2018 52nd Annual Conference on Information Sciences and Systems (CISS)

Head movements are correlated with other measures of visual attention at smaller spatial scales

Brian Hu, Ishmael Johnson-Bey, Mansi Sharma, Ernst Niebur
Zanvyl Krieger Mind/Brain Institute
Johns Hopkins University
Baltimore, Maryland 21218

***Abstract*—Overt visual attention is traditionally studied by recording eye movements under head-fixed viewing conditions. However, during natural visual exploration, both head and eye movements can be used to redirect gaze to new points of interest. In order to better understand the role of head movements in this process, we recorded head movements while subjects explored a set of complex images from five different categories in a virtual reality environment. We found image-category specific differences in head movements, as quantified by the number and duration of head "fixations" (periods of maintained head orientation) as well as the amplitude of head movements. We compared head fixations with several other behavioral measures of attentional selection and with a computational model of bottom-up saliency, using the same set of complex scenes in all experiments. Results show significant positive correlation between head fixations and all other measures of attentional deployment, suggesting that head movements are a readily measurable indicator of overt selective attention at a spatial scale exceeding that of eye movements.**

I. Introduction

Selective attention is one of the crucial mechanisms of perception and cognition. It allows organisms to direct their necessarily limited information processing capabilities to the most relevant sensory inputs gathered in a complex world. An important distinction is between covert attention which is by definition a process that controls information processing resources entirely within the organism, in the absence of immediately observable changes in its appearance, and overt attention. The latter is frequently taken as a synonym for eye movements which, indeed, indicate with high likelihood where attention of the organism is directed at.

Although covert and overt attention can be dissociated [1], it is also true that they are frequently co-localized or at least tightly correlated in space and time [2], [3], even when eye movements are involuntary [4]. Indeed, the most common method of evaluating models of covert selective attention is to use this close relationship between eye movements and covert attention. This method, introduced by Parkhurst et al [5], allows for comparison of predictions of computational models of covert attention, like the classical saliency map model [6], with easily observable eye movements. This approach has been used in a large number of studies, for static and for dynamic scenes (video), and for both human and non-human primates; for a review, see [7].

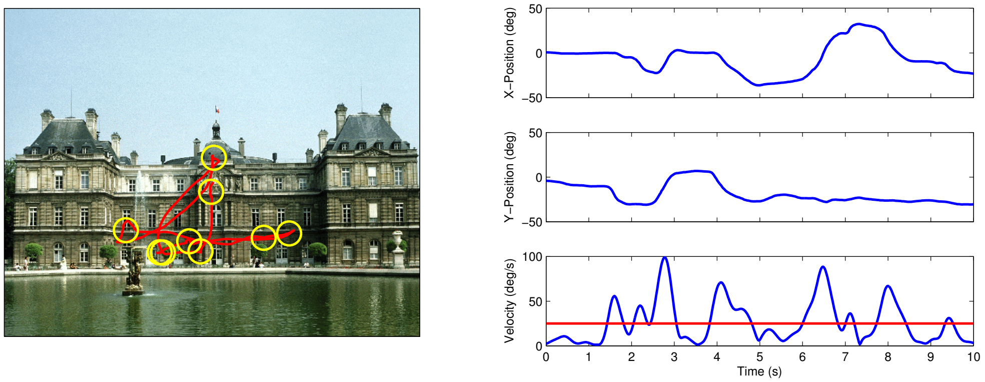
High visual acuity is available only in foveal vision, *i.e.* in a few square degrees out of the over 30,000 in the human visual field (≈ 210° horizontal by 150° vertical, ref. [8]) surrounding the observer. While information certainly can be gathered from other parts of the visual field (the term *covert attention* would be meaningless otherwise), there is no doubt that in many natural situations the information from the foveal area enjoys preferential status. If aligning this area with those parts of the visual input that the organism has determined need closer inspection requires a small change in the center of gaze, this change is naturally made by an eye movement. If information needs to be gathered from a part of the environment that is not visible with the current orientation of the head, the head needs to be turned. If the angle for the head movement is too large, the torso also needs to be turned or moved. Head- and eye-movements are known to be well-integrated in human and non-human primate observers [9], [10]. We therefore consider them as operations serving the same general purpose, but at differents parts of the spectrum of spatial scale: Fast, efficient oculomotor actions direct overt attention to areas within easy reach (see below for a more quantitative statement) of the current center of gaze, while larger scale shifts of attention require movement of the head or even the torso.

