **Figure 5CD**, where the contrast range of the MEDs matched the natural images, the contrast range of the MEDs here were fixed to be [-2.5, 2.5].
F. Example images from these two distributions.
G. The Pearson correlation coefficients between the true image velocities and each kernel's responses in the two synthetic datasets when the contrast range was fixed.

Methods

Fly Husbandry

Flies were grown at 20°C, 50% humidity in 12-hour day/night cycles on a dextrose-based food. Flies used for the behavioral experiment were non-virgin wildtype (*D. melanogaster*: WT: +; +; +) females between 24-72 hour old.

Psychophysics

The fly's turning behavior was measured with the fly-on-a-ball rig, as described in previous studies (Clark et al., 2011; Creamer et al., 2018). The fly was tethered above a ball floating on a cushion of air. The ball served as a treadmill such that the fly could walk and turn while its position and orientation were fixed. The rotational response of the fly was the averaged rotation magnitude of the ball in 1/60s bins with an angular resolution of ~0.5°. Panoramic screens surrounded the fly, covering 270° horizontally and 106° vertically (Creamer et al., 2019). A Lightcrafter DLP (Texas Instruments, USA) projected visual stimulus to the screens with chrome green light (peak 520nm and mean intensity of 100 cd/m^2). The spatial resolution of the projector was around 0.3º and the projector image was updated at 180Hz. The rig's temperature was 34-36°.

Visual stimuli

Visual stimuli varied along the horizontal axis in 5º pixels and were uniform along the vertical dimension. Since the panoramic screen was 270º wide, the horizontal axis was divided into 270/5 = 54 pixels, so the screen was divided into 54 vertical bars.

We used two types of binary stochastic visual stimuli for kernel extraction, a 3-bar-block stimulus type, and a 4-bar-block stimulus type. In the 3(4)-bar-block stimuli, each block contained 3 (4) neighboring vertical bars that flickered white or black independently in space and time. The identical blocks then repeated periodically around the fly. Since there are 54 bars, the entire visual field was divided into 18 (14.5) blocks. Each bar updated its contrast every 1/60th of a second.

We used third-order glider stimuli (Clark et al., 2014; Hu and Victor, 2010) to directly measure the fly's sensitivity to three-point correlations. Third-order glider stimuli are binary patterns of black and white pixels. In each glider stimulus, one can enforce a three-point spatiotemporal correlation. Here we considered only three-point spatiotemporal correlations that involved two neighboring points in space. We described the specific configuration of each glider with a four-parameter scheme, $(\Delta\tau_{31}, \Delta\tau_{21}, L\backslash R, P)$. We defined the 1st point to be the more recent one of the two points sharing a spatial location, while the other point at this spatial location was defined to be the 3rd point. The final point, which was in a position adjacent to the 1st and 3rd points, was defined to be the 2nd point. The temporal interval between the 2nd point and 1st point was denoted as $\Delta\tau_{21}$, and $\Delta\tau_{31}$ was defined similarly. For example, $\Delta\tau_{31} = 1$ means that the 3rd point is 1 frame (16ms) before the 1st point. Although $\Delta\tau_{31}$ is positive by definition, $\Delta\tau_{21}$ can be positive or negative. We used L and R to indicate whether the 1st and the 3rd points are on the left or right of the 2nd point. As detailed previously (Clark et al., 2014; Hu and Victor, 2010), in positive parity gliders ($P = +1$) one or three of these three points are white, whereas negative parity gliders ($P = -1$) have one or three of the points black. We illustrated the configuration of each glider using a "ball-stick" diagram, where the x-axis represents space, the y-axis represents time, time

runs downward, and the plus (minus) sign denotes the polarity of the glider (**Supplementary Figure 4-1A**). Overall, we presented 52 = 13 × 2 × 2 different stimuli: 13 different temporal intervals, each with 2 directions and 2 polarities (**Supplementary. Table 1**).

HRC model

We constructed a classical Hassenstein-Reichardt correlator (HRC) model (Hassenstein and Reichardt, 1956). The output $r_{HRC}(t)$ was defined as

$$r_{HRC}(t) = [s_1 * f_{HRC}][s_2 * g_{HRC}] - [s_2 * f_{HRC}][s_1 * g_{HRC}],$$

where $s_1(t)$ and $s_2(t)$ denote the contrast signals from two spatial locations, * denotes convolution in time, and $f_{HRC}(t)$, $g_{HRC}(t)$ are the temporal filters of the delay line and the non-delay line. In particular,

$$f_{HRC}(t) = t \exp\left(-\frac{t}{\tau_{HRC}}\right),$$

for $t \geq 0$, $f_{HRC}(t) = 0$ for $t < 0$, and

$$g_{HRC}(t) = \frac{d}{dt} f_{HRC}(t),$$

where $\tau_{HRC} = 20$ms.

Modeling the fly's motion computation algorithm with Volterra kernels

We approximated the fly's motion computation algorithm with second- and third-order Volterra kernels. We provide a detailed description of this model in *Section 1* of **Appendix A**. In brief, we discretize space into pixels with $\Delta x = 5°$ resolution, discretize time into time bins with $\Delta t = 1/60$ sec resolution, and index locations in space with an integer subscript ξ. We modeled the response of the fly $r(t)$ as the sum of an array of elementary motion detectors (EMDs) acting at each position in space:

$$r(t) = \sum_{\xi} r_{\xi}(t),$$

where $r_{\xi}(t)$ denotes the response of MED at spatial location ξ. That term is itself the sum of the second-order response $r_{\xi}^{(2)}(t)$ and third-order response $r_{\xi}^{(3)}(t)$:

$$r_{\xi}(t) = r_{\xi}^{(2)}(t) + r_{\xi}^{(3)}(t).$$

The second- and third-order responses are defined as follows (see also **Appendix A**):

$$r_{\xi}^{(2)}(t) = 2 \sum_{\tau_1,\tau_2} K_{LR}^{(2)}(\tau_1,\tau_2) s_{\xi}(t-\tau_1) s_{\xi+1}(t-\tau_2)\,(\Delta t)^2,$$

$$r_{\xi}^{(3)}(t) = 3 \sum_{\tau_1,\tau_2,\tau_3} K_{LRL}^{(3)}(\tau_1,\tau_2,\tau_3) \Big[\Big(s_{\xi}(t-\tau_1) s_{\xi+1}(t-\tau_2) s_{\xi}(t-\tau_3) - s_{\xi+1}(t-\tau_1) s_{\xi}(t-\tau_2) s_{\xi+1}(t-\tau_3)\Big)\Big]\,(\Delta t)^3,$$

where $s_\xi(t)$ and $s_{\xi+1}(t)$ denote the visual inputs that MED at spatial location ξ receives. (In **Figure 2B** and **Supplementary Figure 2-1**, we used $s_L(t)$ and $s_R(t)$ to represent $s_\xi(t)$ and $s_{\xi+1}(t)$). The second-order response $r_\xi^{(2)}(t)$ is the sum of second-order features $s_\xi(t-\tau_1)s_{\xi+1}(t-\tau_2)$ weighted by the second-order kernel $K_{LR}^{(2)}(\tau_1,\tau_2)$. The third-order response $r_\xi^{(3)}(t)$ is the sum of third-order features $s_\xi(t-\tau_1)s_{\xi+1}(t-\tau_2)s_\xi(t-\tau_3)$ and $s_{\xi+1}(t-\tau_1)s_\xi(t-\tau_2)s_{\xi+1}(t-\tau_3)$ weighted by the third-order kernel $K_{LRL}^{(3)}(\tau_1,\tau_2,\tau_3)$. Note that we use the notation $K_{LLR}^{(3)}$ in **Appendix A**, which can be transformed into $K_{LRL}^{(3)}$ by interchanging the position of the second and the third spatial and temporal arguments, i.e. $K_{LRL}^{(3)}(\tau_1,\tau_2,\tau_3)=K_{LLR}^{(3)}(\tau_1,\tau_3,\tau_2)$.

Measuring Volterra kernels with stochastic stimuli and reverse-correlation

To estimate the kernels, we presented stochastic binary stimuli to flies and reverse-correlated the corresponding response with the input. In particular, we estimated the second-order kernel by reverse-correlating the mean-subtracted turning response with the products of two points in space and time (**Appendix A, Section 2, 3**),

$$\widehat{K}_{LR}^{(2)}(\tau_1,\tau_2)=\frac{1}{2}\left(\widehat{K}_{LR-3}^{(2)}(\tau_1,\tau_2)+\widehat{K}_{LR-4}^{(2)}(\tau_1,\tau_2)\right),$$

where $\widehat{K}_{LR-3}^{(2)}(\tau_1,\tau_2)$ is the estimated second-order kernel from the 3-bar-block stimulus and $\widehat{K}_{LR-4}^{(2)}(\tau_1,\tau_2)$ is the estimated second-order kernel from the 4-bar-block stimulus. In particular,

$$\widehat{K}_{LR-3(4)}^{(2)}(\tau_1,\tau_2)=\frac{1}{54}\frac{1}{2\gamma_{stim}^4(\Delta t)^2}\sum_{\xi=1,2,3,(4)}\frac{1}{T}\sum_t r_{turn}(t)s_\xi(t-\tau_1)s_{\xi+1}(t-\tau_2)$$

where $r_{turn}(t)$ is the mean-subtracted response, and γ_{stim} is the magnitude of the contrast in the binary stimulus (**Appendix A, Section 2**). $s_1(t), s_2(t), s_3(t)$ represent the contrasts of 3 independent bars in 3-bar-block stimulus, and $s_{\xi+1}(t-\tau_2)\equiv s_1(t-\tau_2)$ for $\xi=3$. Similarly, $s_1(t)$, $s_2(t)$, $s_3(t)$, $s_4(t)$ represent the contrasts of 4 independent bars in 4-bar-block stimulus, and $s_{\xi+1}(t-\tau_2)\equiv s_1(t-\tau_2)$ for $\xi=4$. We presented 3-bar-block stimulus to 35 flies, and 4-bar-block stimulus to 37 flies, for $T=20$ minutes. All fly kernel estimates were averaged to generate the final kernel estimate. Note that we enforce $\widehat{K}_{LR}^{(2)}(\tau_1,\tau_2)=0$ when $\tau_1=\tau_2$, because we model the visual motion estimator as a mirror-antisymmetric operator (**Appendix A**).

Similarly, we estimated the third-order kernel by reverse-correlating the mean-subtracted turning response with the products of three points in space and time (**Appendix A**).

$$\widehat{K}_{LRL}^{(3)}(\tau_1,\tau_2,\tau_3)=\frac{1}{2}\left(\widehat{K}_{LRL-3}^{(3)}(\tau_1,\tau_2,\tau_3)+\widehat{K}_{LRL-4}^{(3)}(\tau_1,\tau_2,\tau_3)\right),$$

$$\widehat{K}_{RLR}^{(3)}(\tau_1,\tau_2,\tau_3)=\frac{1}{2}\left(\widehat{K}_{RLR-3}^{(3)}(\tau_1,\tau_2,\tau_3)+\widehat{K}_{RLR-4}^{(3)}(\tau_1,\tau_2,\tau_3)\right),$$

where

$$\widehat{K}^{(3)}_{LRL-3(4)}(\tau_1,\tau_2,\tau_3)=\frac{1}{54}\frac{1}{6}\frac{1}{\gamma^6_{stim}(\Delta t)^3}\sum_{\xi=1,2,3,(4)}\frac{1}{T}\sum_t r_{turn}(t)s_\xi(t-\tau_1)s_{\xi+1}(t-\tau_2)s_\xi(t-\tau_3),$$

$$\widehat{K}^{(3)}_{RLR-3(4)}(\tau_1,\tau_2,\tau_3)=\frac{1}{54}\frac{1}{6}\frac{1}{\gamma^6_{stim}(\Delta t)^3}\sum_{\xi=1,2,3,(4)}\frac{1}{T}\sum_t r_{turn}(t)s_{\xi+1}(t-\tau_1)s_\xi(t-\tau_2)s_{\xi+1}(t-\tau_3),$$

and $\tau_1 \neq \tau_3$ (**Appendix Section 2**).

We then enforced mirror anti-symmetry by,

$$\widehat{K}^{(2)}_{LR-sym}(\tau_1,\tau_2)=\frac{1}{2}\left(\widehat{K}^{(2)}_{LR}(\tau_1,\tau_2)-\widehat{K}^{(2)}_{LR}(\tau_2,\tau_1)\right),$$

$$\widehat{K}^{(3)}_{LRL-sym}(\tau_1,\tau_2,\tau_3)=\frac{1}{2}\left(\widehat{K}^{(3)}_{LRL}(\tau_1,\tau_2,\tau_3)-\widehat{K}^{(3)}_{RLR}(\tau_1,\tau_2,\tau_3)\right),$$

$$\widehat{K}^{(3)}_{RLR-sym}(\tau_1,\tau_2,\tau_3)=-\widehat{K}^{(3)}_{LRL-sym}(\tau_1,\tau_2,\tau_3).$$

where $\widehat{K}^{(2)}_{LR-sym}(\tau_1,\tau_2)$, $\widehat{K}^{(3)}_{LRL-sym}(\tau_1,\tau_2,\tau_3)$ and $\widehat{K}^{(3)}_{RLR-sym}(\tau_1,\tau_2,\tau_3)$ are the symmetrized kernels, and we refer to them as the mirror anti-symmetric component.

We next evaluated how much variance in the fly's turning behavior can be explained by the estimated second- and third-order kernels. We presented the same stochastic stimulus sequence to many flies, so we averaged the turning response from different flies, denoted as $\bar{r}_{turn}(t)$ to estimate the true stimulus-associated turning response $r_{stim-driven}(t)$. We predicted the turning response $r_{pred}(t)$ to the same stimulus sequence,

$$r_{pred}(t)=r^{(2)}_{pred}(t)+r^{(3)}_{pred}(t)$$

where $r^{(2)}_{pred}(t)$ $\left(r^{(3)}_{pred}(t)\right)$ is the predicted response from the second-order (third-order) kernel. We calculated $r_{pred}(t)$ using only the anti-symmetric component of the kernel, i.e. $\widehat{K}^{(2)}_{LR-sym}(\tau_1,\tau_2)$ and $\widehat{K}^{(3)}_{LRL-sym}(\tau_1,\tau_2,\tau_3)$. The Pearson correlation between $r_{pred}(t)$ and $\bar{r}_{turn}(t)$ were 0.686 and 0.787 in 3-bar and 4-bar experiments.

