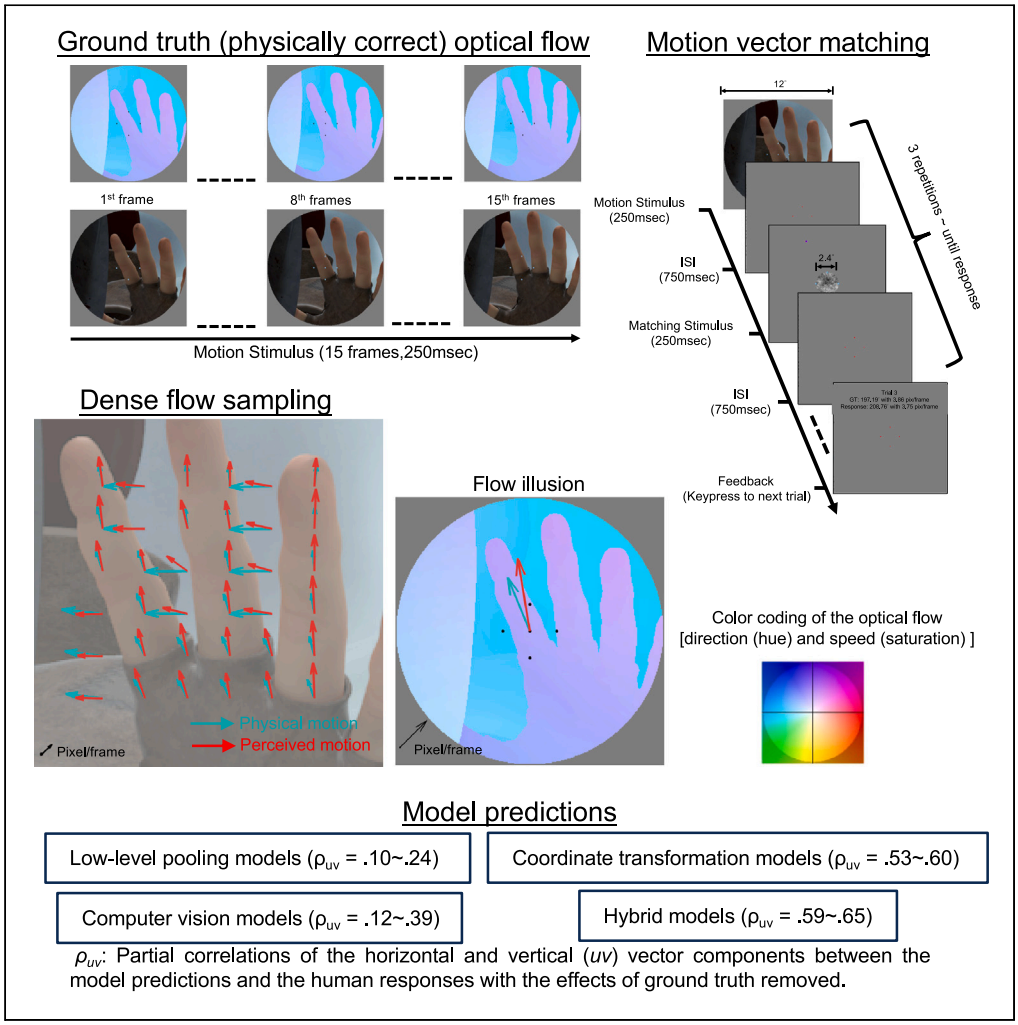
iScience

CellPress
OPEN ACCESS

Article

Psychophysical measurement of perceived motion flow of naturalistic scenes



Yung-Hao Yang, Taiki Fukiage, Zitang Sun, Shin'ya Nishida

shinyanishida@mac.com

Highlights

Human-perceived motion flow of naturalistic movies was psychophysically estimated

The reported flow matched the physical ground truth, with a few notable exceptions

Computational analysis revealed multiple mechanisms causing the flow illusions

They range from low-level spatial pooling to high-level coordinate transformation

Yang et al., iScience *26*, 108307
December 15, 2023
https://doi.org/10.1016/j.isci.2023.108307

iScience

CellPress
OPEN ACCESS

Article

Psychophysical measurement of perceived motion flow of naturalistic scenes

Yung-Hao Yang,[1] Taiki Fukiage,[2] Zitang Sun,[1] and Shin'ya Nishida[1,2,3,*]

SUMMARY

The neural and computational mechanisms underlying visual motion perception have been extensively investigated over several decades, but little attempt has been made to measure and analyze, how human observers perceive the map of motion vectors, or optical flow, in complex naturalistic scenes. Here, we developed a psychophysical method to assess human-perceived motion flows using local vector matching and a flash probe. The estimated perceived flow for naturalistic movies agreed with the physically correct flow (ground truth) at many points, but also showed consistent deviations from the ground truth (flow illusions) at other points. Comparisons with the predictions of various computational models, including cutting-edge computer vision algorithms and coordinate transformation models, indicated that some flow illusions are attributable to lower-level factors such as spatiotemporal pooling and signal loss, while others reflect higher-level computations, including vector decomposition. Our study demonstrates a promising data-driven psychophysical paradigm for an advanced understanding of visual motion perception.

INTRODUCTION

Visual motion perception is one of the most extensively investigated perceptual functions (see[1–4] for reviews). According to the currently prevailing view, the first stage of visual motion processing is the extraction of local motion signals by direction-selective sensors,[5] followed by mutual integration and inhibition of such signals to solve the aperture problem to estimate a spatiotemporal pattern of image motion vectors.[6] The higher-level visual motion processing, including vector analysis,[7] uses this motion vector pattern to recognize object motion in the scene, to control eye and body movements, to self-navigate in the field via optical flow, and to perceive biological motion. The cortical mechanisms underlying these computations have also been extensively studied.[8] Progress in visual motion research has been aided by the use of several artificial stimuli that selectively tap each processing stage; the stimuli include drifting sine-wave gratings,[9] plaids,[10] random-dot kinematograms,[11–13] point-light walkers,[7] and global dot flow patterns.[14]

Recent technical advances have rendered it possible to access large-scale data on neural responses, and to create models that predict the responses to complex stimuli. Interest in visual motion research has thus shifted to how the visual system processes complex visual motion information in dynamic natural environments. For instance, Nishimoto & Gallant[15] measured the cortical functional magnetic resonance imaging (fMRI) responses to movie clips of natural scenes and tested the validity of the models having been proposed to explain human motion perception. Matthis et al.[16] measured visual motions projected on the retina during natural locomotion when studying the role of optical flow in action control in real-world environments.

However, to the best of our knowledge, few attempts have been made to psychophysically measure the pattern of image motion vectors, or optical flow, that human observers really perceive in complex natural or naturalistic scenes. (Note that this paper uses the term "optical flow" to refer to the pattern of image motion vectors in general, including but not limited to the pattern of motion vectors caused by the relative motion between an observer and a scene.) Such perceived optical flow, if measurable, would advance our understanding of visual motion perception. It would be possible to compare human perceptions directly with the physically correct motion vectors, neurophysiologically measured brain responses, and the predictions of the motion models developed in the field of vision science and machine vision.

Psychophysical estimation of a perceived optical flow is challenging because this requires measurements of perceived motion vectors (speed and direction) at many spatiotemporal positions in a dynamic scene. Here, we present a novel measurement procedure, inspired by the gauge-probe task of surface shape estimation.[17] Observers are asked to report perceived vectors by adjusting the speed and direction of a matching noise stimulus. To indicate the target spatiotemporal position during each trial, we superimpose a brief dot probe on the movie stimulus. By changing the target position in a grid-like fashion, we derive a map of the perceived optical flow.

[1]Cognitive Informatics Laboratory, Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
[2]Human Information Science Laboratory, NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, 3-1, Morinosato-Wakamiya, Atsugi, Kanagawa 243-0198, Japan
[3]Lead contact
*Correspondence: shinyanishida@mac.com
https://doi.org/10.1016/j.isci.2023.108307

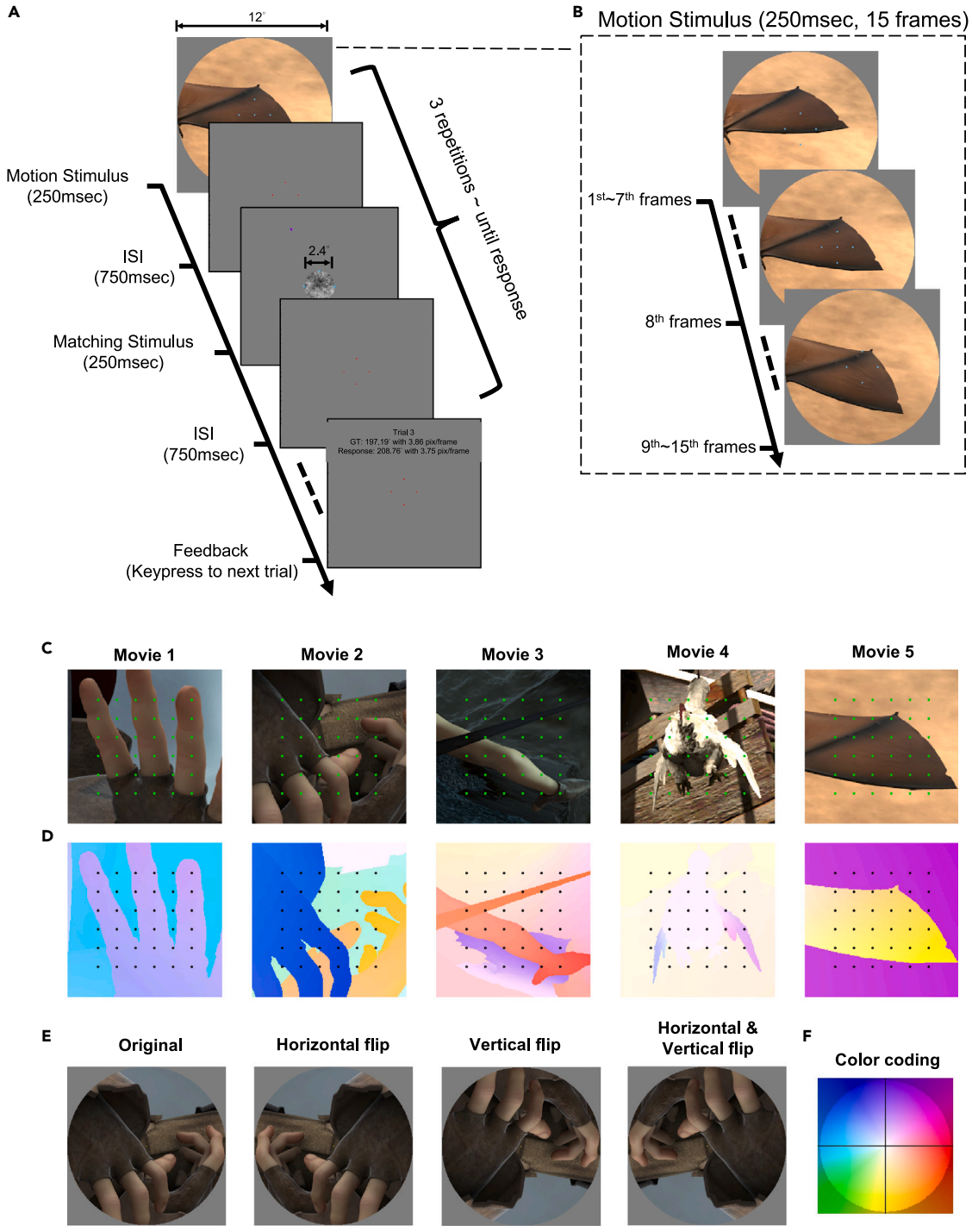


Figure 1. Experimental procedures and visual stimuli

(A) In each trial, motion and matching stimuli were alternatively and repeatedly presented. Each repetition featured a motion stimulus (250 ms), an inter-stimulus interval (ISI, 750 ms), a matching stimulus (250 ms), and an ISI (750 ms).