In the present study, we leverage recent advances in consumer-grade virtual reality (VR) technology to create a novel experimental setup that allows us to record head movements while subjects view natural images in a VR setting. VR environments can render realistic natural images, and give the experimenter tight control over all details of the scene. This is critical for being able to reproduce the experimental setting between subjects and studies. Crucially, the VR environment allows us to use a set of complex visual scenes (adapted to this paradigm as described below) for which we know where human observers direct there attention, using several different experimental methods. Previous work using the same stimuli has shown where observers fixate in these scenes [5], [11], which parts of these scenes they select as being "interesting" with a mouse click [12] and which parts are predicted to be salient by the standard saliency map model [6]. If, as we postulate, head movements are a manifestation of the same underlying attentional selection process at a larger spatial scale, we predict that they should be positively correlated with some or all of these measures.

II. Methods

The experimental methods are only summarized here because they were described previously [13]. We refer the reader

978-1-5386-0579-0/18/$31.00 ©2018 IEEE



(A)

(B)

Fig. 1. Head movement data for one subject on a sample image. (A) Head movement trajectory overlaid on the image. Red lines show the movement trajectory and yellow circles indicate the locations of head "fixations." (B) The x-position, y-position, and velocity of the head are shown for the example movement trajectory (top, middle, and bottom traces, respectively). The red line in the velocity trace shows the threshold used to classify movements as head "fixations".

to that publication for details.

All experimental procedures were approved by the Institutional Review Board of Johns Hopkins University. 27 subjects (15 male; mean age = 20.1 years, SD = 2.7), all neurologically normal and with vision normal or corrected to normal, participated in the experiment. Each subject was fitted with a pair of Google Cardboard VR glasses which was used to display the VR environment and collect head movement data with a custom-designed script. Each subject viewed in the VR environment a total of 70 different scenes (13 images from each of five image categories, with five repeat images, one from each category). All images had a resolution of 640×480 pixels. Four of the categories (buildings, fractals, "old" home interiors and landscapes) were introduced by Parkhurst et al [5] and we added an additional category ("new" home interiors) for reasons explained below. Image categories were chosen to provide subjects with a variety of scenes that are ecologically important. The scenes also differ in semantic content (*e.g.* fractals are devoid of meaning), which likely leads to differences in the allocation of top-down visual attention. We chose these images because previous studies have extensively characterized eye movements (fixations) that humans make in these scenes [5], [11]. Furthermore, a large study with hundreds of subjects determined which portions of these scenes were considered subjectively interesting [12]. The fifth, additional set of ("new") home interior images was collected from the internet and used in our experiment because the original home interior images were digitized from photographs and, as a consequence, were often perceived as blurry when rendered in the VR environment.

Subjects were seated on a stationary (non-swiveling) chair in a quiet room to minimize the influence of body movements and noise disturbances. Each image presentation began with a 1-second, small view of the image to give subjects an overview (gist) of the whole image. This allowed participants to (consciously or subconsciously) decide which portions of the whole image needed to be scrutinized in greater detail once the full-resolution image became available. This then immediately zoomed into a large-scale, immersive image. The small image subtended approximately 30 degrees in the horizontal direction and 23 degrees in the vertical direction, while the full-size image subtended approximately 116 degrees in the horizontal direction and 100 degrees in the vertical direction. Subjects viewed the images through the field of view of the VR glasses, a square aperture with side length 74 degrees. As a result, the full size image was larger than the visible portion of the VR environment by more than a factor of two in surface area, and the head had to be moved to see parts of the image in the periphery. Participants were instructed to visually explore the images and were told that they would be asked about image contents. No explicit mention of head movements was made. After each image viewing, subjects were asked to describe the scene concisely in one sentence. Viewing time was set to 10 seconds for each image, with unlimited time for image description. Each experiment lasted approximately one hour.