If fly turning responses are driven only by visual stimuli, are mirror anti-symmetric, and use only second- and third-order correlations, then the measured second- and third-order kernel would explain all the variance in the stimulus-driven turning responses, and the Pearson correlation between $r_{pred}(t)$ and $r_{stim-driven}(t)$ should be one. However, our measured kernels only explained about half of the variance. There are several potential reasons for this. First, the turning responses of flies appeared very noisy, making it difficult to estimate the true stimulus-driven response. That is, $\bar{r}_{turn}(t)$ was a poor estimation of $r_{stim-driven}(t)$. Second, we wanted to minimally extend the canonical second-order motion detector while being able to account for light-dark asymmetric visual processing, so we added only one more term, the third-order kernel. However, the fly might respond to higher-order spatiotemporal correlations in visual inputs, and our model did not capture them.

Representing the second- and third-order kernels with impulse response format

To better understand and visualize the extracted kernels, we rearranged the elements in the kernels such that we could interpret kernels as the *impulse response* to a pair (triplet) of contrasts. This is analogous to the *impulse response* to a single contrast change at one point in a linear system.

Before rearrangement, the rows (columns) of the second-order kernel represent the temporal argument τ_1 (τ_2) in the matrix $\widehat{K}^{(2)}_{LR-sym}(\tau_1, \tau_2)$. After rearrangement, the rows correspond to the time since the more recent point, and the columns represent different temporal intervals between the two points, with negative intervals meaning that the right point is more recent than the left point (**Supplementary Figure 4D**) (Salazar-Gatzimas et al., 2016). We denote this new format as

$$\widehat{K}^{(2)}_{impulse}(\tau, \Delta\tau_{21}) \equiv \widehat{K}^{(2)}_{LR-sym}(\tau_1, \tau_1 + \Delta\tau_{21}),$$

where $\Delta\tau_{21} = \tau_2 - \tau_1$ and $\tau = \min(\tau_1, \tau_1 + \Delta\tau_{21})$. Because we have enforced mirror anti-symmetry in $\widehat{K}^{(2)}_{LR-sym}$, the columns of $\widehat{K}^{(2)}_{impulse}(\tau, \Delta\tau_{21})$ are anti-symmetric around $\Delta\tau_{21} = 0$. We interpreted the columns of $\widehat{K}^{(2)}_{impulse}(\tau, \Delta\tau_{21})$ as the impulse response of the fly to a pair of adjacent contrast changes separated by $\Delta\tau_{21}$ in time.

Similarly, before rearrangement, the three dimensions of the third-order kernel represent the three temporal arguments of $\widehat{K}^{(3)}_{LRL-sym}(\tau_1, \tau_2, \tau_3)$. Once rearranged, we define

$$\widehat{K}^{(3)}_{impulse}(\tau, \Delta\tau_{21}, \Delta\tau_{31}) \equiv \widehat{K}^{(3)}_{LRL-sym}(\tau_1, \tau_2 + \Delta\tau_{21}, \tau_1 + \Delta\tau_{31}),$$

where $\tau = \min(\tau_1, \tau_2 + \Delta\tau_{21}, \tau_1 + \Delta\tau_{31})$, $\Delta\tau_{21} = \tau_2 - \tau_1$, and $\Delta\tau_{31} = \tau_3 - \tau_1$. Rows again represent the time since the last point, the columns represent the temporal distance between the more recent point on the left and the sole right point, and the third tensor dimension represents the temporal distance between two left points (**Figure 4B**). For this third-order kernel, we also summed along the rows for 0.75 seconds to define the summed kernel strength (**Figure 4C**),

$$\widehat{K}^{(3)}_{summed}(\Delta\tau_{21}, \Delta\tau_{31}) = \sum_{\tau<0.75} \widehat{K}^{(3)}_{impulse}(\tau, \Delta\tau_{21}, \Delta\tau_{31}).$$

Testing the significance of the measured third-order kernel with "null kernels"

We tested the significance of the measured kernel with synthetic null kernels (**Figure 4C**). We shifted the stimulus with 100 random temporal offsets (the offset was at least 2 seconds long), reverse-correlated these shifted stimuli with responses, and generated 100 synthetic null kernels. The 100 kernels extracted from the misaligned stimulus and response were used to test the significance of the real kernel. We calculated the summed kernel strength of these 100 null kernels, and built the null distribution of summed kernel strength and performed two-tailed z-test. We tested kernel strength in the region of the kernels:$\tau_3 - \tau_1$ from 0 to 250ms, and $\tau_2 - \tau_1$ from – 250 to 250ms, which equaled $528 = 16 * 33$ kernel strengths in total. There are 43 significant ($p < 0.05$) responses, and around 23% of the total significant points (10 in total) aggregated when $|\tau_1 - \tau_2| < 83ms, |\tau_3 - \tau_1| < 83ms$. Therefore, we further simplified our kernel by setting third-order kernel elements to zero when $|\tau_1 - \tau_2| \geq 83ms$ or $|\tau_3 - \tau_1| \geq 83ms$, and denoted the "cleaned" kernel as $\widehat{K}^{(3)}_{LRL-sym-clean}(\tau_1, \tau_2, \tau_3)$. To be consistent, we also set elements of the second-order kernel to zero when $|\tau_1 - \tau_2| > 83ms$, and denoted it as $\widehat{K}^{(2)}_{LR-sym-clean}(\tau_1, \tau_2)$.

The exact p values for the summed third-order strength (**Figure 4C**) were shown from left bar to the right bar in the following table from top to bottom.

$\Delta\tau_{31} = 16ms$	$\Delta\tau_{31} = 33.3ms$	$\Delta\tau_{31} = 50ms$
.4328	.7506	.3948
.9784	.0514	.1870
.7381	.0587	.0904
.3340	.9650	.5661
.4435	.6801	.6273
.2154	.8081	.8081
.0645	.0942	.0641
.0008	.7077	.0015
<.0001	.3014	.9246
<.0001	.0823	.9375
.8873	<.0001	.0149
.0177	.3112	.0001
.1011	.0487	.2435
.2565	.3086	.5983
.6328	.5057	.7153
.1037	.4815	.3233
.3677	.2075	.6670

Comparing the measured third-order kernel with the glider responses

We measured the fly's sensitivity to three-point correlations using a suit of third-order glider stimuli. Overall, we presented 52 = 13 × 2 × 2 different stimuli (See Visual stimuli in **Methods**, **Table 1, Supplementary Figure 4-1A**). Each glider stimulus elicited sustained turning responses (Clark et al., 2014), so we averaged the response over time and denote it as $r^{glider}_{(\tau_{31},\tau_{21},L/R,P)}$, where the subscript specifies the stimulus type. Since we assume the fly's motion computation is mirror anti-symmetric, we subtracted responses to the pairs of gliders with different directions but with the same temporal interval and polarity, and denote it as $r^{glider}_{(\tau_{31},\tau_{21},P)}$,

$$r^{glider}_{(\tau_{31},\tau_{21},P)} = \frac{1}{2}\left(r^{glider}_{(\tau_{31},\tau_{21},L,P)} - r^{glider}_{(\tau_{31},\tau_{21},R,P)}\right) \; or \; \frac{1}{2}\left(r^{glider}_{(\tau_{31},\tau_{21},R,P)} - r^{glider}_{(\tau_{31},\tau_{21},L,P)}\right).$$

We plotted 18 out of 26 averaged responses in **Supplementary Figure 4-1B**.

In **Table 1**, we listed the number of flies tested for each glider ($n_{P=1}$ is the number of flies tested with positive gliders, $n_{P=-1}$ with negative gliders), and the p-values of Student t-tests, which were tested against zero response ($p_{P=1}$ is the significance level for positive gliders, $p_{P=-1}$ for negative gliders).

The measured third-order kernel and the measured glider responses should both reflect the fly's sensitivity to three-point correlations. To test agreement between these two measurements, we used the measured third-order kernel to predict the fly's responses to glider stimuli. We made the prediction by summing the "diagonal line" of the third-order kernel. Specifically, we found the predicted response to specific third-order gliders by summing over all elements in the kernel with the same temporal differences as the glider:

$$r^{(3)-pred}_{(\tau_{31},\tau_{21})} = 54 * 6 \sum_{\tau} K^{(3)}_{LRL-sym}(\tau, \tau+\tau_{21}, \tau+\tau_{31})\,(\Delta t)^3$$

The constant of $54 * 6$ takes into consideration the spatial summation all 54 putative MEDs and all six parts of third-order kernel in one EMD (**Appendix A**). Since the third-order kernel is agnostic to the polarity of the three-point correlations and reflected only the average of the fly's sensitivity to positive and negative correlations, we averaged the responses of positive and negative gliders.

$$r^{glider-ave}_{(\tau_{31},\tau_{21})} = \frac{1}{2}\left(r^{glider}_{(\tau_{31},\tau_{21},1)} - r^{glider}_{(\tau_{31},\tau_{21},-1)}\right).$$

We then compare the $r^{glider-ave}_{(\tau_{31},\tau_{21})}$ with $r^{(3)-pred}_{(\tau_{31},\tau_{21})}$ (**Figure 4D**).

Table 1. Statistics of responses to third-order glider stimuli with different spatiotemporal structures

index	$\Delta\tau_{31}\ (16ms)$	$\Delta\tau_{21}\ (16ms)$	$(n_{P=1}, n_{P=-1})$	$(p_{P=1}, p_{P=-1})$
1	1	0	(18, 12)	(0.0003, <0.0001)
2	2	0	(35,29)	(0.0299, 0.0003)
3	3	0	(14,8)	(0.4218, 0.0092)
4	4	0	(14,8)	(0.0201, 0.3552)
5	1	1	(18,13)	(<0.0001, <0.0001)
6	2	2	(35,30)	(0.0026, < 0.0001)
7	3	3	(14,9)	(0.2323, 0.0875)
8	4	4	(14,9)	(0.6778, 0.1700)
9	1	-1	(8,8)	(<0.0001, 0.0044)
10	2	1	(8,8)	(0.0041, 0.0617)
11	1	2	(8,8)	(0.0044, 0.6713)
12	3	1	(21,21)	(0.5396, 0.4470)
13	3	2	(21,21)	(0.1042, 0.0203)

Comparing the temporal structure of the second- and third-order kernels

To compare the temporal structure of the two kernels, we first rearranged and combined the elements in the third-order kernel into a 2-dimensional representation and then rearrange the second-order kernel with impulse response format. Specifically, we rearranged the elements in the third-order kernel into $\widehat{K}^{(3)}_{aligned-LRL}(\tau, \Delta\tau_{LR}, \Delta\tau_{LL})$, where τ corresponds to the time since the most recent point, $\Delta\tau_{LR}$ is the average time difference between left points and the right point, and $2\Delta\tau_{LL}$ is the temporal separation between the two left points (**Supplementary Figure 4-2A** *left*). In particular,

$$\widehat{K}^{(3)}_{aligned-LRL}(\tau, \Delta\tau_{LR}, \Delta\tau_{LL}) \equiv \widehat{K}^{(3)}_{LRL-sym}(\tau_2 + \Delta\tau_{LR} - \Delta\tau_{LL}, \tau_2, \tau_2 + \Delta\tau_{LR} + \Delta\tau_{LL})$$

where $\tau = \min(\{\tau_2 + \Delta\tau_{LR} - \Delta\tau_{LL}, \tau_2, \tau_2 + \Delta\tau_{LR} + \Delta\tau_{LL}\})$. We similarly defined

$$\widehat{K}^{(3)}_{aligned-RLR}(\tau, \Delta\tau_{RL}, \Delta\tau_{RR}) \equiv \widehat{K}^{(3)}_{RLR-sym}(\tau_2 + \Delta\tau_{RL} - \Delta\tau_{RR}, \tau_2, \tau_2 + \Delta\tau_{RL} + \Delta\tau_{RR}),$$

where $\Delta\tau_{RL}$ is the average time difference between right points and the left point, and $2\Delta\tau_{RR}$ is the temporal separation between the two right points (**Supplementary Figure 4-2A** *right*). Finally, we summed over the within-point time differences ($\Delta\tau_{RR}$ and $\Delta\tau_{LL}$), and summed these two pieces to obtain a matrix (**Supplementary Figure 4-2B**),

$$\widehat{K}^{(3)}_{align-2D}(\tau, \Delta\tau_{RL}) = \sum_{2\Delta\tau_{LL}<9\Delta t} \widehat{K}^{(3)}_{aligned-LRL}(\tau, -\Delta\tau_{RL}, \Delta\tau_{LL}) + \sum_{2\Delta\tau_{RR}<9\Delta t} \widehat{K}^{(3)}_{aligned-RLR}(\tau, \Delta\tau_{RL}, \Delta\tau_{RR}),$$

where $\Delta t = 1/60$ sec.

$\widehat{K}^{(3)}_{align-2D}(\tau, \Delta\tau_{RL})$ has rows and columns that are conceptually comparable to those of $\widehat{K}^{(2)}_{impulse}(\tau, \Delta\tau_{21})$, as rows represent times since the most recent point and columns describe the temporal distance between right and left points. However, in $\widehat{K}^{(2)}_{impulse}$ the columns are spaced by 16.67ms **Supplementary Figure 4-2D**), whereas in $\widehat{K}^{(3)}_{align-2D}$ the columns are spaced by 8.33ms (**Supplementary Figure 4-2B**). This results from the fact that $\Delta\tau_{21}$ is an integer number of frames in $\widehat{K}^{(2)}_{impulse}$, whereas $2\Delta\tau_{RL}$ is an integer number of frames in $\widehat{K}^{(3)}_{align-2D}$. We thus averaged two neighboring elements in $\widehat{K}^{(3)}_{align-2D}$ (**Supplementary Figure 4-2C**), so that it has the same resolution as the $\widehat{K}^{(2)}_{impulse}$.

