Oculo-retinal dynamics can explain the perception of minimal recognizable configurations

Liron Zipora Gruber[a], Shimon Ullman[b], and Ehud Ahissar[a,1]

[a]Department of Neurobiology, Weizmann Institute of Science, Rehovot 76100, Israel; and [b]Department of Computer Science and Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel

Edited by Norma Graham, Columbia University, New York, NY, and approved July 15, 2021 (received for review November 17, 2020)

Natural vision is a dynamic and continuous process. Under natural conditions, visual object recognition typically involves continuous interactions between ocular motion and visual contrasts, resulting in dynamic retinal activations. In order to identify the dynamic variables that participate in this process and are relevant for image recognition, we used a set of images that are just above and below the human recognition threshold and whose recognition typically requires >2 s of viewing. We recorded eye movements of participants while attempting to recognize these images within trials lasting 3 s. We then assessed the activation dynamics of retinal ganglion cells resulting from ocular dynamics using a computational model. We found that while the saccadic rate was similar between recognized and unrecognized trials, the fixational ocular speed was significantly larger for unrecognized trials. Interestingly, however, retinal activation level was significantly lower during these unrecognized trials. We used retinal activation patterns and oculomotor parameters of each fixation to train a binary classifier, classifying recognized from unrecognized trials. Only retinal activation patterns could predict recognition, reaching 80% correct classifications on the fourth fixation (on average, ∼2.5 s from trial onset). We thus conclude that the information that is relevant for visual perception is embedded in the dynamic interactions between the oculomotor sequence and the image. Hence, our results suggest that ocular dynamics play an important role in recognition and that understanding the dynamics of retinal activation is crucial for understanding natural vision.

active vision | eye movements | fixational drift | closed-loop perception | neural code

The mechanisms underlying visual acquisition are not yet understood. In natural conditions, humans perceive the world around them using continuous eye movements. Yet, the relevance of ocular dynamics to visual perception, and specifically to object recognition, is not known. One factor that supports irrelevance is the success of artificial algorithms for visual recognition that are based on static image snapshots (1–6), therefore ignoring ocular dynamics while preserving the similarity to other biological processes (7–13). Yet, importantly, artificial algorithms have been tested so far only on a limited set of perceptual tasks and suffer from yet unresolved difficulties (e.g., refs. 14 and 15). Given this gap, we have designed experiments to test the role of ocular dynamics in human visual recognition.

Traditionally, object recognition has been tested in the laboratory using briefly presented, flashed images. With flashed images, answering the question of whether eye movements are involved in object recognition, and how, is challenging. In the current study, we used a set of images whose recognition was shown to require continuous looking. This set is composed of MIRC (minimal recognizable configurations) images and subMIRC images (14). A MIRC is defined as an image patch that can be reliably recognized by human observers and which is minimal in that further reduction in either size or resolution makes the patch typically unrecognizable. A subMIRC is thus defined as an image patch created by a further minimal reduction in either size or resolution of a MIRC, rendering it typically unrecognizable (see details and the full set of images in *SI Appendix*, Fig. S1). The original MIRC study showed that human recognition could not be replicated by any visual recognition algorithm (14). Importantly for the current context, it was shown that recognizing these partial images takes time, typically over 2 s (16). This is in contrast to the recognition of full images, which is accomplished within short presentation times of typically less than 300 ms (17). In our experiments, we presented relatively small images (3 × 3 degrees in size), which can be captured almost entirely by the foveal region of the retina and whose perception, thus, should not depend on integrating several foveal foci.

As the eyes are never still, when we continuously look at an image, the flow of visual information to our brains results from the interaction of eye movements with the image (18–24). The kinematics of eye movements have been studied extensively. Studies show that from the point of view of motion kinematics, almost every section of ocular trajectory can be classified as a saccade or a fixational period in which the latter is dominated by drift motions (13, 18, 25–31). According to this kinematic classification, fixation on a moving target, such as during smooth pursuit or optokinetic response, is considered a fixational period. Saccades and fixations have been suggested to be controlled differently and to play different roles in visual perception (21–23, 25–29, 32, 33). Yet, both have been implicated as potentially playing major roles in visual acquisition, which makes them candidates for contributing to the process enabling the recognition of MIRC images.

The relatively long duration of MIRC recognition allows a prolonged iterative process, possibly combining bottom-up and top-down components of the visual system (34–36) as well as

Significance

Humans move their eyes continuously to scan their environment. Yet, the role of eye movements in object recognition is not known. In this work, we recorded eye movements of participants attempting to recognize images that are just above and below the threshold of human recognition. To assess the contribution of retinal dynamics, we modeled the activation patterns resulting from the continuous interactions of eye movements with the viewed image. We then trained a classifier to differentiate recognized from unrecognized trials. We show that recognition could be classified only when the continuous interactions between eye movements and the image were used. We suggest that vision is mediated by continuous interactions between eye movements and the environment, resulting in dynamic oculo-retinal coding.

Author contributions: L.Z.G., S.U., and E.A. designed research; L.Z.G. performed research; S.U. contributed new reagents/analytic tools; L.Z.G. analyzed data; and L.Z.G. and E.A. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: ehud.ahissar@weizmann.ac.il.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2022792118/-/DCSupplemental.

Published August 20, 2021.

controlling oculo-retinal dynamics (37, 38), that is, the dynamics that link ocular motion and retinal activations via closed-loop interactions. Oculo-retinal dynamics is dictated primarily by ocular dynamics, image properties, and retinal filtering. In our experiments, image properties were given, ocular dynamics was recorded, and retinal filtering was modeled using commonly accepted retinal models (39, 40). We tested whether we could predict recognition and, if possible, recognition timing using the trial-by-trial modeled retinal output. We found that, indeed, oculo-retinal dynamics can account for the behavioral characteristics of MIRC recognition.

Results

Relevance of Eye Movements to Recognition. A total of 20 healthy participants participated in three experimental sessions, 10 trials in each. Across the sessions, each participant viewed three versions of each of 10 images: full, MIRC, and subMIRC (see *Methods*). In each trial, participants viewed the image version for 3 s and then shifted their gaze to a location indicating whether they did or did not recognize the object shown in the image.

The full images were recognized at 100% of the trials as expected (Fig. 1*A*, black). The recognition rates of MIRCs (80 ± 4%) and subMIRCs (24 ± 4%) seen by participants for the first time (Fig. 1*A*, blue and red, respectively) replicated the behavioral results reported previously (14). The MIRC–subMIRC recognition gap was also evident for the individual images; for nine out of 10 pairs of image versions, there was a >50% difference in the recognition rate (*SI Appendix*, Fig. S2). As may be expected from the fact that our images were presented in a relatively small size, we did not find any tendency to gaze at specific image coordinates and did not find any difference between the distributions of gaze locations in trials in which the image was recognized or not (we have created visit-rate heat maps for all trials of each image and found the 5, 10, and 20% most-visited regions of interest [ROIs]). MIRCs maps did not have significantly more visited ROIs than subMIRC maps (permutation tests, all *P*s > 0.05).

