Distinct Roles for Eye and Head Movements in Selecting Salient Image Parts during Natural Exploration

Wolfgang Einhäuser,[a] Frank Schumann,[b] Johannes Vockeroth,[c] Klaus Bartl,[c] Moran Cerf,[d] Jonathan Harel,[e] Erich Schneider,[c] and Peter König[b]

[a] *Department of Neurophysics, Philipps-University, Marburg, Germany*

[b] *Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany*

[c] *Clinical Neurosciences, University of Munich Hospital, Munich, Germany*

[d] *Computation and Neural Systems, California Institute of Technology, Pasadena, California, USA*

[e] *Division of Electrical Engineering, California Institute of Technology, Pasadena, California, USA*

Humans adjust gaze by eye, head, and body movements. Certain stimulus properties are therefore elevated at the gaze center, but the relative contribution of eye-in-head and head-in-world movements to this selection process is unknown. Gaze- and head-centered videos recorded with a wearable device (EyeSeeCam) during free exploration are reanalyzed with respect to responses of a face-detection algorithm. In line with results on low-level features, it was found that face detections are centered near the center of gaze. By comparing environments with few and many true faces, it was inferred that actual faces are centered by eye and head movements, whereas spurious face detections ("hallucinated faces") are primarily centered by head movements alone. This analysis suggests distinct contributions to gaze allocation: head-in-world movements induce a coarse bias in the distribution of features, which eye-in-head movements refine.

Key words: **eye movements; face; natural scenes; real-world behavior; salience**

Introduction

The question as to which stimulus properties control the direction of human gaze under real-world conditions has been a subject of research for decades.[1–6] While early studies demonstrated a profound influence of task on fixation behavior,[1,2] many later studies assessed the role of stimulus features on driving attention.[3–6] Typically, such studies focus on low-level features (color, luminance, orientation) and measure their elevation at the gaze center or incorporate them into a model of saliency[7] to predict fixated locations.[6] Recently, however, the role of higher-level scene structure such as faces has received increasing interest[8] and attention models have partly been reinterpreted as preattentive models of a scene's object content.[9] Whereas studies focusing on stimulus features typically present photographs and record eye movements with the head fixed, studies with less restricted settings usually focus on a specific task, such as making tea, preparing food,[10,11] washing hands,[12] or throwing and catching a ball.[13] By using the novel EyeSeeCam recording setup[14] we are able to pursue a complementary strategy: we record large amounts of data without a specific task,

Address for correspondence: Wolfgang Einhäuser, FB Physik – Neurophysik, Renthof 7, 35032 Marburg, Germany. wet@physik.uni-marburg.de

Basic and Clinical Aspects of Vertigo and Dizziness: Ann. N.Y. Acad. Sci. 1164: 188–193 (2009).
doi: 10.1111/j.1749-6632.2008.03714.x © 2009 New York Academy of Sciences.

asking observers to perform natural free exploration. Here we reanalyze previously recorded data[15] with respect to face detections and address the question to what extent eye-in-head and head-in-world movements play a distinct role in gaze allocation.

Methods

Recordings

We use the EyeSeeCam recording setup[14] (Fig. 1A) for simultaneous recordings of gaze- and head-centered videos during free exploration. Most of the data have been used earlier,[15] but none of the analysis with respect to face detection has been reported elsewhere. For the purpose of the present study, we separated the recordings into nine groups (hereafter called "environments") on the basis of true face content. This was done by visual inspection prior to all analysis and not changed thereafter. We estimated the fraction of video frames that contains true faces by manual inspection of 100 randomly chosen frames in each environment. Five of the nine environments are recorded outdoors and include (1) a crowded Munich shopping street with plenty of people present (16 min, 100% with faces), (2) city squares and streets in Munich that are populated (84 min, 80%) (3) similar squares and streets with fewer people present (25 min, 41%), (4) a forest, park, and a residential area with very few people around (181 min, 12%), (5) a Californian beach and desert in winter, with virtually no encounters with other people (81 min, 2%). Indoor environments include (6) a conference hall during a poster session of the 2007 Society for Neuroscience meeting (21 min, 95%), (7) the same conference hall at the end of the day (2 min, 33%), (8) a hospital, where people typically occur at far distance (65 min, 28%), (9) as well as indoor environments with few people: the main building of a university, the office of one of the authors, and a Munich art museum (102 min, 24%).