We then summed both $\widehat{K}^{(3)}_{align-2D}$ and $\widehat{K}^{(2)}_{impulse}$ in each column, and we rescaled the two summed kernels so that the norm of each is 1 (**Figure 4E**). For visualization purpose, we also flipped the sign of the summed $\widehat{K}^{(3)}_{align-2D}$.

Comparing the second- and third-order kernels with singular value decomposition (SVD)

We factorized the $\widehat{K}^{(3)}_{align-2D}$ and $\widehat{K}^{(2)}_{impulse}$ into the products of a set of basis vectors with SVD (**Supplementary Figure 4-2EFG**),

$$\widehat{K}^{(2)}_{impulse} = U^{(2)} \cdot \Sigma^{(2)} \cdot {V^{(2)}}^{T},$$

$$\widehat{K}^{(3)}_{align-2D} = U^{(3)} \cdot \Sigma^{(3)} \cdot {V^{(3)}}^{T},$$

where $U^{(i)}, \Sigma^{(i)}, V^{(i)}$, are the left-singular vectors, singular values and right-singular vectors of the *i*th-order kernel. We use $u_1^{(2)}, v_1^{(2)}$ ($u_1^{(3)}, v_1^{(3)}$) to denote the left and right singular vectors corresponding to the largest singular values. For visualization purpose, in **Supplementary Figure 4-2G,** we flipped the sign of $v_1^{(3)}$, and we rescaled $u_1^{(2)}, v_1^{(2)}, u_1^{(3)}, v_1^{(3)}$ such that the norm of each is 1.

Extracting kernels of various motion detectors that were optimized to predict image velocity in natural scenes.

We characterized four other motion detectors (**Supplementary Figure 4-3)** with Volterra kernels (Fitzgerald and Clark, 2015). These motion detectors have various physiological plausible structures and were optimized to predict image velocities in natural scenes. We fed the same stochastic binary stimuli sequence to these motion detectors, collected the corresponding responses, and extracted the second and third-order kernels using reverse-correlation. In **Figure 4F**, we presented the summed kernel strength of the third-order kernels. Note that we only presented several examples, and the spatiotemporal arguments of these examples are represented graphically.

Natural scene dataset

We used a natural scene dataset (Meyer et al., 2014), which contains 421 panoramic luminance-calibrated naturalistic 2-dimensional pictures. Each picture has 927 * 251 pixels and subtends 360° horizontally and 97.4° vertically, so that the spatial resolution is ~0.30º/pixel. In our study, we used 1-dimensional images, which were single rows from the 2-dimensional pictures. Therefore, there were $105671 = 421 * 251$ images in the dataset. We refer to the 2-dimensional scenes as pictures or photographs, and refer to the 1-dimensional slices as images.

Preprocessing photographs

To simulate the photoreceptor, we converted the luminance pictures into contrast pictures with a blurring step and contrast computation step.

First, to simulate the spatial resolution of the fly's ommatidia, we blurred the original photograph (**Supplementary Figure 1-1B**), denoted by I, with a 2-dimensional Gaussian filter $f_{blur}(x, y)$.

$$I_{blur}(x, y) = I \otimes f_{blur} = \sum_{u,v} I(x + u, y + v) f_{blur}(u, v),$$

$$f_{blur}(u, v) = \frac{1}{2\pi\lambda_{blurr}^2} \exp\left(-\frac{u^2 + v^2}{2\lambda_{blur}^2}\right), u_- \le u \le u_+, v_- \le v \le v_+.$$

where $\otimes$ denotes cross-correlation. The filter extends to $\pm 3\lambda_{blur}$, i.e. $u_+ = u_+ = |v_-| = |v_-| = 3\lambda_{blur}$, λ_{blur} is related to full-width-at-half-maximum (FWHM) by $\lambda_{blur} = \frac{\mathrm{FWHM}_{blur}}{2\sqrt{2ln2}}$ and we chose $\mathrm{FWHM}_{blur} = 5.3°$ (Stavenga, 2003). The original picture covers the full circular range horizontally, but only 97.4° vertically. When the range of the filter extended beyond the vertical boundary of the picture, we padded the picture by "vertical reflection". This reflection padding was also used when we calculated the local mean-luminance.

Second, we converted the luminance signals in the blurred photograph to contrast signals (**Supplementary Figure 1-1CD**) (Fitzgerald and Clark, 2015),

$$c(x, y) = \frac{I_{blur}(x, y) - I_{mean}(x, y)}{I_{mean}(x, y)},$$

$$I_{mean}(x, y) = I_{blur} \otimes f_{local-mean},$$

where $c(x, y)$ is the contrast at each location (x, y). I_{mean} is the local mean luminance, which is the averaged luminance weighted by a 2-dimensional Gaussian spatial filter $f_{local-mean}$. The length scale of $f_{local-mean}$ can be equivalently described by $\lambda_{local-mean}$ and $\mathrm{FWHM}_{local-mean}$, where $\lambda_{local-mean} = \frac{\mathrm{FWHM}_{local-mean}}{2\sqrt{2ln2}}$. We swept $\mathrm{FWHM}_{local-mean}$ from 10° to 75° (**Supplementary Figure 3-2ABC**).

Alternatively, we also computed the local mean luminance over time instead of over space in **Supplementary Figure 3-2FG**,

$$I_{mean}(x, y, t) = \sum_{u} I_{blur-time}(x, y, t) f_{local-mean-time}(t - u),$$

where the local mean luminance was the averaged luminance signals over time, with the temporal filter $f_{local-mean-time}(t) = \exp(-t/\tau_{local-mean})$, where $f_{local-mean-time}(t)$ is normalized to have a sum of 1. We swept $\tau_{local-mean}$ from 10ms to 500ms (**Supplementary Figure 3-2FG**).

Depending on the parameters in local mean computation, we had 20 natural scene datasets, including 14 datasets whose local mean luminance was computed statically and 6 dynamically. Unless specified, we used the natural scene dataset with contrast images preprocessed statically with $\text{FWHM}_{local-mean} = 25°$.

Eliminating the higher-order structure of natural scene ensemble

We created a synthetic image dataset where we effectively preserved only the second-order structure of natural scenes ensemble and eliminated the higher-order structure. We viewed each 1D image as a random vector **X** determined by $\text{P}_{\text{natural}}(\boldsymbol{X})$, where $\boldsymbol{X}$ is a n-dimensional vector (n=927 for 927 pixels). The covariance matrix of $\text{P}_{\text{natural}}(\boldsymbol{X})$ determines the point variance and pairwise spatial correlations in natural scenes. We intended to construct a Gaussian distribution $\text{P}_{\text{synthetic}}(\boldsymbol{X})$ such that its covariance matrix is the same as $\text{P}_{\text{natural}}(\boldsymbol{X})$. In this way, the image ensemble sampled from $\text{P}_{\text{synthetic}}(\boldsymbol{X})$ will contain the same second-order statistics as the natural scene ensemble, but it would lack the higher-order statistics present in the natural scene ensemble.

Because the pixels in images are horizontally translational invariant, the covariance matrix of $\text{P}_{\text{natural}}(\boldsymbol{X})$ is a circulant matrix and can be diagonalized by a discrete Fourier transform. Therefore, we constructed $\text{P}_{\text{synthetic}}(\boldsymbol{X})$ and generated the synthetic dataset in the frequency domain (Bialek, 2012). Operationally, we first performed a discrete Fourier transform (with fft function in Matlab v2018a, RRID:SCR_001622) on each 1-dimensional image in the natural scene dataset and obtained its Fourier domain representation $\boldsymbol{y}^{(i)} = (y_{k_1}^{(i)}, y_{k_2}^{(i)}, \dots, y_{k_n}^{(i)})$, where (i) denotes the ith images, and k_n denotes the k_nth Fourier component. We then calculated the average power of frequency k_n, denoted as $\sigma_{k_n}^2$, where $\sigma_{k_n}^2 = \frac{1}{m}\sum_i^m \left(y_{k_n}^{(i)}\right)\left(y_{k_n}^{(i)}\right)^*$, and $*$ denotes complex conjugate. At each frequency k_n, we built two Gaussian distributions $G_{Re} \sim \mathcal{N}(0, \frac{\sigma_{k_n}^2}{2})$ and $G_{Im} \sim \mathcal{N}(0, \frac{\sigma_{k_n}^2}{2})$. To generate one synthetic image, we sampled two real numbers from these two distributions as the real and imaginary part of Fourier component of the synthetic image at frequency k_n. In the end, we performed an inverse Fourier transform to the sampled Fourier components to gain a synthetic image in the spatial domain. In total, we generated 1000 high-order-structure-eliminated synthetic images, and refer to this synthetic image dataset to as the dataset in which higher-order structure was eliminated.

Computing and manipulating statistics of individual 1-dimensional images

The natural scene dataset was comprised of an ensemble of heterogeneous images, and the statistics of different images can be drastically different from each other. Thus, we considered each 1-dimensional image to have its own contrast distribution, $P_{pixel}^{(i)}(X)$, where i indexes the image and the contrast of each pixel in an image as an independent sample of the random variable X. For each natural image, we computed its sample mean, sample variance, sample skewness, and sample kurtosis of pixels, and show the histogram of these statistics (**Supplementary Figure 5AB** and **Figure 5B**).

We generated 14=2+10+2 synthetic image datasets to mimic various statistical properties of natural scenes (**Table 2**). These 14 datasets differ in three parameters: 1) the contrast range; 2) whether the contrast skewness was constrained; and 3) the specific value of the imposed skewness when the skewness was constrained. We conceptually justify this image synthesis method in **Appendix B**, and here we focus on methodological details. For every image in the natural scene dataset, we generated a corresponding image in each synthetic image dataset. This involved two steps.

In step one, we determined all relevant image statistics and generated the corresponding maximum entropy distribution (MED). Operationally, for each individual image, we found its contrast range, $[c_{min}, c_{max}]$, the largest contrast magnitude, $\delta c = \max(|c_{min}|, |c_{max}|)$, and the sample mean, c_μ. We then derived the contrast range specified in Table 2, binned the range into N discrete levels, calculated the contrast frequency at each level to estimate $P_{pixel}^{(i)}(X)$, and estimated the contrast variance c_{σ^2} and skewness c_γ from this estimated distribution. Finally, we solved the MED (**Appendix B**) with the constrained statistics specified in **Table 2**. In the "imposed skewness" column, "NA" means that the skewness was not constrained in the MED.

In step two, we generated the synthetic image. The solved MED captures the pixel statistics but does not capture any spatial correlations between pixels. Therefore, we decided to interpolate between sampled pixels to coarsely mimic the spatial correlations. Operationally, for an individual image, we calculated its spatial correlation function and found the cutoff distance, Δx_α, where the correlation falls below $\alpha = 0.2$. We then sampled contrast values independently from the solved MED and placed these contrasts Δx_α pixels apart. Finally, we up-sampled the low-resolution image to a high-resolution image using the **resample** function in Matlab v2018a (RRID:SCR_001622).

Note that MED-3 through MED-7 are theoretically related to MED-8 through MED-12 by contrast inversion. We thus generated synthetic datasets MED-8 to MED-12 by simply inverting the contrast of images in dataset MED-3 to MED-7.

Table 2. Parameters for synthetic image datasets.

Index of dataset	imposed mean	imposed variance	imposed skewness	contrast range	discrete levels
MED-1	c_μ	c_{σ^2}	NA	$[c_\mu - \delta c, c_\mu + \delta c]$	32
MED-2	c_μ	c_{σ^2}	c_γ	$[c_\mu - \delta c, c_\mu + \delta c]$	32
MED-3	0	c_{σ^2}	1.25	$[-\delta c, \delta c]$	32
MED-4	0	c_{σ^2}	1	$[-\delta c, \delta c]$	32
MED-5	0	c_{σ^2}	0.75	$[-\delta c, \delta c]$	32
MED-6	0	c_{σ^2}	0.5	$[-\delta c, \delta c]$	32
MED-7	0	c_{σ^2}	0.25	$[-\delta c, \delta c]$	32
MED-8	0	c_{σ^2}	-0.25	$[-\delta c, \delta c]$	32
MED-9	0	c_{σ^2}	-0.5	$[-\delta c, \delta c]$	32
MED-10	0	c_{σ^2}	-0.75	$[-\delta c, \delta c]$	32
MED-11	0	c_{σ^2}	-1	$[-\delta c, \delta c]$	32
MED-12	0	c_{σ^2}	-1.25	$[-\delta c, \delta c]$	32
MED-13	0	c_{σ^2}	NA	$[-2.5, 2.5]$	512
MED-14	0	c_{σ^2}	c_γ	$[-2.5, 2.5]$	512

Simulating naturalistic motion with natural scenes

To simulate the full-field motion signals induced by self-rotation in the natural environment, we rigidly translated images at various horizontal velocities. For each trial, we randomly chose one 1-dimensional image from the dataset. We had 35 image datasets in total, including 20 natural scene datasets preprocessed with different parameters and 15 synthetic image datasets, we therefore built 35 naturalistic motion datasets. Where all images were sampled from a particular image dataset in each motion dataset. We sampled the velocity in two ways. First, we sampled it from a Gaussian distribution with zero mean and standard deviation of 114°/s This standard deviation is the measured standard deviation of the spontaneous rotational turning speed of freely walking flies. In **Supplementary Figure 3-1**, we varied the standard deviation of the Gaussian distribution from 32°/s to 512°/s. Second, we selected an image velocity at discrete values ranging from 0°/s to 1000°/s with a 10°/s interval. Given an image-velocity pair, we rigidly move this image with this velocity for one second. The temporal resolution is 60Hz.