(B) A probe dot was flashed in the middle (the 8th frame) of the motion stimulus presentation (250 ms, 15 frames) and observers were required to report the motion vector at the location/time of the flash by adjusting the speed and direction of the matching stimulus. After at least three repetitions, the observers could terminate the trial when satisfied with their settings. A feedback display was then presented.

(C) The 36 probed locations (1° spacing, green dots) of each movie clip.

Figure 1. *Continued*

(D) Ground truth motion vectors of each movie using a color-coding scheme[43] where hue indicates the motion direction and saturation the speed (normalized by the maximum speed of each five clips) as shown in (F).

(E) Each movie clip was presented through a circular aperture (12° in diameter) in one of four flip modes: Original orientation, horizontal flip, vertical flip, and horizontal and vertical flip. See also Videos S1–S6.

Objective evaluation of human performance requires a comparison with the physically correct optical flow, or the ground truth flow. It is easy to know the ground truth optical flow for a synthesized movie given that the physical model generating the movie is known. On the other hand, knowing the ground truth flow is difficult for movies shooting natural scenes with standard cameras. We therefore used a synthesized, naturalistic movie dataset, the MPI Sintel Flow Dataset,[18] that has been widely used in the field of computer vision for training and evaluation of optical flow models.

Having applied our new measurement method to the Sintel Dataset, we found that human observers could reproduce the speed and direction of a motion vector at a specified spatiotemporal location close to the ground truth at many locations. We also found, however, that the estimated perceived flow at some locations deviated systematically from the ground truth in stimulus-dependent manners, which we termed "flow illusions." We explored how much these illusions were explained by spatiotemporal pooling of local motion signals, by visual motion models motivated by biological computations, by computer vision models engineered for optical flow estimation, and by models simulating coordinate transformation in high-level human processing. We found that a small portion of the "illusions" were explained by retinotopic optical flow estimation mechanisms, but others reflected high-level computations, including vector decomposition. Our study demonstrates the strengths and limitations of existing models that predict visual motion flow perceptions when viewing naturalistic scenes and that estimation of human-perceived optical flow maps of a variety of natural scenes will advance our understanding of human visual motion processing.

RESULTS

In the main experiment, we measured the human-perceived motion vectors of the optical flows in five movie clips selected from a high-frame-rate version of the MPI Sintel Flow Dataset.[18,19] The Dataset includes the ground truth optical flow data (Figure 1D) that can be directly compared to human responses.

The human-perceived optical flow map was measured by a local vector-matching method (Figure 1 and Video S1). In each trial, pairs of target motion clips and matching noise stimuli were repeatedly presented. The observers were asked to adjust the direction and speed of the matching stimulus to reproduce the local motion vector in the target movie clip at the location indicated by a flashed probe (Figures 1A and 1B). The probe location was selected from a 6 × 6 grid placed on the region of interest (Figures 1C and 1D). Each movie clip was presented in one of four flip modes (Figure 1E); this allowed us to decompose the human response errors into two components. One component of errors is defined on the display coordinates and determined by the relationship between the observer and the display. The other component is defined on the image coordinates, the directions of which flip as the image flips. Our principal interest was the latter type of human error.

Accuracy evaluation

We evaluated the accuracy of our novel method from two perspectives. One is vector-matching accuracy, which refers to how accurately the observers can reproduce the perceived vector. One might think that this accuracy might be low since we used a "time-saving" adjustment method, instead of more strict psychophysical procedures. To evaluate the basic vector-matching accuracy, we analyzed the human responses to a random-dot kinematogram (Figure 2A, where the magnitudes of the horizontal and vertical (uv) vector components were analyzed; see Figure S1 for analyses of the direction and speed components, separately). The results showed a good, though not perfect, agreement between the reported motion and the ground truth motion (R^2 = 0.901; Equation 6). Since the target and matching stimuli were similar noise patterns in this condition, the task was straightforward. The results indicate that our procedure can provide the estimation of the human-perceived vector with this level of accuracy under an optimal condition. On the other hand, the vector-matching accuracy for Sintel Dataset was much worse (Figure 2B, R^2 = 0.643, See also Figure S2 for direction and speed components). The performance reduction could be ascribed to the stimulus-dependent errors we will consider in the following section, but one might suspect this was because the flash-probing accuracy of our task was low.

The flash-probing accuracy refers to how accurately in space and time the observers can localize local vectors specified by the flashed probe. It is known that various phenomena including flash lag[20] and flash drag[21] can cause (apparent) spatiotemporal misalignments between continuously moving and flashing patterns. It might be therefore hard for human observers to accurately localize the target vector indicated by the flash probe. Random-dot kinematogram cannot be used for evaluation of this aspect of human accuracy since it is spatiotemporally uniform. We analyzed the Euclidean distance of the endpoints between the perceived vectors and the ground truth vectors (which we call endpoint error; see Equation 7) for the MPI Sintel movie clips. If the human observers reported the perceived vector at the probed point, the endpoint error should be smallest when we compare the human response with the ground truth at the probed spatial location and in the probed frame, rather than when we compare the human response with the ground truth vectors in its neighbors. The results indicated that the minimum endpoint error best agreed with the ground truth local motions at spatiotemporal locations very close to the probe (Figure 2C, see also Figures S3; and S4 for spatiotemporal distributions of the endpoint errors). This indicates that the flash-probing accuracy of our procedure is high enough to measure the perceived vector at the probed point (at least for the temporally smooth motion clips that we used).

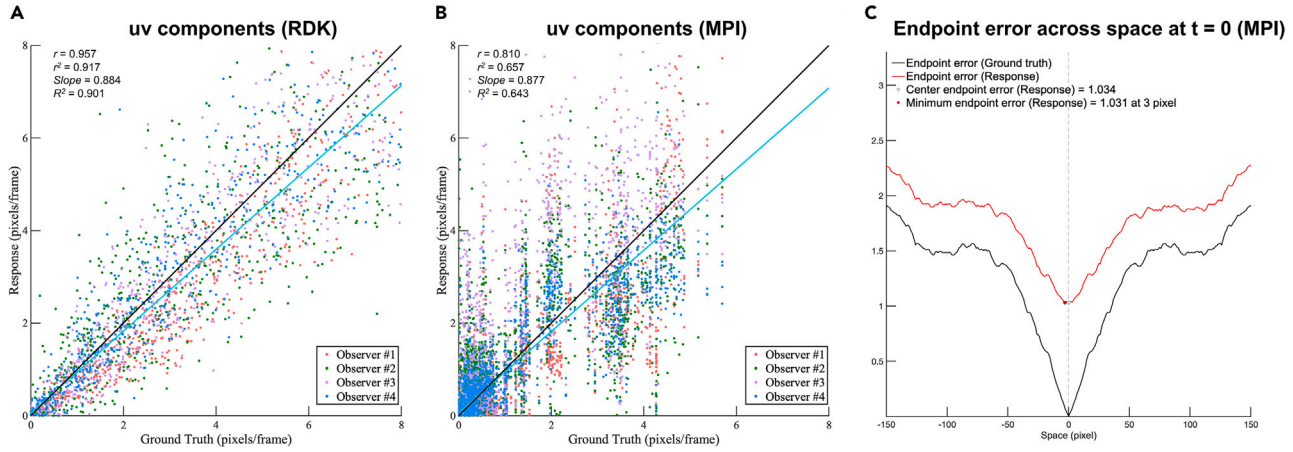


Figure 2. Evaluation of two aspects of accuracy

(A) Vector-matching accuracy is shown by a scatterplot of responses versus the ground truths for the horizontal and vertical (*uv*) vector components based on the results of a preliminary experiment using a random-dot kinematogram. Each point indicates a response of one observer in one trial. Different colors refer to different observers. *r*, r^2: Pearson correlations and the squares; slope: the slope of the linear regression line; R^2: the coefficient of determination with respect to the ideal response (i.e., the ground truth). The index values were computed based on the signed *uv* values; the Figure plots the absolute (unsigned) *uv* values for visualization. See the separate plots of the direction and speed components in Figure S1 and Data S1.

(B) A similar scatterplot of the results of the main experiment using the MPI Sintel dataset. Vector-matching accuracy is reduced by flow illusions. See also Figure S2 and Data S2.

(C) Flash-probing accuracy as shown by the spatiotemporal distribution of the endpoint differences in the display coordinates of the MPI Sintel Flow Dataset. The red line indicates the endpoint differences between the human response to a probe and the physical optical flows at surrounding locations. The observer responses were closest to the ground truths at the probed locations, indicating that observers reported motions at probed locations rather accurately. The black line indicates the within-stimulus similarity, thus the endpoint differences between the ground truths at the probed location and the physical optical flows at surrounding locations. See also Figures S1–S4.

Human perceived flow: An overview

Next, consider the pattern of human perceived flow. Figure 3 shows the response flow map for each Sintel movie clip. We averaged the response vectors over the four flip conditions after unflipping them into the original image coordinates. This was for revealing the response biases associated with the stimulus pattern while controlling for the biases associated with the relationship between the observer and the display. For instance, even if an observer had a bias to report stronger rightward motions than leftward motions, it was canceled by our procedure. The results of the four observers were averaged. The pattern of results was generally similar across the observers (see Figure S5; and Table S1).

In Figure 3, the response flow vector (red arrow) agrees with the ground truth vector (green arrow) at certain probed locations. The endpoint error averaged over all probe points is 0.928 pixels or 1.11 arcmin (see the first row of Table 1), suggesting that the observers reliably reproduced certain flow patterns in naturalistic scenes. However, the reported vector deviated from the ground truth vectors at many other locations in various ways. For example, large errors are found in the gaps between fingers in Video S1, in the hand on the right in Video S2, in the dark background surrounding the arm in Video S3, in the body of the rooster in Video S4, and in the wing of the bat in Video S5. These deviations explain why the agreement between the reported and ground truth vectors was less in the main Sintel experiment than in the preliminary random-dot kinematogram experiment (compare Figures 2A and 2B).

How to compare human perceived flow and model predictions?

The patterns of systematic deviation between the human vectors and ground truth vectors, which we term "flow illusions", would be expected to provide valuable information on human visual motion processing. We used the term "illusion" instead of "errors" simply because we believe "flow illusions" must be functionally meaningful as are many other illusions. Some components of flow illusions must be classifiable into the known categories of motion illusion, but it is not readily obvious only from visual inspection of the pattern.

