A Model of Full-body Kinematics-based Visual Attention in Daily-Life Tasks

J. Alex Harston (j.harston17@imperial.ac.uk)
Dept of Bioengineering, Imperial College London
London, UK, SW7 2AZ

Chaiyawan Auepanwiriyakul (ca1216@ic.ac.uk)
Dept of Computing, Imperial College London
London, UK, SW7 2AZ

A. Aldo Faisal (a.faisal@imperial.ac.uk)
Dept of Bioengineering & Dept of Computing, Imperial College London
London, UK, SW7 2AZ

Abstract:

Visual attention and motor actions are intrinsically linked and tightly spatiotemporally coupled in real-world behavior, and yet very few studies of natural gaze behavior account for the dynamics of the body, thereby missing a fundamental aspect of the perception-action loop. To address this, we experimentally capture whole body kinematics and time-synced gaze in a natural, high-dimensional task, to investigate the influence of motor actions on gaze behavior. We use a combination of linear and nonlinear autoregressive models with exogenous body input to assess the predictive power of prior gaze and motor dynamics on future gaze location, and find that our nonlinear model significantly outperforms previous linear models for predicting natural gaze dynamics, and that incorporating whole body kinematic information into our model significantly improves gaze prediction performance versus simple gaze autoregression. Incorporating this body information into visual saliency models helps improve our understanding of visuomotor interactions in the real world.

Keywords: neuroscience; saliency; gaze; kinematics; behavior

Introduction

How perception is linked to action is a fundamental question in neuroscience - much remains unknown about the mechanisms of visuomotor behavior in real-world environments. Eye movements are tightly linked to our behavior and cognition through the allocation of overt and covert attention (Land & Furneaux, 1997), and foveal constraints on image resolution result in eye movements providing a direct proxy of attention (Hendrickson & Yuodelis, 1984) or 'attentional spotlight' over our environment. The specific sequence of these serial targets of attention must therefore facilitate the ongoing cognitive demands of the task at hand (Treisman & Gelade, 1980).

Until relatively recently, research into gaze behavior was restricted to laboratory settings due to the complexity and non-portability of eye tracking equipment. Whilst work over the last two decades has shifted the focus onto gaze behavior in natural everyday tasks and contexts, this has generally been restricted to static scene viewing, which only represents a very small subset of gaze behavior. Such work in real-world saliency has demonstrated that the control of where we look is based overwhelmingly on the location of information required by ensuing action sub-goals, and as such is spatiotemporally locked with body actions (Epelboim et al., 1995; Hayhoe et al., 2003; Hayhoe et al., 2003; Land and Furneaux, 1997; Land et al., 1999; Patla and Vickers, 1997; Pelz and Canosa, 2001, Schütz et al., 2011; Tatler et al., 2011). If we are to understand from this mounting body of evidence that gaze behavior exists as part of an embodied system (Sprague et al. 2007), then to truly understand natural gaze behavior we must design experimental methods to capture and incorporate this embodiment, rather than remove it.

Modelling the link between perception and action has historically proven difficult due to the lack of high-resolution body movement information. Given our understanding that eye-movements are embedded in a rich visuomotor repertoire, we must now develop data-driven models to predict eye-movements, not only from a visual saliency perspective, but from an embodied perspective, placing perception into the context and actions of the body, i.e. motor behavior. In this regard, we present an embodied methodology here to capture sensory inputs and motor outputs in natural behavior, rather than constrain them. Sensory inputs are recorded using a head mounted eye-tracker, scene camera and microphone, whilst simultaneously recording skeletal motor outputs through motion

tracking 66 degrees-of-freedom (DOF) in the body to capture and quantify unconstrained behavior. This allows us to test a simple hypothesis of embodied saliency, namely whether it is possible to predict eye-movements directly from the movement dynamics of the body.

Methods

We used a portable eye-tracker (SMI Eye Tracking Glasses 120Hz, Sensomotoric Instruments, Teltow Germany) in combination with a portable full body motion capture suit, measuring 66 degrees of freedom (DOF) from the body using 17 inertial measurement units at 60Hz (XSENS MVN) (Figure 1A). Experiments were filmed with a static video camera, as well as with the integrated egovideo from the eye tracker camera. Subjects (n=7) were asked to perform a cooking task, in this case cooking an omelette, using a standardized neurorehabilitative rubric in a working hospital kitchen environment (Charing Cross Hospital, London, UK). (Figure 1B).