To eliminate any potential left-right asymmetry in naturalistic motion datasets, if we moved an image rightward at a certain speed, we always simulated a paired trial in which the same image was flipped horizontally and moved leftward with the same speed.

Calculating the responses of different motion detectors to naturalistic motion signals

Our study concerns two motion detectors: the HRC and the measured kernels. For the HRC, we created an array of 53 overlapping HRC elementary motion detectors which extended 270° horizontally. We calculated the instantaneous HRCs' outputs at the end of each trial and averaged them across space to get the model's output r_{HRC}. For the measured kernels, we calculated the instantaneous output of the kernels at the end of each trial, including the second-order response, $r^{(2)}$, the third-order response, $r^{(3)}$, and the total response, $r = r^{(2)} + r^{(3)}$.

In **Table 3**, we listed the simulation parameters for each figure, including the motion detector, the image dataset, the velocity distribution, and the number of trials (image-velocity pairs). If the dataset is natural scenes, we also listed its preprocessing parameter $FWHM_{local-mean}$. If the velocity distribution is Gaussian, we listed its standard deviation.

For example, in Figure 1C, we predicted the motion estimates of the HRC to the naturalistic motion. The naturalistic motion was created with images sampled from the natural scene dataset that has preprocessing parameter $FWHM_{local-mean} = 25°$, and created with velocity sampled from the Gaussian distribution that has standard deviation of 114°/s. We simulated 8000 independent trials, and had 16000 trials after enforcing left-right symmetry.

Table 3. Natural scene datasets for naturalistic motion simulations.

	Image Dataset ($FWHM_{cont}/\tau_{cont}$)	Velocity Distribution (σ_{vel}, or discrete values)	Number of Trials	Motion Detector
Fig. 1C. *green*	natural scene (25°)	Discrete [0:10:1000]	1000 each velocity	HRC
Fig. 1C. *purple*	synthetic-higher-order structure eliminated	Discrete [0:10:1000]	1000 each velocity	HRC
Fig. 3	natural scene (25°)	Gaussian (114)	10000	Fly
Supp. Fig. 3-1	natural scene (25°)	Gaussian (32, 64, 128, 256,	8000 each velocity distribution	Fly

		512)		
Supp. Fig. 3-2DE	natural scene (10°~75°)	Gaussian (114)	8000 each $FWHM_{local-mean}$	Fly
Supp. Fig. 3-2FG	natural scene (10~500ms)	Gaussian (114)	8000 each $\tau_{local-mean-time}$	Fly
Fig. 5CDE *green*, Supp. Fig. 5-3A	synthetic-MED-1	Gaussian (114)	8000	Fly
Fig. 5CDE *brown* Supp. Fig. 5-3B	synthetic-MED-2	Gaussian (114)	8000	Fly
Fig. 5 FGH Supp. Fig. 5-3CD	synthetic-MED 5-14	Gaussian (114)	8000 each dataset	Fly
Supp. Fig. 5G *blue*	synthetic-MED-13	Gaussian (114)	8000	Fly
Supp. Fig.5G *red*	synthetic-MED-14	Gaussian (114)	8000	Fly

Evaluating the performance of a motion detector.

We denote the true velocity of the image as v_{img} and the response of a motion detector as v_{est}. We assessed the accuracy of the motion estimation using two metrics. First, we measured the Pearson correlation ρ between v_{img} and v_{est},

$$\rho = \frac{\mathrm{cov}(v_{est}, v_{img})}{\sqrt{\mathrm{var}(v_{est})}\sqrt{\mathrm{var}(v_{img})}},$$

where the variances and covariances are evaluated across any of the naturalistic motion datasets introduced above. To estimate the uncertainty in ρ induced by finite sample sizes, we randomly separated all independently-simulated trials into 10 groups, calculated the Pearson correlation for each group, and estimated the *SEM* of the Pearson correlation across the groups. Second, when the image velocity was sampled at discrete values, we measured the variance of v_{est} conditional on each possible image velocity, $\mathrm{var}(v_{est}|v_{img} = v_0)$.

Evaluating the improvement added by the third-order response

To evaluated how much improvement was added by the third-order response to the second-order response, we calculated the relative Pearson correlations: $\mathrm{improvement} = \frac{\rho^{(2+3)} - \rho^{(2)}}{\rho^{(2)}}$. As in *Evaluating the performance of a motion detector*, to estimate the uncertainty induced by finite sample sizes, we separated all trials into 10 groups, calculated the Pearson correlation for each group, calculated the improvements in each group, and estimated the *SEM* of the improvements across the groups.

Assessing the empirical weighting of the second-order and third-order responses

We modeled the image velocity as a linear combination of the second-order and third-order responses

$$v_{img} = \alpha^{(2)} r^{(2)} + \alpha^{(3)} r^{(3)}$$

and estimated the optimal weighting coefficients, $(\hat{\alpha}^{(2)}, \hat{\alpha}^{(3)})$, using ordinary least square regression. We calculated $\rho_{best} = \frac{\mathrm{cov}(r_{best}, v_{img})}{\sqrt{\mathrm{var}(r_{best})}\sqrt{\mathrm{var}(v_{img})}}$, where $r_{best} = \hat{\alpha}^{(2)} r^{(2)} + \hat{\alpha}^{(3)} r^{(3)}$. We computed the relative weighting coefficient as $w = \hat{\alpha}^{(2)} / \hat{\alpha}^{(3)}$.

Calculating the residual of second-order response

The second-order response can be viewed as a function of the image velocity, as well as noise that depends on image structure:

$$r^{(2)} = \beta^{(2)} v_{img} + \epsilon^{(2)}.$$

We estimated the noise term as

$$\hat{\epsilon}^{(2)} = r^{(2)} - \hat{\beta}^{(2)} v_{img},$$

where $\hat{\beta}^{(2)}$ minimized the residual and $\hat{\epsilon}^{(2)}$ denotes the estimated residual (noise).

Data and Code Availability

All code and data produce these figures are available at:
https://github.com/ClarkLabCode/ThirdOrderKernelCode for code, and
https://datadryad.org/review?doi=doi:10.5061/dryad.7jm87bt for data.

Appendix A: Modeling the fly's optomotor turning response with Volterra kernels.

The visual motion system of the fly can generally be considered as a nonlinear mapping from spatiotemporal visual stimuli, denoted $s(x,t)$, to behavioral turning responses, denoted $r(t)$. In general, any time-invariant nonlinear system can be written with a Volterra series,

$$r(t) = H^{(0)} + H^{(1)}[s(x,t)] + H^{(2)}[s(x,t)] + H^{(3)}[s(x,t)] + \cdots,$$

where each $H^{(n)}$ is a convolutional operator,

$$H^{(n)}[s(x,t)] = \int d\eta_1 \int d\tau_1 \cdots \int d\eta_n \int d\tau_n K^{(n)}(\eta_1, \ldots, \eta_n, \tau_1, \ldots, \tau_n) s(\eta_1, t-\tau_1) \cdots s(\eta_n, t-\tau_n),$$

η_i are spatial coordinates, τ_i are temporal coordinates, and $K^{(n)}$ is termed the nth-order Volterra kernel (Marmarelis and McCann, 1973; Schetzen, 1980). The system is fully specified by these kernels. In this section, we describe how we use this framework to model the fly's visual system. We make several simplifying assumptions about the fly's visual system.

Assumption 1. The system is third-order. This implies, $K^{(n)}(\eta_1, \ldots, \eta_n, \tau_1, \ldots, \tau_n) = 0$ for $n > 3$. This choice aims to account for the system's potential ability to asymmetrical process light-dark signals while minimally extending the canonical second-order algorithm.

Assumption 2. The system is spatially-invariant. This implies that

$$K^{(n)}(\eta_1, \ldots, \eta_n, \tau_1, \ldots, \tau_n) = K^{(n)}(\eta_1 + \delta\eta, \ldots, \eta_n + \delta\eta, \tau_1, \ldots, \tau_n).$$

Assumption 3. The system consists of a spatial array of elementary motion detectors (EMDs), and each EMD responds to inputs that are close in space. Since the spatial distance between neighboring ommatidia in fruit fly is around 5°, we discretize space into pixels with $\Delta x = 5°$ resolution, and we use integer indices to denote the relative spatial location. We further assume that the fly EMD does not respond to interactions between points that are spaced more than Δx apart,

$$K^{(n)}(\eta_1, \ldots, \eta_n, \tau_1, \ldots, \tau_n) = 0, |\eta_i - \eta_j| > \Delta x.$$

This modeling simplification is frequently made in fly neuroscience (Buchner, 1976), but modern receptive field measurements in individual motion detectors in Drosophila suggest that they integrate over a wider field of view (Leonhardt et al., 2016; Salazar-Gatzimas et al., 2016).

Assumption 4. The system is mirror anti-symmetric. This means that if $r(t)$ is the system's response to stimulus $s(x,t)$, then the system will respond to $s(-x,t)$ with response $-r(t)$. This is an important assumption for motion processing systems, and we will discuss its implications for Volterra kernels later in this section.

Given the first three assumptions, we model the fly's turning response $r(t)$ as

$$r(t) = \sum_{\xi=1}^{M} \left(r_\xi^{(0)}(t) + r_\xi^{(1)}(t) + r_\xi^{(2)}(t) + r_\xi^{(3)}(t) \right),$$

where

$$r_\xi^{(0)}(t) = r_0,$$

$$r_\xi^{(1)}(t) = \int d\tau_1 K^{(1)}(\tau_1) s\big(x_\xi, t-\tau_1\big) + \int d\tau_1 K^{(1)}(\tau_1) s\big(x_\xi + \Delta x, t-\tau_1\big),$$

$$r_\xi^{(2)}(t) = \sum_{\eta_1=(0,\Delta x)} \sum_{\eta_2=(0,\Delta x)} \iint d\tau_1 d\tau_2 K^{(2)}(\eta_1, \eta_2, \tau_1, \tau_2) s\big(x_\xi + \eta_1, t-\tau_1\big) s(x_\xi + \eta_2, t-\tau_2),$$

$$r_\xi^{(3)}(t) = \sum_{\eta_1=(0,\Delta x)} \sum_{\eta_2=(0,\Delta x)} \sum_{\eta_3=(0,\Delta x)} \iiint d\tau_1 d\tau_2 d\tau_3 K^{(3)}(\eta_1, \eta_2, \eta_3, \tau_1, \tau_2, \tau_3) s\big(x_\xi + \eta_1, t-\tau_1\big) s\big(x_\xi + \eta_2, t-\tau_2\big) s\big(x_\xi + \eta_3, t-\tau_3\big).$$

The response $r(t)$ is the sum of responses contributed by the ith-order kernel at different spatial locations ξ, denoted with $r_\xi^{(i)}(t)$. $K^{(1)}(\tau_1), K^{(2)}(\eta_1, \eta_2, \tau_1, \tau_2)$, and $K^{(3)}(\eta_1, \eta_2, \eta_3, \tau_1, \tau_2, \tau_3)$ approximates how each EMD maps the inputs $s(x_\xi, t)$ and $s(x_\xi + \Delta x, t)$ into its output. $M = 54$ is the number of putative EMD units.

Next, we will simplify the notations for $K^{(2)}(\eta_1, \eta_2, \tau_1, \tau_2)$ and $K^{(3)}(\eta_1, \eta_2, \eta_3, \tau_1, \tau_2, \tau_3)$, and discuss symmetries in these kernels.

The second-order kernel $K^{(2)}(\eta_1, \eta_2, \tau_1, \tau_2)$ is a four-dimensional tensor, with 2 dimensions in time and 2 dimensions in space. Since $|\eta_1 - \eta_2| \le \Delta x$, the values can be taken are limited 2 discrete points, and all possible combinations of (η_1, η_2) are (L, L), (L, R), (R, L) and (R, R), where L (R) denotes left (right). We replace the spatial arguments with subscripts, and the 4-dimensional $K^{(2)}(\eta_1, \eta_2, \tau_1, \tau_2)$ can be rewritten with four 2-dimensional kernels,

$$K^{(2)}(\eta_1, \eta_2, \tau_1, \tau_2) = \begin{cases} K_{LL}^{(2)}(\tau_1, \tau_2) \\ K_{LR}^{(2)}(\tau_1, \tau_2) \\ K_{RL}^{(2)}(\tau_1, \tau_2) \\ K_{RR}^{(2)}(\tau_1, \tau_2) \end{cases}.$$

In $K_{\eta_1,\eta_2}^{(2)}(\tau_1, \tau_2)$, the first (second) temporal argument is related to the spatial location represented by the first (second) subscript. For example, $K_{LR}^{(2)}(\tau_1, \tau_2)$ means $K^{(2)}(\eta_1 = L, \eta_2 = R, \tau_1, \tau_2)$.

There are three types of symmetries in these $K_{\eta_1,\eta_2}^{(2)}(\tau_1, \tau_2)$. First, since the system is translationally-invariant, $K_{LL}^{(2)}(\tau_1, \tau_2) = K_{RR}^{(2)}(\tau_1, \tau_2)$. Second, $K_{LR}^{(2)}(\tau_1, \tau_2)$ describes how the system responds to $s_L(t-\tau_1)s_R(t-\tau_2)$, and $K_{RL}^{(2)}(\tau_2, \tau_1)$ describes how the system responds to $s_R(t-\tau_2)s_L(t-\tau_1)$. Because multiplication operator is commutative, i.e. $s_L(t-\tau_1)s_R(t-\tau_2) = s_R(t-\tau_2)s_L(t-\tau_1)$, one should have $K_{LR}^{(2)}(\tau_1, \tau_2) = K_{RL}^{(2)}(\tau_2, \tau_1)$.

Therefore, the second-order response of an EMD can be simplified as,

$$r_\xi^{(2)}(t) = \iint d\tau_1 d\tau_2 [2K_{LR}^{(2)}(\tau_1, \tau_2) s_\xi(t-\tau_1) s_{\xi+1}(t-\tau_2) + K_{LL}^{(2)}(\tau_1, \tau_2)(s_\xi(t-\tau_1) s_\xi(t-\tau_2) + s_{\xi+1}(t-\tau_1) s_{\xi+1}(t-\tau_2))],$$

where we replace the spatial arguments x in stimulus with discrete subscripts ξ as well.