To seek the mechanisms underlying the flow illusions, we compared the human response to the predictions of visual processing models, including models that approximate signal pooling by lower-level visual processing, computer vision models developed for optical flow estimation, coordinate transformation models that approximate biologically plausible higher-level visual processing, and hybrid models that combine both low-level and high-level visual processing (See model descriptions in the STAR methods for more details). The model type was not uniform. Some use ground truth information, and others are image-computable. Some include parameter fitting, but others do not. This was because the main purpose of comparing these models was not to find the best model to explain the human data, but to reveal various kinds of mechanisms contributing to the flow illusions.

As shown in Table 1, to quantitatively evaluate the performance of each model, we used predicted vectors to compute: (1) the average endpoint error by reference to the ground truths averaged over all 180 probed locations; (2) the average endpoint error by reference to

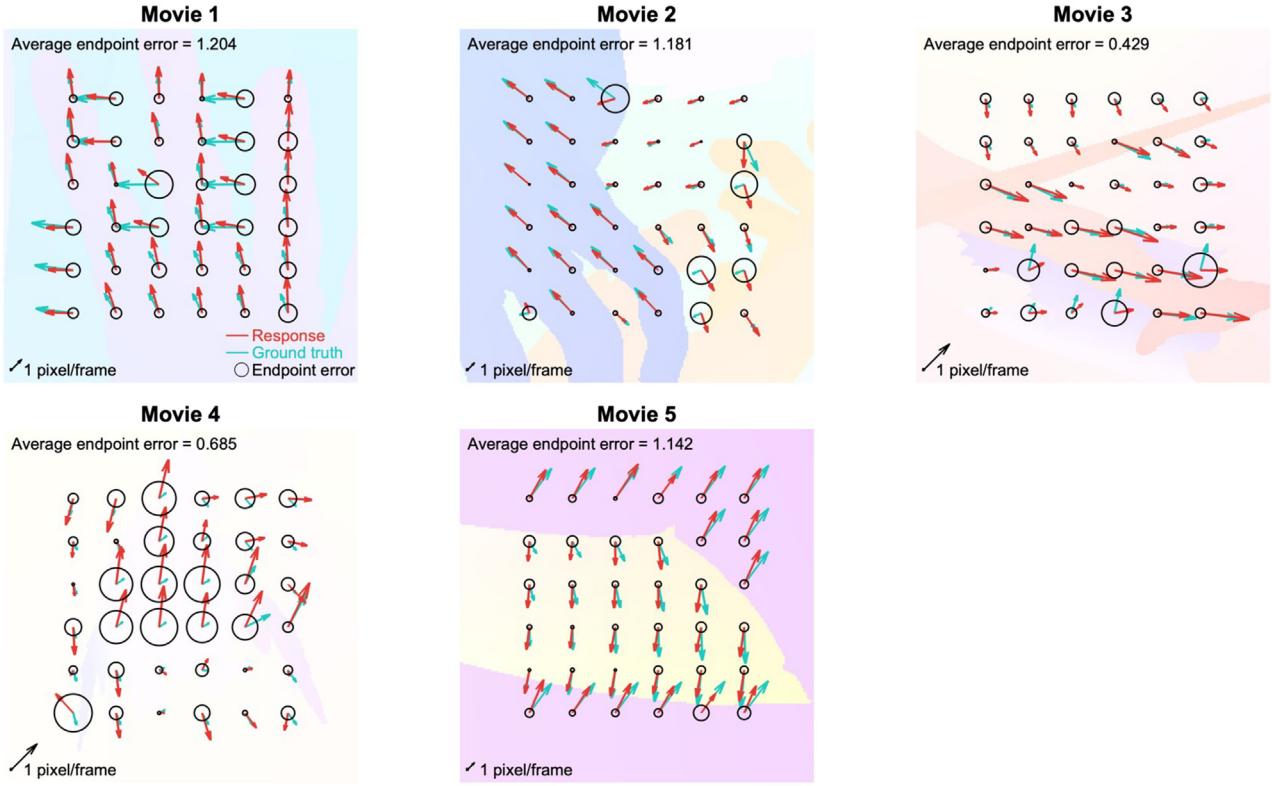


Figure 3. The perceived vector map for the MPI Sintel Flow Dataset

The response flow map using the image coordinates of the MPI Sintel Flow Dataset. At each location, the green arrow denotes the ground truth, the red arrow the human response, and the diameter of the black circle the endpoint error. In each panel, the average endpoint error of the 36 movie locations is shown in the top left corner. As the vector length was normalized within each movie, the spatial scale is shown by an arrow (of length 1 pixel/frame) in the bottom left corner. See also Figure S5, Table S1 and Data S3.

the human responses; (3) the partial correlations (ρ) between the human-perceived horizontal and vertical (uv) vector components with the effects of ground truth removed; and, (4) the average response consistency index (RCI, see Equations 10.1, 10.2, 10.3) over the probed locations. Except for the computer vision models, we estimated the free model parameters that best explained the response data (see Table S2 for the best-fit parameters for each model). To ensure fair comparisons across models with different numbers of free parameters, we also computed index values via 2-fold cross-validation and used (principally) these values for model evaluation.

Of the indices shown in Table 1, two are the most informative in terms of the extent of overall agreement between the model predictions and human responses. One is the partial correlation of uv components between the model predictions and human responses (ρ_{uv}). As the Pearson correlation between the model predictions and human responses (r_{uv}) could be associated with the common variable, the ground truth, ρ_{uv} evaluates the correlation between the two while excluding the effect of the ground truth. The reason to use the uv components was to evaluate the direction and speed components together (See Table S3 for the Pearson correlations and partial correlations between model predictions and human responses/ground truth for direction and speed).

The other informative index is the spatially averaged RCI. The RCI is an index of consistency between model predictions and the human responses, and takes a value between −1 and +1. The RCI becomes positive and approaches +1 as the prediction becomes closer to the human response than to the ground truth, approaches zero when the model prediction becomes closer to the ground truth, and becomes negative when the model predicts an error in the opposite direction (see the STAR methods). Although the absolute value of this index was small due to the multiplication of the three terms, we found a strong positive correlation between the average RCI and the partial correlation (Figure 4). An advantage of the RCI compared to the partial correlation is that the RCI can be separately computed for each location, as shown in Figures 5 and S11–S26. In addition, our analysis suggests that the average RCI is less affected by the small number of outliers than the partial correlation.

Figure 4 shows the distributions of the partial correlation in uv components (ρ_{uv}) and the spatially averaged RCI, computed using the bootstrapping method (for computer vision models with no free parameter) and via 2-fold cross-validation (for the other models with free parameters). See Figure S6 for the distribution of each index. For statistical testing of the differences in index values, we checked whether zero was included in the 95% confidence intervals of the differences between each model pair computed by the bootstrapping method (for computer models, Figures S7 and S9) or the cross-validation method (other models, Figures S8 and S10).

Table 1. Predictive performances of the models

	Models	*Average endpoint error (Ground truth)*	*Average endpoint error (Response)*	*Partial correlation ρ_{uv}*	*Average RCI*
	Ground truth	0.000	0.928	NaN	NaN
Low-level pooling models	Spatial Pooling	0.342 (**0.343**)	0.868 (**0.876**)	0.285 (**0.239**)	0.016 (**0.015**)
	Temporal Pooling	0.315 (**0.324**)	0.863 (**0.890**)	0.214 (**0.101**)	0.011 (**0.010**)
	Spatiotemporal Pooling	0.347 (**0.357**)	0.867 (**0.890**)	0.310 (**0.214**)	0.016 (**0.015**)
Computer vision models	Farneback	1.963	2.024	0.267	0.039
	FFV1MT	1.282	1.479	0.309	0.043
	FlowNet 2.0	0.472	0.935	0.386	0.034
	StaRFlow	1.150	1.415	0.117	0.008
	RAFT	0.261	0.889	0.344	0.026
Coordinate transformation models	Translation (per-movie)	0.622 (**0.625**)	0.719 (**0.742**)	0.564 (**0.534**)	0.075 (**0.072**)
	Translation-Rotation-Scaling (per-movie)	0.692 (**0.714**)	0.602 (**0.651**)	0.652 (**0.589**)	0.098 (**0.093**)
	Translation (per-object)	0.710 (**0.729**)	0.599 (**0.658**)	0.656 (**0.573**)	0.106 (**0.098**)
	Translation-Rotation-Scaling (per-object)	0.801 (**0.875**)	0.468 (**0.619**)	0.781 (**0.597**)	0.136 (**0.121**)
Hybrid models	FlowNet 2.0 & Translation (per-movie)	0.777 (**0.789**)	0.752 (**0.776**)	0.603 (**0.587**)	0.086 (**0.083**)
	FlowNet 2.0 & Translation-Rotation-Scaling (per-movie)	0.870 (**0.892**)	0.625 (**0.675**)	0.684 (**0.648**)	0.119 (**0.114**)
	FlowNet 2.0 & Translation (per-object)	0.825 (**0.868**)	0.594 (**0.664**)	0.689 (**0.634**)	0.121 (**0.113**)
	FlowNet 2.0 & Translation-Rotation-Scaling (per-object)	0.808 (**0.873**)	0.455 (**0.598**)	0.804 (**0.653**)	0.138 (**0.123**)

Endpoint error (Ground truth): endpoint error between the model prediction and the ground truth. endpoint error (Response): endpoint error between the model prediction and the human response. ρ_{uv}: Partial correlations of the *uv* components between the model predictions and the human responses with the effects of ground truth removed. RCI: Response Consistency Index. For the fitting functions, in addition to values estimated for models fitted to all data, the medians of the estimated values computed via two-fold cross-validation are shown in bold within brackets. NaN (Not a Number) was due to the ground truth as "predicted vectors" and thus invalid operation in partial correlation and RCI. See also Tables S3–S4.

Using the map of RCIs, we visualized how well the models explained the human response errors at each movie probe location (see Figure 5 for the representative four models; Figures S11–S26 for all models). To compare the similarities of model prediction patterns, we computed the correlations of the spatial RCI patterns between each model pair (Table S4).

Low-level pooling models

Consider first the effects of signal pooling. Previous studies have shown that the spatial resolution of human visual motion processing is low. For example, the detection of spatial modulations in optical flow is most sensitive at spatial frequencies lower than 1 c/° and is difficult at frequencies higher than 1 c/°.[22,23] The sensitivity function is shifted to the lower frequency for motion detection in comparison with luminance detection. The low spatial resolution of motion processing may be one of the factors that explain why perceived vectors deviate from ground truths. Although many neural mechanisms at multiple processing stages may contribute to the low spatial resolution of visual motion perception,[13,24,25] we approximated them by a 2D Gaussian pooling of the ground truth vectors and estimated the amplitude and sigma of the Gaussian weighting function that best explained the human response. The reason the pooling model was run on the ground truth flow, not on vectors estimated from the movie clips by some algorithm, was to evaluate the contribution of signal pooling to human errors as independently as possible from the other factors.

The standard deviation of the best-fit function was found to be ~5 pixels (Table S2). Small but positive cross-validated partial correlations with the human response (ρ_{uv} = 0.239) and a positive averaged RCI (0.015) indicate that the spatial pooling model can explain a small portion of the response errors. In agreement with our expectation that spatial pooling would be associated with deviations from the ground truths at points close to motion boundaries, the points with high RCIs indeed lie near object borders (Figure 5A). In addition, when points distant and close to borders were separately analyzed (see Figure S27 for the definition of border points), the agreement between the model and the responses was evident for only the border points (Table S5).

