VISGRAF Technical Report

Eye Movements and Stereoscopic Depth Perception

From oculomotor control to active depth inference

Abstract. Stereoscopic depth perception is not a passive comparison of two retinal images. It is an active process in which saccades select targets, vergence aligns the two foveae, pursuit and stabilization reflexes preserve useful retinal information, and fixational movements maintain sensitivity during attempted steady gaze. This report organizes these movements by function and control regime, derives the relationship among fixation distance, vergence, and binocular disparity, and describes how correspondence and disparity signals support relative and metric depth judgments. It then connects these mechanisms to combined saccade–vergence behavior, vergence–accommodation coupling, and the design of stereoscopic and extended-reality systems. The resulting view is a closed perception–action loop: gaze changes the sensory evidence from which depth is estimated, and depth errors in turn drive gaze.

Luiz Velho – VISGRAF Laboratory, IMPA

GPT – Open.AI (collaborator)

September 2026

Contents

1 Introduction **1**

2 Functional taxonomy of eye movements **1**

2.1 Saccades: target acquisition 1

2.2 Pursuit and retinal-image stabilization 3

2.3 Vergence: binocular alignment in depth 3

2.4 Fixational movements 3

3 Stereo geometry around fixation **3**

4 From correspondence to disparity **4**

5 The active fixation–depth loop **5**

6 Coupled control mechanisms **6**

6.1 Version and vergence 6

6.2 Vergence and accommodation 6

6.3 Feedback and feed-forward distinctions 6

7 Implications for XR and active vision **6**

7.1 Geometric and temporal consistency 7

7.2 Comfort and focus cues 7

7.3 Eye tracking and foveated computation 7

8 A computational interpretation **8**

9 Scope, qualifications, and conclusions **8**

1 Introduction

Human stereopsis begins with two horizontally displaced views of the world, but binocular image differences alone do not explain stable three-dimensional perception. The eyes continuously redirect and stabilize the retinal images. Those actions determine which features fall on the foveae, which disparities are available to cortical mechanisms, and which depth is treated as the current reference. Eye movements are therefore part of the computation of depth, not only a motor response after depth has been recovered.

This report develops that claim at three mutually dependent levels:

1. **Oculomotor function:** saccades, smooth pursuit, vestibulo-ocular and optokinetic responses, vergence, and fixational movements have different objectives and dynamics [1, 2].
2. **Stereo computation:** matching left- and right-eye structure produces binocular disparity, which specifies depth relative to fixation [3, 4].
3. **Active control:** target selection, version, vergence, accommodation, and image stabilization operate as coupled feed-forward and feedback processes rather than a strictly serial pipeline.

The central distinction is between *gaze direction* and *gaze depth*. Conjugate or version movements change the direction of both eyes together. Disjunctive or vergence movements change the angle between the visual axes and therefore the distance at which they intersect. Natural gaze shifts usually contain both components.

Central idea

A saccade places a selected target near the foveae; vergence reduces its binocular disparity; residual disparities around that fixation encode relative depth. The process repeats as attention and gaze sample the scene.

2 Functional taxonomy of eye movements

Table 1 summarizes the movements most relevant to stereoscopic vision. Numerical ranges are deliberately approximate: amplitude, task, predictability, age, and measurement method all affect the observed dynamics.

2.1 Saccades: target acquisition

Saccades are rapid gaze shifts. Their peak velocity rises with movement amplitude, a relationship known as the *main sequence* [5]. During a typical saccade, visual sensitivity is reduced and spatial localization is transiently altered; "vision is suppressed" is therefore a useful shorthand but not a literal absence of visual processing. Because the movement is too fast for continuous retinal feedback to reshape its trajectory substantially, the flight is commonly modeled as ballistic. Feedback still matters before launch, through prediction, and after landing, through corrective saccades and adaptation.

For stereopsis, the decisive effect is foveation. Peripheral evidence and attention select a target; the saccade brings its projections toward the two foveae, where acuity and disparity discrimination are best. When the target also changes in depth, the two eyes need not execute identical rotations.