To test whether the scanning eye movements are necessary for the recognition of MIRCs, we ran two pilot sessions, each with five participants viewing the set of 10 MIRCs. We prevented the scanning of the images by either stabilizing the image on the retina (see *Methods*; five participants) or by instructing participants to fixate on a fixational cross at the center of the image (five participants). The recognition rates in these cases dropped to 30 ± 8% and 32 ± 5%, respectively (Fig. 1*A*, dark blue). These results revealed the importance of scanning eye movements for recognizing MIRCs.

The question we ask here is: Can we find acquisition variables that correlate with single trial recognition? So far, such variables could not be found in the images themselves; first, the same images are sometimes recognized and sometimes not by the different subjects. Second, no computer-based classifier was found so far to discriminate between MIRCs and subMIRCs (14, 41). We thus turned to look at the other major components of the acquisition process—ocular kinematics and retinal activation. For this aim, we pulled together all trials of partial images, including both the MIRC and the subMIRC session for each participant (see *Methods*), and classified them according to recognition. Altogether, there were 251 trials in which a partial image was recognized ("recognized trials") and 149 trials in which a partial image was not recognized ("unrecognized trials"; Fig. 1*B*).

Ocular Kinematics. To analyze ocular kinematics, each 3-s scanning pattern was divided into saccade and fixation periods (see *Methods*). We compared the kinematic behavior measured during recognized and unrecognized trials (e.g., Fig. 2 *A* and *B*). The saccadic rate was not significantly different between recognized and unrecognized trials ($P > 0.2$, two-tailed Student's *t* test, Fig. 2*C*). Accordingly, when comparing the mean fixation duration, no difference was found between the groups ($P > 0.2$, Kolmogorov–Smirnov [KS] test, Fig. 2*D*). In contrast, the mean drift speed and amplitude during fixation were higher for unrecognized trials ($P < 0.05$, KS test, Fig. 2 *E* and *F*). This is consistent with other cases in which challenging visual conditions induce an increase in the fixation drift speed (32). We have verified, as was done in ref. 32, that the changes we observed in the mean drift speed could not be explained by differences in saccadic kinematics. Specifically, no significant difference was found in saccadic amplitude, saccadic speed, saccadic peak speed, or saccadic duration between the two sets of trials. Note, that in order not to lose temporal information, the ocular speeds were computed here with minimal low-pass filtering (32) (see *Methods*). While precluding a direct comparison of absolute speed values with

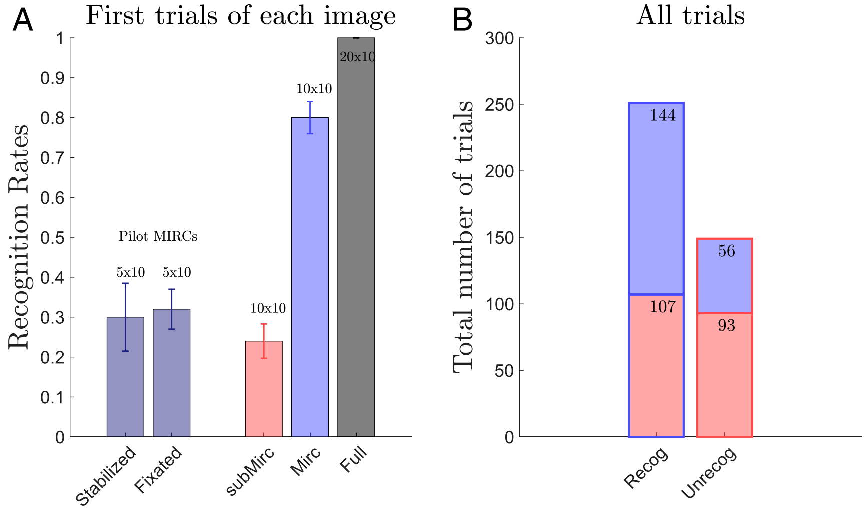

Fig. 1. Recognition rates. (*A*) Mean recognition rates for the two pilot sessions and three experimental sessions. In the pilot sessions (dark blue), five participants in each session viewed 10 MIRCs, either while the images were stabilized on the retina using real-time gaze following or while fixating on a cross in the center of the images. In the experimental sessions, subMIRC (red) and MIRC (blue) recognition rates were calculated only for the first time participants viewed each image (whether in its subMIRC or MIRC version), replicating the behavioral results reported in ref. 14 (10 participants × 10 trials for each of the partial images; all 20 participants × 10 trials for the full images). Error bars represent the SEMs. (*B*) Total number of recognized and unrecognized trials, of all trials of partial images (20 participants × 20 trials), divided to those with subMIRCs (red) and MIRCs (blue).