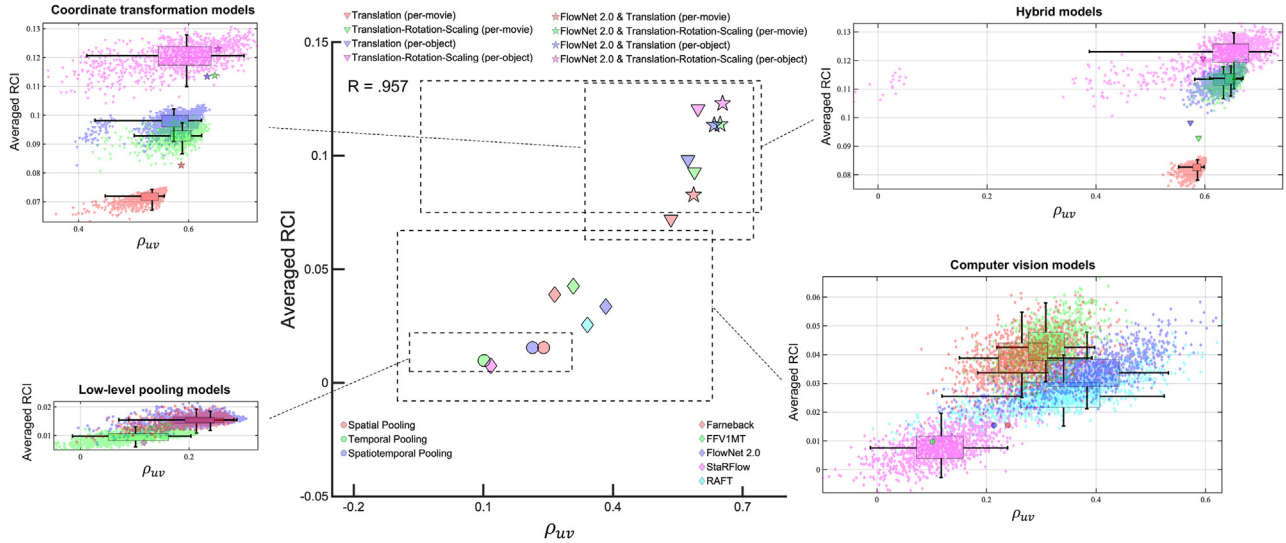


Figure 4. Scatterplots of the partial correlations of the *uv* components (ρ_{uv}) versus the average RCIs

In the center panel, each point represents the median performance index distribution computed by the bootstrapping method for the computer vision models and by the 2-fold cross-validation for the fitting models. The other four panels show the performance index distributions of 1,000 sampling sets and 2D boxplots with the (25% percentile, median, 75% percentile) ranges (boxes) and the 95% confidence intervals (whiskers). See also Figures S6–S10.

The temporal resolution of human visual motion processing is also known to be low. This is suggested by poor detection of temporal change in speed (acceleration/deceleration)[26] as well as integration of direction signal over a few hundred milliseconds or longer.[27] It is thus likely that the observers integrate motion flow information across multiple frames. To evaluate the effect of temporal pooling on flow illusion, we performed 1D Gaussian pooling of the ground truth motion vectors over time and estimated the best amplitude, mean, and sigma of the Gaussian weighting function that explained the human response. The (cross-validated) partial correlations (ρ_{uv}) and averaged RCI were 0.101 and 0.010, respectively. When the spatial and temporal pooling were combined into a 3D Gaussian spatiotemporal pool, the power afforded in terms of explaining the human response was similar to that of the spatial pooling model (ρ_{uv} = 0.214, average RCI = 0.015).

The cross-validation-based statistical tests revealed that ρ_{uv} did not differ significantly among the three (spatial, temporal, and spatiotemporal) pooling models (Figure S8), while the average RCI was lower for the temporal pooling model than the other two models (Figure S10). Also, there were high correlations among the spatial RCI patterns (Table S4), which indicates that the three pooling models might explain similar aspects of human response errors. It should be noted that the movie clips tested here include only temporally smooth optical flows, for which the effects of temporal pooling could be similar to spatial pooling; the results might be different if stimuli featuring large temporal changes had been used.

In summary, we found evidence that signal pooling explains some but limited aspects of flow illusions.

Computer vision models for optical flow estimation

Next, we compared the human responses to the outputs of five computer vision models developed for optical flow estimation. Farnebäck[28] is a conventional model that estimates dense optical flow. FFV1MT[29] is a biologically motivated model that includes a feedforward, primary visual cortex (V1)-middle temporal area (MT) structure. FlowNet 2.0,[30] StaRFlow,[31] and RAFT[32] are convolutional neural net models based on supervised learning; they employ different architectures when seeking to improve ground truth flow estimations. Specifically, FlowNet 2.0 features multiple subnetworks, StaRFlow a spatiotemporal recurrent architecture, and RAFT iterative refinement of optical flow (See model descriptions in the STAR methods for details).

Table 1 indicates that the RAFT model output was the closest to the human response in terms of the endpoint error, but this was because the model output was also the closest to the ground truth. When comparing the partial correlation and the average RCI, the power of explaining the human response errors was highest for FlowNet 2.0 (ρ_{uv} = 0.386, average RCI = 0.034), but only slightly lower for the other models except for StaRFlow. To statistically compare the performance among the computer vision models, we used bootstrapping, not cross-validation, since the models have no free parameter to fit our data. The results indicated that ρ_{uv} did not differ significantly between FlowNet 2.0, RAFT, Farnebäck, and FFV1MT (Figure S7), but the average RCI was higher for FFV1MT than RAFT (Figure S9).

The RCI spatial map of FlowNet 2.0 (Figure 5B) reveals at which probe locations the model and humans make similar errors. Most such locations lie where dark objects make movements that differ from those of adjacent regions (see the lower right corner of Video S2, and the lower part of Video S3). Estimating correct vectors at such points may be difficult for both computer vision models and humans because of a lack of image information (signal loss).

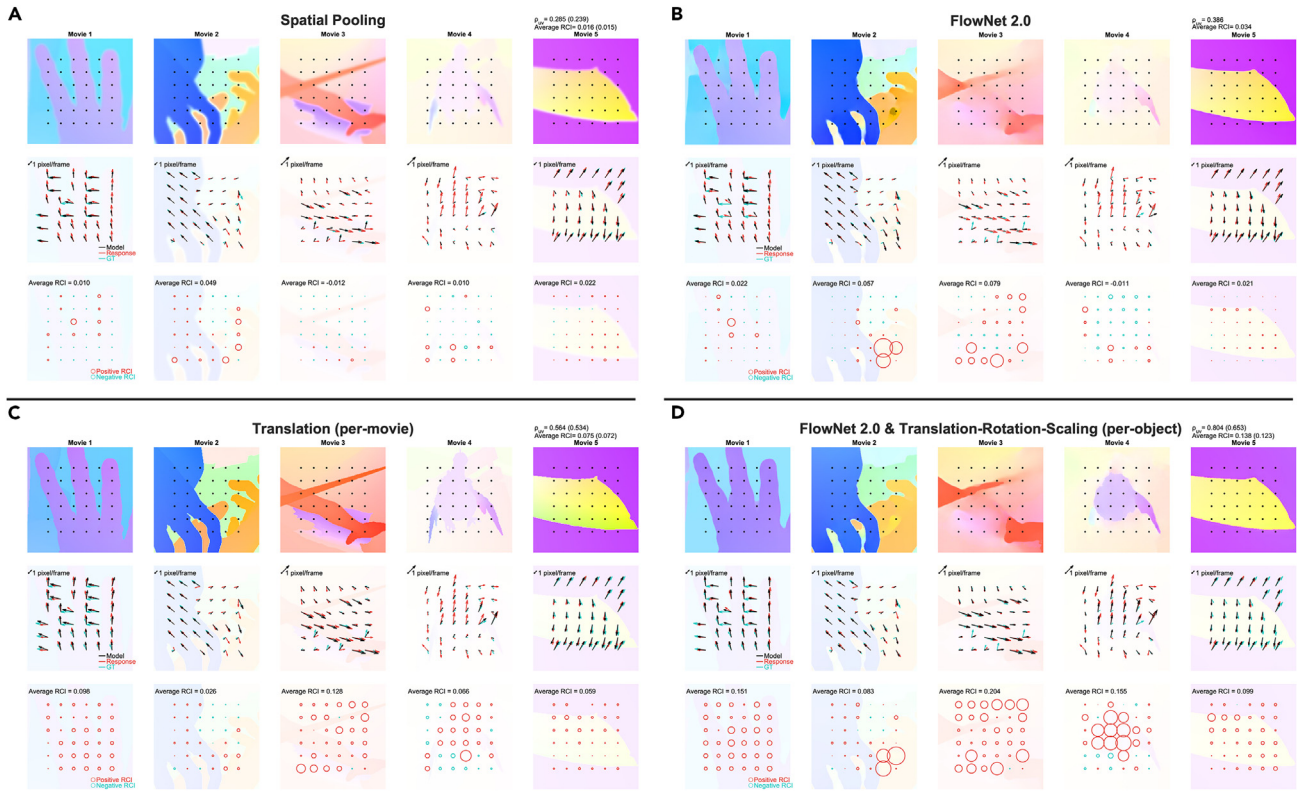


Figure 5. The RCIs of the representative models

Spatial maps showing the extent of consistency between human responses and model predictions for (A) a pooling model (spatial pooling), (B) a computer vision model (FlowNet2.0), (C) a coordinate transformation model (Translation [per-movie]), and (D) a hybrid model (FlowNet 2.0 & Translation-Rotation-Scaling [per-object]). In each panel, the first row is the color-coded visualization of the motion vectors in the optical flows predicted by each model. The second row depicts the motion vectors in each location; the black arrow denotes the model-predicted response, the red arrow denotes the human response, and the green arrow denotes the ground truth. The plotting scale (the length of a 1 pixel/frame vector) is shown by an arrow in the top left corner of each panel. The diameters of the circles in the third row indicate the RCIs, with red circles indicating positive values (the deviations from the ground truth were consistent between the model predictions and the human responses) and green circles indicating negative values (the deviations from the ground truths were inconsistent). The average RCIs of the 36 locations in each movie are shown in the top left corners. See also Figures S11–S26 and Data S3 for the results of all models tested.

The pattern of RCI spatial maps is similar among the computer vision models. The correlation exceeded 0.8 between Farnebäck and FFV1MT and between FlowNet 2.0 and RAFT (Figures S14–S18; and Table S4). This suggests that these models share the mechanisms explaining flow illusions.

When compared to the pooling models, cross-validation-based tests on ρ_{uv} (Figure S8) and the average RCI (Figure S10) indicated that the explanatory power for the human response was significantly higher for FFV1MT, FlowNet 2.0, and RAFT than for any pooling models. This may be partially because the computer vision models implicitly include a pooling process. In agreement with this idea, we found positive correlations in the RCI spatial maps between the computer vision models and the pooling models (Table S4).