To derive the consequences of mirror anti-symmetry, note that the second-order kernel is locally sensitive to two discrete points in space, $s(x,t) = [s_L(t), s_R(t)]$. Mirror-anti-symmetry states that if the system's response to $s_L(t) = a(t)$ and $s_R(t) = b(t)$ is $r(t)$, then the system's response to the reflected stimulus, $s_L(t) = b(t)$ and $s_R(t) = a(t)$, is $-r(t)$, where $a(t)$ and $b(t)$ are arbitrary functions of time. The system's response when $s_L(t) = a(t)$ and $s_R(t) = b(t)$ is

$$r^{(2)}_{\xi,ab}(t) = \iint d\tau_1 d\tau_2 \left[K^{(2)}_{LL}(\tau_1,\tau_2)\big(a(t-\tau_1)a(t-\tau_2) + b(t-\tau_1)b(t-\tau_2)\big) + 2K^{(2)}_{LR}(\tau_1,\tau_2)a(t-\tau_1)b(t-\tau_2)\right],$$

and the response when $s_L(t) = b(t)$ and $s_R(t) = a(t)$ is

$$r^{(2)}_{\xi,ba}(t) = \iint d\tau_1 d\tau_2 \left[K^{(2)}_{LL}(\tau_1,\tau_2)\big(b(t-\tau_1)b(t-\tau_2) + a(t-\tau_1)a(t-\tau_2)\big) + 2K^{(2)}_{LR}(\tau_1,\tau_2)b(t-\tau_1)a(t-\tau_2)\right].$$

If $r^{(2)}_{\xi,ba}(t) = -r^{(2)}_{\xi,ab}(t)$ for any $a(t)$ and $b(t)$, then one must have

$$K^{(2)}_{LL}(\tau_1,\tau_2) = -K^{(2)}_{LL}(\tau_1,\tau_2) = 0,$$

and

$$K^{(2)}_{LR}(\tau_1,\tau_2) = -K^{(2)}_{LR}(\tau_2,\tau_1).$$

Thus, the general expression for the second-order response of the EMD further simplifies to

$$r^{(2)}_{\xi}(t) = 2\iint d\tau_1 d\tau_2 K^{(2)}_{LR}(\tau_1,\tau_2)s_{\xi}(t-\tau_1)s_{\xi+1}(t-\tau_2).$$

We analyzed the third-order kernel in a similar manner. It is a 6-dimensional tensor, with 3 dimensions in time and 3 dimensions in space. Its spatial arguments are also limited to two points, and all possible combinations of (η_1,η_2,η_3) are (L,L,L), (R,R,R), (L,L,R), (L,R,L), (R,L,L), (R,R,L), (R,L,R) and (L,R,R). We replace the spatial arguments with subscripts, and represent the 6-dimensional third-order kernel with eight 3-dimensional kernels,

$$K^{(3)}(\eta_1,\eta_2,\eta_3,\tau_1,\tau_2,\tau_3) = \begin{cases} K^{(3)}_{LLL}(\tau_1,\tau_2,\tau_3) \\ K^{(3)}_{RRR}(\tau_1,\tau_2,\tau_3) \\ K^{(3)}_{LLR}(\tau_1,\tau_2,\tau_3) \\ K^{(3)}_{LRL}(\tau_1,\tau_2,\tau_3) \\ K^{(3)}_{RLL}(\tau_1,\tau_2,\tau_3) \\ K^{(3)}_{RRL}(\tau_1,\tau_2,\tau_3) \\ K^{(3)}_{RLR}(\tau_1,\tau_2,\tau_3) \\ K^{(3)}_{LRR}(\tau_1,\tau_2,\tau_3). \end{cases}$$

As was the case of the second-order kernel, in $K^{(3)}_{\eta_1,\eta_2,\eta_3}(\tau_1,\tau_2,\tau_3)$, the first (second, third) temporal argument is related to the spatial location denoted with the first (second, third) subscript. For example, $K^{(3)}_{LLR}(\tau_1,\tau_2,\tau_3) = K^{(3)}(\eta_1 = L, \eta_2 = L, \eta_3 = R, \tau_1, \tau_2, \tau_3)$.

As before, these eight blocks are redundant. From spatial-invariance and the commutative property of multiplication, one finds

$$K^{(3)}_{LLL}(\tau_1,\tau_2,\tau_3) = K^{(3)}_{RRR}(\tau_1,\tau_2,\tau_3),$$

$$K^{(3)}_{LLR}(\tau_1,\tau_2,\tau_3) = K^{(3)}_{LRL}(\tau_1,\tau_3,\tau_2) = K^{(3)}_{RLL}(\tau_3,\tau_1,\tau_2),$$

$$K^{(3)}_{RRL}(\tau_1,\tau_2,\tau_3) = K^{(3)}_{RLR}(\tau_1,\tau_3,\tau_2) = K^{(3)}_{LRR}(\tau_3,\tau_1,\tau_2).$$

For each 3D tensor, at least two out of the three spatial arguments are the same, so there is symmetry within the tensor. For example,

$$K^{(3)}_{LLR}(\tau_1,\tau_2,\tau_3) = K^{(3)}_{LLR}(\tau_2,\tau_1,\tau_3)$$

and

$$K^{(3)}_{LRR}(\tau_1,\tau_2,\tau_3) = K^{(3)}_{LRR}(\tau_1,\tau_3,\tau_2).$$

Following this logic, the third-order response of an EMD can be written as

$$\begin{aligned} r^{(3)}_{\eta}(t) = \iiint d\tau_1 d\tau_2 d\tau_3 \Big[& 3K^{(3)}_{LLR}(\tau_1,\tau_2,\tau_3) s_{\xi}(t-\tau_1) s_{\xi}(t-\tau_2) s_{\xi+1}(t-\tau_3) \\ & + 3K^{(3)}_{RRL}(\tau_1,\tau_2,\tau_3) s_{\xi+1}(t-\tau_1) s_{\xi+1}(t-\tau_2) s_{\xi}(t-\tau_3) \\ & + K^{(3)}_{LLL}(\tau_1,\tau_2,\tau_3) \Big(s_{\xi}(t-\tau_1) s_{\xi}(t-\tau_2) s_{\xi}(t-\tau_3) \\ & + s_{\xi+1}(t-\tau_1) s_{\xi+1}(t-\tau_2) s_{\xi+1}(t-\tau_3) \Big) \Big]. \end{aligned}$$

One can again apply the definition of mirror anti-symmetry to find

$$K^{(3)}_{LLL}(\tau_1,\tau_2,\tau_3) = 0,$$

$$K^{(3)}_{LLR}(\tau_1,\tau_2,\tau_3) = -K^{(3)}_{RRL}(\tau_1,\tau_2,\tau_3).$$

Thus, with mirror anti-symmetry, the third-order response of an EMD can be simplified as,

$$\begin{aligned} r^{(3)}_{\eta}(t) = 3 \iiint d\tau_1 d\tau_2 d\tau_3 K^{(3)}_{LLR}(\tau_1,\tau_2,\tau_3) \Big[\Big(& s_{\xi}(t-\tau_1) s_{\xi}(t-\tau_2) s_{\xi+1}(t-\tau_3) \\ & - s_{\xi+1}(t-\tau_1) s_{\xi+1}(t-\tau_2) s_{\xi}(t-\tau_3) \Big) \Big]. \end{aligned}$$

We have shown how mirror anti-symmetry manifest in the second and third-order kernels. In a similar way, for the zero- and first-order kernel, mirror anti-symmetry implies that,

$$r_0 = 0,$$

$$K^{(1)}(\tau_1) = 0.$$

Intuitively, r_0 term is the response of the EMD when there is no visual input. A reasonable motion detector should not be biased to turn left or right when there is no visual input. The first-order kernel

describes how the response is influenced by stimulus at a single spatial location, which should not give any motion information, thus $K^{(1)}(\tau_1) = 0$.

In summary, in our model, the turning response is

$$r(t) = \sum_{\xi} r_{\xi}(t) = \sum_{\xi} r_{\xi}^{(2)}(t) + r_{\xi}^{(3)}(t),$$

$$r_{\xi}^{(2)}(t) = 2 \iint d\tau_1 d\tau_2 K_{LR}^{(2)}(\tau_1, \tau_2) s_{\xi}(t - \tau_1) s_{\xi+1}(t - \tau_2),$$

$$r_{\xi}^{(3)}(t) = 3 \iiint d\tau_1 d\tau_2 d\tau_3 K_{LLR}^{(3)}(\tau_1, \tau_2, \tau_3) \Big[\Big(s_{\xi}(t - \tau_1) s_{\xi}(t - \tau_2) s_{\xi+1}(t - \tau_3) - s_{\xi+1}(t - \tau_1) s_{\xi+1}(t - \tau_2) s_{\xi}(t - \tau_3) \Big) \Big],$$

where $r_{\xi}(t)$ is the response of each EMD at spatial location ξ. In the experiments, we discretized the time into bins of size $\Delta t = 1/60$ seconds, so the integrals become summations, and we write the local responses as

$$r_{\xi}^{(2)}(t) = 2 \sum_{\tau_1, \tau_2} K_{LR}^{(2)}(\tau_1, \tau_2) s_{\xi}(t - \tau_1) s_{\xi+1}(t - \tau_2) \, (\Delta t)^2,$$

$$r_{\xi}^{(3)}(t) = 3 \sum_{\tau_1, \tau_2, \tau_3} K_{LLR}^{(3)}(\tau_1, \tau_2, \tau_3) \Big[\Big(s_{\xi}(t - \tau_1) s_{\xi}(t - \tau_2) s_{\xi+1}(t - \tau_3) - s_{\xi+1}(t - \tau_1) s_{\xi+1}(t - \tau_2) s_{\xi}(t - \tau_3) \Big) \Big] (\Delta t)^3.$$

Appendix B: Manipulating image statistics with maximum entropy distributions

To investigate how the skewness of the image contrast distribution influences motion detection, we want to develop a method for generating synthetic images with specified skewness values. In Section 1, we formalize this image synthesis problem and provide a rationale for the maximum entropy method. In Section 2, we derive formulae for finding maximum entropy distributions (MED) with constrained lower-order moments. In Section 3, we discuss how the contrast range sets implicit constraints on the MED.

1. Motivation

We consider each 1-dimensional natural image to have its own contrast distribution, $P_{pixel}^{(i)}(X)$, such that the contrast of each pixel in the i^{th} image is an independent sample from $P_{pixel}^{(i)}(X)$. We would like to generate a synthetic contrast distribution, $P_{syn}^{(i)}(X)$, such that $P_{syn}^{(i)}(X)$ shares certain low-order contrast statistics with $P_{pixel}^{(i)}(X)$. More specifically, we want $P_{syn}^{(i)}(X)$ and $P_{pixel}^{(i)}(X)$ to have matched means, variances, and/or skewness levels. However, it's important to recognize that $P_{syn}^{(i)}(X)$ is ambiguously determined by its lower-order moments, because many distributions share finite set of moments. Nevertheless, among all of these qualified distributions, the distribution with maximal entropy is unique. The maximum entropy requirement thus implicitly specifies all unconstrained statistics. Maximum entropy distributions are also conceptually appealing because they have the least statistical structure consistent with the set of chosen constraints. Here we use maximum entropy distributions to generate synthetic images with specific statistics by sampling pixel contrasts from $P_{syn}^{(i)}(X)$. By manipulating the statistics that define $P_{syn}^{(i)}(X)$, we are able to manipulate the contrast statistics of the synthetic image.

2. Solving maximum entropy distribution by minimizing the free energy function.

To illustrate the structure of the maximum entropy distribution (MED) with constrained moments, we concretely consider the example when the MED is constrained to have a specific mean, variance, and skewness. We refer to this distribution as the mean-variance-and-skewness constrained MED. Since constraining the mean, variance, and skewness is equivalent to constraining the first, second and third moments, we equivalently refer to this distribution as the third-order MED.

The entropy of a random variable X is

$$H(X) = \mathrm{E}[-\log P(X)],$$

where $P(X)$ is the probability distribution of X, and E[ ]is the expectation operator over $P(X)$. In particular, $\mathrm{E}[\ ] = \int dx P(X = x)\,[\ ]$. By definition, the third-order MED is the probability distribution, $P^*(X)$, such that

$$P^*(X) = \underset{P(X)}{\operatorname{argmax}}\, H(X),$$

subject to,

$$\mathrm{E}\left[X^i\right] = \mu_i, i = 1,2,3.$$

where μ_1, μ_2 and μ_3 are the first-, second-, and third moments. To optimize $P^*(X)$, one can introduce Lagrange multipliers, $\lambda_0, \lambda_1, \lambda_2$ and λ_3, to enforce the normalization of the probability distribution and the three moment constraints. The problem is thus transformed into extremizing

$$L(P(X),\lambda_0,\lambda_1,\lambda_2,\lambda_3) = \mathrm{E}[-\log P(X)] + \lambda_0\left(\int dx P(X=x) - 1\right) + \sum_{i=1,2,3}\lambda_i\left(\mathrm{E}\left[X^i\right] - \mu_i\right).$$

Note that $L(P(X),\lambda)$ is a functional of $P(X)$, and a necessary condition for $P(X)$ to extremize L is

$$0 = \frac{\delta L}{\delta P(X=x)} = -(\log P(X=x) + 1) + \lambda_0 + \lambda_1 x + \lambda_2 x^2 + \lambda_3 x^3.$$

This implies that third-order MED has the form

$$P(X=x) = \exp(-1 + \lambda_0 + \lambda_1 x + \lambda_2 x^2 + \lambda_3 x^3),$$

and the Lagrange multipliers must be solved to satisfy $\int dx P(X=x) = 1$, and $\mathrm{E}\left[X^i\right] = \mu_i, i = 1,2,3$.