Head movements were divided into "head fixations" (referred to below sometimes simply as "fixations;" context will indicate whether head or eye fixations are meant) and interfixations by using a velocity threshold of 25 degrees/second. Although this threshold value is somewhat arbitrary, other studies have reported using similar thresholds between 15-25 degrees/second [9], [10], [14], [15]. Movements below the threshold were classified as fixations, and the centroid of recorded movements during fixation periods was used as the fixation center. We analyzed the head movements in terms of the following metrics: number of fixations, fixation duration, and amplitude of inter-fixation movements. We tested for significant differences in head movements across image categories using ANOVA, and we corrected for multiple comparisons using the Bonferroni correction.

In order to create smoothed maps of the head movements for comparison with eye fixations [5] and interest points [12], we recorded the location of each head fixation center and convolved these binary maps with a Gaussian with $\sigma = 27$ pixels. Before convolution, both eye and head fixation locations were weighted by the duration of the fixation. Interest maps were computed from the first interest points only, see ref [12] for details. To compute the marginal fixational density maps (FDMs) used to compare head movements with eye movements and interest points, we first averaged all smoothed head movement maps from one image category (across all subjects and images), while subtracting out the average smoothed head movement maps from all other categories. This controls for spatial biases common across categories, including the center bias. In each image category's marginal FDM, positive values represent regions fixated more than the grand average across categories, and negative values represent regions fixated less then the grand average across categories. More details about this approach can be found in ref [16].

All data and code associated with this paper can be found online at: https://github.com/brianhhu/VR_HeadMovements.

III. Results

A. Previous results: Head movement kinematics in a VR environment

We have described the basic kinematics of head movements in VR environments, using the identical stimuli used here, in a previous study [13]. Figure 1 shows an example head movement trajectory for one subject, including the raw head movement data that was used to complete our analysis. We found that, similar to eye fixations, head movements follow approximately a "main sequence," with both peak velocity and the duration of "head fixations" (defined in analogy to eye fixations) being proportional to their duration. Furthermore, most head movements occur along the cardinal directions, with a strong preponderance of horizontal over vertical movements.

B. Comparison with other correlates of visual attention

Jeck et al [17] pointed out that traditional correlation measures between maps like those considered here lead to a systematic underestimation of map similarity because of the finite data size. They describe a resampling technique that corrects for this error. Applying their methods on maps of eye fixations [5], interest point selections [12], and a saliency map model [6], we found that head movements were significantly correlated with each of the other measures of visual attention. As we propose that head movements are a coarse measure of visual attention, we chose a relatively large bin size of 128×128 pixels when comparing the maps, which effectively turns each map into 6×8 resolution. Head movements were correlated with fixations, $r = 0.59$, significantly above the null hypothesis ($p = 6.85 \times 10^{-14}$). Head movements were also correlated with interest point selections, $r = 0.69$, allowing us to reject the null hypothesis of absence of correlations between these two measures of attention ($p = 3.86 \times 10^{-5}$). Finally, head movements were correlated with computed saliency maps, $r = 0.40$, again rejecting the null hypothesis ($p = 3.13 \times 10^{-3}$). Our results are summarized in Figure 2.

C. Marginal Fixation Density Maps

In the next step, we examined differences between image categories using the marginal FDMs (see Methods). We applied this approach to the eye movement data, interest point selections, head movement data, and saliency maps computed using low-level image features computed using a classical bottom-up model of visual attention [6] (Figure 3). In these maps, shown in Figure 3, warmer colors indicate areas in the image category that were either fixated (eye and head data) or selected (interest point data) more often compared to the other image categories, or where the computed saliency for this category exceeded that for others. Although the marginal FDMs for the eye fixations and head fixations differ, there are also some similarities. For example, for the building images, subjects tended to fixate more on the lower half of the image and less on the upper half. This trend is also seen in the head movement maps, as more of the fixations fall on the lower half of the image compared to the upper half. Saliency maps also show a strong bias towards the lower half of the image for the building images, but show less qualitative agreement on the other image categories. Additionally, fractal images tend to show the strongest center bias with more activity in the center of the image, and this is captured in the eye movements, interest points, and head movements. We also note that the head marginal FDMs are more concentrated near the center of the image, while for both the eye fixation and interest points, the marginal FDMs are more uniformly distributed across the image.