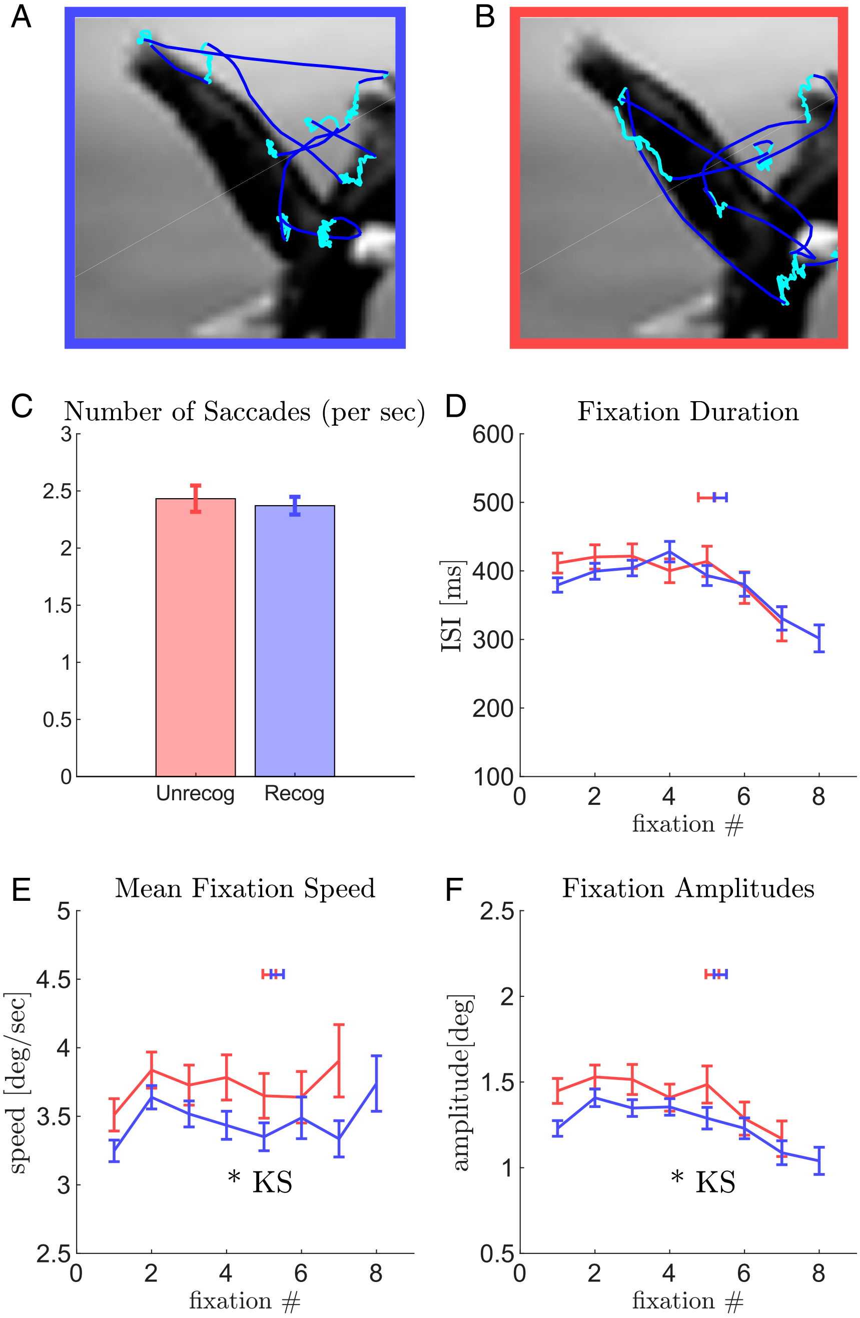

Fig. 2. Motor parameters. (*A*) Example of a scanning path of a recognized MIRC of an eagle. Identified saccades (dark blue) and fixation periods (light blue). (*B*) Same as *A* for an unrecognized trial of a different participant viewing the same eagle MIRC. (*C*) Mean number of saccades per seconds for recognized (blue) and unrecognized (red) trials, error bars represent the SEs, no significant difference was found ($P > 0.05$, two-tailed Student's *t* test). (*D*) Mean inter saccadic interval (i.e., fixation duration) per fixation number (sequential from trial onset) for recognized (blue) and unrecognized (red) trials. Error bars represent SEs. No significant difference was found ($P > 0.1$, KS test). The blue and red horizontal bars above the curves denote the mean ± STD number of fixations in a trial. (*E*) Same as *D* for mean drift speed ($P < 0.05$, KS test). (*F*) Same as *D* for mean drift amplitude ($P < 0.05$, KS test). *, indicates a significant difference between the compared distributions.

studies that used substantial low-pass filtering, the minimal filtering used here did not impair direct comparisons of ocular speed values between different conditions in the present study (32).

Retinal Activation. Eye movements induce dynamic retinal coding (23, 39, 40, 42–44). In order to evaluate the difference in visual acquisition between different trials, we created a dynamical model describing the visual acquisition process. The model assumes that the output of the retina is determined by the spatiotemporal interactions between the stationary (for 3 s) image and the continuous ocular motion (Fig. 3*A*). We modeled the activations of retinal ganglion cells (RGCs) using commonly accepted spatiotemporal filters (see *Methods* and Fig. 3*A*) and assessed their informative value. We removed redundant activation patterns (45) and then used only those modeled RGCs (mRGCs) whose correlation with the mean mRGC activation was <0.5 (termed informative mRGC, see *Methods* and the example in Fig. 3*B*).

Acquisition Dynamics. Consistent with previous reports (32, 46), the mean speed of the eye changed during the fixational pause, starting with relatively high speeds and converging to a lower, steady-state speed (Fig. 3*C*). Consistent with the increase in the mean speed of the eye during fixation in unrecognized trials (Fig. 2*E*), the steady-state ("target," see *Methods*) speed that the eye converged to within each fixational pause was higher for unrecognized trials (for t > 100 ms, $P < 0.05$, KS test, Fig. 3*C*). In contrast, our retinal model revealed that the mean retinal activation was higher for recognized trials ($P < 0.05$, KS test; Fig. 3*D*). And similar to the dynamics of ocular speed, also the within-pause ongoing activation of the retina converged to more or less steady target values, with the target value for recognized trials being larger than that for unrecognized trials (for t > 100 ms, $P < 0.05$, KS test, Fig. 3*E*). The ongoing retinal activation described in this paper is the residual activation after subtracting the mean retinal activation (see *Methods*). This subtraction results in an initial dip (Fig. 3*E*), reflecting the dynamics of the temporal filter applied (Fig. 3*A* and *Methods*).

Acquisition Correlates of Recognition. As shown above, both ocular speeds and retinal activations exhibited differences in their dynamics during recognized and unrecognized trials. To test whether any of these dynamic variables can predict image recognition, we trained a binary support vector machine (SVM) classifier using the different variables and tested it using a leave-one-out method (see *Methods*). Training the SVM using the instantaneous activation of the retina (vectors of the ongoing mean activation values sampled at 125 Hz along each fixational pause, e.g., Fig. 3*B*, black curve), resulted in classifying correctly 0.81 ± 0.02 of the trials (Fig. 4*A*, blueish curve). Specifically, this highest percent of correct classifications was achieved when training the model on the fourth fixation data (though it was also above chance level for the first, sixth, and seventh fixations). Similar results were obtained when using an alternative representation of retinal activation, a representation based on the eigenvectors of a functional principal component analysis [FPCA (47), see *Methods*]. Classifying these representations also yielded a fixation-dependent performance, with the first and fourth fixations yielding the highest success levels (0.61 ± 0.02 and 0.56 ± 0.01, respectively).