An alternative to solving the nonlinear constraint-satisfaction equations for the Lagrange multipliers is to find $\boldsymbol{\lambda}^* = (\lambda_1,\lambda_2,\lambda_3)$ such that

$$\boldsymbol{\lambda}^* = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}}\, F(\boldsymbol{\lambda}),$$

where

$$F(\boldsymbol{\lambda}) = \log\left(\mathrm{Z}(\boldsymbol{\lambda})\right), \mathrm{Z}(\boldsymbol{\lambda}) = \int dx Q(x,\boldsymbol{\lambda}), Q(x,\boldsymbol{\lambda}) = \exp(\lambda_1(x-\mu_1) + \lambda_2(x^2-\mu_2) + \lambda_3(x^3-\mu_3)).$$

Here $Q(x,\boldsymbol{\lambda})$ can be thought of an unnormalized probability distribution for X, $\mathrm{Z}(\boldsymbol{\lambda})$ is the normalizing factor, and $F(\boldsymbol{\lambda})$ is the log of this normalizing factor. Readers familiar with statistical mechanics will recognize $F(\boldsymbol{\lambda})$ as the free energy function and $\mathrm{Z}(\boldsymbol{\lambda})$ as the partition function. It turns out that the third-order MED is

$$P^*(X) = \frac{1}{\mathrm{Z}(\boldsymbol{\lambda}^*)} Q(x,\boldsymbol{\lambda}^*).$$

This distribution manifestly has the right functional form, so to prove that it is the third-order MED we merely need to show that all constraints are satisfied by the distribution. Indeed, the constraint that $\int dx P(X=x) = 1$ is trivially satisfied because the partition function $\mathrm{Z}(\boldsymbol{\lambda}^*)$ normalizes $Q(x,\boldsymbol{\lambda}^*)$. Since $\boldsymbol{\lambda}^*$ minimizes the free energy, we also know

$$\frac{\partial F(\boldsymbol{\lambda}^*)}{\partial \lambda_i} = 0.$$

This derivative is easily evaluated as

$$\frac{\partial F(\boldsymbol{\lambda})}{\partial \lambda_i} = \frac{\partial \log(\mathrm{Z}(\boldsymbol{\lambda}))}{\partial \lambda_i} = \frac{1}{\mathrm{Z}(\boldsymbol{\lambda})}\frac{\partial \mathrm{Z}(\boldsymbol{\lambda})}{\partial \lambda_i} = \frac{1}{\mathrm{Z}(\boldsymbol{\lambda})}\frac{\partial \int dx Q(x,\boldsymbol{\lambda})}{\partial \lambda_i} = \frac{1}{\mathrm{Z}(\boldsymbol{\lambda})}\int dx \frac{\partial Q(x,\boldsymbol{\lambda})}{\partial \lambda_i}$$
$$= \frac{1}{\mathrm{Z}(\boldsymbol{\lambda})}\int dx Q(x,\boldsymbol{\lambda})\left(x^i - \mu_i\right) = E_\lambda\left[X^i\right] - \mu_i,$$

where $E_\lambda[\ \ ]$ is the expectation operator over $P_\lambda(x) = Q(x,\boldsymbol{\lambda})/\mathrm{Z}(\boldsymbol{\lambda})$. Consequently,

$$\frac{\partial F(\boldsymbol{\lambda}^*)}{\partial \lambda_i} = 0 \Longrightarrow E_{\lambda^*}\left[X^i\right] = \mu_i,$$

and minimizing free energy provides parameters that guarantee that the constraints are satisfied.

This free energy formulation is computationally convenient because $F(\boldsymbol{\lambda})$ is a convex function that can be easily minimized using powerful techniques from convex optimization. To see this, note that the matrix of second derivatives is,

$$\frac{\partial^2 F(\boldsymbol{\lambda})}{\partial\lambda_i\partial\lambda_j} = \frac{\partial}{\partial\lambda_i}\frac{1}{\mathrm{Z}(\boldsymbol{\lambda})}\frac{\partial \mathrm{Z}(\boldsymbol{\lambda})}{\partial\lambda_j} = -\left(\frac{1}{\mathrm{Z}(\boldsymbol{\lambda})}\right)^2\frac{\partial \mathrm{Z}(\boldsymbol{\lambda})}{\partial\lambda_i}\frac{\partial \mathrm{Z}(\boldsymbol{\lambda})}{\partial\lambda_j} + \frac{1}{\mathrm{Z}(\boldsymbol{\lambda})}\frac{\partial^2 \mathrm{Z}(\boldsymbol{\lambda})}{\partial\lambda_i\partial\lambda_j}.$$

We showed above that

$$\frac{1}{\mathrm{Z}(\boldsymbol{\lambda})}\frac{\partial \mathrm{Z}(\boldsymbol{\lambda})}{\partial\lambda_i} = E_\lambda\left[x^i - \mu_i\right] \Longrightarrow -\left(\frac{1}{\mathrm{Z}(\boldsymbol{\lambda})}\right)^2\frac{\partial \mathrm{Z}(\boldsymbol{\lambda})}{\partial\lambda_i}\frac{\partial \mathrm{Z}(\boldsymbol{\lambda})}{\partial\lambda_j} = -E_\lambda\left[x^i - \mu_i\right]E_\lambda\left[x^j - \mu_j\right].$$

Similarly, we next note that

$$\frac{1}{\mathrm{Z}(\boldsymbol{\lambda})}\frac{\partial^2 \mathrm{Z}(\boldsymbol{\lambda})}{\partial\lambda_i\partial\lambda_j} = \frac{1}{\mathrm{Z}(\boldsymbol{\lambda})}\frac{\partial^2}{\partial\lambda_i\partial\lambda_j}\int dx Q(x,\boldsymbol{\lambda}) = \frac{1}{\mathrm{Z}(\boldsymbol{\lambda})}\int dx\frac{\partial^2 Q(x,\boldsymbol{\lambda})}{\partial\lambda_i\partial\lambda_j}$$
$$= \frac{1}{\mathrm{Z}(\boldsymbol{\lambda})}\int dx Q(x,\boldsymbol{\lambda})\left(x^i - \mu_i\right)\left(x^j - \mu_j\right) = E_\lambda\left[\left(x^i - \mu_i\right)\left(x^j - \mu_j\right)\right]$$

Therefore, these two terms together imply that

$$\frac{\partial^2 F(\boldsymbol{\lambda})}{\partial\lambda_i\partial\lambda_j} = E_\lambda\left[\left(x^i - \mu_i\right)\left(x^j - \mu_j\right)\right] - E_\lambda\left[x^i - \mu_i\right]\mathrm{E}\left[x^j - \mu_j\right] = \mathrm{Cov}_\lambda\left[x^i - \mu_i, x^j - \mu_j\right].$$

Since the second derivative of $F(\boldsymbol{\lambda})$ is a covariance matrix for all values of $\boldsymbol{\lambda}$**,** and all covariance matrices are positive semi-definite, this implies $F(\boldsymbol{\lambda})$ is a convex function.

Similarly, one can find the maximum entropy distribution (MED) with constrained mean and variance by minimizing another free energy function:

$$\boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_2^*) = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}}\, F(\lambda_1, \lambda_2),$$

where

$$F(\boldsymbol{\lambda}) = \log\left(\mathrm{Z}(\boldsymbol{\lambda})\right), \mathrm{Z}(\boldsymbol{\lambda}) = \int dx Q(x,\boldsymbol{\lambda}), Q(x,\boldsymbol{\lambda}) = \exp(\lambda_1(x-\mu_1) + \lambda_2(x-\mu_2)),$$

and the maximum entropy distribution will be

$$P^*(X) = \frac{1}{\mathrm{Z}(\boldsymbol{\lambda}^*)}Q(x,\boldsymbol{\lambda}^*).$$

We refer this distribution as the mean-and-variance-constrained MED or second-order MED.

In practice, we solved each MED by numerically minimizing its associated free energy function with the **fminunc** function in MATLAB v2018a.

3. Contrast bounds set implicit constraints on the MED.

In the previous section, we emphasized that the free energy, F, depends on the Lagrange multipliers parameters $\boldsymbol{\lambda}$, because optimizing F over these parameters allowed us to find the corresponding MED. However, the free energy also depended on the constrained moments, $\boldsymbol{\mu}$, which entered the formulae as a set of fixed parameters. Moreover, the free energy depends on the state space of the random variable,

which is the set of values that X can take. For continuous variables, this manifests as the integral bound, $[a, b]$, in the partition function, $\mathrm{Z}(\boldsymbol{\lambda}) = \int_a^b dx Q(x, \boldsymbol{\lambda})$. Therefore, we generally expect

$$\boldsymbol{\lambda}^* = \boldsymbol{\lambda}^*(\boldsymbol{\mu}, a, b) = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} F(\boldsymbol{\lambda}|\boldsymbol{\mu}, a, b).$$

Since X represents the pixel contrast, we refer to $[a, b]$ as the contrast range and a or b separately as contrast bounds. In this section, we discuss in detail how $\boldsymbol{\lambda}^*$ depends on $[a, b]$. Since $[a, b]$ influences the MED, we say $[a, b]$ set an implicit constraint on the MED.

A natural choice of the contrast range is the entire real line. For example, when one constrains the mean and variance of a distribution and defines X on the entire real line, then the MED is the Gaussian distribution. Moreover, by rewriting the probability density of the Gaussian

$$P(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

in a slightly different form

$$P(X = x) = \frac{e^{-1/2}}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\mu}{\sigma^2}(x-\mu) - \frac{1}{2\sigma^2}(x^2 - \sigma^2 - \mu^2)\right)$$

we may identify the Lagrange multipliers that minimize the free energy as, $\lambda_1 = \frac{\mu}{\sigma^2}$, $\lambda_2 = -\frac{1}{2\sigma^2}$.

In the Gaussian distribution, the probability density drops to zero very fast as X goes to positive or negative infinity, because $\lambda_2 x^2$ is a large negative number in either case. However, when one wants to find a MED whose highest-order constraint is an odd-order moment, a finite contrast bound becomes necessary. This is because the sign of $\lambda_{2n+1} x^{2n+1}$ depends on the sign of x for any natural number n, thereby causing the probability density to approach zero on one half of the real line and to diverge on the other half. To illustrate this point concretely, consider the first-order MED that has a constrained mean. The first-order MED must take the form

$$P(X = x) = \frac{1}{Z} \exp(\lambda_1(x - \mu_1)).$$

In this distribution, $\exp(\lambda_1(x - \mu_1))$ blows up when x either goes to $+\infty$ or $-\infty$ (unless $\lambda_1 = 0$). In any case, no normalizing factor Z could satisfy $\int_{-\infty}^{+\infty} \frac{1}{Z} \exp(\lambda_1(x - \mu_1)) = 1$. Therefore, the first-order MED is generally ill-defined on the real line.

We thus sought to compute the dependence of the first-order MED on its contrast range. The parameters in the MED minimize the free energy

$$\lambda^* = \underset{\lambda}{\operatorname{argmin}} F(\lambda|a, b, \mu_1) = \underset{\lambda}{\operatorname{argmin}} \log \int_a^b dx \exp(\lambda(x - \mu_1)).$$

Without loss of generality, we can shift the x-axis such that the contrast range is symmetric around 0. Setting $x' = x - \frac{1}{2}(b + a)$, we find

$$\lambda^* = \underset{\lambda}{\operatorname{argmin}}\ F(\lambda|b', \mu') = \underset{\lambda}{\operatorname{argmin}} \log \int_{-b'}^{b'} dx' \exp\big(\lambda(x' - \mu')\big),$$

where $b' = \frac{1}{2}(b-a)$ and $\mu' = \mu_1 - \frac{1}{2}(b+a)$.

This optimization problem for the first-order MED can be solved analytically. First, note that the partition function is

$$Z(\lambda) = \int_{-b'}^{b'} dx' \exp\big(\lambda(x'-\mu')\big) = \begin{cases} 2b' & \lambda = 0 \\ \frac{1}{\lambda}\left(\mathrm{e}^{\lambda(b'-\mu')} - \mathrm{e}^{\lambda(-b'-\mu')}\right) & \lambda \neq 0 \end{cases}.$$

This partition function is continuous at $\lambda = 0$. To see this, note that

$$\lim_{\lambda\to 0} Z(\lambda) = \lim_{\lambda\to 0} \frac{\mathrm{e}^{\lambda(b'-\mu')} - \mathrm{e}^{\lambda(-b'-\mu')}}{\lambda} = \lim_{\lambda\to 0} \frac{(b'-\mu')\mathrm{e}^{\lambda(b'-\mu')} - (-b'-\mu')\mathrm{e}^{\lambda(-b'-\mu')}}{1} = 2b' = Z(0),$$

where we used L'Hôpital's rule to evaluate the limit. Therefore the free energy,

$$F(\lambda|b',\mu') = \begin{cases} \log(2b') & \lambda = 0 \\ \log\left(\mathrm{e}^{\lambda(b'-\mu')} - \mathrm{e}^{\lambda(-b'-\mu')}\right) - \log(\lambda) & \lambda \neq 0 \end{cases},$$

is also continuous at $\lambda = 0$. Furthermore, the partition function is differentiable at $\lambda = 0$. In particular,

$$\begin{aligned} \frac{\mathrm{dZ}}{d\lambda}|_{\lambda=0} &= \lim_{h\to 0} \frac{\mathrm{Z}(h) - \mathrm{Z}(0)}{h} = \lim_{h\to 0} \frac{\frac{1}{h}\left(\mathrm{e}^{h(b'-\mu')} - \mathrm{e}^{h(-b'-\mu')}\right) - 2b'}{h} \\ &= \lim_{h\to 0} \frac{\mathrm{e}^{h(b'-\mu')} - \mathrm{e}^{h(-b'-\mu')} - 2b'h}{h^2} \\ &= \lim_{h\to 0} \frac{(b'-\mu')\mathrm{e}^{h(b'-\mu')} - (-b'-\mu')\mathrm{e}^{h(-b'-\mu')} - 2b'}{2h} \\ &= \lim_{h\to 0} \frac{(b'-\mu')^2\mathrm{e}^{h(b'-\mu')} - (-b'-\mu')^2\mathrm{e}^{h(-b'-\mu')}}{2} = -2b'\mu', \end{aligned}$$

where we again used L'Hôpital's rule. Consequently, the derivative of the free energy at $\lambda = 0$ is

$$\frac{\mathrm{d}F}{d\lambda}|_{\lambda=0} = \frac{1}{Z}\frac{dZ}{d\lambda}|_{\lambda=0} = \frac{-2b'\mu'}{2b'} = -\mu'.$$

The derivative for $\lambda \neq 0$ is straightforward to evaluate, and we find

$$\frac{\mathrm{d}F}{d\lambda}|_{\lambda\neq 0} = \frac{(b'-\mu')\mathrm{e}^{\lambda(b'-\mu')} - (-b'-\mu')\mathrm{e}^{\lambda(-b'-\mu')}}{\mathrm{e}^{\lambda(b'-\mu')} - \mathrm{e}^{\lambda(-b'-\mu')}} - \frac{1}{\lambda} = -\mu' + b'\frac{\mathrm{e}^{\lambda b'} + \mathrm{e}^{-\lambda b'}}{\mathrm{e}^{\lambda b'} - \mathrm{e}^{-\lambda b'}} - \frac{1}{\lambda}.$$

Therefore, the final formula is

$$\frac{F(\lambda|b',\mu')}{d\lambda} = \begin{cases} -\mu' & \lambda = 0 \\ -\mu' + b'\frac{\mathrm{e}^{\lambda b'} + \mathrm{e}^{-\lambda b'}}{\mathrm{e}^{\lambda b'} - \mathrm{e}^{-\lambda b'}} - \frac{1}{\lambda} & \lambda \neq 0 \end{cases}.$$

We determine λ^* by setting this expression equal to zero.