Table 1: Functional taxonomy of ocular movements in binocular viewing.

Mechanism	Coordination	Primary role	Stereo relevance
Saccade	Mostly conjugate; can be disconjugate	Rapidly redirects the foveae to a selected location	Acquires candidate correspondences and changes the reference fixation; sensitivity is reduced around the movement
Smooth pursuit	Mostly conjugate	Keeps a moving target near the fovea	Preserves correspondence for moving surfaces and supports motion-in-depth tracking
VOR	Compensatory	Counters head rotation and translation	Stabilizes binocular input during self-motion; interacts with vergence for near targets
OKN	Conjugate slow phase plus resetting quick phase	Responds to sustained large-field retinal slip	Stabilizes the scene during prolonged visual motion
Vergence	Disjunctive	Aligns the two visual axes in depth	Drives fixation disparity toward a fusible value and supplies an extraretinal distance cue
Microsaccades, drift, tremor	Small fixational motion	Maintains and corrects attempted fixation	Modulates fine spatial and disparity signals; prevents a perfectly static retinal input
Accommodation	Lens response, not an eye rotation	Focuses the retinal image	Neurally coupled to vergence and normally provides a consistent distance cue

2.2 Pursuit and retinal-image stabilization

Smooth pursuit reduces the image velocity of a moving target and depends strongly on visual motion and prediction [6]. The vestibulo-ocular reflex (VOR) counters head motion with short latency, whereas the optokinetic response (OKN) responds to sustained large-field image motion. Their common contribution to stereo is to keep the left- and right-eye signals sufficiently stable for correspondence and disparity estimation. During translational head motion, the required compensation is depth dependent, so stabilization and depth estimation cannot be separated cleanly.

2.3 Vergence: binocular alignment in depth

Convergence rotates the visual axes inward for nearer targets; divergence reduces that rotation for farther targets. Fusional vergence is driven by binocular disparity error and is naturally described as a feedback controller. Vergence is slower than a same-amplitude saccade in isolation, but natural combined gaze shifts can contain fast disjunctive components [7].

Motor-command copies and proprioceptive signals make vergence an extraretinal cue to viewing distance. This cue is most informative at near distances and is not a perfect, image-independent range sensor. The brain combines it with accommodation, motion parallax, perspective, familiar size, and other cues.

2.4 Fixational movements

The eyes never become perfectly stationary. Microsaccades, slow drift, and high-frequency tremor continually transform the retinal input. These motions help counter neural adaptation, correct fixation error, and shape fine spatial signals [8, 9]. Their exact causal contribution to stereo hyperacuity remains an active research topic; it is safest to say that fine depth judgments occur in the presence of, and are modulated by, fixational motion.

3 Stereo geometry around fixation

Let the eye centers be separated by an interocular baseline b, with a symmetric fixation point at distance Z_f on the head-centered median plane. The total vergence angle is

$$V_f = 2\arctan\left(\frac{b}{2Z_f}\right). \tag{1}$$

For distances large relative to the baseline, $V_f \approx b/Z_f$ radians. Thus the angle changes rapidly in near space and only weakly at long range.

Now consider a second point on the same median axis at distance Z. Its angular disparity relative to fixation is exactly

$$\delta = 2\arctan\left(\frac{b}{2Z}\right) - 2\arctan\left(\frac{b}{2Z_f}\right), \tag{2}$$

under the convention that crossed disparity is positive. In the small-angle regime,

$$\delta \approx b\left(\frac{1}{Z} - \frac{1}{Z_f}\right), \qquad Z \approx \frac{1}{Z_f^{-1} + \delta/b}. \tag{3}$$

Equation (3) makes three facts explicit:

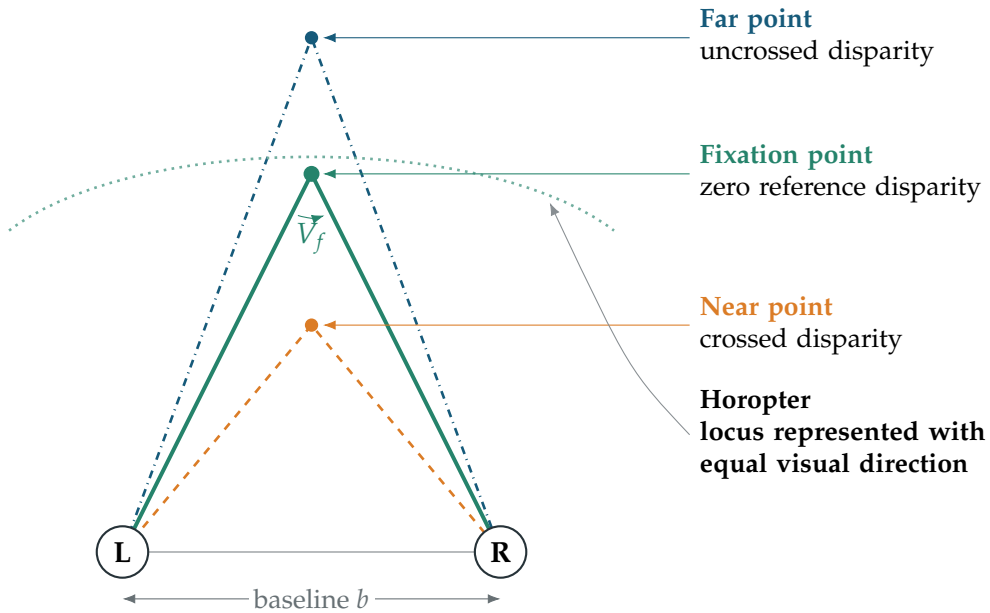


Figure 1: Symmetric binocular geometry. Labels are separated from the rays so the vergence angle, fixation point, and disparity sign remain legible. The drawn horopter is schematic; the empirical horopter depends on viewing conditions.

- $Z < Z_f$ produces positive, crossed disparity; $Z > Z_f$ produces negative, uncrossed disparity.
- disparity is fundamentally a *relative* depth signal around the current fixation;
- converting disparity into distance requires a scale, supplied partly by fixation distance or vergence and by other depth cues.

Depth sensitivity is nonlinear. For a fixed disparity uncertainty, the first-order propagation

$$\left|\frac{\partial Z}{\partial \delta}\right| \approx \frac{Z^2}{b} \tag{4}$$

shows why metric depth from disparity becomes progressively less precise with distance.

The theoretical horopter is the locus of points with the same binocular visual direction as fixation. A finite neighborhood around it can be fused into a single percept; this is commonly associated with Panum's fusional area. Its width is not a universal constant: it varies with eccentricity, spatial frequency, contrast, exposure, and stimulus structure [10]. Outside the fusion range, diplopia can occur and can itself provide coarse ordinal depth information.

4 From correspondence to disparity

Before disparity can be interpreted, the visual system must establish which left-eye and right-eye image structures arise from the same scene element. This is the stereo correspondence problem. Random-dot stereograms demonstrate that stereoscopic form can emerge without identifiable monocular objects [3]; correspondence therefore operates on image structure below the level of object recognition.

Computational accounts often invoke similarity, epipolar geometry, uniqueness, continuity, ordering, and multiscale regularity. These are useful models, not inviolable laws of perception. Repetitive texture, occlusions, transparency, specularities, and half-occluded regions create genuine ambiguity. Coarse-to-fine processing expands the capture range while fine-scale mechanisms preserve precision.

Neurons sensitive to binocular disparity appear in primary visual cortex and in multiple extrastriate areas. Classic recordings described tuned-zero, near, far, and tuned inhibitory response classes [11]; contemporary accounts emphasize a distributed population code rather than a single labeled depth channel [4]. Importantly, a neural response to retinal disparity is not by itself equivalent to a conscious estimate of scene depth. Eye position, context, surface organization, motion, and priors all influence the percept.

5 The active fixation–depth loop

Natural stereoscopic perception can be organized as a recurrent loop rather than a one-pass pipeline (Figure 2).