Coordinate transformation models

The mechanisms included in the pooling models and computer vision optic-flow models can explain some, but only a portion of the deviations of human responses from the ground truth flow ($\rho_{uv} \leq 0.386$). What kinds of mechanisms are additionally necessary to explain the remaining deviations? Since the models considered so far process optical flow on the retinal image coordinates, as does early visual motion processing in humans, they cannot explain the illusions occurring at non-retinotopic processing. Nevertheless, certain visual motion illusions, including induced motion[33,34] and biological motion,[7] may be associated with higher-level vector analysis that transforms optical flows from image coordinates to world- or object-centered coordinates. Specifically, vector analysis decomposes element movements into a global common movement and local relative movements[7,35,36] It is also known that deformations in perceived vectors are produced by mutual repulsion/attraction between adjacent objects, such as motion contrast,[37] direction repulsion,[38] and motion capture.[39] To evaluate the contributions of these higher-level optical flow illusions to human motion perception when viewing naturalistic scenes, we fitted the following four transformation models to our data (See model descriptions in the STAR methods for details).

In the Translation (per-movie) model, a common translation vector is subtracted from the ground truths at all points in the same movie clip. The translation parameters that best explain the human response were estimated by fitting. This is one of the simplest descriptive models of induced motion based on vector analysis,[7,35,36] where a motion pattern of multiple objects is decomposed into a global common translation of the movie scene and local relative motions. We assumed that the observers completely ignored or underestimated the global translation component when reporting the perceived motion. Using the ground truth as the input makes the model biologically implausible, but enables us to purely estimate the contribution of coordinate transformation to flow illusion.

In the Translation-Rotation-Scaling (per movie) model, which tests more complex coordinate transformations, a common rotation (in direction) and rescaling (of speed), in addition to a common translation, are applied to the ground truths of all points in the same movie clip. This model can cope with cases where the common motion includes a global rotation and/or a global scale change, in addition to a global translation.

The remaining two models were identical to the first two, except that different parameters were estimated for different parts of the scene. For this analysis, 36 points in each movie were classified into 2–3 groups depending on the object layer to which each point belongs (see Figure S27 for how to define object layers). Separate parameter-fitting for different objects can capture flow distortions among multiple objects that cannot be captured by global transformations, such as motion repulsion[37,38] and attraction.[39] In the Translation (per-object) model, a common translation vector is subtracted from the ground truths at all points in the same object layer of each movie. In the Translation-Rotation-Scaling (per-object) model, a common translation, a rotation in direction, and a rescaling of speed are applied to the ground truths at all points in the same object layer of each movie. The last one is the most complex model with the largest number of free parameters. In all cases, the transformation parameters for each movie or object layer that best fit the human responses were estimated (see Table S2).

These four models are not independent. The first model is a special case of the other three models, while the second and the third models are special cases of the fourth model. The model with a smaller number of parameters may not be sufficient to describe the coordinate transformation effects in our movie clips, but the model with a larger number of parameters may overfit the data. We used cross-validation to compare the models with different numbers of free parameters.

The four models captured the pattern of human errors rather well (ρ_{uv} = 0.534–0.597, average RCI = 0.072–0.121). Compared to the low-level pooling and computer vision models, the coordinate transformation models showed significantly higher partial correlations (Figure S8) and average RCIs (Figure S10). Also, the RCI spatial maps differed from those of the low-level models, which indicates that the transformation models can explain such flow illusion components that the low-level models cannot explain. For example, the coordinate transformation models well-predicted the pattern of human errors in Video S4 (Figures 5C and S19–S22), where induced motion appeared to alter the perceived motion of a flying chicken.

The successful fitting of coordinate transformation models to human response data suggests that human perception of naturalistic movies is affected not only by lower-level effects but also by higher-level factors such as induced motion and vector decomposition. A comparison of the absolute values of the partial correlations and average RCI also suggests that the higher-level factors dominate the low-level factors in flow illusions. There were, however, caveats. Global performance indices accumulating local effects over many locations could overestimate the effects of global factors in comparison with local factors. Furthermore, even though we used cross-validation, the transformation models were optimized to explain the human data by parameter fitting, while computer vision models were made without the knowledge of human response.

When the four translation models were compared, statistical tests on the average RCIs (Figure S10) indicated that model complexity increased the explanatory power of the response data; the Translation-Rotation-Scaling models outperformed the Translation models, and per-object models outperformed per-movie models. However, statistical tests on ρ_{uv} (Figure S8) indicated that only the difference between the Translation (per-movie) model and the Translation-Rotation-Scaling (per-movie) model attained statistical significance. Note also that the RCI spatial maps of the four models (Figures S19–S22) are similar (see also Table S4), suggesting that the mechanisms underlying the perceived errors predicted by these models may be shared or overlap significantly.

These results indicate that the simplest transformation model that computes local relative motions by subtracting a single global translation for each movie can explain a significant portion of flow illusions seen in our movie clips, but more complex models that can also cope with global rotation, scale change, and/or inter-object interactions can explain the human response better. A significant proportion of the flow illusions humans perceive in naturalistic movies can be ascribed to induced motion based on vector decomposition with inter-object interactions.

Hybrid models

If we are correct in stating that computer vision models (that include signal pooling) reflect lower-level visual motion processing and transformation models reflect higher-level motion processing, with each explaining different aspects of the human errors, it would be expected that combinations of such models would predict human errors better than either alone. We, therefore, computed the outputs of coordinate transformation models after changing the inputs from ground truth flows to the optical flows estimated by FlowNet 2.0, which evidenced the highest ρ_{uv} of all five tested models. As expected, we found that the ρ_{uv} values and average RCIs for the hybrid models were better than those of the computer vision models and the original transformation models employing ground truths. The increases in both indices (Figures S8 and S10) attained statistical significance for the two per-movie models, and the increase in average RCI attained statistical significance for the Translation (per-object) model. Although neither reached the statistical significance for the Translation-Rotation-Scaling (per-object) model,

many free parameters gave this model the ability to describe some aspects of flow illusions FlowNet 2.0 explains (see positive correlation of the two models in Table S4). In summary, the results support our interpretation that both lower-level and higher-level mechanisms contribute to the flow illusions that humans perceive in naturalistic movie clips, and the hybrid model consisting of a high-performance retinotopic motion flow algorithm and a coordinate transformation with a reasonable complexity is a promising architecture to explain the human-perceived optical flow.

DISCUSSION

Traditionally, vision scientists have sought to understand the mechanisms of human visual perception by analyzing relationships between subjective perceptual experiences and the responses of neural mechanisms and/or the predictions of computational models. Recent technical advances have enabled researchers to access large-scale data on neural responses and model behaviors for complex inputs including natural stimuli. Thus, data on human visual perception should become big and rich. Here, we psychophysically measured the human-perceived motion vectors in the optical flows of dynamic naturalistic movie clips, using a novel method based on motion matching and flash probing. We found that the perceived flows of naturalistic stimuli deviated from the ground truths in ways that depended on the stimulus structures ("flow illusions"). By comparing the human-perceived flows to the predictions of a variety of models, we concluded that the illusions were attributable to both lower-level factors, such as spatiotemporal pooling and signal loss, and higher-level factors, such as inter-object interactions and coordinate transformations resulting from vector analysis. Earlier studies showed that such factors produced specific illusions when certain experimental stimuli were presented to the observers, but we are the first to show that a combination of such factors explains the human-perceived optic flows of naturalistic scenes. Our method provides a way to scale up the data on subjective visual experience, paving the way for a leap in human motion perception research.

To the best of our knowledge, the dense maps of human-perceived optical flow of naturalistic scenes have not been reported previously. There are several reasons. First, appropriate stimuli are lacking. When exploring how the visual system processes a stimulus of the external world embedded in images, one needs to know the ground truth of the variable. Many movies with natural scenes are available, but the ground truth of the optical flows is not easy to derive. To overcome this problem, the computer vision community uses synthesized movies for training and evaluation of models, and the MPI Sintel Dataset is one of the most popular among such resources. We thus used this dataset.

The second difficulty is the workload. When estimating human-perceived flows, perceived motion vectors must be measured at many points. In psychophysical experiments, such measurements cannot be made in parallel, but must be made separately for each location. Use of a strict psychophysical procedure (e.g., estimation of the point of perceptual equality via a forced-choice paired comparison of direction or speed employing the method of Constant) would be very time-consuming. Thus, we derived a quick (thus practical) psychophysical method: motion vector matching by a method of adjustment. We limited the number of points to 180, sampled at 1° intervals within 5 × 5° regions of five movie clips.

One contribution of the present study is the direct comparison of human-perceived flows with the flows predicted by computer vision models. Four of the five models that we tested, including FlowNet2.0, similarly explained some aspects of human response errors caused principally by input signal loss and spatial pooling. However, the agreement between the model prediction and human response was not high. This is not because the models knew the ground truth. The training image dataset of the machine learning models did not include the test stimuli we used. On the other hand, we did not fine-tune the parameters and/or structure of the cutting-edge computer vision models to make them close to humans.

Our basic motivation for comparing human motion perception with cutting-edge computer vision algorithms that are optimized to estimate retinotopic optic flows as close as possible to the ground truth of input images was not to test their plausibility as the model of human motion processing, but to see which components of flow illusion were shared by humans and the models. The shared components are likely to reflect a general computational difficulty in determining motion correspondence (produced by signal loss and the aperture problem, for example), rather than the specific characteristics of the processing mechanism.

The limited capacities of the computer vision models in terms of explaining human response errors may reflect differences in the computational goals. The models that we tested are aimed at estimating local image shifts accurately between frames in image coordinates. Human visual motion processing is aimed not only at estimating optical flows in retinal image coordinates but also at estimating object motions in appropriate coordinate frames, self-motions relative to the environments, and relative object depths as revealed by motion parallax. Based on optic flows estimated via lower-level processing, higher-level processing attains biologically meaningful goals when viewing complex natural stimuli, while produces motion illusions when viewing simple artificial stimuli. The fact that transformation models explain human responses well supports this view.

The ultimate but unachieved goal of this project is to develop an image-computable model that can account for the perceived human flows for arbitrary visual inputs. One promising model architecture is a hybrid model in which optical flows estimated by image-based motion detectors are fed to higher-level processing, including coordinate transformation. The question then is how to predict the behaviors of the higher-level processing (e.g., coordinate transformation parameters) directly from the input image sequence. One strategy is the development of a computational model for high-level human motion processing.[35,36] An alternative is a data-driven approach; artificial neural networks could be trained using human response data. In either case, it is necessary to scale up the data size, i.e., collecting more human responses at higher spatiotemporal densities with a broader range of stimuli.

We expect our study will usefully connect psychophysics and neuroscience, since our data can be directly compared to the cortical activities of humans or animals measured via fMRI, optical imaging, or other methods, while observers watch the same movie clips.

Complex naturalistic stimuli and simple artificial stimuli are both useful for understanding visual processing.[40] In our case, several visual illusions having been studied using artificial stimuli can explain a significant proportion of flow illusions arising during the perception of naturalistic movies. This is reasonable based on the idea that many of the visual illusions arising when viewing simple artificial images should reflect visual processing that attempts to achieve higher-level visual functions (e.g., perceptual constancy) in complex natural scenes. Extending this idea, it could be argued that visualization and modeling of complex human perceptions of natural scenes, which we are challenging, is the ultimate way to understand the computational goals and mechanisms producing the conventional illusions triggered by simple artificial stimuli.