Most simply, if $\mu' = 0$, then $\lambda^* = 0$ minimizes the free energy, and the MED is a uniform distribution. This result is intuitive. Without an explicit mean constraint, the zeroth-order MED with a finite contrast bound is a uniform distribution with zero mean, and the requirement that the distribution have zero mean is already satisfied.

If $\mu' \neq 0$, we need to solve

$$0 = -\mu' + b' \frac{e^{\lambda^* b'} + e^{-\lambda^* b'}}{e^{\lambda^* b'} - e^{-\lambda^* b'}} - \frac{1}{\lambda^*}.$$

We rearrange this expression to find

$$\frac{\mu'}{b'} = \frac{e^{\lambda^* b'} + e^{-\lambda^* b'}}{e^{\lambda^* b'} - e^{-\lambda^* b'}} - \frac{1}{\lambda^* b'} = \coth(\lambda^* b') - \frac{1}{\lambda^* b'},$$

where $\coth(x)$ is the hyperbolic cotangent function. If we define $f(x) = \coth(x) - \frac{1}{x}$, then we can rewrite above equation as

$$\frac{\mu'}{b'} = f(\lambda^* b'),$$

and we get

$$\lambda^* = \frac{1}{b'} f^{-1}\left(\frac{\mu'}{b'}\right).$$

From the above equation, we observe that λ^* depends only on $\frac{\mu'}{b'}$ and $\frac{1}{b'}$. From **Appendix Figure 1A**, we see that λ^* monotonically increases with $\frac{\mu'}{b'}$. Intuitively, $\frac{\mu'}{b'}$ sets the relative scale of μ' in the contrast range and the degree of asymmetry of the distribution. For example, when μ' is zero, we know that λ^* is zero and the MED is perfectly symmetric. As μ' deviates from 0, λ^* deviates from zero and the distribution distorts asymmetrically from the uniform distribution to satisfy the non-zero mean. As the ratio of μ' and b' grows larger, λ^* has to deviate more from 0 and the distribution becomes increasingly asymmetric. When $\frac{\mu'}{b'}$ is small, λ^* is roughly linear in $\frac{\mu'}{b'}$. Also note that there is a simple proportionality between λ^* and $\frac{1}{b'}$. This dependence is easily understood by the requirement that $\lambda^* x$ is dimensionless.

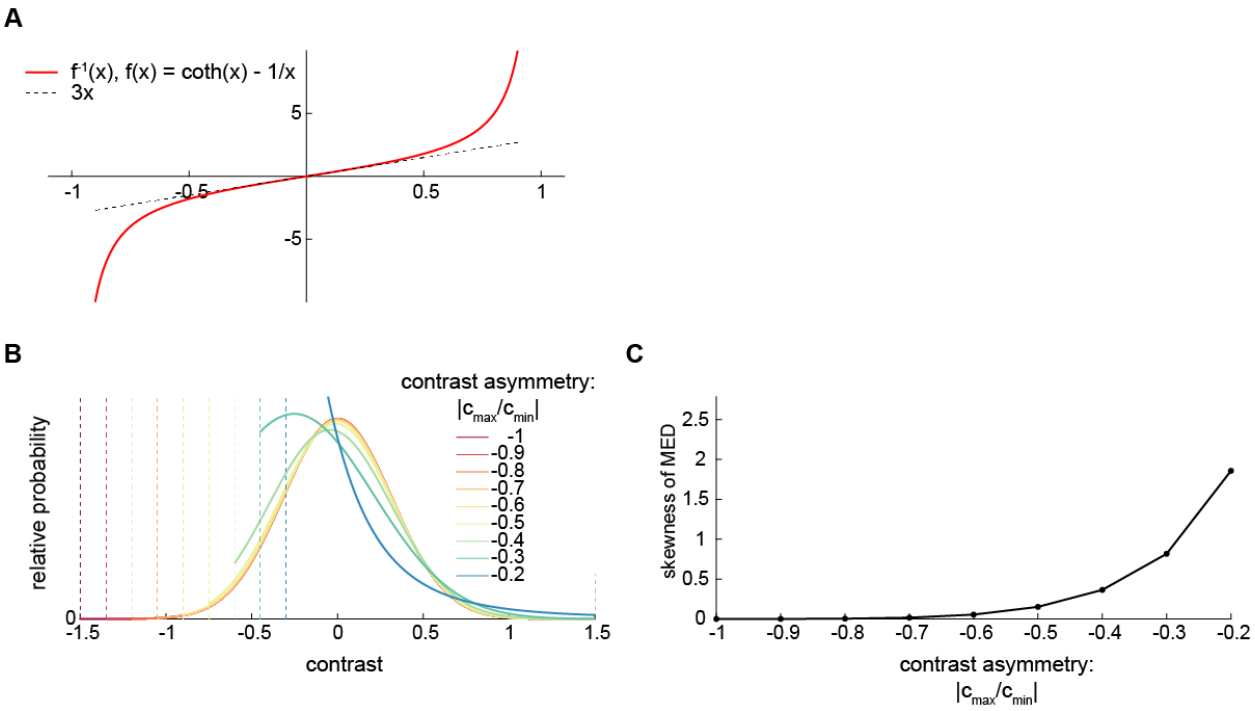
The mean-and-variance-constrained MED also depends on the finite contrast range. To simplify the discussion, we again set the zero point of x-axis such that the contrast range is symmetric around 0. Thus, our task is to find λ_1, λ_2 that minimize the free energy function,

$$\boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_2^*) = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} F(\lambda_1, \lambda_2 | \mu_1, \mu_2, b) = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \log \int_{-b}^{b} dx \exp(\lambda_1(x - \mu_1) + \lambda_2(x^2 - \mu_2)),$$

Where μ_1 and μ_2 are the shifted first-order and second-order moments imposed on the MED, and $[-b, b]$ is the contrast range. As was the case for the first-order MED, when $\mu_1 = 0$, the distribution is symmetric and $\lambda_1^* = 0$. Furthermore, as the contrast bound becomes large, the distribution approaches Gaussian distribution. However, when $\mu_1 \neq 0$, the distribution has to distort asymmetrically to achieve the non-zero mean, with large distortions occurring when the mean is within a few standard deviations of the boundary (**Appendix Figure 1B**). As a result, even though we did not require the distribution to have a specific skewness, the second-order MED can have non-zero skewness (**Appendix Figure 1C**).

These mathematical considerations lead to the design of our numerical methods. Operationally, for each individual image, we found its contrast range, $[c_{min}, c_{max}]$, and the largest contrast magnitude, $\delta c = \max(|c_{min}|, |c_{max}|)$. Then we constructed a symmetric contrast range around the constrained mean μ_1,

and computed the MED on $[\mu_1 - \delta c, \mu_1 + \delta c]$. This avoided induced skewness in the mean-and-variance-constrained MED. For consistency, we used the same symmetric contrast range for the mean-variance-and-skewness-constrained MED. Since the contrast bound implicitly influences the shape of the MED and different individual images have different contrast ranges, this method takes guidance from natural scenes and uses *heterogenous* contrast ranges that match the contrast range of each original natural scene. However, this heterogeneity introduces variability across images that could potentially impact the estimation accuracy we observed. Therefore, we also simulated secondary datasets where we used a common symmetric bound for all synthetic images. We chose this bound such that we could successfully solve the MED for the mean, variance, and skewness levels of most natural images.



Appendix Figure 1. The shape of MED depends on the contrast range

A. The shape of function $f^{(-1)}(x)$, where $f(x) = \coth(x) - 1/x$.
B. The mean-variance-constrained MED depends on the contrast range. The mean is constrained to be 0. The upper bound of the contrast range is fixed at 1.5, and the ratio of the lower-bound to upper bound changes from -1 to -0.2. As the contrast range becomes more asymmetric, the MED becomes more asymmetric.
C. When the contrast range is asymmetric, the mean-variance-constrained MED has induced non-zero skewness.

Acknowledgments

The authors thanks Ann Hermundstad and Emily Cooper for comments on the manuscript. JEF was supported by the Swartz Foundation and the Howard Hughes Medical Institute. DAC and this research were supported by NIH R01EY026555, NIH P30EY026878, NSF IOS1558103, a Searle Scholar Award, a Sloan Fellowship in Neuroscience, the Smith Family Foundation, and the E. Matilda Ziegler Foundation.

Citations

Adelson, E.H., and Bergen, J.R. (1985). Spatiotemporal energy models for the perception of motion. Josa a *2*, 284-299.

Arenz, A., Drews, M.S., Richter, F.G., Ammer, G., and Borst, A. (2017). The temporal tuning of the Drosophila motion detectors is determined by the dynamics of their input elements. Current Biology *27*, 929-944.

Badwan, B.A., Creamer, M.S., Zavatone-Veth, J.A., and Clark, D.A. (2019). Dynamic nonlinearities enable direction opponency in Drosophila elementary motion detectors. Nature neuroscience *22*, 1318.

Balasubramanian, V., and Sterling, P. (2009). Receptive fields and functional architecture in the retina. The Journal of physiology *587*, 2753-2767.

Balboa, R.M., and Grzywacz, N.M. (2003). Power spectra and distribution of contrasts of natural images from different habitats. Vision research *43*, 2527-2537.

Behnia, R., Clark, D.A., Carter, A.G., Clandinin, T.R., and Desplan, C. (2014). Processing properties of ON and OFF pathways for Drosophila motion detection. Nature *512*, 427.

Berger, A.L., Pietra, V.J.D., and Pietra, S.A.D. (1996). A maximum entropy approach to natural language processing. Computational linguistics *22*, 39-71.

Bialek, W. (2012). Biophysics: searching for principles (Princeton University Press).

Bialek, W., and van Steveninck, R.R. (2005). Features and dimensions: Motion estimation in fly vision. arXiv preprint q-bio/0505003.

Borst, A., and Egelhaaf, M. (1989). Principles of visual motion detection. Trends in neurosciences *12*.

Borst, A., and Helmstaedter, M. (2015). Common circuit design in fly and mammalian motion vision. Nature Neuroscience *18*, 1067.

Buchner, E. (1976). Elementary movement detectors in an insect visual system. Biological cybernetics *24*, 85-101.

Bülthoff, H., and Götz, K.G. (1979). Analogous motion illusion in man and fly. Nature *278*, 636.

Burkhardt, D.A., Fahey, P.K., and Sikora, M.A. (2006). Natural images and contrast encoding in bipolar cells in the retina of the land-and aquatic-phase tiger salamander. Visual neuroscience *23*, 35-47.

Card, G., and Dickinson, M.H. (2008). Visually mediated motor planning in the escape response of Drosophila. Current Biology *18*, 1300-1307.

Carey, J.R., Papadopoulos, N., Kouloussis, N., Katsoyannos, B., Müller, H.-G., Wang, J.-L., and Tseng, Y.-K. (2006). Age-specific and lifetime behavior patterns in Drosophila melanogaster and the Mediterranean fruit fly, Ceratitis capitata. Experimental gerontology *41*, 93-97.

Chander, D., and Chichilnisky, E. (2001). Adaptation to temporal contrast in primate and salamander retina. Journal of Neuroscience *21*, 9904-9916.

Chichilnisky, E., and Kalmar, R.S. (2002). Functional asymmetries in ON and OFF ganglion cells of primate retina. Journal of Neuroscience *22*, 2737-2747.

Clark, D.A., Bursztyn, L., Horowitz, M.A., Schnitzer, M.J., and Clandinin, T.R. (2011). Defining the computational structure of the motion detector in Drosophila. Neuron *70*, 1165-1177.

Clark, D.A., and Demb, J.B. (2016). Parallel computations in insect and mammalian visual motion processing. Current Biology *26*, R1062-R1072.

Clark, D.A., Fitzgerald, J.E., Ales, J.M., Gohl, D.M., Silies, M.A., Norcia, A.M., and Clandinin, T.R. (2014). Flies and humans share a motion estimation strategy that exploits natural scene statistics. Nature neuroscience *17*, 296.

Creamer, M.S., Mano, O., and Clark, D.A. (2018). Visual Control of Walking Speed in Drosophila. Neuron.