D. Category-Specific Head Movement Metrics

We compared differences in head movement metrics to see if they were sensitive to the image category. Specifically, we examined differences in head movements based on the number of head fixations, head fixation duration, the number of head inter-fixation movements, and head inter-fixation movement amplitude. A one-way ANOVA revealed significant main effect of image category on the number of fixations, $F(4, 120) = 10.01, p = 5.03 \times 10^{-7}$. The number of fixations was significantly higher for the old home interiors compared to the other image categories (all p-values < 0.05). A one-way ANOVA revealed significant main effect of image category on fixation duration, $F(4, 120) = 10.96$, $p = 1.32 \times 10^{-7}$. Fixation duration was significantly lower for the old home interiors compared to the other image categories (all p-values < 0.05). Furthermore, fixation duration for fractals was significantly higher compared to landscapes. Inter-fixation movement amplitude was significantly higher for the old home interiors compared to buildings and fractals (all p-values < 0.05). Inter-fixation movement amplitudes for landscapes and new home interiors were not statistically different from the other image categories (all p-values > 0.05).

We found significant differences in head movement metrics across image categories, suggesting that subjects used different exploration strategies for all images that they viewed. In particular, we found that viewing of old home interior images resulted in significantly different head movement metrics

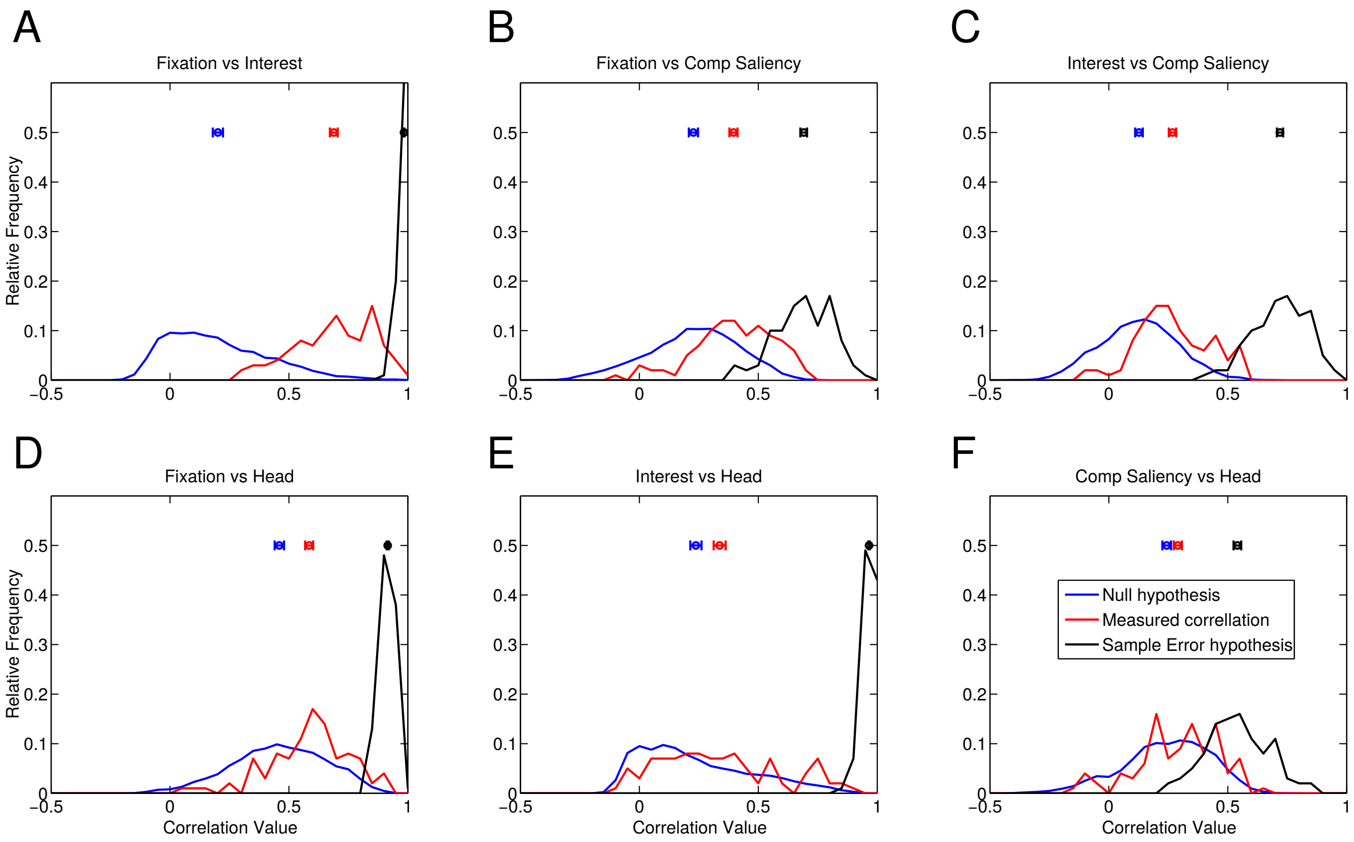


Fig. 2. Aggregate results of natural scene analysis at 6 x 8 resolution. Each subplot shows a distribution of measured correlations between two types of maps compared against the null hypothesis and sample error hypothesis. Means of each distribution are shown above the histograms, with error bars indicating standard error given the 100 images used. Most error bars are smaller than the markers used. (A) Fixation and Interest maps. (B) Fixation and Computed saliency maps generated from Itti et al. (1998). (C) Interest and saliency maps. (D) Fixation and head movement maps. (E) Interest and head movement maps. (F) Computed saliency and heaed movement maps. All measured averages are significantly above the null hypothesis ($p < 0.05$). All measured averages are below the sample error hypothesis ($p < 0.05$). The legend in panel F applies to all panels.