Limitations of the study

The human-perceived flows we reported here were those measured with the specific stimulus and procedural parameters we chose, and slightly different perceived flows may be obtained when different parameters are used. This is an issue one should consider when comparing our psychophysical data to neural and behavioral data collected under different viewing conditions. We measured perceived motion vectors at single spatiotemporal points while directing each observer's gaze and attention to those points. The perceived flow map of a collection of independent local measurements may significantly differ from a neural response map that is recorded in parallel at a time. Our procedure might underestimate context effects. In addition, the use of a random-dot motion field as a matching stimulus may direct the observer's attention more to textural motion flow than to high-level object trajectories. Our current dataset is limited in the variation of movie clips and the number of observers.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Apparatus
  - Stimuli
  - Design and procedures
  - Model descriptions
  - Parameter-fitting models
  - Computer vision models
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Coefficient of determination for the ideal response model R^2
  - Endpoint error
  - Comparison of direction
  - Partial correlation ρ
  - Response consistency index
  - Cross-validation
  - Bootstrapping

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108307.

ACKNOWLEDGMENTS

This research was supported by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) Grant Numbers JP20H00603 and JP20H05957.

AUTHOR CONTRIBUTIONS

S.N. conceived the idea. Y.Y., T.F., and S.N. designed experiments. Y.Y. performed experiments. Z.S. tested computer vision models. All authors contributed substantially to data analysis. Y.Y. and S.N. drafted the original manuscript. All authors reviewed the draft and revised it critically in terms of intellectual content.

DECLARATION OF INTERESTS

T.F. and S.N. were employed by Nippon Telegraph and Telephone Corporation. The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: April 3, 2023
Revised: August 9, 2023
Accepted: October 20, 2023
Published: October 23, 2023

REFERENCES

1. McCool, C.H., and Britten, K.H. (2008). Cortical Processing of Visual Motion. In The Senses: A Comprehensive Reference, *2* (Elsevier Inc), pp. 157–187.
2. Burr, D., and Thompson, P. (2011). Motion psychophysics: 1985–2010. Vis. Res. *51*, 1431–1456. https://doi.org/10.1016/J.VISRES.2011.02.008.
3. Nishida, S. (2011). Advancement of motion psychophysics: Review 2001–2010. J. Vis. *11*, 11. https://doi.org/10.1167/11.5.11.
4. Nishida, S., Kawabe, T., Sawayama, M., and Fukiage, T. (2018). Motion Perception: From Detection to Interpretation. Annu. Rev. Vis. Sci. *4*, 501–523. https://doi.org/10.1146/annurev-vision-091517-034328.
5. Adelson, E.H., and Bergen, J.R. (1985). Spatio-temporal energy models for the perception of motion. J. Opt. Soc. Am. *2*, 284–299.
6. Simoncelli, E.P., and Heeger, D.J. (1998). A model of neuronal responses in visual area MT. Vis. Res. *38*, 743–761. https://doi.org/10.1016/s0042-6989(97)00183-1.
7. Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. Percept. Psychophys. *14*, 201–211. https://doi.org/10.3758/BF03212378.
8. Wallisch, P., and Movshon, J.A. (2019). Responses of neurons in macaque MT to unikinetic plaids. J. Neurophysiol. *122*, 1937–1945. https://doi.org/10.1152/jn.00486.2019.
9. Kelly, D.H. (1982). Fourier components of moving gratings. Behav. Res. Methods *14*, 435–437. https://doi.org/10.3758/BF03203305.
10. Adelson, E.H., and Movshon, J.A. (1982). Phenomenal coherence of moving visual patterns. Nature *300*, 523–525. https://doi.org/10.1038/300523a0.
11. Julesz, B. (1971). Foundations of Cyclopean Perception (University of Chicago Press).
12. Braddick, O. (1974). A short-range process in apparent motion. Vis. Res. *14*, 519–527.
13. Newsome, W.T., and Paré, E.B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). J. Neurosci. *8*, 2201–2211. https://doi.org/10.1523/JNEUROSCI.08-06-02201.1988.
14. Warren, W.H., and Hannon, D.J. (1988). Direction of self-motion is perceived from optical flow. Nature *336*, 162–163. https://doi.org/10.1038/336162a0.
15. Nishimoto, S., and Gallant, J.L. (2011). A Three-Dimensional Spatiotemporal Receptive Field Model Explains Responses of Area MT Neurons to Naturalistic Movies. J. Neurosci. *31*, 14551–14564. https://doi.org/10.1523/jneurosci.6801-10.2011.
16. Matthis, J.S., Muller, K.S., Bonnen, K.L., and Hayhoe, M.M. (2022). Retinal optic flow during natural locomotion. PLoS Comput. Biol. *18*, e1009575. https://doi.org/10.1371/journal.pcbi.1009575.
17. Koenderink, J.J., Van Doorn, A.J., and Kappers, A.M. (1992). Surface perception in pictures. Percept. Psychophys. *52*, 487–496.
18. Butler, D.J., Wulff, J., Stanley, G.B., and Black, M.J. (2012). A Naturalistic Open Source Movie for Optical Flow Evaluation. In European Conf. on Computer Vision (ECCV), pp. 611–625.
19. Janai, J., Güney, F., Wulff, J., Black, M.J., and Geiger, A. (2017). Slow Flow: Exploiting High-Speed Cameras for Accurate and Diverse Optical Flow Reference Data. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1406–1416.
20. Nijhawan, R. (1994). Motion extrapolation in catching. Nature *370*, 256–257. https://doi.org/10.1038/370256b0.
21. Whitney, D., and Cavanagh, P. (2000). Motion distorts visual space: Shifting the perceived position of remote stationary objects. Nat. Neurosci. *3*, 954–959. https://doi.org/10.1038/78878.
22. Nakayama, K., Silverman, G.H., MacLeod, D.I., and Mulligan, J. (1985). Sensitivity to shearing and compressive motion in random dots. Perception *14*, 225–238. https://doi.org/10.1068/P140225.
23. Watson, A.B., and Eckert, M.P. (1994). Motion-contrast sensitivity: Visibility of motion gradients of various spatial frequencies. J. Opt. Soc. Am. *11*, 496–505.
24. Burr, D.C., and Ross, J. (1982). Contrast sensitivity at high velocities. Vis. Res. *22*, 479–484. https://doi.org/10.1016/0042-6989(82)90196-1.
25. Amano, K., Edwards, M., Badcock, D.R., and Nishida, S. (2009). Adaptive pooling of visual motion signals by the human visual system revealed with a novel multi-element stimulus. J. Vis. *9*, 4.1–425. https://doi.org/10.1167/9.3.4.
26. Werkhoven, P., Snippe, H.P., and Toet, A. (1992). Visual processing of optic acceleration. Vis. Res. *32*, 2313–2329. https://doi.org/10.1016/0042-6989(92)90095-Z.
27. Burr, D.C., and Santoro, L. (2001). Temporal integration of optic flow, measured by contrast and coherence thresholds. Vis. Res. *41*, 1891–1899. https://doi.org/10.1016/S0042-6989(01)00072-4.
28. Farnebäck, G. (2003). Two-Frame Motion Estimation Based on Polynomial Expansion. In Image Analysis. SCIA 2003. Lecture Notes in Computer Science, 2749, J. Bigun and T. Gustavsson, eds. (Springer).
29. Solari, F., Chessa, M., Medathati, N.K., and Kornprobst, P. (2015). What can we expect from a V1-MT feedforward architecture for optical flow estimation? Signal Process. Image Commun. *39*, 342–354. https://doi.org/10.1016/j.image.2015.04.006.
30. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2016). FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. Preprint at arXiv *1*. https://doi.org/10.48550/arXiv.1612.01925.
31. Godet, P., Boulch, A., Plyer, A., and Besnerais, G.L. (2020). *STaRFlow*: A spatiotemporal recurrent cell for lightweight multi-frame optical flow estimation. Preprint at arXiv *1*. https://doi.org/10.48550/arXiv.2007.05481.
32. Teed, Z., and Deng, J. (2020). RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. Preprint at arXiv *1*. https://doi.org/10.48550/arXiv.2003.12039.
33. Duncker, K. (1938). Induced motion. In A source book of Gestalt psychology, W.D. Ellis, ed. (Kegan Paul, Trench, Trubner & Company), pp. 161–172. https://doi.org/10.1037/11496-012.
34. Wallach, H. (1959). The perception of motion. Sci. Am. *201*, 56–60. https://doi.org/10.1038/scientificamerican0759-56.
35. Bill, J., Gershman, S.J., and Drugowitsch, J. (2022). Visual motion perception as online hierarchical inference. Nat. Commun. *13*, 7403. https://doi.org/10.1038/s41467-022-34805-5.
36. Gershman, S.J., Tenenbaum, J.B., and Jäkel, F. (2016). Discovering hierarchical motion structure. Vis. Res. *126*, 232–241. https://doi.org/10.1016/j.visres.2015.03.004.
37. Tynan, P., and Sekuler, R. (1975). Simultaneous motion contrast: Velocity, sensitivity and depth response. Vis. Res. *15*, 1231–1238. https://doi.org/10.1016/0042-6989(75)90167-4.
38. Marshak, W., and Sekuler, R. (1979). Mutual repulsion between moving visual targets. Science *205*, 1399–1401. https://doi.org/10.1126/science.472756.
39. Ramachandran, V.S., and Cavanagh, P. (1987). Motion capture anisotropy. Vis. Res. *27*, 97–106. https://doi.org/10.1016/0042-6989(87)90146-5.
40. Rust, N.C., and Movshon, J.A. (2005). In praise of artifice. Nat. Neurosci. *8*, 1647–1650. https://doi.org/10.1038/nn1606.
41. Van Rossum, G., and Drake, F.L. (2009). Python 3 Reference Manual. Scotts Valley (CreateSpace).

42. Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J.K. (2019). PsychoPy2: experiments in behavior made easy. Behav. Res. Methods *51*, 195–203. https://doi.org/10.3758/s13428-018-01193-y.

43. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., and Szeliski, R. (2011). A Database and Evaluation Methodology for Optical Flow. Int. J. Comput. Vis. *92*, 1–31.

44. Cavanagh, P., and Mather, G. (1989). Motion: The long and short of it. Spat. Vis. *4*, 103–129. https://doi.org/10.1163/156856889X00077.

45. Cavanagh, P. (1992). Attention-based motion perception. Science *257*, 1563–1565. https://doi.org/10.1126/science.1523411.

46. Lu, Z.L., and Sperling, G. (2001). Three-systems theory of human visual motion perception: Review and update. J. Opt. Soc. Am. Opt Image Sci. Vis. *18* (9), 2331–2370.

47. Solari, F., Chessa, M., Medathati, N.V.K., and Kornprobst, P.. Matlab code for "FFV1MT: A V1-MT feedforward architecture for optical flow estimation". https://senselab.med.yale.edu/modeldb/showModel.cshtml?model=181035&file=/FFV1MT_code/#tabs-2.

48. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Series B Stat Methodol *36*, 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x.