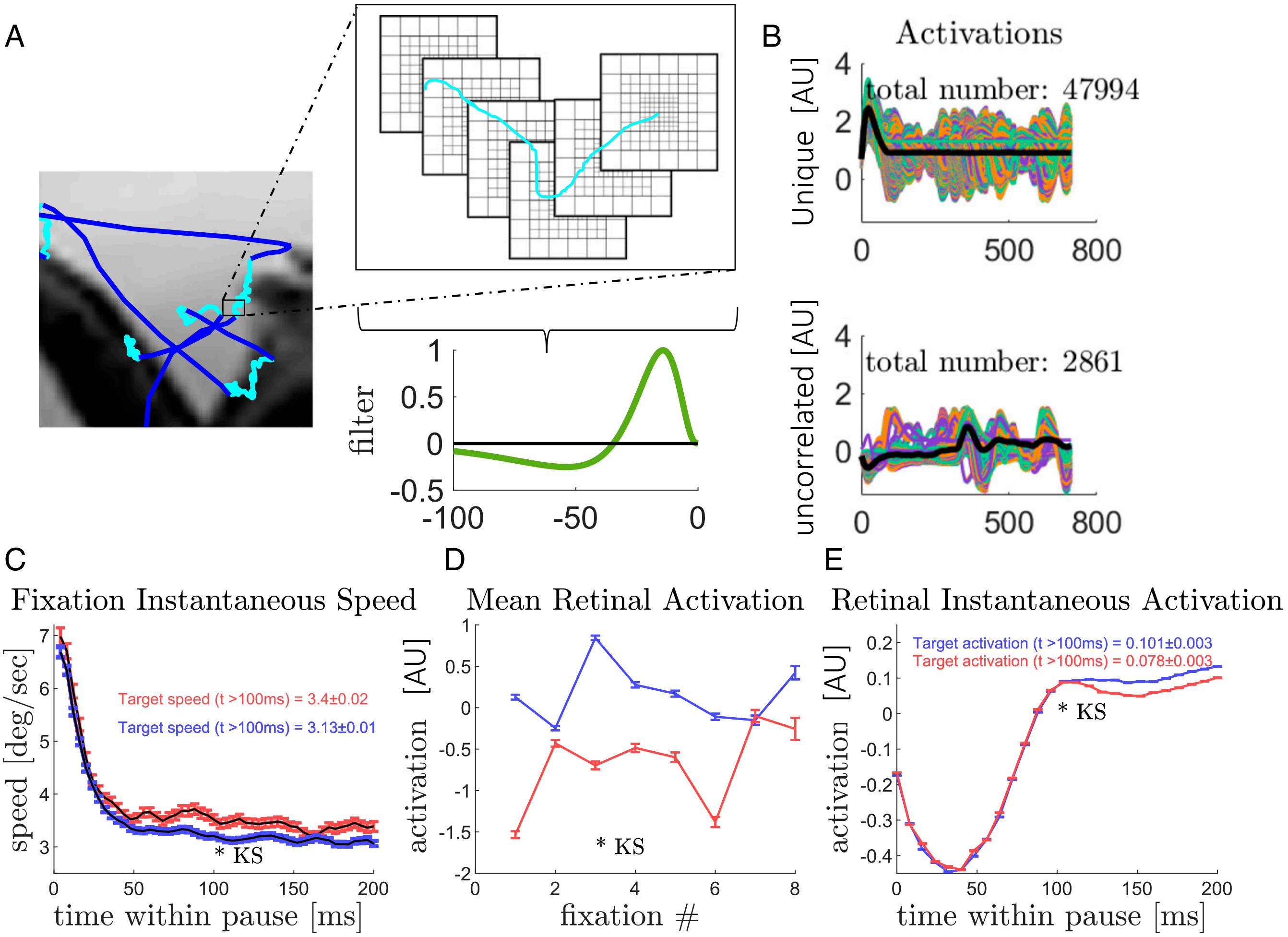

Fig. 3. The retinal model and visual acquisition parameters. (*A*) A schematic diagram of the retinal model: zoom in to a scanning path of an example trial; the retinal mosaic following the eye's trajectory (note the increasing receptive fields sizes when moving away from the center); the temporal filter used for each cell. For spatial and temporal aspects of the retinal model, see *Methods*. (*B*) Activation dynamics of the modeled cells during a fixational pause in the example trial. (*Upper*) A total of 47,994 unique activations (out of ∼160,000; see *Methods*). The mean activation is shown in black. (*Lower*) A total of 2,861 uncorrelated activations (see *Methods*). Mean activation, which is now defined as Retinal Activation, is shown in black. (*C*) The convergence of within-pause instantaneous fixation (drift) speeds, averaged across all fixations (that lasted over 100 ms) from recognized (blue) and unrecognized (red) trials (target speeds are the mean speeds for t > 100 ms, $P < 0.05$, KS test). (*D*) Mean retinal activation per fixation number, averaged across all fixations from recognized (blue) and unrecognized (red) trials ($P < 0.05$, KS test). (*E*) Same as *C* for the within-pause instantaneous retinal activation (target activations, t > 100 ms, $P < 0.05$, KS test). *, indicates a significant difference between the compared distributions; AU, arbitrary units.

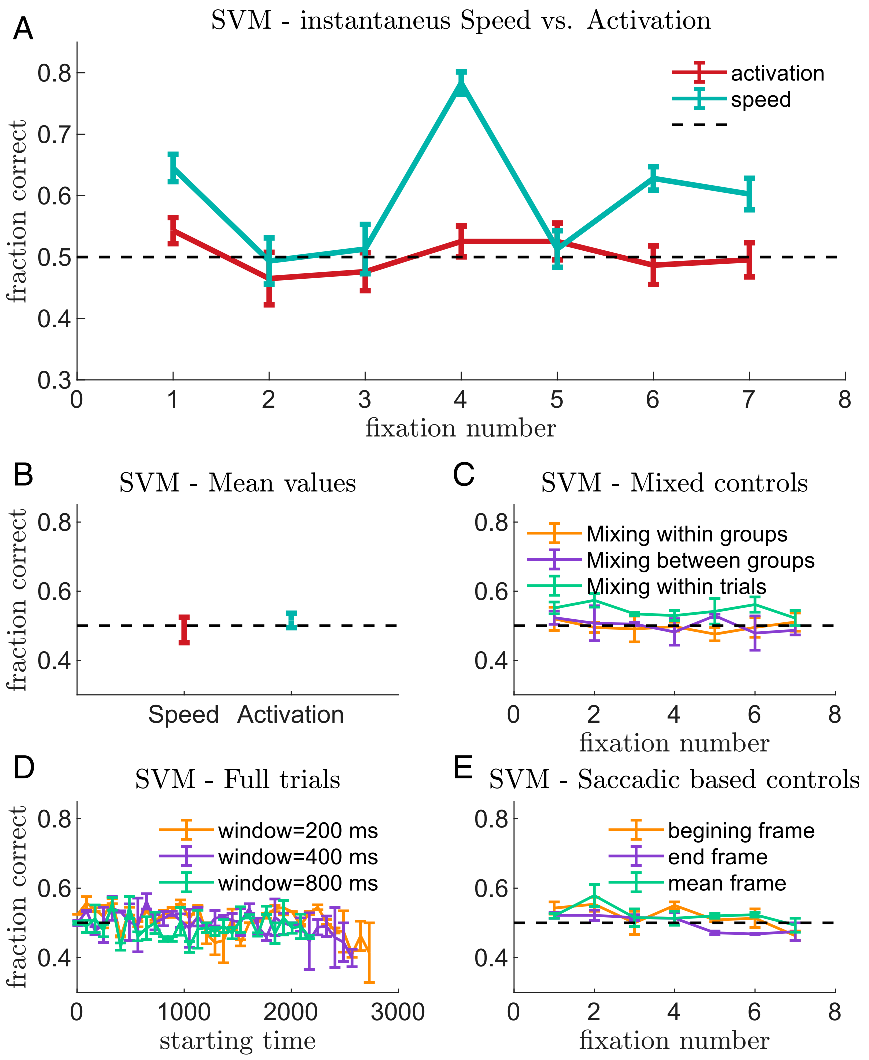

Fig. 4. SVM classification. (*A*) Percent of correct classifications of a binary SVM classifier trained in a leave-one-out method to classify recognized from unrecognized trials. The SVM was trained on the within-pause instantaneous retinal cell activations (turquoise, Fig. 3*E*) and on the within-pause instantaneous fixation speeds (red, Fig. 3*C*), each training per fixation number. Error bars represent SEs between 10 repetitions of the training (see *Methods*). (*B*) The SVM was trained on the mean per fixation retinal cells activation along the trial (turquoise, Fig. 3*D*) and on the mean per fixation speeds along the trial (red, Fig. 2*E*). Error bars represent SEs between 10 repetitions of the training (see *Methods*). (*C*) Same as *A*, the SVM was trained on the within-pause instantaneous retinal cells activations using artificially mixed trials. Mixing movements within trials (green, see *Methods*), mixing movements and images between the two classes (purple), and mixing movements and images within the two classes (orange). (*D*) Same as *A*, the SVM was trained on different durations of activations along each trial with 80-ms lags in starting times, ignoring the classification to saccades and fixations (time windows of –200 ms in orange, 400 ms in purple, 800 ms in green). (*E*) Same as *A*, the SVM was trained on all retinal cell activations per "frame," a specific point in time, either the end of fixation (purple), the beginning of fixation (orange), or an average frame of the entire fixation (green).