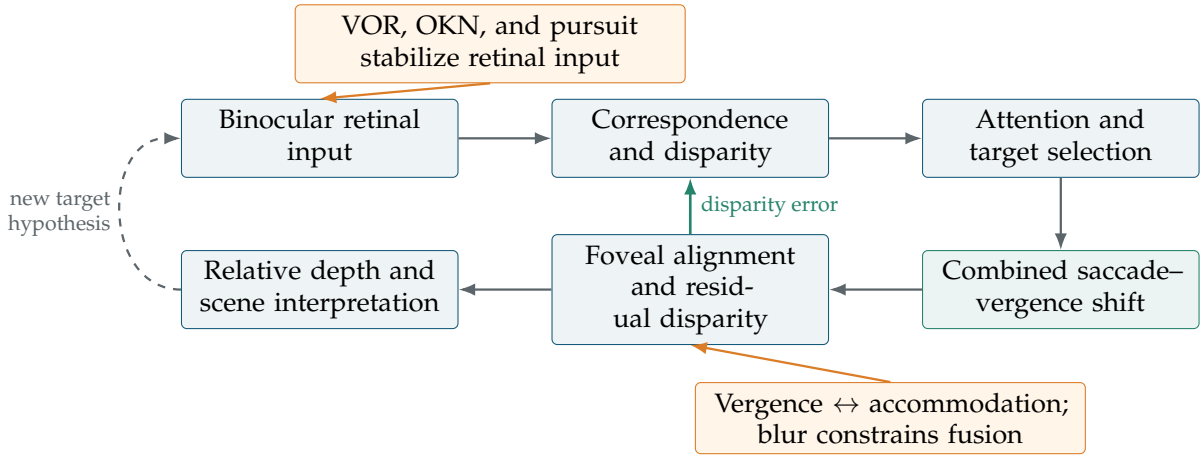


Figure 2: Perception–action organization of stereoscopic viewing. Green denotes the principal disparity-driven feedback path; orange denotes stabilization and focus mechanisms that constrain the binocular signal.

1. **Candidate selection.** Peripheral contrast, motion, task goals, and sometimes coarse disparity mark a potentially useful target.
2. **Gaze transfer.** A version command redirects gaze while a vergence command changes gaze depth. The movements may overlap in time.
3. **Fusion and refinement.** Disparity-driven vergence reduces the target's fixation error; corrective movements handle residual direction error.
4. **Local depth readout.** Once the target is fused, neighboring disparities describe surfaces in front of and behind the fixation reference.
5. **Recurrent sampling.** A new attentional decision moves the high resolution sampling window elsewhere.

Successive fixations provide complementary samples, but the phrase "the brain stitches a complete metric depth map" should be treated as a computational metaphor. Perception is stable across saccades without requiring a pixel-dense, globally precise internal reconstruction. The retained representation is likely task dependent and combines object, surface, spatial, and sensorimotor information.

6 Coupled control mechanisms

6.1 Version and vergence

Hering's classical description decomposes binocular motion into a conjugate version command and a disjunctive vergence command. Algebraically, horizontal eye positions can be written

$$E_L = H + \frac{V}{2}, \qquad E_R = H - \frac{V}{2}, \tag{5}$$

where H is the conjugate component and V the vergence component, subject to a chosen sign convention. This decomposition remains valuable for measurement and control even though monocular pathways and unequal saccades show that biological coordination is more flexible than two perfectly shared commands.

Combined direction-and-depth shifts are not simply a completed saccade followed by an isolated slow vergence response. Their trajectories overlap, and vergence kinematics change in the presence of a saccade. However, the interpretation that the saccadic system always "facilitates the vergence command" is debated; some data can be explained by shared target selection and nonlinear plant or control dynamics rather than a single multiplicative boost [7].

6.2 Vergence and accommodation

The near triad links convergence, accommodation, and pupil constriction. Neural cross-links allow accommodative demand to influence vergence and vergence demand to influence accommodation. In the natural world, disparity and focus cues usually indicate compatible distances. Conventional stereoscopic displays break that relationship: screen blur and accommodation specify the display surface, whereas binocular disparity may specify a virtual object in front of or behind it.