Observers

Seven volunteers (age 25–40, 5 male) participated in the study. All had normal or corrected-to-normal vision and were accustomed to wearing the setup. With the exception of environment (7) each environment contains data from at least two observers. In environment (1), observers were instructed to interact with people; for all other environments, observers were asked to behave "naturally." All procedures conformed with national and institutional guidelines for experiments with human subjects and with the Declaration of Helsinki.

Face Detection

The location of faces was determined by using the Viola Jones face detector.[16] In line with Ref. 8, the implementation of Intel's "OpenCV" library[17] was used with standard settings and default training set. Note that systematic differences between this standard set and our recordings preclude statements on the quality of the algorithm per se, which is, however, not the interest of the present study.

Results

For head-centered and gaze-centered videos, we determine face locations by using a standard face-detection algorithm,[16] returning location and the size of the rectangular bounding box of each face in a given video frame (Fig. 1B). The large amount of data prohibits manual verification of the detector's results. Qualitatively, we observe that detector misses (false negatives, a true face is not detected) often result from partial occlusion, faces at a large distance, or from nonfrontal viewing angles. Generally, false-positives (detection of a non-face as a face) seem more abundant in our material. As a rough estimate for false-positives, we find that 33.7% of head-centered frames in the desert/beach environment (5) elicit a detector response, while only about 2% contain a true