In contrast to the success in classifying retinal activations, the use of instantaneous speed (vectors of the ongoing ocular speed values sampled at 125 Hz along each fixational pause) resulted in a chance-level classification for all fixations (Fig. 4*A*, red curve). Similarly, the use of mean speed per fixation or mean activation per fixation (vectors of the mean values of either activation or speed per fixation), variables that exhibited significantly different values between recognized and unrecognized trials (Figs. 2*E* and 3*D*), also resulted in a chance-level classification (Fig. 4*B*). Hence, the only variable that was predictive of visual recognition in a trial was the modeled retinal activation.

The modeled retinal activation reflected the spatiotemporally filtered versions of the dynamic interactions between ocular motion and the image. As such, it makes use of more information than the ocular motion alone. To test whether the modeled retinal activation allowed better classification merely due to its larger number of information sources (ocular + image versus ocular alone), we created artificially mixed activations, which maintained the number of information sources without the exact motor–sensory interactions (Fig. 4*C*). Thus, we computed the activations that could be generated when taking the ocular movement from one trial and the image from another and used them to train the SVM. These shuffled interactions resulted in a chance-level classification (Fig. 4*C*, orange, shuffling within the recognized/unrecognized groups; purple, shuffling between the groups). We further tested whether the specific dynamics within a fixation pause is crucial for the classification. To achieve this, we shuffled the movements within each pause and calculated the new activations thus created (see *Methods* and Fig. 4*C*, green curve). This shuffling achieved above chance-level classification (highest for the second fixation, which resulted in classifying correctly 0.57 ± 0.02 of the trials), which means that the general oculo-retinal dynamics within each fixation pause was also slightly predictive of visual recognition.

We further checked whether we could predict recognition using the ongoing activations of the retina, ignoring the separation to fixational pauses. We computed the ongoing activation of the retina during each full 3-s trial and trained the SVM using the activations of different time windows along the entire trial, without preclassifying it to saccades and fixations. The biological separation to individual pauses was found important, as this training also resulted in a chance-level classification for all window sizes and starting times used (Fig. 4*D*). Finally, we also controlled for the possibility that saccade-triggered activations, and not the ongoing activations generated along the entire fixational pause, are sufficient for the recognition based classification. To test that, we trained the SVM using the entire retinal activation (400 × 400 "frame") generated when landing on a new saccadic target (Fig. 4*E*, orange) or just before leaving that target (Fig. 4*E*, purple). We also checked whether the mean "frame" of a fixational pause can be used as a predictor (Fig. 4*E*, green). Similar to all previous controls, these saccadic snapshot based trainings were not successful (Fig. 4*E*).

Discussion

In this work, we demonstrate that correlates of visual recognition can be found in the dynamic sensory activations that result from fine ocular motor–sensory interactions. Using MIRC and sub-MIRC images (14), which were found to be just above and below human recognition thresholds, respectively, we showed that recognition could be accounted for by the dynamics of retina-like activations resulting from continuous motor–sensory (oculo-retinal) visual interactions. These interactions were modeled here as the convolutions between eye movements, image contrasts, and retinal spatiotemporal filters. This result stands in contrast to the inability of previous attempts to find such correlates based on image contrasts alone (14).

We first replicated, using 20 participants, the recognition rates reported in the original MIRC study, which used thousands of subjects (14) (Fig. 1*A*), demonstrating the robustness of this threshold phenomena. Then, as we were interested in recognition correlates, we pulled together all trials and classified them by their recognition reports (Fig. 1*B*). Comparing the oculomotor variables revealed that while the saccadic rate (and hence also the mean durations of fixational pauses, Fig. 2 *C* and *D*) were similar for recognized and unrecognized trials, the mean ocular speed within the fixational pauses (and thus also the amplitude of the pause) were lower for recognized trials (Fig. 2 *E* and *F*). Thus, while the task difficulty (48) and the images were similar, scanning dynamics differed between recognized and unrecognized trials.

In order to assess the possible effect of these differences in ocular dynamics on visual processing and visual recognition, we used a dynamical model for retinal activation (39, 40) that convolves ocular motion with external images and retinal filtering (Fig. 3*A*). Our model integrates the moment-to-moment retinal motion, and not just its statistics (40, 44), as an informative feature to be used by the visual system. Our results show that predicting recognition in a trial-by-trial manner was achievable only using these modeled dynamics of retinal activations within each fixational pause (Fig. 4*A*). Neither the images alone (14) nor the oculomotor variables alone (Fig. 4 *B* and *C*) could predict recognition.

The model we used consisted of identical retinal-like cells, differing only in their receptive field locations and sizes. This is of course valid only as a first approximation, as the human retina is known to possess different kinds of cells (49, 50). Nevertheless, for the purpose of the current work, which is testing the dependence of visual recognition on motor–sensory dynamics, the use of a single-cell type proved to be sufficient (40, 51, 52).

Our results suggest that visual perception is based on the continuous activation of retinal cells during each entire fixational pause (26, 32, 53–56). Alternatively, visual perception might be based primarily on snapshots of retinal activations that are induced by each postsaccadic landing (57–60). We thus tested the possibility that saccade-triggered activations are sufficient for recognition detection. This alternative failed in predicting visual recognition in our task (Fig. 4*E*). Another alternative to the use of the continuous activation during fixational pauses is that the visual system only uses the statistics of fixational eye movements (40). To test this alternative, we tried to predict recognition based on shuffled data, detaching the specific movements from the image they were originally scanning; this attempt failed as well (Fig. 4*C*).

Thus, at least in our task, only the entire activation patterns during fixational pauses could account for recognition. The next question we asked was: Is the separation to individual pauses crucial? Could the visual system simply process the entire retinal dynamics along an entire trial continuously, ignoring the separation to individual fixational pauses? The answer was negative—applying our dynamical model continuously throughout the trials, ignoring the saccades-fixations classification, resulted in a chance-level classification as well (Fig. 4*D*). This finding provides a possible function for the well-known peri-saccadic suppression phenomenon (61, 62); resetting the activity in some circuits of the visual system around saccades (63) may facilitate the processing of oculo-retinal interactions in the new fixational location. At the system level, this result supports a functional separation between motor–sensory-motor loops controlling the saccades and those controlling the ocular drift (33). Furthermore, since we model here only foveal activations, the resetting suggestion may not be relevant to circuits processing peripheral vision, circuits that are capable of fast postsaccadic reaction (64).