compared to the other image categories. This difference may be due to the fact that the old home interiors were the lowest resolution images among all image categories. While others have studied changes in eye fixations when viewing low-resolution images compared to high-resolution images [18], very little is known about how image resolution affects head movements.

IV. Discussion

Selective attention is a crucial component of perception and cognition, allowing adaptive behavior in the face of overwhelming amounts of raw sensory data[1] Much work has been devoted over the last decades to compare the predictions of models of selective attention with behavioral data. Nearly all of this work is concerned with selections expressed by eye movements which, by definition, are limited to the instantaneous field of vision. It is, however, clear, that the rotation of the eyes in their orbits selection is not the only possible selection mechanism. Following ideas originating in the pre-motor theory of selective attention [20], we suggest that shifts of attention can occur from the smallest scales, *i.e.* from within locations projected onto different parts of the foveola [21] all the way to movements of the torso [22], [23]. We propose that attentional selection makes use of all of these affordances and that the underlying mechanisms for moving these different effectors are the closely related. We here focus on one intermediate scale, movements of the head, which thus reflect shifts of attention just like eye movements, except at a larger spatial scale.

Although when required the eyes can deviate from the orientation of the head by a large amount, eye position during natural vision is strongly coupled to head position. Stahl [24] found that in a paradigm where both head and eye movements were possible to localize illuminated LEDs in the environment, normal human subjects tend to avoid large eye movements and replace them instead with head movements. Over a small range around a fixed head position (median 13°, mean 18° horizontal, smaller for vertical), subjects only made eye movements while larger deviations usually involved

[1] The numbers of somatosensory and visual afferents in humans are each on the order of 10^6. Elementary application of Shannon's information theory shows that the channel capacities of these two sensory systems alone are on the order of 10^9 bits per second [19], an amount enormously higher than what can be expected to be processed in detail by the brain. Judicial selection of the instantaneously relevant sensory data is therefore of the highest importance.