We gathered time-synced gaze and full body motion data from 7 healthy subjects with perfect or corrected vision, with an average trial length of 16 minutes. Suit sensors and eye-tracking glasses were calibrated using standard manufacturer procedures and were recorded simultaneously on the same device. Time alignment of the data streams was performed post-hoc using Chronoviz software and body data interpolated to match gaze sampling frequency. We subsequently created several different autoregressive models to test predictivity. We initially took a linear approach, using a linear Vector AutoRegression with eXogenous input model (LinVARX) where *y* represents the gazepoint coordinates from the eye tracker at time *t* :

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{k=1}^{r} \beta_k x_{t-k} + \epsilon_t$$

We also utilise a nonlinear autoregressive model termed 'DeepVARX', developed using the TensorFlow and Keras libraries, consisting of an autoencoder model with simultaneous convolved exogenous body input (see Figure 1C). Models were trained and predictions made both in open loop (Figure 1D, 1E) (only predicting 1 data point ahead at a time) and closed loop (predicting multiple timepoints ahead)). We performed leave-one-subject-out crossvalidation to minimize overfitting and increase inter-subject generalizability of results.

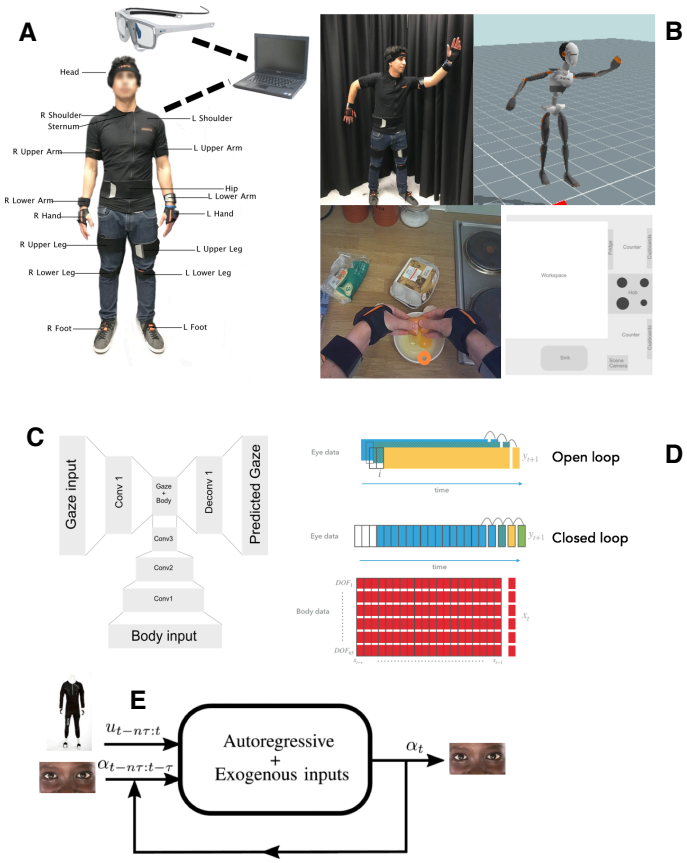

Figure 1: A) Layout of recording system with with sensors located on suit. B) (Top) Full reconstruction of body kinematics with wireless motion tracking suit at 60Hz. (Bottom Left) Egovideo during task with gazepoint overlaid. (Bottom Right) Layout of experimental workspace with camera recording location. C) DeepVARX architecture showing convolutional layers. D) Open Loop and Closed Loop methodologies with exogenous body input. E) Autoregressive models combine prior gaze and body data to predict future gazepoints.