Two aspects of our results call for further theoretical and empirical explorations. First, while showing that the modeled dynamics of retinal activations can predict visual recognition with high accuracy, our data cannot provide insights about the actual dynamical representations of the images, about the differences between recognizable and unrecognizable dynamics, or about the necessity of a separate reafferent coding channel (51, 52) for image identification. Second, we showed that the recognition potential was maximized at the fourth fixational pause. This result is consistent with the indications, in other mammals, that perceptual convergence takes about four motor–sensory interaction cycles (65–67) as well as with the typical recognition time in previous MIRC experiments (16). Yet, the mechanism underlying such convergence is not yet known. The explorations of these intriguing aspects require targeted empirical designs. Importantly, the consistency of our correlation-based and FPCA-based methods (see *Results*) together with the independence of the percent of informative retinal cells on the fixation number when using our correlation-based method (*SI Appendix*, Fig. S3) suggest that the fluctuations in success level across fixations along the trial reflect a dynamical, closed-loop process whose controlled variables (32, 37, 68) include visual information.

Our results were obtained with near-threshold stimuli. Yet, their conclusions are valid for all forms of natural vision. Specifically, these results suggest that computational models of primate vision should take into account, and be tested against, dynamic retinal outputs—the outputs dictated by the interactions between eye movements and external images. Our own hypothesis is that visual recognition results from a closed-loop convergence process (32, 37). The convergence dynamics exhibited by our data seem to divide to two levels: at a lower-level, a drift-based process that converges within each individual fixational pause, and at a higher-level, a saccade-based process that converges within approximately four saccades. Our results further suggest that the recognition potential does not increase monotonically across saccades. Rather, it starts on average with a positive potential that then decreases to chance level before reaching its maximum during a later (in this case the fourth) fixation. While the mechanism underlying the nonmonotonous behavior of recognition potential along this process is not yet clear, we suggest that it is part of a circular process attempting to coordinate neuronal and ocular processes and speculate that the drop of this potential after the fourth fixation reflects the loss of such coordination after the perceptual decision was made.

Methods

Participants. A total of 30 healthy participants with normal vision at the ages 23 to 36 y old (13 males) participated in either a pilot experiment session (five in each of the two conditions) or in three experimental sessions (10 in each of the two conditions). All participants were given a full and detailed explanation about the eye tracker device and the behavioral task and were paid for their participation (50 Israeli new shekel, ∼12 US dollars, per hour). Informed written consents were obtained from all participants in accordance with the approval of the Institutional Review Board of the Weizmann Institute of Science for this project.

Experimental Setup. The experiment took place in a darkened and quiet room where subjects sat in front of a high-resolution, fast computer screen (VPixx, 1920 × 1080, 120 Hz). The movements of the dominant eye were recorded using EyeLink II at 250 Hz [which is sufficient for tracking drift eye movements (69)]. Subjects sat 1 m away from the screen and placed their chin on a chinrest to reduce head movements.

Stimuli Used. Three versions of each image were used: car door, bicycle, eagle, glasses, eye, fly, horse, airplane, ship, and suit. Each image had a full image version, a MIRC version (which was found to be recognized in most trials), and a subMIRC version (which was found to be recognized in minimum 50% less trials than its MIRC). All images were taken from ref. 14. Following this study, we have also verified that the difference in recognition between the MIRC versions and the subMIRC versions cannot be explained by simple image parameters (no significant difference was found between the groups; *SI Appendix*, Fig. S4).

Experimental Design. The experiments took place in a darkened and quiet room where subjects sat in front of a high-resolution, fast computer screen (VPixx, 1920 × 1080, 120 Hz). In the pilot session, each condition had 10 trials, showing each of the MIRC versions of the images. In the first condition, the image was stabilized on the retina using a gaze-contingent display with which the image was locked to the participant's gaze [update rate was 100 Hz (26, 32)]. In the second condition, a fixation cross was displayed at the center of each image, and the participants were instructed to fixate on it throughout the trial. In each trial, participants clicked to start, fixated on a fixation cross for 2 s, viewed an image for 3 s, and then chose a "YES/NO" answer by shifting their gaze on the screen, reporting whether they did or did not recognize the object in the image. Each experimental condition had three sessions, 10 trials in each. The two different experimental conditions differed in the order of the sessions. Condition 1: subMIRCs, MIRCs, full images. Condition 2: MIRCs, full images, subMIRCs. All images were 3 × 3 visual degrees. In order to validate correct object recognition, each participant was asked, after the session, to report all objects that he/she remembers. No participant reported any false object name (we have considered the following answers as correct ones: bird = eagle, tie = suit).

Eye-Movement Processing. A velocity-based algorithm (modified from ref. 70) was used for detecting all saccades and fixations. We used the following threshold parameters for saccade detection: 16 deg/s minimal peak velocity and 0.3 deg minimal amplitude. Each detected saccade and each fixation pause were visually examined to verify the quality of saccadic detection. Fixation periods between saccades were analyzed only if they lasted at least 30 ms. For the analysis of within-pause instantaneous speed, only fixation periods that lasted at least 100 ms were used. The instantaneous fixation speed was calculated as the derivative of the raw eye position signal (32) and smoothed using a moving window of three samples (12 ms). The target speed for each fixation was defined as the mean of the speed between 100 ms and end of pause.

Retinal Model. We built a model of a 3 × 3 visual degrees retina that consisted of 400 × 400 cells based on the spatial properties of a typical human retina (39) and the commonly assumed spatiotemporal filtering properties of foveal neurons (40, 52, 71). For estimating the number of cells, we assumed a linear increase in the spacing between them, starting from 0.5 arcmin at the fovea, up to 1.6 arcmin at 4° eccentricity as well as a corresponding linear increase in receptive field diameters. Thus, for each modeled cell, we defined the size of its receptive field (RF) as the number of pixels that it is sensitive to, depending on its distance from the center of the gaze. We then used the following temporal filtering (40) (Fig. 3*A*) to calculate each cell activation:

$$\mathrm{activation}_i = \mathrm{RF}_i\_\mathrm{grayScale\_value} \otimes \left(\frac{t^n}{{\mathrm{T}_1}^{n+1}} e^{-\frac{t}{\mathrm{T}_1}} - \mathrm{R}\frac{t^n}{{\mathrm{T}_2}^{n+1}} e^{-\frac{t}{\mathrm{T}_2}}\right),$$

with $\mathrm{T}_1 = 5$ ms, $\mathrm{T}_2 = 15$ ms, $n = 3$, $R = 0.8$, and t from −100 ms till the current time.

The gain of each element is determined by the first term in the right-hand side of the activation equation. This term ($\mathrm{RF}_i\_\mathrm{grayScale\_value}$) reflects the mean gray scale value of the pixels contained in the RF. This value is then being convolved with the time filter (second term in the right-hand side of the equation, see also Fig. 3*A*).

Following previous modeling efforts (40, 52), we thus model foveal RGCs with a significant biphasic temporal filter (71) and without surround components (72). This model is a generic one, likely not fully matching specific individual RGCs while primarily capturing the generic pattern of their temporal dynamics. For each trial, we moved this array of retinal cells across the presented image according to the ocular trajectory recorded at that trial (down sampled to 125 Hz). This resulted in activation dynamics for each of the 400 × 400 cells, composing together a 3 s "movie" describing the modeled retinal activation during a trial. Unless mentioned otherwise, the model assumed a reset of retinal activation following each saccade.