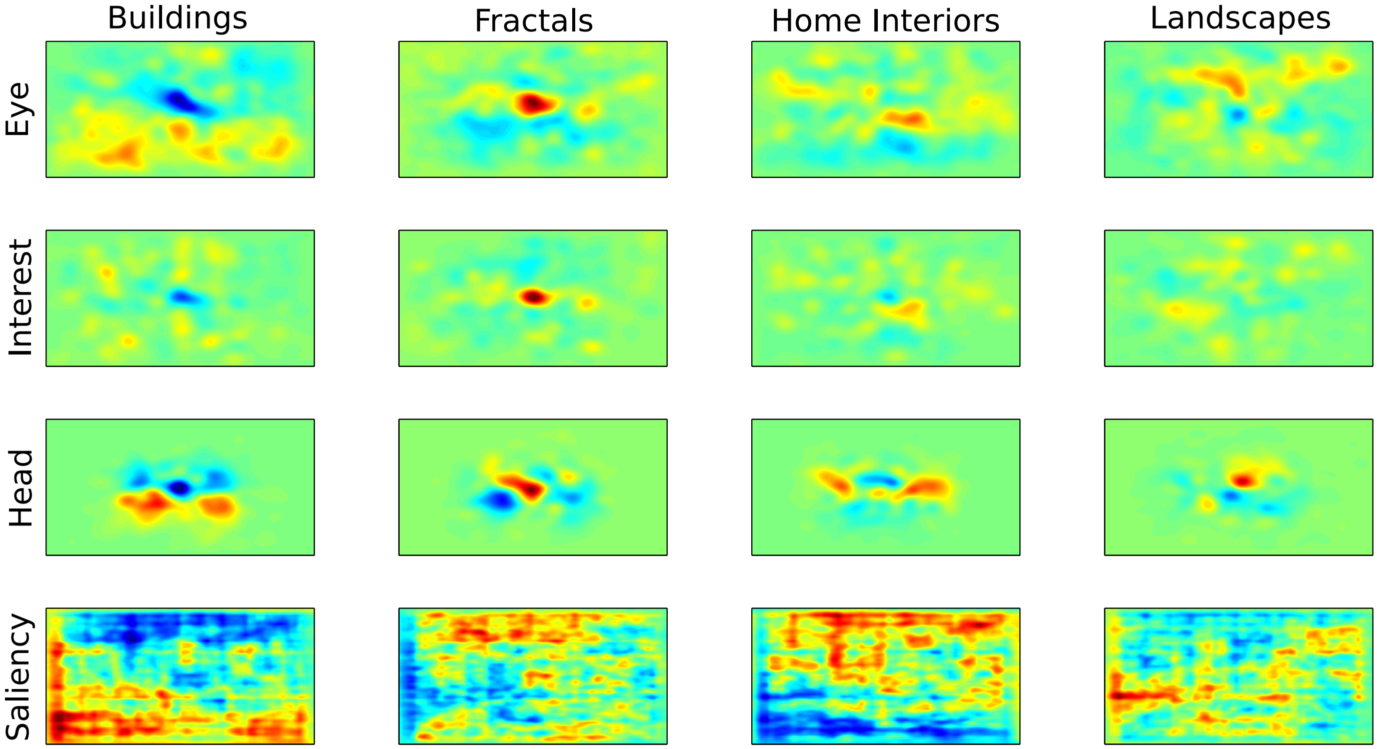


Fig. 3. Marginal fixation density maps (FDMs) across image categories (columns) and modalities (rows). The eye FDMs were calculated using the fixation data from all subjects and images in [5]. The interest FDMs were calculated using the first interest point selection from the data in [12]. The head FDMs were calculated based on the head movements recorded during the VR experiment. The saliency FDMs were based on the saliency maps of all images in each category, calculated using a bottom-up visual attention model [6]. Within each modality, warmer colors indicate image regions that had either more fixations (eye and head), more selections (interest points), or higher activity (saliency) compared to the other image categories.