This *vergence–accommodation conflict* can slow visual performance, reduce stereoacuity, distort perceived depth, and increase discomfort [12, 13]. Individual differences, duration, content, and display geometry matter, so conflict is a risk factor rather than a deterministic prediction of symptoms.

6.3 Feedback and feed-forward distinctions

Fusional vergence is a canonical closed-loop response: a disparity estimate provides an error signal that the motor system attempts to reduce. A saccade is largely preprogrammed during its brief flight, but the saccadic system is not globally open loop. Prediction shapes the command; visual and proprioceptive feedback evaluate the landing; learning adapts later commands. Pursuit combines visual feedback with strong predictive components, while the VOR uses inertial signals and is recalibrated by retinal slip. The useful distinction is therefore the *dominant feedback timescale*, not a rigid binary classification.

7 Implications for XR and active vision

The physiology translates into concrete design constraints for head-mounted and stereoscopic systems.

7.1 Geometric and temporal consistency

- Render each view from the tracked eye position and use the viewer's interpupillary distance when accurate scale is important.
- Keep vertical disparity and left–right timing errors small; they consume fusion capacity without representing intended scene depth.
- Minimize motion-to-photon latency. Head motion accompanied by stale imagery conflicts with VOR-based stabilization.
- Avoid abrupt disparity steps at high-contrast fixation targets. A content transition can demand a vergence response before fusion is established.

7.2 Comfort and focus cues

Comfort cannot be specified by one disparity limit, because fusional range varies with the observer and stimulus. A practical system should manage disparity over time, preserve scale consistency, and place high-priority content near a comfortable binocular working range. Varifocal, multifocal, and light-field architectures attempt to reduce focus-cue error by changing or sampling the focal distance; each trades optical complexity, field of view, resolution, latency, and calibration burden [14].

7.3 Eye tracking and foveated computation

Eye tracking exposes both an opportunity and a measurement problem. Gaze-contingent rendering can concentrate samples near the fovea, but a binocular system must distinguish version from vergence and account for tracker bias. Small angular errors can produce large inferred depth errors through Equation (4). During saccades, prediction and conservative transition regions are preferable to assuming an instantaneous, perfectly known fixation. Vergence estimates are often more useful as noisy state variables within a temporal filter than as direct per-frame range measurements.

Table 2: Mapping biological functions to engineering concerns.

Biological function	Engineering analogue	Failure mode
Saccadic target acquisition	Gaze prediction; foveated region relocation	Visible quality transition, incorrect foveal center, post-saccadic blur
Vergence alignment	Binocular camera/display calibration	Scale error, fusion difficulty, unstable virtual depth
VOR stabilization	Low-latency head tracking and reprojection	World-locked content appears to swim or lag
Accommodation coupling	Varifocal/multifocal/light-field optics	Discomfort and slower refocusing under cue conflict
Fixational motion	High-resolution gaze estimation and temporal sampling	Treating noisy gaze as a static ground-truth point
Correspondence	Stereo matching and confidence estimation	False matches near occlusion, repetition, or specularity

8 A computational interpretation

For an active stereo agent, the biological account suggests maintaining a state that includes at least head pose, left and right eye orientation, fixation hypothesis, disparity confidence, and target identity. A compact cycle is:

$$\begin{aligned}\text{predict sensory evidence} &\rightarrow \text{select an informative target} \\ &\rightarrow \text{execute version + vergence} \rightarrow \text{update depth and uncertainty.}\end{aligned} \tag{6}$$

This framing avoids three common modeling errors:

1. treating binocular depth as independent of eye pose;
2. treating gaze as a single ray when the two visual axes may not intersect exactly because of fixation disparity and measurement noise;
3. treating a depth estimate as equally reliable at all ranges, despite the quadratic sensitivity in Equation (4).

A probabilistic estimator can represent these uncertainties explicitly. Rather than triangulating from one measured vergence angle, it can fuse binocular image evidence, motor state, temporal continuity, and scene priors. In robotics and XR, the most useful biological lesson is not to copy every neural mechanism, but to couple sensing and action: choose gaze changes that improve the next depth estimate.