Assessment of retinal information. Retinal activations are often highly redundant (45). In order to avoid the overdominance of specific retinal patterns, we applied the following selection of cells for processing. First, we used only unique activations (i.e., when exact duplicates of cell activations were found across the retina, only one of them was used). Second, we used only activations whose Pearson correlation with the mean retinal activation of all cells along the trial was <0.5 (choosing 0.5 as a threshold enabled using 5% of the cells on average; *SI Appendix*, Fig. S3. Other threshold choices resulted in a similar distribution of informative cells along the trial (*SI Appendix*, Fig. S3). Third, we subtracted the mean activation pattern from each activation pattern (Fig. 3*B*). The target activation for each fixation was defined similarly to the target speed as the mean activation of the eye between 100 ms after pause onset and the end of the pause.

FPCA. Retinal information was also assessed as the first principal component of an FPCA transformation (47) of all unique retinal activations for each fixation. Briefly, FPCA projects functional data to an eigenfunction basis that explains more variation than any other basis expansion. We used the MATLAB implementation FPCA.m taken from the Principal Analysis by Conditional Expectation (PACE) package.

SVM Classification. For classification, we trained and tested a binary SVM using MATLAB implementations "fitcsvm.m" and "predict.m." The two possible classes were "recognized" and "unrecognized," For each feature (speed, activations, frames, etc.), we used a leave-two-out method (one out from each class) and computed the percent of correct classifications. We used three types of kernels (linear, Fourier, and Gaussian) and present the results of the most successful one, the linear kernel. As the "recognized" class was larger, we ran 10 repetitions of the leave-two-out process, each time using a different subgroup of the smaller "unrecognized" class. Error bars in the figure represent the SE between these repetitions. For the shuffled controls, the same process was done using the artificial activation created by using movements from one trial with image from another. To create shuffling within a trial, we computed the derivative of the eye movement along each pause (i.e., the instantaneous speed). We then shuffled this vector of speeds and computed the activation created by this artificial movement (which only preserved the statistical properties of the natural speeds and not those of the natural accelerations or those of the power spectrum in general). For the saccadic-based control, we trained the SVM using a full "frame" of activations (400 × 400 cell activations at a specific time). For the saccadic-based frame at the beginning of a fixation, we used the activation frame at the second time sample after a saccade. For the frame at the end of a fixation, we used the activation frame at the one before last time sample before a saccade. For a mean frame, we calculated the mean activations of the entire fixation pause.

Data Availability. Anonymized MATLAB code data have been deposited in GitHub (https://github.com/lirongruber/Oculo-retinal-dynamics-can-explain-the-perception-of-minimal-recognizable-configurations).

https://doi.org/10.1073/pnas.2022792118

ACKNOWLEDGMENTS. We thank Rafi Malach and Michal Irani for invaluable advice and Michal Ramot and David Burr for reviewing the manuscript. This project has received funding from the European Research Council under the European Union Horizon 2020 Research and Innovation Programme (Grant Agreement No. 786949). E.A. holds the Helen Diller Family Professorial Chair of Neurobiology. S.U. is supported by the D. Dan and Betty Kahn Foundation Collaborative Research and the Minerva Foundation.