head motions. Similar results were obtained in experimental paradigms closer to ours in which human participants were exposed to virtual reality environments. Sitzmann and collaborators [15] found that mean gaze direction relative to the head orientation $\approx 13°$, in close agreement with another VR experiment [25] where mean was $\approx 11°$. Even less deviation of eye direction from head direction was found [26] when eye and head movements were recorded while participants walked in a natural environment: the deviation of the eye orientation relative to the head orientation followed roughly a bell curve with half-width of $\approx \pm 8°$. These results also agree with the well-known center bias of visual observers, typically described in the frame of visual stimuli presented *e.g.* on a screen [5], [27], [28] but demonstrated also in the reference frame of the head [26].

These numbers, all in reasonably good agreement, led us to hypothesize that head movement is a valid predictor of the location where visual attention is deployed, at least within a resolution in the range given by these numbers. If that is the case, and if attention is a unified mechanism over several scales, one expects that the different measures of attention are positively correlated. Existing data did, however, not allow for direct testing of this prediction since, to our knowledge, correlations between behavioral measures of attentional selection on the same visual input data have never been determined. We fill this gap by quantitatively comparing the attentional selection process in complex scenes predicted in a computational model [6] with the behavior of freely acting humans using eye fixations [5], [11] and conscious interest selection [12] with the coarse attentional selection given by head movements (this study). We found significant positive correlations between all these measures. These results strongly support that attentional selection processes at different spatial scales are closely related and may be part of the same underlying mechanism, and that head movements are indicative of attentional selection at scales exceeding that of eye movements, probably beyond $\approx 10° - 15°$. Within this range, head direction can be used as a good first approximation of gaze direction, because in many cases people are fixating in the center of the head-centred field of view. It is a small step, though not addressed in this study, to hypothesize that torso movements are the next step in this progression, covering even larger angles over which head movements are impossible or uncomfortable.

V. Conclusion

We developed a novel experimental setup that allowed us to record head movements of human participants as they viewed natural images in a VR environment. Notably, we displayed the same set of images that have been used in several studies to understand visual saliency in terms of eye movements and of interest point selections. This allowed us to quantitatively compare head movements with previous measures of visual attention. We found significant, positive correlation between head fixations and other measures of visual attention at a coarse spatial scale and also qualitative agreement in the marginal fixation density maps across image categories. Our results give insight into the role of head movements during natural visual exploration, and provide evidence that head movements may also be a marker for visual attention. While our study was performed on healthy subjects, it may also be useful to quantitatively measure head movements in clinical populations, e.g. patients with autism or stroke patients. Head movements could potentially be used as a biomarker that could help clinicians treat these patients. An interesting extension of our work would be simultaneous recording of eye and head movements.

Acknowledgment

This work is supported by the National Institutes of Health under grants R01EY027544 and R01DA040990. We thank Danny Jeck for helpful discussions.

References

[1] M. I. Posner, "Orienting of attention," *Quart. J. Exp. Psychol.*, vol. 32, pp. 3–25, 1980.

[2] J. Hoffman and B. Subramaniam, "The Role of Visual Attention in Saccadic Eye Movements," *Perception and Psychophysics*, vol. 57, no. 6, pp. 787–795, 1995.

[3] H. Deubel and W. X. Schneider, "Saccade target selection and object recognition: Evidence for a common attentional mechanism," *Vision research*, vol. 36, no. 12, pp. 1827–1837, 1996.

[4] M. S. Peterson, A. F. Kramer, and D. E. Irwin, "Covert shifts of attention precede involuntary eye movements," *Attention, Perception, & Psychophysics*, vol. 66, no. 3, pp. 398–405, 2004.

[5] D. Parkhurst, K. Law, and E. Niebur, "Modelling the role of salience in the allocation of visual selective attention," *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.

[6] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based fast visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, November 1998.

[7] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[8] H. M. Traquair, *Clinical perimetry*. Kimpton, 1938.