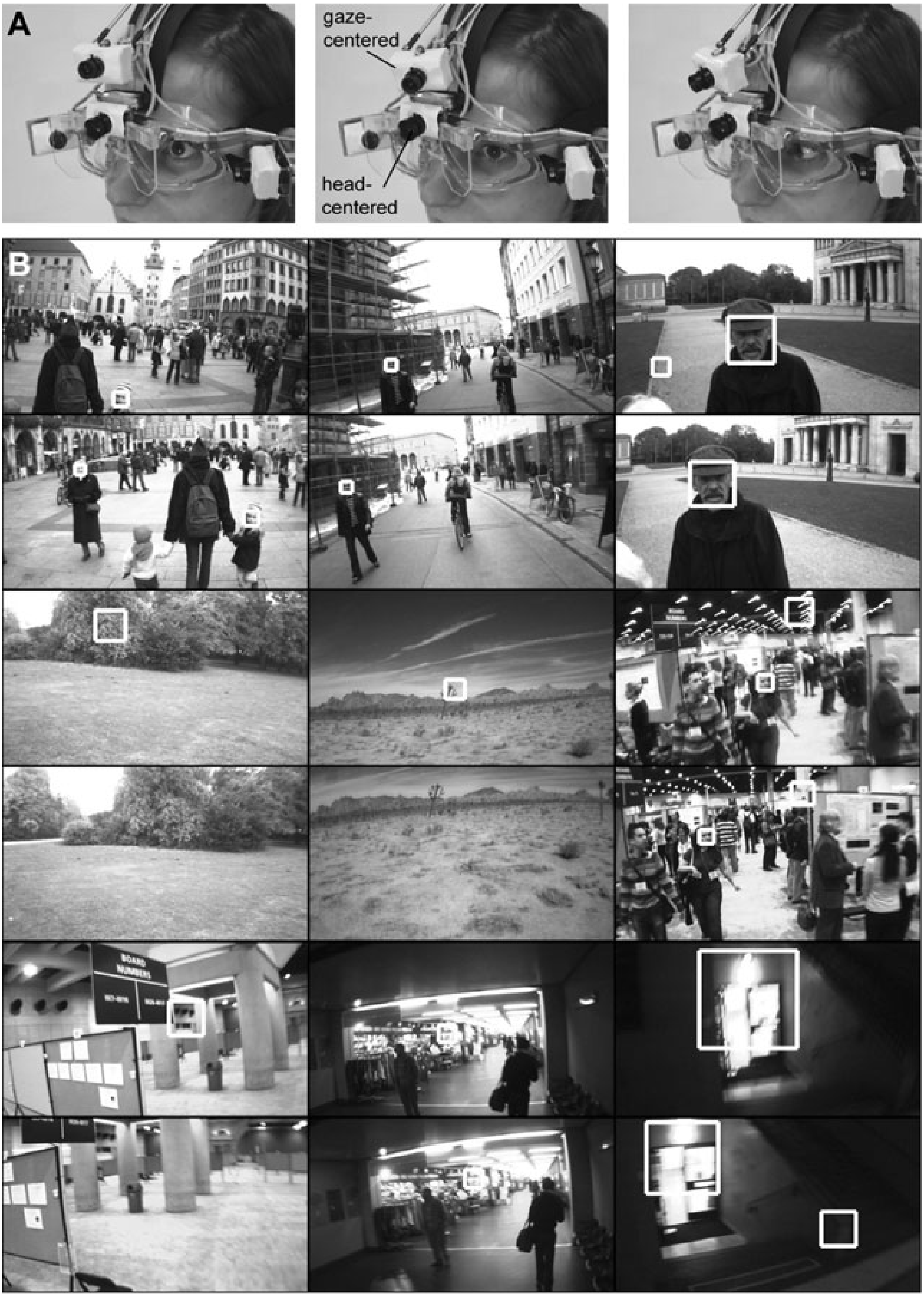

Figure 1. (**A**) The EyeSeeCam recording setup. The head camera is fixed in head coordinates, while the pivotable-gaze camera is automatically aligned to the direction of gaze in real-time. (Adapted from Schumann *et al.*[15]) (**B**) Representative frame of each environment (environment (1), upper left; (3), upper right; (9), lower right).*Top:* Gaze-centered recordings. *Bottom.* head-centered recordings. *White boxes* denote face detections.

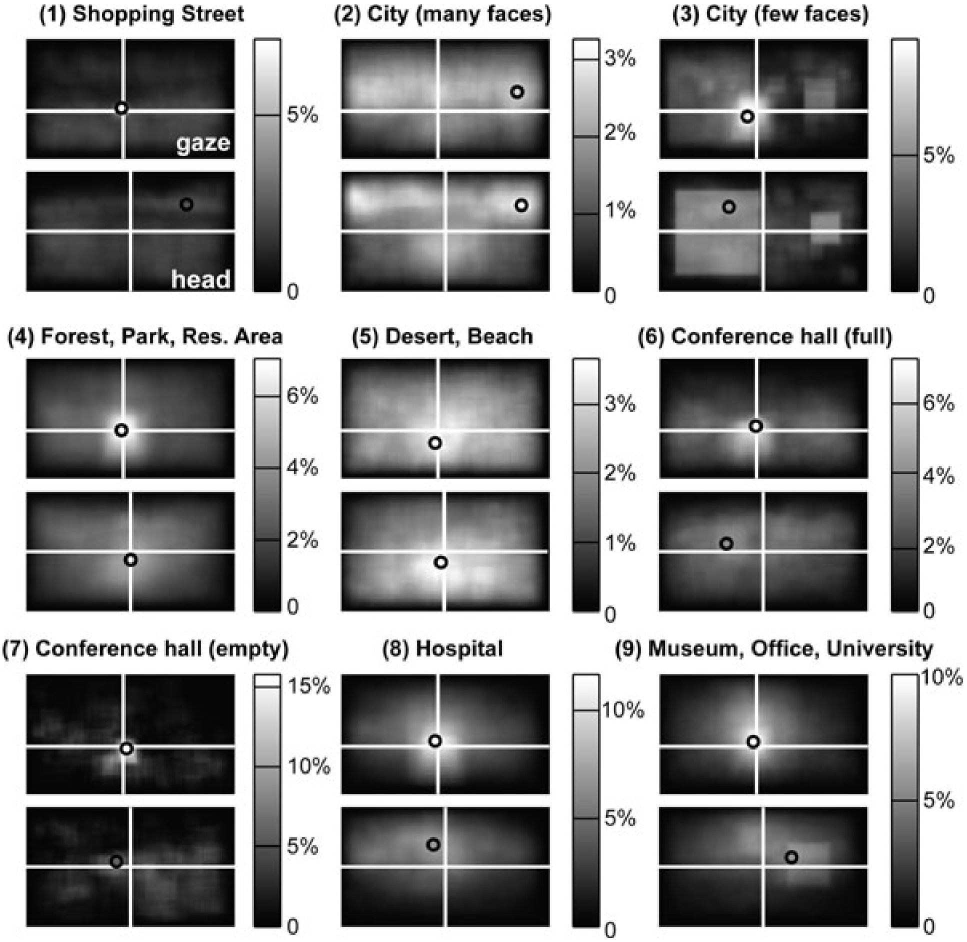

Figure 2. Face-detection maps for the 9 different environments. Each map represents the full field of view of the cameras (64° × 41°). In each panel, gaze-centered recordings are on top, head-centered on bottom, *black circles* denote location of peak. Crossed lines mark calibrated center-of-gaze (constant offset of about 5° from the image center) and center-of-head camera, respectively. Scales are identical for both maps of each environment, and report the percentage of face detections that cover the respective location.

face. Consequently, in environments with low face content, we can safely assume that false-positives dominate the face-detector responses.