49. Efron, B. (1979). Bootstrap methods: Another look at the jackknife. Ann. Stat. *7*, 1–26.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw data	OSF storage	https://doi.org/10.17605/OSF.IO/BU7PD
Software and algorithms		
MATLAB 2021a		RRID:SCR_001622
Python	http://www.python.org/	RRID:SCR_008394
Psychopy	http://www.psychopy.org	RRID:SCR_006571
High-Speed Sintel Dataset	https://www.cvlibs.net/projects/slow_flow/	N/A
Other		
Videos S1–S6	This manuscript (supplemental information)	Videos S1–S6

RESOURCE AVAILABILITY

Lead contact

Further information and any related requests should be directed to and will be fulfilled by the lead contact, Shin'ya Nishida (shinyanishida@mac.com)

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Raw data derived from human samples, ground truth, and all models have been deposited at Open Science Framework and are publicly available as of the date of publication. The DOI is listed in the key resources table.
- This paper does not report the original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Three of the authors and a naive laboratory member (all males aged 22 to 39 years old) participated in the experiments. All observers had normal or corrected-to-normal vision and gave informed consent before the study. This study was approved by the Research Ethics Committee of the Graduate School of Informatics, Kyoto University (approval no. KUIS-EAR-2020-003). The experiments were performed in accordance with the Declaration of Helsinki, except for preregistration.

METHOD DETAILS

Apparatus

All experimental codes were written in Python 3[41] using PsychoPy.[42] Visual stimuli were presented on the display of a 13-inch MacBook Pro (resolution 1,440 × 900, refresh rate 60 Hz, gamma 2.2, background gray luminance 150 cd/m^2, viewing distance 57 cm) in a normally illuminated room, or on a monitor (Eizo CG303W, 1,920 × 1,200, 60 Hz, 2.2, 45 cd/m^2, 72 cm) controlled by a desktop Windows PC in a darkened chamber. For each setup, the viewing distance was adjusted so that 50 pixels subtended a visual angle of 1° to the observer.

Stimuli

To evaluate the basic vector-matching accuracy, we first measured the human responses to Random-Dot Kinematogram (RDK) with spatio-temporally uniform optical fields in a preliminary experiment. The RDK stimulus consisted of 5,000 white and black dots at a density ratio of 50-50 in a 600-pixels (12°) circular aperture on a gray background. Each dot was 3 pixels (3.6 arcmins) in diameter, and the total dot density was ~20%. Throughout the experiment, four dots (each five pixels in diameter) were presented 60 pixels above, below, left, and right of the display center (as position markers) to indicate the location of the display center to which attention was to be paid during the task. The four position-marker dots and the probe dot were red. Note that this traditional random-dot pattern differs from that of the Brownian matching-noise stimulus (see design and procedures).

In the main MPI Sintel experiment, we used a higher-frame-rate version[19] of the MPI Sintel dataset[18] originally rendered at 1,008 frames per second (FPS) and a resolution of 2,048 × 872 pixels. We selected five short movie clips that contained relatively complicated spatial structures, multiple objects, and backgrounds. We presented these clips at 60 FPS after magnifying the clip image to 4,098 × 1,744 pixels before clipping it to fit a 600-pixels (12°) circular aperture, which had a sufficient viewing angle containing multiple objects and contents to avoid aperture problem. This also enabled spatial sampling of reasonably dense perceived flow in a local image area while suppressing long-range jumps between adjacent frames, for which human motion detection may rely more on high-level feature tracking than low-level motion detection.[44–46] In each clip, we defined 36 probed locations on a six-by-six grid with a spacing of 1° (see Figure 1C). The grid location was chosen so that the 36 probed points covered a variety of figure objects and the background. In each trial, the target motion clip was relocated such that the to-be-probed location came to the center of the aperture (Videos S2–S6). Unlike the uniform optical fields in the RDK stimulus, the MPI Sintel dataset included images with varying chromatic and brightness characteristics. To render the flashed probe dot and the surrounding four position marker dots equally salient, regardless of the movie context, we computed the HSV values of the dot locations in the movie clip and imparted the reverse HSV values to the dots. Each movie clip was presented under four flipped conditions: no flip (original orientation), horizontal flip, vertical flip, and horizontal and vertical flips (Figure 1E). This allowed us to decompose the two types of response error. One is an error defined by the display coordinates, thus determined by the relationship between the observer and the display, not by the image pattern. For example, an observer may tend to overestimate rightward motion or over-report vectors located left of the flashed probe. One can visualize errors of this type by averaging the response errors of the four flipped conditions without unflipping them to the original coordinates (e.g., Figure S3). We analyzed the responses on display coordinates to evaluate the accuracy of our method. The other type is response errors defined on the image coordinates whose directions flip together with the image flip. Most visual motion illusions (e.g., induced motions) are determined by the relative relationships of image components, and we thus analyzed responses in the image coordinates by averaging the response errors of the four flipped conditions after unflipping them to the original coordinates (e.g., Figures 3, 5, and S4).

Design and procedures

In the preliminary RDK experiment, the speed and direction were randomly and uniformly selected from 1 – 10 pixels per frame (PPF) and 0 – 360° respectively. All dots moved coherently in one direction at a constant speed within each trial. There were two blocks, each with 144 trials. Each observer underwent 288 trials. On the other hand, in the main MPI Sintel experiment, the five movie clips were tested in separate blocks, each consisting of 144 trials (36 flash-probed locations × four flip conditions). Each observer performed 720 trials.

Both preliminary RDK and the main MPI Sintel experiments were conducted with the same procedure; We used a method of adjustment to measure the motion vector that the observer perceived at a specific spatiotemporal location in a short video clip (Figure 1A and Video S1). During each trial, a target motion stimulus and a matching stimulus were repeatedly presented on a uniform gray screen. One cycle of repetition consisted of the target clip (250 ms, 15 frames), an interstimulus interval (ISI, 750 ms), the matching stimulus (250 ms), and an ISI (750 ms). The target clip was presented at the center of the display in a circular aperture. In the middle of the target clip presentation (i.e., in the 8^{th} frame of 15 frames, Figure 1B), a probe dot (five pixels in diameter) was flashed at the display center for 1 frame (16.67 ms). The matching stimulus was a circular Brownian ($1/f^2$) noise field (120 pixels in diameter, 100% peak contrast) presented at the center of the display. The broadband Brownian noise aided observers to perceive a wide range of image speeds without aliasing attributable to under-sampling in time. To reduce the abrupt stimulus onset/offset of the target and matching stimuli, the stimulus contrast was linearly increased during the first half of the presentation, attained full contrast in the eighth frame, and decreased linearly in the latter half of the presentation. Observers were asked to reproduce the motion vector at the location/time indicated by the target probe by matching the speed and direction of the stimulus using the mouse cursor. There was a purple dot (five pixels in diameter) freely moveable within a central circular area (600 pixels in diameter). Depending on the distance of the mouse cursor from the display center (from 0 to 300 pixels in radius), the speed of the matching stimulus varied logarithmically from 0 to 20 pixels per frame (PPF). When satisfied with the matching, the observers terminated the trial by clicking the mouse, but early termination before completion of three repetition cycles was not accepted (See Video S1).

At the end of each trial, the speed and direction of the target and the matching stimuli of the trial were shown as numbers on the display. One difficulty associated with the method of adjustment was the setting of an appropriate criterion for trial termination. It was likely that too-early termination would sacrifice accuracy, but that longer observation would not necessarily improve performance. The main purpose of feedback was to help observers develop appropriate response strategies. This input should not be used as training feedback by the observers because each location and flip presentation was only presented once.

Model descriptions

We used the least-squares fitting method to determine the best-fit parameters that minimized the summed endpoint error of human optical flow (see Table S2 for all best-fit parameters). The numbers of free parameters were 2 (spatial pooling), 3 (temporal pooling), 4 (spatiotemporal pooling), 0 (computer vision models), 10 (Translation [per-movie], 2 × 5 movies), 20 (Translation-rotation-scaling [per-movie], 4 × 5 movies), 24 (Translation [per-object], 2 × 12 objects). and 48 (Translation-rotation-scaling [per-object], 4 × 12 objects). All model predictions can be found in Datas S3 and S4. Since the numbers of free parameters differed among the models, to evaluate statistically the explanatory powers of the models with fair comparisons, the values estimated via 2-fold cross-validation and bootstrapping were also calculated (see quantification and statistical analysis for details).

Parameter-fitting models

Spatial Pooling

This fitting function spatially aggregates local ground truth flow vector fields (u_{GT}, v_{GT}) to obtain a predicted perceived vector (u_{est},v_{est}). A 2D Gaussian kernel $g_S(x, y)$ parameterized by the amplitude (A) and sigma (σ_S) is used to weight the ground truth vectors surrounding the probed location (x_P, y_P). We did not consider the offset between the center of the Gaussian kernel and the probed location because the endpoint errors were symmetrically distributed between the human response and the flow vector field in 2D space (Figure S3C). Therefore, the vector components (u_{est}, v_{est}) for each (x_P, y_P) were defined as in Equation 1:

$$\begin{cases} u_{est}(x_P, y_P) = \sum_x \sum_y g_S(x - x_P, y - y_P) u_{GT}(x, y) \\ v_{est}(x_P, y_P) = \sum_x \sum_y g_S(x - x_P, y - y_P) v_{GT}(x, y) \end{cases}, \quad \text{(Equation 1)}$$

where:

$$g_S(x, y) = A \exp\left\{ - \left(\frac{x^2}{2{\sigma_S}^2} + \frac{y^2}{2{\sigma_S}^2} \right) \right\}.$$

To reduce computational complexity, we constrained the spatial extent of the 2D Gaussian kernel to −30 to +30 pixels on both the x and y axes, with the probe at 0.

When separately analyzing the effects of spatial pooling on border and non-border areas, we categorized 55 of the 180 probed locations as border points, as indicated by the bracketed numbers in Figure S27, depending on whether a window of $\pm 2.5\sigma_S$ around that point included more than one object layer or not, where σ_S was the best-fit sigma (5.025pixels).

Temporal Pooling

This fitting function temporally aggregates ground truth flow vectors at the probed location. A 1D Gaussian kernel $g_T(t)$ parameterized by the amplitude (A), mean (μ_T), and sigma (σ_T) is used to weight the ground truth vectors from $t = -7$ to +7 frames ($t = 0$ refers to the probed frame) as shown in Equation 2:

$$\begin{cases} u_{est}(x_P, y_P) = \sum_t g_T(t) u_{GT}(x_P, y_P, t) \\ v_{est}(x_P, y_P) = \sum_t g_T(t) v_{GT}(x_P, y_P, t) \end{cases}, \quad \text{(Equation 2)}$$

where:

$$g_T(t) = A \exp\left(- \frac{(t - \mu_T)^2}{2{\sigma_T}^2} \right).$$

Spatiotemporal Pooling

This fitting function spatiotemporally aggregates the ground truth flow vectors. A 3D Gaussian kernel $g_{ST}(t)$ parameterized by the amplitude (A), spatial sigma (σ_S) temporal mean (μ_T), and temporal sigma (σ_T) is used to weight the ground truth vectors as in Equation 3:

$$\begin{cases} u_{est}(x_P, y_P) = \sum_x \sum_y \sum_t g_{ST}(x - x_P, y - y_P, t) u_{GT}(x, y, t) \\ v_{est}(x_P, y_P) = \sum_x \sum_y \sum_t g_{ST}(x - x_P, y - y_P, t) v_{GT}(x, y, t) \end{cases}, \quad \text{(Equation 3)}$$

where:

$$g_{ST}(x, y) = A \exp\left\{ - \left(\frac{x^2}{2{\sigma_S}^2} + \frac{y^2}{2{\sigma_S}^2} + \frac{(t - \mu_T)^2}{2{\sigma_T}^2} \right) \right\}.$$

We constrained the spatial extent of the Gaussian kernel to −30 to +30 pixels on both the x and y axes, and the temporal extent to −7 to +7 frames.