9 Scope, qualifications, and conclusions

Several statements in informal accounts need qualification. Vergence contributes an absolute-distance cue but is neither perfectly accurate nor independent of visual context. Saccades are ballistic over short intervals but participate in an adaptive feedback system over longer intervals. Panum's area is stimulus dependent, not a fixed geometric band. Fixational movements accompany fine stereo judgments, but their precise functional role cannot be reduced to a single mechanism. Finally, stable perception across gaze shifts does not require a literal dense depth map.

With those qualifications, the overall mechanism is clear. Conjugate movements select and stabilize image content; vergence establishes a depth reference; correspondence mechanisms estimate disparity; extraretinal and monocular cues help scale and disambiguate the result; and recurrent gaze shifts gather the evidence needed for the current task. Stereoscopic perception is therefore best understood as an active, coupled system in which movement both depends on and creates information about three-dimensional structure.

References

[1] R. John Leigh and David S. Zee. *The Neurology of Eye Movements*. Oxford University Press, New York, 5 edition, 2015. doi: 10.1093/med/9780199969289.001.0001.

[2] Mohit Agarwal, John L. Ulmer, Tushar Chandra, Andrew P. Klein, Leighton P. Mark, and Suyash Mohan. Imaging correlates of neural control of ocular movements. *European Radiology*, 26(7):2193–2205, 2016. doi: 10.1007/s00330-015-4004-9.

[3] Bela Julesz. Binocular depth perception without familiarity cues. *Science*, 145(3630):356–362, 1964. doi: 10.1126/science.145.3630.356.

[4] Bruce G. Cumming and Gregory C. DeAngelis. The physiology of stereopsis. *Annual Review of Neuroscience*, 24:203–238, 2001. doi: 10.1146/annurev.neuro.24.1.203.

[5] A. Terry Bahill, Michael R. Clark, and Lawrence Stark. The main sequence, a tool for studying human eye movements. *Mathematical Biosciences*, 24(3–4):191–204, 1975. doi: 10.1016/0025-5564(75)90075-9.

[6] Richard J. Krauzlis. Recasting the smooth pursuit eye movement system. *Journal of Neurophysiology*, 91(2):591–603, 2004. doi: 10.1152/jn.00801.2003.

[7] Olivier A. Coubard. Saccade and vergence eye movements: A review of motor and premotor commands. *European Journal of Neuroscience*, 38(10):3384–3397, 2013. doi: 10.1111/ejn.12356.

[8] Susana Martinez-Conde, Stephen L. Macknik, and David H. Hubel. The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience*, 5(3):229–240, 2004. doi: 10.1038/nrn1348.

[9] Michele Rucci and Jonathan D. Victor. The unsteady eye: An information-processing stage, not a bug. *Trends in Neurosciences*, 38(4):195–206, 2015. doi: 10.1016/j.tins.2015.01.005.

[10] Ian P. Howard and Brian J. Rogers. *Perceiving in Depth, Volume 2: Stereoscopic Vision*. Oxford University Press, New York, 2 edition, 2012. doi: 10.1093/acprof:oso/9780199764150.001.0001.

[11] Gian F. Poggio and Burkhart Fischer. Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *Journal of Neurophysiology*, 40(6):1392–1405, 1977. doi: 10.1152/jn.1977.40.6.1392.

[12] David M. Hoffman, Ahna R. Girshick, Kurt Akeley, and Martin S. Banks. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision*, 8(3):33.1–33.30, 2008. doi: 10.1167/8.3.33.

[13] Martin S. Banks, Jenny C. A. Read, Robert S. Allison, and Simon J. Watt. Stereoscopy and the human visual system. *SMPTE Motion Imaging Journal*, 121(4):24–43, 2012. doi: 10.5594/j18173.

[14] Gregory Kramida. Resolving the vergence–accommodation conflict in head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics*, 22(7):1912–1931, 2016. doi: 10.1109/TVCG.2015.2473855.