To quantify the location of face detections in gaze-centered and head-centered coordinates, we compute a face map for each frame. In this map, each pixel gets assigned the value corresponding to the number of face-detection bounding boxes overlapping with it. For each environment these maps are summed and divided by the number of total face detections. The resulting map represents the fraction of face detections that occur at this particular location in either head-centered or gaze-centered coordinates (Fig. 2).

We compare the location of the peaks in face maps between head-centered and gaze-centered coordinates. The gaze-centered face maps of all but one environment peak within 0.5° to 4.4° of the center of gaze. The exception is formed by environment (2) (city, many faces), for which a broad band of face detections is observed at the horizontal midline, without a clear peak in the horizontal direction (Fig. 2, top middle). For the remaining eight environments, the mean distance from the center is

2.0° ± 1.3°. A comparison of this number with the cameras' full field of view (64° × 41°) suggests that gaze typically centers on face detections, irrespective of the environment. In head-centered coordinates, face detections are less centered than in gaze-centered coordinates. When again excluding environment (2), the mean distance between the center and peak of the face map amounts to 9.8° ± 6.1°. This is significantly larger than in gaze-centered coordinates ($P = .003$, t-test). In conclusion, although face detections are already close to the center of head-centered maps, eye-in-head movements tend to further facilitate centering face detections in retinal (i.e., gaze-centered) coordinates.

In addition to being more centered, visual inspection suggests that peaks in gaze-centered maps are typically more pronounced than in head-centered maps. We quantify this by measuring the height of the peak in each map. Since the maps are normalized by the total amount of face detections, the peak would be 100% if all faces would fall on the same location (in practice slightly lower, due to frames with multiple detections). The lower bound is given by randomly putting face detections of the same size as the real data at random locations in each frame ("baseline map").[a] We find this lower bound to be 3.2% ± 0.9% (mean ± SD over those nine environments) for head-centered and 3.2% ± 1.1% for gaze-centered maps. In all environments, the peaks in both the head-centered map and the gaze-centered map exceed the lower bound, averaging to 4.9% ± 1.6% and 8.3% ± 4.0%, respectively. These means of true peak heights are significantly above the baseline mean (t-test: $P = .01$ and $P = .002$, respectively). This rules out that the observed peaks are an artifact of boundary effects. With the exception of environment (2) and (5), peaks are higher in gaze-centered than in head-centered maps. Across all nine environments, this difference is significant ($P = .03$, t-test). This implies that head-centered face maps already exhibit a significant peak, but peaks are more pronounced in gaze coordinates than in a head-centered coordinate frame. The differences between peak location and peak height in gaze-centered as compared to head-centered maps suggest that face detections are predominantly centered by eye-in-head rather than by head-in-world movements.

In environment (5) (desert/beach) there are faces in about 2% of the frames, in contrast to false-positive face detections in about 34% of the head-centered and 26% of the gaze-centered frames. This implies that false-positives dominate the face map. Interestingly, for this environment, the gaze-centered and head-centered maps are remarkably similar (Fig. 2). The peak height is 3.6% for gaze- and 3.7% for head-centered coordinates (baseline: 2.8% for both), and faces are even slightly more centered in head (4.0°) than in gaze (4.4°). This is in sharp contrast to the outdoor environment with most true faces (1), whose maps are nearly uniform in head-centered coordinates (peak height: 2.3%, compared to baseline of 1.8%), but sharply peaked in gaze-centered coordinates (7.2%). This indicates a distinct role of eye-in-head as compared to head-in-world movements: false-positives are already centered by head movements with little effect of eye movements on top; true face detections, in contrast, experience an additional refinement as a consequence of eye movements.