Coordinate transformations

The translation transformation decomposes the ground truth vectors, including the global translation vector (u_0,v_0), to parameters that fit the human response. The predicted perceived vector for each position is computed as shown in Equation 4:

$$\begin{cases} u_{est} = u_{GT} - u_0 \\ v_{est} = v_{GT} - v_0 \end{cases}. \quad \text{(Equation 4)}$$

The translation-rotation-scaling transformation includes additional scaling (s) and rotation (α) parameters that affine-transforms ground truth vectors to fit the human response as shown in Equation 5:

$$\begin{cases} u_{est} = (s\cos\alpha)u_{GT} + (s\sin\alpha)v_{GT} - u_0 \\ v_{est} = (-s\sin\alpha)u_{GT} + (s\cos\alpha)v_{GT} - v_0 \end{cases}. \quad \text{(Equation 5)}$$

Per-object transformation models

The per-object transformation models served as proxies of optical flow object-based processing. We manually defined object layers for each movie. In these models, we classified the 36 probed locations in each movie into layers with two or three objects and backgrounds based on their boundaries and the unique motion flows. The object layers are numbered as shown in Figure S27. Video S1 contained a hand and background layers. Video S2 contained left hand, right hands, and background layers. Video S3 contained arm, animal, and background layers. Video S4 contained a rooster and background layers. Video S5 contained a bat's wing and background layers.

Computer vision models

Farnebäck

This model uses a polynomial expansion transform to estimate the speed and direction of dense optical flows by approximating neighboring pixels between two frames based on the classic constant brightness assumption ($I(x+u\delta t,y+v\delta t,t+\delta t) = I(x,y,t)$) and global flow smoothness constraints.[28] We estimated the displacements at five levels of the image pyramid via three iterations per level, using a 15 × 15 average window size and a Gaussian weighting function of SD = 1.1 for averaging over the neighborhoods. Each pixel neighborhood used to find polynomial expansions contained five pixels.

FFV1MT

This model is biologically inspired and includes V1 motion energy estimations employing spatiotemporal Gabor filters and normalization, static nonlinear pooling of feedforward V1 responses in the MT layer, and a velocity estimate derived by decoding a linear weighted average of the MT response. The model also includes non-linear filtering of the MT response to handle spatial flow discontinuities more effectively. We selected −2∼2 subframes from the probed frame using the parameters of the original study[29] (Table 1 in [29]) and the code on the group site.[47]

FlowNet 2.0

This is a convolutional neural network-based motion model that includes iterative refinement via multiple applications of FlowNetSimple (to extract motion vectors directly from two stacks of paired images) and FlowNetCorr (to extract motion vectors from a correlation layer for comparison of two separate images), to handle large displacements. FlowNet-SD is used to fine-tune the model with a focus on small displacements, and a final fusion network estimates motion flows.[30] We trained the FlowNet 2.0 model using the Flyingchair, ChairsSDHom, and 3DFlyingthings datasets.

StaRFlow

This is a temporally dynamic, convolutional neural network-based motion model that includes spatiotemporal recurrent cells to generate multi-frame optical flow estimations and handle occlusion information.[31] We trained this model on the MPI Sintel dataset (without experimental stimuli). Four continuous frames covered each training iteration with a batch size of 8, and the model attained convergence after 50,000 iterations.

RAFT

This deep network motion model includes a feature encoder to extract features from two frames, a context encoder to extract context features from the first frame, a 4D correlation layer that computes the visual similarities of all pairs of pixel features at the 1/8 level, and an updating Gated recurrent unit to refine optical flow iteratively.[32] Following the original study, we used the official model pre-trained on Flying chair and KITTI, and then fine-tuned the MPI Sintel dataset (without experimental stimuli) over 100,000 iterations with a batch size of 8.

QUANTIFICATION AND STATISTICAL ANALYSIS

To compare the human responses with model predictions, we initially averaged the four human responses and the data for the four flip conditions in the original image coordinates, thus obtaining 180 data points (i.e., five movies × 36 positions) to contrast with vectors predicted by each model. To evaluate model performance (summarized in Table 1), we computed average endpoint error, partial correlations, and the Response consistency index (RCI) to evaluate how the models predicted human response errors at each location.

Coefficient of determination for the ideal response model R^2

R^2 refers to the proportion of variance in the human response ($Resp_i$) that can be predicted by the ground truth as shown in Equation 6:

$$R^2 = 1 - \frac{\sum_i (Resp_i - GT_i)^2}{\sum_i (Resp_i - \overline{Resp})^2}. \quad \text{(Equation 6)}$$

Endpoint error

The endpoint error is the Euclidean distance between the ground truth (u_{GT}, v_{GT}) and human response (u_{Resp}, v_{Resp}) motion vectors as shown in Equation 7:

$$EPE = \sqrt{(u_{Resp} - u_{GT})^2 + (v_{Resp} - v_{GT})^2}. \quad \text{(Equation 7)}$$

Comparison of direction

When we compared correlations in direction (r and ρ), we evaluated the shortest angular distance of the angle of the response, or the model ($\theta_{Resp,or\ model}$), from the angle of the ground truth (θ_{GT}), by adjusting the phase of $\theta_{Resp,or\ models}$ as shown in Equation 8:

$$\theta_{resp\ or\ model} = \theta_{resp\ or\ model} - 2\pi, while\ (\theta_{resp\ or\ model} - \theta_{GT}) > \pi$$

$$\theta_{resp\ or\ model} = \theta_{resp\ or\ model} + 2\pi, while\ (\theta_{resp\ or\ model} - \theta_{GT}) < -\pi. \quad \text{(Equation 8)}$$

Partial correlation ρ

For each model, we calculated partial correlation coefficients between the predicted outputs and the human response by controlling for the effect of ground truth to determine which fitting functions/computer vision models best explained the pattern of deviations in the human response from the ground truth as shown in Equation 9:

$$\rho_{model} = r_{resp\ model \bullet GT} = \frac{r_{resp\ model} - r_{resp\ GT} r_{model\ GT}}{\sqrt{1 - r^2_{resp\ GT}}\sqrt{1 - r^2_{model\ GT}}}. \quad \text{(Equation 9)}$$

Response consistency index

The RCI is the product of three terms, $A \bullet B \bullet C$, where:

$$A = \frac{\left|\overrightarrow{GR}\right|}{\left|\overrightarrow{OG}\right| + \left|\overrightarrow{OR}\right|}. \quad \text{(Equation 10.1)}$$

$$B = \frac{\overrightarrow{GR} \bullet \overrightarrow{GM}}{\left|\overrightarrow{GR}\right|\left|\overrightarrow{GM}\right|}. \quad \text{(Equation 10.2)}$$

$$C = 0.5\left(\frac{\left|\overrightarrow{GM}\right| - \left|\overrightarrow{RM}\right|}{\left|\overrightarrow{GM}\right| + \left|\overrightarrow{RM}\right|} + 1\right) = \frac{\left|\overrightarrow{GM}\right|}{\left|\overrightarrow{GM}\right| + \left|\overrightarrow{RM}\right|} \quad \text{(Equation 10.3)}$$

and G, R, and M, are the endpoints of the ground truth, response, and model prediction vectors respectively, and O is the origin. A highlights points of interest (i.e., those evidencing flow illusions). A is 0 when the ground truth and the response are in perfect agreement ($\left|\overrightarrow{GR}\right| = 0$) but 1 when they completely disagree ($\left|\overrightarrow{GR}\right| = \left|\overrightarrow{OG}\right| + \left|\overrightarrow{OR}\right|$). B specifies the direction similarity (the cosine of the direction difference angle) between the response error relative to the ground truth ($\overrightarrow{GR}$) and the model error relative to the ground truth ($\overrightarrow{GM}$). B is 1 when agreement is perfect, and −1 when there is complete disagreement. C compares the distance between the model prediction and the ground truth ($\left|\overrightarrow{GM}\right|$) and the distance between the model prediction and the response ($\left|\overrightarrow{RM}\right|$). C is 1 when the prediction agrees with the response ($\left|\overrightarrow{RM}\right| = 0$) but 0 when the model agrees with the ground truth ($\left|\overrightarrow{GM}\right| = 0$). When the terms are combined, the RCI becomes positive and approaches 1 at locations where the model predicts conspicuous flow illusions; the RCI becomes negative at locations where the model predicts illusions in the opposite direction. This index is similar to the partial correlation between the model predictions and responses with the effects of ground truth removed ($r_{dir} = 0.886$, $r_{spd} = 0.909$ and $r_{uv} = 0.974$ when calculating the correlation between mean RCI and non-cross-validated partial

correlations across 16 models), but the index evaluates model predictability at each location with integrating the direction and speed components.

Cross-validation

2-fold cross-validation[48] was employed to evaluate model fitting statistically. During each cross-validation repetition, the 180 probed locations were split into two halves, thus subsets X and X-tilde, in a constrained random manner (see below). The model predictions for X were computed based on the best-fit parameters for X-tilde, and the model predictions for X-tilde were computed based on the best-fit parameters for X. Evaluation index values such as ρ_{uv} were computed from the combined set of predictions for all 180 points. This was repeated 1,000 times to estimate the population distribution of each index value. Such analysis enables the direct comparison of models with different numbers of free parameters. For models with no free parameters, the estimated index values were always the same (we thus used a bootstrapping method to evaluate computer vision models statistically).

For the statistical comparison of the index values between models, the same cross-validation was used to compute the population distribution of differences in each index value between each pair of models. To exclude the effects of random subset selection on variation in the index value difference, the same sets of X and X-tilde were used for all models during every cross-validation repetition. If the 95% confidence interval of the index value difference did not include zero, we regarded the difference between the models as statistically significant (alpha = 0.05, two-sided).

During cross-validation of the per-movie and per-object transformation models, we added a constraint to the random selection of subset X such that the locations were split into two halves within each movie and within each object layer (or approximately so when the number of points was odd). Except when otherwise noted, we used such constrained subset selection for all models and the same sets of X and X-tilde for comparisons across the models, as noted above. We found that this constraint exerted only very minor effects on the estimated index values.

Bootstrapping

A bootstrapping method[49] was employed to estimate the population distributions of performance index values for each model and during statistical comparison of the models, the distributions of the differences in index values between each model pair. The fitting models employed the best-fit parameters for the response data at all probed locations (without cross-validation). During each of 1,000 repetitions, we re-sampled 180 locations from the original 180 locations with replacement, and computed the index values of each model and the differences among models. As during cross-validation, when comparing the models, we used the same re-sampling sets. Re-sampling was random with the constraint that the number of re-sampled data points was the same as the original number of locations within each object layer. We found that this constraint exerted very minor effects on the estimated index values.