Creamer, M.S., Mano, O., Tanaka, R., and Clark, D.A. (2019). A flexible geometry for panoramic visual and optogenetic stimulation during behavior and physiology. Journal of neuroscience methods *323*, 48-55.

Dayan, P., Abbott, L.F., and Abbott, L. (2001). Theoretical neuroscience: computational and mathematical modeling of neural systems.

DeAngelis, B.D., Zavatone-Veth, J.A., and Clark, D.A. (2019). The manifold structure of limb coordination in walking Drosophila. eLife *8*, e46409.

Dickinson, M.H., Farley, C.T., Full, R.J., Koehl, M., Kram, R., and Lehman, S. (2000). How animals move: an integrative view. science *288*, 100-106.

Dror, R.O., O'carroll, D.C., and Laughlin, S.B. (2001). Accuracy of velocity estimation by Reichardt correlators. JOSA A *18*, 241-252.

Euler, T., Detwiler, P., and Denk, W. (2002). Direction-Selective Calcium Signals in Starburst Cell Dendrites. Investigative Ophthalmology & Visual Science *43*, 4763-4763.

Fisher, Y.E., Silies, M., and Clandinin, T.R. (2015). Orientation selectivity sharpens motion detection in Drosophila. Neuron *88*, 390-402.

Fitzgerald, J.E., and Clark, D.A. (2015). Nonlinear circuits for naturalistic visual motion estimation. Elife *4*, e09123.

Fitzgerald, J.E., Katsov, A.Y., Clandinin, T.R., and Schnitzer, M.J. (2011). Symmetries in stimulus statistics shape the form of visual motion estimators. Proceedings of the National Academy of Sciences *108*, 12909-12914.

Freifeld, L., Clark, D.A., Schnitzer, M.J., Horowitz, M.A., and Clandinin, T.R. (2013). GABAergic lateral interactions tune the early stages of visual processing in Drosophila. Neuron *78*, 1075-1089.

Geisler, W.S. (2008). Visual perception and the statistical properties of natural scenes. Annu Rev Psychol *59*, 167-192.

Gjorgjieva, J., Clopath, C., Audet, J., and Pfister, J.-P. (2011). A triplet spike-timing–dependent plasticity model generalizes the Bienenstock–Cooper–Munro rule to higher-order spatiotemporal correlations. Proceedings of the National Academy of Sciences *108*, 19383-19388.

Götz, K.G., and Wenking, H. (1973). Visual control of locomotion in the walking fruitflyDrosophila. Journal of comparative physiology *85*, 235-266.

Gruntman, E., Romani, S., and Reiser, M.B. (2018). Simple integration of fast excitation and offset, delayed inhibition computes directional selectivity in Drosophila. Nature neuroscience *21*, 250.

Haag, J., Arenz, A., Serbe, E., Gabbiani, F., and Borst, A. (2016). Complementary mechanisms create direction selectivity in the fly. Elife *5*, e17421.

Hassenstein, B., and Reichardt, W. (1956). Systemtheoretische analyse der zeit-, reihenfolgen- und vorzeichenauswertung bei der bewegungsperzeption des rüsselkäfers chlorophanus. Zeitschrift für Naturforschung B *11*, 513-524.

Hermundstad, A.M., Briguglio, J.J., Conte, M.M., Victor, J.D., Balasubramanian, V., and Tkačik, G. (2014). Variance predicts salience in central sensory processing. Elife *3*, e03722.

Hu, Q., and Victor, J.D. (2010). A set of high-order spatiotemporal stimuli that elicit motion and reverse-phi percepts. Journal of vision *10*, 9-9.

Hu, Y., Trousdale, J., Josić, K., and Shea-Brown, E. (2014). Local paths to global coherence: cutting networks down to size. Physical Review E *89*, 032802.

Jaynes, E.T. (1957). Information theory and statistical mechanics. Physical review *106*, 620.

Jin, J., Wang, Y., Lashgari, R., Swadlow, H.A., and Alonso, J.-M. (2011). Faster thalamocortical processing for dark than light visual targets. Journal of Neuroscience *31*, 17471-17479.

Julesz, B. (1962). Visual pattern discrimination. IRE transactions on Information Theory *8*, 84-92.

Julész, B., Gilbert, E., Shepp, L., and Frisch, H. (1973). Inability of humans to discriminate between visual textures that agree in second-order statistics—revisited. Perception *2*, 391-405.

Julész, B., Gilbert, E.N., and Victor, J.D. (1978). Visual discrimination of textures with identical third-order statistics. Biological Cybernetics *31*, 137-140.

Katsov, A.Y., and Clandinin, T.R. (2008). Motion processing streams in Drosophila are behaviorally specialized. Neuron *59*, 322-335.

Kremkow, J., Jin, J., Komban, S.J., Wang, Y., Lashgari, R., Li, X., Jansen, M., Zaidi, Q., and Alonso, J.-M. (2014). Neuronal nonlinearity explains greater visual spatial resolution for darks than lights. Proceedings of the National Academy of Sciences *111*, 3170-3175.

Leong, J.C.S., Esch, J.J., Poole, B., Ganguli, S., and Clandinin, T.R. (2016). Direction selectivity in Drosophila emerges from preferred-direction enhancement and null-direction suppression. Journal of Neuroscience *36*, 8078-8092.

Leonhardt, A., Ammer, G., Meier, M., Serbe, E., Bahl, A., and Borst, A. (2016). Asymmetry of Drosophila ON and OFF motion detectors enhances real-world velocity estimation. Nature neuroscience *19*, 706.

Maheswaranathan, N., Kastner, D.B., Baccus, S.A., and Ganguli, S. (2018). Inferring hidden structure in multilayered neural circuits. PLoS computational biology *14*, e1006291.

Mano, O., and Clark, D.A. (2017). Graphics processing unit-accelerated code for computing second-order wiener kernels and spike-triggered covariance. PloS one *12*, e0169842.

Marmarelis, P.Z., and McCann, G.D. (1973). Development and application of white-noise modeling techniques for studies of insect visual nervous system. Kybernetik *12*, 74-89.

Marmarelis, P.Z., and Naka, K.-I. (1972). White-noise analysis of a neuron chain: an application of the Wiener theory. Science *175*, 1276-1278.

Marmarelis, V.Z. (2004). Nonlinear dynamic modeling of physiological systems, Vol 10 (John Wiley & Sons).

Marr, D., and Poggio, T. (1976). From understanding computation to understanding neural circuitry.

Mazade, R., Jin, J., Pons, C., and Alonso, J.-M. (2019). Functional Specialization of ON and OFF Cortical Pathways for Global-Slow and Local-Fast Vision. Cell reports *27*, 2881-2894. e2885.

McIntosh, L., Maheswaranathan, N., Nayebi, A., Ganguli, S., and Baccus, S. (2016). Deep learning models of the retinal response to natural scenes. Paper presented at: Advances in neural information processing systems.

Meyer, H.G., Schwegmann, A., Lindemann, J.P., and Egelhaaf, M. (2014). Panoramic high dynamic range images in diverse environments.

Movshon, J.A., Adelson, E.H., Gizzi, M., and Newsome, W.T. (1985). The analysis of moving visual patterns. In Pattern recognition mechanisms, C.C.a.R.G.a.C. Gross, ed. (Vatican Press), pp. 117-151.

Nitzany, E., Menda, G., Shamble, P.S., Golden, J.R., Hu, Q., Hoy, R.R., and Victor, J. (2017). Neural computations combine low-and high-order motion cues similarly, in dragonfly and monkey. bioRxiv, 240101.

Nitzany, E.I., Loe, M.E., Palmer, S.E., and Victor, J.D. (2016). Perceptual interaction of local motion signals. Journal of vision *16*, 22-22.

Nitzany, E.I., and Victor, J.D. (2014). The statistics of local motion signals in naturalistic movies. Journal of vision *14*, 10-10.

Olshausen, B.A., and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature *381*, 607.

Pandarinath, C., Victor, J.D., and Nirenberg, S. (2010). Symmetry breakdown in the ON and OFF pathways of the retina at night: functional implications. Journal of Neuroscience *30*, 10006-10014.

Pfister, J.-P., and Gerstner, W. (2006). Triplets of spikes in a model of spike timing-dependent plasticity. Journal of Neuroscience *26*, 9673-9682.

Poggio, T., and Reichardt, W. (1973). Considerations on models of movement detection. Kybernetik *13*, 223-227.

Portilla, J., and Simoncelli, E.P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. International journal of computer vision *40*, 49-70.

Potters, M., and Bialek, W. (1994). Statistical mechanics and visual signal processing. Journal de Physique I *4*, 1755-1775.

Ratliff, C.P., Borghuis, B.G., Kao, Y.-H., Sterling, P., and Balasubramanian, V. (2010). Retina is structured to process an excess of darkness in natural scenes. Proceedings of the National Academy of Sciences *107*, 17368-17373.

Ravi, S., Ahn, D., Greschner, M., Chichilnisky, E., and Field, G.D. (2018). Pathway-specific asymmetries between ON and OFF visual signals. bioRxiv, 384891.

Richards, W.A. (1982). Lightness scale from image intensity distributions. Applied Optics *21*, 2569-2582.

Rust, N.C., Schwartz, O., Movshon, J.A., and Simoncelli, E.P. (2005). Spatiotemporal elements of macaque v1 receptive fields. Neuron *46*, 945-956.

Sagdullaev, B.T., and Mccall, M.A. (2005). Stimulus size and intensity alter fundamental receptive-field properties of mouse retinal ganglion cells in vivo. Visual neuroscience *22*, 649-659.

Salazar-Gatzimas, E., Agrochao, M., Fitzgerald, J.E., and Clark, D.A. (2018). The Neuronal Basis of an Illusory Motion Percept Is Explained by Decorrelation of Parallel Motion Pathways. Current Biology.

Salazar-Gatzimas, E., Chen, J., Creamer, M.S., Mano, O., Mandel, H.B., Matulis, C.A., Pottackal, J., and Clark, D.A. (2016). Direct measurement of correlation responses in Drosophila elementary motion detectors reveals fast timescale tuning. Neuron *92*, 227-239.

Salisbury, J.M., and Palmer, S.E. (2016). Optimal prediction in the retina and natural motion statistics. Journal of Statistical Physics *162*, 1309-1323.

Sandler, R.A., and Marmarelis, V.Z. (2015). Understanding spike-triggered covariance using Wiener theory for receptive field identification. Journal of vision *15*, 16-16.

Sanes, J.R., and Zipursky, S.L. (2010). Design principles of insect and vertebrate visual systems. Neuron *66*, 15-36.

Schetzen, M. (1980). The Volterra and Wiener theories of nonlinear systems.

Schneidman, E., Berry II, M.J., Segev, R., and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. Nature *440*, 1007.

Schwartz, O., Pillow, J.W., Rust, N.C., and Simoncelli, E.P. (2006). Spike-triggered neural characterization. Journal of vision *6*, 13-13.

Simoncelli, E.P., and Heeger, D.J. (1998). A model of neuronal responses in visual area MT. Vision research *38*, 743-761.

Simoncelli, E.P., and Olshausen, B.A. (2001). Natural image statistics and neural representation. Annual review of neuroscience *24*, 1193-1216.

Sinha, S.R., Bialek, W., and van Steveninck, R.R. (2018). Optimal local estimates of visual motion in a natural environment. arXiv preprint arXiv:181211878.

Sjöström, P.J., Turrigiano, G.G., and Nelson, S.B. (2001). Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. Neuron *32*, 1149-1164.

Song, S., Sjöström, P.J., Reigl, M., Nelson, S., and Chklovskii, D.B. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. PLoS biology *3*, e68.

Srinivasan, M.V., Laughlin, S.B., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. Proc R Soc Lond B *216*, 427-459.

Stavenga, D. (2003). Angular and spectral sensitivity of fly photoreceptors. II. Dependence on facet lens F-number and rhabdomere type in Drosophila. Journal of Comparative Physiology A *189*, 189-202.

Takemura, S.-y., Nern, A., Chklovskii, D.B., Scheffer, L.K., Rubin, G.M., and Meinertzhagen, I.A. (2017). The comprehensive connectome of a neural substrate for 'ON'motion detection in Drosophila. Elife *6*, e24394.

Tkačik, G., Prentice, J.S., Victor, J.D., and Balasubramanian, V. (2010). Local statistics in natural scenes predict the saliency of synthetic textures. Proceedings of the National Academy of Sciences *107*, 18149-18154.

Trousdale, J., Hu, Y., Shea-Brown, E., and Josić, K. (2012). Impact of network structure and cellular response on spike time correlations. PLoS computational biology *8*, e1002408.

Victor, J.D., and Conte, M.M. (2012). Local image statistics: maximum-entropy constructions and perceptual salience. JOSA A *29*, 1313-1345.

Wienecke, C.F., Leong, J.C., and Clandinin, T.R. (2018). Linear summation underlies direction selectivity in Drosophila. Neuron *99*, 680-688. e684.

Wiener, N. (1966). Nonlinear problems in random theory. Nonlinear Problems in Random Theory, by Norbert Wiener, pp 142 ISBN 0-262-73012-X Cambridge, Massachusetts, USA: The MIT Press, August 1966(Paper), 142.

Yang, H.H., St-Pierre, F., Sun, X., Ding, X., Lin, M.Z., and Clandinin, T.R. (2016). Subcellular imaging of voltage and calcium signals reveals neural processing in vivo. Cell *166*, 245-257.

Zacarias, R., Namiki, S., Card, G.M., Vasconcelos, M.L., and Moita, M.A. (2018). Speed dependent descending control of freezing behavior in Drosophila melanogaster. Nature communications *9*, 3697.

Zaghloul, K.A., Boahen, K., and Demb, J.B. (2003). Different circuits for ON and OFF retinal ganglion cells cause different contrast sensitivities. Journal of Neuroscience *23*, 2645-2654.

Zylberberg, J. (2017). Untuned But Not Irrelevant: The Role of Untuned Neurons In Sensory Information Coding. BioRxiv, 134379.