Discussion

In line with earlier work in head-fixed settings,[8] our analysis shows that true and spurious face detections are elevated at the center of gaze. We interpret the environment-dependent differences between head and gaze maps as false-positives driving rough gaze-allocation only up to several degrees retinal eccentricity through head-in-world movements, whereas finer gaze-allocation is accomplished by

[a] Note that the baseline map is not entirely uniform due to boundary effects. Therefore the baseline also serves as control that centering of the peak is not an artifact of such boundary effects: when the peaks in the baseline map are substantially lower than in any true map, there is no evidence for such an artifact.

eye-in-head movements and only for true faces. These results are consistent with earlier data on low-level features that showed a sharpening and centering of peaks already present in head-centered coordinates by eye-in-head movements.[15] On the basis of our present analysis, we may furthermore speculate that eye-in-head movements refine gaze only after an additional stage of processing. In this view, candidates for face locations are determined in the periphery and centered by head movements. Only if the candidates are confirmed as true faces do finer gaze allocation follows by eye-in-head movements. At the present stage, it cannot be decided whether spurious face detections ("hallucinated faces") themselves or the low-level features correlated with them attract attention and gaze. Notwithstanding this exciting open issue for future psychophysical investigation, our study highlights the distinct roles of eye and head movements during free exploration, and therefore the importance of recording gaze-centered statistics in conditions when eye, head, and body can move freely.

Acknowledgments

We gratefully acknowledge the support by IST-027268-POP (Perception On Purpose, to P.K.), by the Bavaria California Technology Center (BaCaTeC, to W.E. and E.S.), and by Bayerische Forschungsstiftung (DOK-88-07 to J.V.). This work was financially supported in part within the DFG excellence initiative research cluster Cognition for Technical Systems – CoTeSys.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Buswell, G.T. 1935. *How People Look at Pictures. A Study of the Psychology of Perception in Art*. The University of Chicago Press. Chicago, IL.
2. Yarbus, A.L. 1967. *Eye Movements and Vision*. Plenum Press. New York.
3. Mannan, S.K., K.H. Ruddock & D.S. Wooding. 1997. Fixation sequences made during visual examination of briefly presented 2D images. *Spat. Vision* **11:** 157–178.
4. Reinagel, P. & A.M. Zador. 1999. Natural scene statistics at the centre of gaze. *Network* **10:** 341–350.
5. Einhäuser, W. & P. König. 2003. Does luminance-contrast contribute to a saliency map for overt visual attention? *Eur. J. Neurosci.* **17:** 1089–1097.
6. Peters, R.J., A. Iyer, L Itti & C. Koch. 2005. Components of bottom-up gaze allocation in natural images. *Vision Res.* **45:** 2397–2416.
7. Itti, L. & C. Koch. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* **40:** 1489–1506.
8. Cerf, M., J. Harel, W. Einhäuser & C. Koch. 2008. Predicting human gaze using low-level saliency combined with face detection. *NIPS* **20:** 241–248.
9. Elazary, L. & L. Itti. 2008. Interesting objects are visually salient. *J. Vision* **8:** 1–15.
10. Land, M.F., N. Mennie & J. Rusted. 1999. The roles of vision and eye movements in the control of activities of daily living. *Perception* **28:** 1311–1328.
11. Land, M.F. & M. Hayhoe. 2001. In what ways do eye movements contribute to everyday activities? *Vision Res.* **41:** 3559–3565.
12. Pelz, J.B. & R. Canosa. 2001. Oculomotor behaviour and perceptual strategies in complex tasks. *Vision Res.* **41:** 3587–3596.
13. Hayhoe, M., N. Mannie, B. Sullivan & K. Gorgos. 2005. The role of internal models and prediction in catching balls. *Proceedings of the American Association for Artificial Intelligence*. Fall 2005.
14. Schneider, E., T. Villgrattner, J. Vockeroth, *et al.* 2008. EyeSeeCam: An eye movement-driven head camera for the examination of natural visual exploration. *Ann. N. Y. Acad. Sci.* **1431:** In press.
15. Schumann, F., W. Einhäuser, J. Vockeroth, *et al.* 2008. Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *J. Vision* **8:** 12.1–17.
16. Viola, P. & M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. *Comput, Vis. Pattern Recog.* **1:** 511–518.
17. Bradski, G., A. Kaehler & V. Pisarevsky. 2005. Learning-based computer vision with Intel's open source computer vision library. *Intel Technol. J.* **9:** 119–130.