1. N. Kriegeskorte, Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).
2. D. L. K. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
3. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
4. M. Huh, P. Agrawal, A. A. Efros, What makes ImageNet good for transfer learning? arXiv [Preprint] (2016). https://arxiv.org/abs/1608.08614 (Accessed 10 December 2016).
5. O. Russakovsky *et al.*, ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
6. J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks? arXiv [Preprint] (2014). https://arxiv.org/abs/1411.1792 (Accessed 6 November 2014).
7. M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999).
8. S. R. Kheradpisheh, M. Ghodrati, M. Ganjtabesh, T. Masquelier, Humans and deep networks largely agree on which kinds of variation make object recognition harder. *Front. Comput. Neurosci.* **10**, 92 (2016).
9. J. Kubilius, S. Bracci, H. P. Op de Beeck, Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.* **12**, e1004896 (2016).
10. L. Z. Gruber, A. Haruvi, R. Basri, M. Irani, Perceptual dominance in brief presentations of mixed images: Human perception vs. deep neural networks. *Front. Comput. Neurosci.* **12**, 57 (2018).
11. C. F. Cadieu *et al.*, Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* **10**, e1003963 (2014).
12. D. L. K. Yamins *et al.*, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
13. S. M. Khaligh-Razavi, N. Kriegeskorte, Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
14. S. Ullman, L. Assif, E. Fetaya, D. Harari, Atoms of recognition in human and computer vision. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2744–2749 (2016).
15. A. Azulay, Y. Weiss, Why do deep convolutional networks generalize so poorly to small image transformations? arXiv [Preprint] (2018). https://arxiv.org/abs/1805.12177 (Accessed 19 November 2018).
16. H. Benoni, D. Harari, S. Ullman, What takes the brain so long: Object recognition at the level of minimal images develops for up to seconds of presentation time. arXiv [Preprint] (2020). https://arxiv.org/abs/2006.05249 (Accessed 9 June 2020).
17. J. J. DiCarlo, D. Zoccolan, N. C. Rust, How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
18. R. M. Steinman, J. Z. Levinson, The role of eye movement in the detection of contrast and spatial detail. *Rev. Oculomot. Res.* **4**, 115–212 (1990).
19. H. B. Barlow, Eye movements during fixation. *J. Physiol.* **116**, 290–306 (1952).
20. A. L. Yarbus, *Eye Movements and Vision* (Plenum, New York, 1967).
21. R. W. Ditchburn, *Eye-Movements and Visual Perception* (Clarendon Press, Oxford, 1973).
22. M. Rucci, E. Ahissar, D. Burr, Temporal coding of visual space. *Trends Cogn. Sci.* **22**, 883–895 (2018).
23. E. Ahissar, A. Arieli, Figuring space by time. *Neuron* **32**, 185–201 (2001).
24. M. Leszczynski, C. E. Schroeder, The role of neuronal oscillations in visual active sensing. *Front Integr. Neurosci.* **13**, 32 (2019).
25. D. Noton, L. Stark, Scanpaths in eye movements during pattern perception. *Science* **171**, 308–311 (1971).
26. S. C.-H. Yang, M. Lengyel, D. M. Wolpert, Active sensing in the categorization of visual patterns. *eLife* **5**, e12215 (2016).
27. A. Meermeier, S. Gremmler, M. Lappe, The influence of image content on oculomotor plasticity. *J. Vis.* **16**, 17 (2016).
28. M. Rucci, J. D. Victor, The unsteady eye: An information-processing stage, not a bug. *Trends Neurosci.* **38**, 195–206 (2015).
29. R. J. Krauzlis, L. Goffart, Z. M. Hafed, Neuronal control of fixation and fixational eye movements. *Philos Trans R Soc B Biol Sci.* **372**, 20160205 (2017).
30. E. Ahissar, S. Ozana, A. Arieli, 1-D vision: Encoding of eye movements by simple receptive fields. *Perception* **44**, 986–994 (2015).
31. Z. M. Hafed, R. J. Krauzlis, Similarity of superior colliculus involvement in microsaccade and saccade generation. *J. Neurophysiol.* **107**, 1904–1916 (2012).
32. L. Z. Gruber, E. Ahissar, Closed loop motor-sensory dynamics in human vision. *PLoS One* **15**, e0240660 (2020).
33. E. Ahissar, A. Arieli, M. Fried, Y. Bonneh, On the possible roles of microsaccades and drifts in visual perception. *Vision Res.* **118**, 25–30 (2016).
34. G. Ben-Yosef, L. Assif, S. Ullman, Full interpretation of minimal images. *Cognition* **171**, 65–84 (2018).
35. Y. Holzinger, S. Ullman, D. Harari, M. Behrmann, G. Avidan, Minimal recognizable configurations elicit category-selective responses in higher order visual cortex. *J. Cogn. Neurosci.* **31**, 1354–1367 (2019).
36. S. Hochstein, M. Ahissar, View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron* **36**, 791–804 (2002).
37. E. Ahissar, E. Assa, Perception as a closed-loop convergence process. *Elife* **5**, 1–26 (2016).
38. E. Ahissar, A. Arieli, Seeing via miniature eye movements: A dynamic hypothesis for vision. *Front. Comput. Neurosci.* **6**, 89 (2012).
39. A. Roorda *et al.*, Adaptive optics scanning laser ophthalmoscopy. *Opt. Express* **10**, 405–412 (2002).
40. X. Pitkow, H. Sompolinsky, M. Meister, A neural computation for visual acuity in the presence of eye movements. *PLoS Biol.* **5**, e331 (2007).
41. S. Srivastava, G. Ben-Yosef, X. Boix, Minimal images in deep neural networks: Fragile object recognition in natural images. arXiv [Preprint] (2019). https://arxiv.org/abs/1902.03227 (Accessed 8 February 2019).
42. K. Donner, S. Hemilä, Modelling the effect of microsaccades on retinal responses to stationary contrast patterns. *Vision Res.* **47**, 1166–1177 (2007).
43. B. P. Ölveczky, S. A. Baccus, M. Meister, Segregation of object and background motion in the retina. *Nature* **423**, 401–408 (2003).
44. M. Aytekin, J. D. Victor, M. Rucci, The visual input to the retina during natural head-free fixation. *J. Neurosci.* **34**, 12701–12715 (2014).
45. J. L. Puchalla, E. Schneidman, R. A. Harris, M. J. Berry, Redundancy in the population code of the retina. *Neuron* **46**, 493–504 (2005).
46. C.-Y. Chen, Z. M. Hafed, Postmicrosaccadic enhancement of slow eye movements. *J. Neurosci.* **33**, 5375–5386 (2013).
47. J. O. Ramsay, B. W. Silverman, *Functional Data Analysis* (Springer-Verlag, New York, 2004).
48. X. Gao, H. Yan, H.-J. Sun, Modulation of microsaccade rate by task difficulty revealed through between- and within-trial comparisons. *J. Vis.* **15**, 1–15 (2015).
49. F. Soto *et al.*, Efficient coding by midget and parasol ganglion cells in the human retina. *Neuron* **107**, 656–666.e5 (2020).
50. S. A. Baccus, B. P. Ölveczky, M. Manu, M. Meister, A retinal circuit that computes object motion. *J. Neurosci.* **28**, 6807–6817 (2008).
51. A. G. Anderson, K. Ratnam, A. Roorda, B. A. Olshausen, High-acuity vision from retinal image motion. *J. Vis.* **20**, 34 (2020).
52. Y. Burak, U. Rokni, M. Meister, H. Sompolinsky, Bayesian model of dynamic image stabilization in the visual system. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 19525–19530 (2010).
53. C. Wolf, A. C. Schütz, Trans-saccadic integration of peripheral and foveal feature information is close to optimal. *J. Vis.* **15**, 1 (2015).
54. D. C. Burr, Temporal summation of moving images by the human visual system. *Proc R Soc London Ser B Biol Sci.* 211, 321–39 (1981).
55. R. J. Watt, Scanning from coarse to fine spatial scales in the human visual system after the onset of a stimulus. *J. Opt. Soc. Am. A* **4**, 2006–2021 (1987).
56. M. Rucci, R. Iovin, M. Poletti, F. Santini, Miniature eye movements enhance fine spatial detail. *Nature* **447**, 851–854 (2007).
57. M. Gur, Space reconstruction by primary visual cortex activity: A parallel, non-computational mechanism of object representation. *Trends Neurosci.* **38**, 207–216 (2015).
58. M. Rolfs, Attention in active vision: A perspective on perceptual continuity across saccades. *Perception* **44**, 900–919 (2015).
59. J. J. Clark, "Spatial attention and saccadic camera motion" in *Proceedings 1998 IEEE International Conference on Robotics and Automation (Cat. No. 98CH36146)* (IEEE, 1998), pp. 3247–3252.
60. E. Ganmor, M. S. Landy, E. P. Simoncelli, Near-optimal integration of orientation information across saccades. *J. Vis.* **15**, 8 (2015).
61. M. R. Diamond, J. Ross, M. C. Morrone, Extraretinal control of saccadic suppression. *J. Neurosci.* **20**, 3449–3455 (2000).
62. J. Knöll, P. Binda, M. C. Morrone, F. Bremmer, Spatiotemporal profile of peri-saccadic contrast sensitivity. *J. Vis.* **11**, 15 (2011).
63. F. Bremmer, M. Kubischik, K.-P. Hoffmann, B. Krekelberg, Neural dynamics of saccadic suppression. *J. Neurosci.* **29**, 12374–12383 (2009).
64. N. Guyader, A. Chauvin, M. Boucart, C. Peyrin, Do low spatial frequencies explain the extremely fast saccades towards human faces? *Vision Res.* **133**, 100–111 (2017).
65. G. Horev *et al.*, Motor-sensory convergence in object localization: A comparative study in rats and humans. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **366**, 3070–3076 (2011).
66. P. M. Knutsen, M. Pietr, E. Ahissar, Haptic object localization in the vibrissal system: Behavior and performance. *J. Neurosci.* **26**, 8451–8464 (2006).
67. J. Voigts, D. H. Herman, T. Celikel, Tactile object localization by anticipatory whisker motion. *J. Neurophysiol.* **113**, 620–632 (2015).
68. R. S. Marken, Controlled variables: Psychology as the center fielder views it. *Am. J. Psychol.* **114**, 259–281 (2001).
69. N. R. Bowers, A. E. Boehm, A. Roorda, The effects of fixational tremor on the retinal image. *J. Vis.* **19**, 8 (2019).
70. Y. S. Bonneh *et al.*, Motion-induced blindness and microsaccades: Cause and effect. *J. Vis.* **10**, 22 (2010).
71. E. J. Chichilnisky, R. S. Kalmar, Functional asymmetries in ON and OFF ganglion cells of primate retina. *J. Neurosci.* **22**, 2737–2747 (2002).
72. S. J. Schein, Anatomy of macaque fovea and spatial densities of neurons in foveal representation. *J. Comp. Neurol.* **269**, 479–505 (1988).