ScanDy: Simulating Realistic Human Scanpaths in Dynamic Real-World Scenes

Nicolas Roth[1, 2], **Martin Rolfs**[1, 3, 4], **and Klaus Obermayer**[1, 2, 4]

[1]Exzellenzcluster Science of Intelligence, Technische Universität Berlin, Marchstraße 23, 10587 Berlin, Germany
[2]Institut für Softwaretechnik und Theoretische Informatik, Technische Universität Berlin, Marchstraße 23, 10587 Berlin, Germany
[3]Institut für Psychologie, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489 Berlin, Germany
[4]Bernstein Center for Computational Neuroscience Berlin, Philippstraße 13, 10115 Berlin, Germany

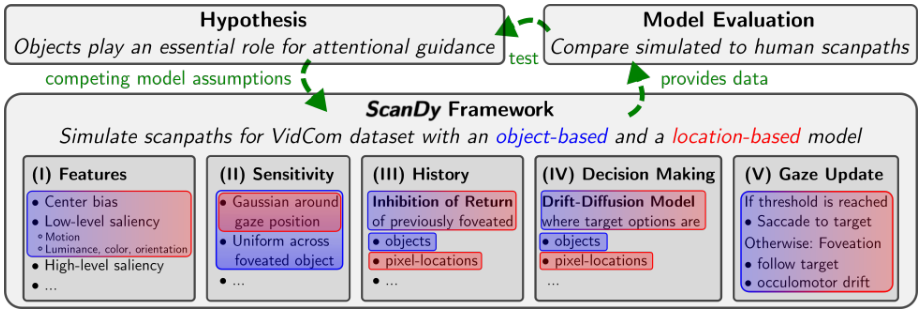

Fig. 1. Overview of the *ScanDy* framework, which is designed to quantitatively test various hypotheses about eye movement behavior in dynamic real-world scenes. We demonstrate the potential of this framework by evaluating an hypothesis on the importance of object-based attention. Blue and red indicate the different implementations of each module (I-V) for the competing object-based and location-based model.

How we move our eyes fundamentally shapes how we perceive the world around us, and vice versa. Psychophysical experiments have uncovered multiple mechanisms that underlie our gaze behavior, like bottom-up salience, inhibition of return, or object based attention. To what extent these effects actually influence visual exploration behavior when observing ecologically valid scenes is a long-standing debate. We present *ScanDy*, a modular computational framework for scanpath simulation under free-viewing conditions. In contrast to other models, we simultaneously predict where and when the eyes move in dynamic real-world scenes.

Figure 1 shows the modular architecture of *ScanDy* and how it can be used to assess the importance of objects for exploration behavior. Saccadic gaze shifts (V) are modeled as a sequential *decision making* (IV) process between potential targets. Each decision depends on (I) the guiding *features* in the scene, (II) the visual *sensitivity* depending on the current gaze position, and (III) the previous scanpath *history*. We construct two competing models within this framework, a location-based model, in which saccades target pixel-locations and objects do not play any role, and a object-based model, in which target selection is based on semantic objects. Both models use the same frame-wise feature maps as input, in this case based on low-level features (motion, color, luminance, and orientation) (1) and an anisotropic Gaussian to account for the central viewing bias (2). The decrease of visual sensitivity to the target with eccentricity is approximated with a Gaussian, while in the object-based model, the part falling within the currently foveated object is replaced by uniform sensitivity. The feature and sensitivity maps are multiplied to a frame-wise evidence measure that is accumulated over time for each potential target. This evidence is combined with an inhibition of return factor which is set to one if a target is foveated and linearly decreases over time. The result for each target is then used as its drift rate in a drift-diffusion model (DDM). As soon as a target reaches a fixed DDM threshold, the gaze position is shifted. Otherwise, gaze remains on the current target, resulting in fixation or smooth pursuit depending on the target's movement.

The implemented mechanisms have a small number of well-interpretable parameters. These are fitted using evolutionary optimization (3), such that the simulated scanpaths for the VidCom dataset (4) reproduce the fixation duration and saccade amplitude distributions of the ground-truth eye-tracking data. The performance of each model is then evaluated based on the spatial and temporal fixation behavior. By analyzing the functional aspects of the predicted scanpaths (5), we find that the object-based model accurately reflects the balance of detection, inspection, and revisits of objects, while the location-based model primarily explores the background of the scene. This reveals the importance of objects for attentional guidance, confirming our hypothesis.

We hope that this framework will motivate new hypotheses on attention allocation in dynamic real-world scenes, and help to bridge the gap between theoretical knowledge about attentional mechanisms and actual viewing behavior.

REFERENCES

1. J. Molin, R. Etienne-Cummings, and E. Niebur. How is motion integrated into a proto-object based visual saliency model? In *49th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2015.
2. A. Clarke and B. Tatler. Deriving an appropriate baseline for describing fixation behaviour. *Vision research*, 102:41–51, 2014.
3. C. Cakan, N. Jajcay, and K. Obermayer. neurolib: a simulation framework for whole-brain neural mass modeling. *Cognitive Computation*, pages 1–21, 2021.
4. Z. Li, S. Qin, and L. Itti. Visual attention guided bit allocation in video compression. *Image and Vision Computing*, 29(1):1–14, 2011.
5. M. Linka and B. de Haas. Detection, inspection and re-inspection: A functional approach to gaze behavior towards complex scenes. *Journal of Vision*, 21(9):1971–1971, 2021.