[9] E. G. Freedman and D. L. Sparks, "Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys," *Journal of neurophysiology*, vol. 77, no. 5, pp. 2328–2348, 1997.

[10] M. K. McCluskey and K. E. Cullen, "Eye, head, and body coordination during large gaze shifts in rhesus monkeys: movement kinematics and the influence of posture," *Journal of neurophysiology*, vol. 97, no. 4, pp. 2976–2991, 2007.

[11] D. Parkhurst and E. Niebur, "Scene Content Selected by Active Vision," *Spatial Vision*, vol. 16, no. 2, pp. 125–54, 2003.

[12] C. Masciocchi, S. Mihalas, D. Parkhurst, and E. Niebur, "Everyone knows what is interesting: Salient locations which should be fixated," *Journal of Vision*, vol. 9, no. 11, pp. 1–22, October 2009, pMC 2915572.

[13] B. Hu, I. Johnson-Bey, M. Sharma, and N. E., "Head Movements During Visual Exploration of Natural Images in Virtual Reality," in *51st Annual Conference on Information Systems and Sciences IEEE-CISS*. IEEE Press, March 2017.

[14] Y. Fang, R. Nakashima, K. Matsumiya, I. Kuriki, and S. Shioiri, "Eye-head coordination for visual cognitive processing," *PloS one*, vol. 10, no. 3, p. e0121035, 2015.

[15] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, and G. Wetzstein, "Saliency in VR: How do people explore virtual environments?" *ArXiv e-prints*, Dec. 2016.

[16] T. P. O'Connell and D. B. Walther, "Dissociation of salience-driven and content-driven spatial attention to scene category with predictive decoding of gaze patterns," *Journal of vision*, vol. 15, no. 5, pp. 20–20, 2015.

[17] D. M. Jeck, M. Qin, H. Egeth, and E. Niebur, "Attentive pointing in natural scenes correlates with other measures of attention," *Vision Research*, vol. 135, pp. 54–64, 2017, nIHMSID: NIHMS868593.

[18] T. Judd, F. Durand, and A. Torralba, "Fixations on low-resolution images," *Journal of Vision*, vol. 11, no. 4, pp. 14–14, 2011.

[19] J. Singer, "Information theory and the human visual system," *JOSA*, vol. 49, no. 6, pp. 639_1–640, 1959.

[20] G. Rizzolatti, L. Riggio, I. Dascola, and C. Umiltá, "Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention," *Neuropsychologia*, vol. 25, pp. 31–40, 1987.

[21] M. Poletti, M. Rucci, and M. Carrasco, "Selective attention within the foveola," *Nature Neuroscience*, vol. 20, no. 10, pp. 1413–1417, 2017.

[22] J. D. Grubb and C. L. Reed, "Trunk orientation induces neglect-like lateral biases in covert attention," *Psychological Science*, vol. 13, no. 6, pp. 553–556, 2002.

[23] M. Mon-Williams, S. Sheehan, A. D. Wilson, and G. P. Bingham, "Head-torso coordination and overt shifts in attention," *Journal of Vision*, vol. 9, no. 8, pp. 837–837, 2009.

[24] J. S. Stahl, "Amplitude of human head movements associated with horizontal saccades," *Experimental brain research*, vol. 126, no. 1, pp. 41–54, 1999.

[25] T. Kollenberg, A. Neumann, D. Schneider, T.-K. Tews, T. Hermann, H. Ritter, A. Dierker, and H. Koesling, "Visual search in the (un) real world: how head-mounted displays affect eye movements, head movements and target detection," in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM, 2010, pp. 121–124.

[26] T. Foulsham, E. Walker, and A. Kingstone, "The where, what and when of gaze allocation in the lab and the natural environment," *Vision research*, vol. 51, no. 17, pp. 1920–1931, 2011.

[27] T. Foulsham and G. Underwood, "What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition," *J Vis*, vol. 8, pp. 1–17, 2008.

[28] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision Res.*, vol. 45, pp. 643–659, Mar 2005.