Results

Comparing the results of the linear model with the nonlinear, we find that in both open loop and closed loop the nonlinear model (DeepVARX) significantly outperforms the linear (Figure 2A, 2B). Whilst open loop prediction significantly outperforms closed loop overall, for online prediction systems, we would require predictions further ahead in time than 1/120th of a second, so closed loop is the more salient method in this regard. Incorporating body dynamics as an exogenous control signal (gaze and body dynamics predicting gaze) into our model improves closed loop rollahead prediction performance versus simple gaze autoregression (gaze dynamics predicting gaze), gaze

with white noise exogenous input, or gaze with head inputs alone, to an average of 1.5 seconds (Figure 2C). The models were trained both using body joint angles and angular velocities. We find that angular velocities hold significantly more predictive power than joint angles (Figure 2D).

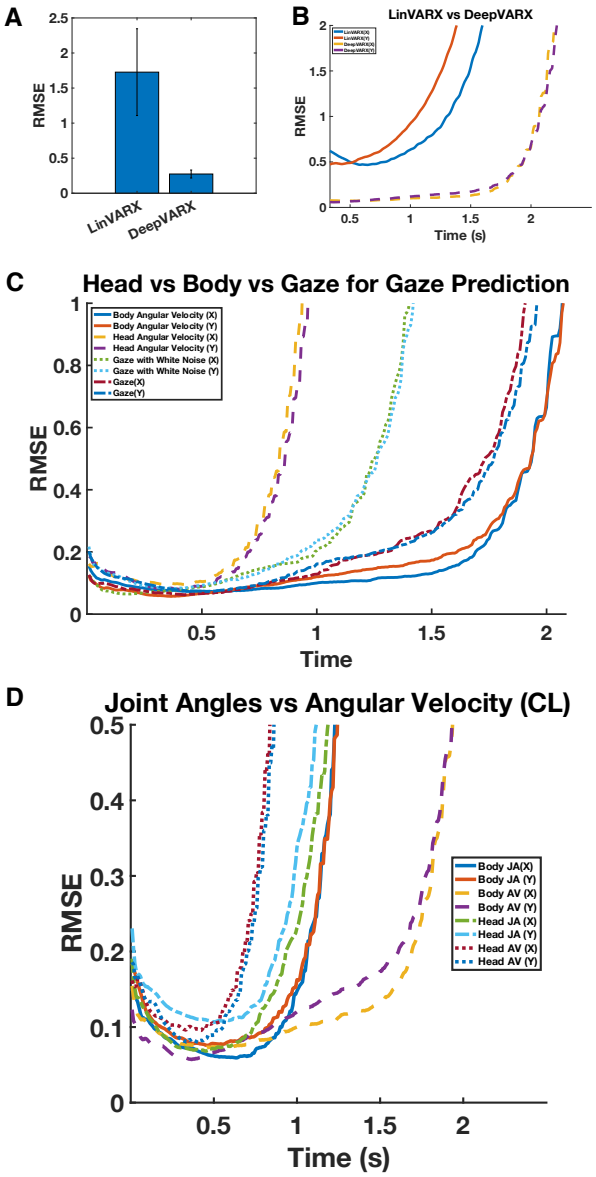

Figure 2: A) The nonlinear DeepVARX model significantly outperforms the linear VARX model in gaze prediction in open loop. B) The same result applies in closed loop prediction, where the nonlinear model predicts significantly better than the linear model. C) Average crossvalidated rollahead prediction of gaze with various exogenous control signals. Incorporating body angular velocities into prediction improves prediction performance of the nonlinear model, versus simple gaze autoregression. Additionally, gaze with exogenous white noise and gaze with exogenous head input perform significantly poorer, demonstrating the predictive power of the whole body, as opposed to just the head. D) Body angular velocities (AV) provide superior prediction performance versus joint angles (JA) or head angular velocity.

Discussion

Our results indicate the importance of the role of whole body kinematics in predicting gaze. This is likely due to the body dynamics holding task information implicitly. It is a possibility that angular velocities are the more useful input for gaze prediction due to the additional speed information contained within, versus simple joint conformation. It remains to be explored whether individual joints or particular subsets of joint combinations contribute more to this predictive power. We are working to refine and validate our model further and increase cohort size.

Acknowledgments

We would like to thank and acknowledge the support of the EPSRC for financial support on this project. This research was also supported by eNHANCE under the European Union's Horizon 2020 research and innovation programme, Grant Agreement No. 644000.

References

Epelboim, J., Steinman, R. M., Kowler, E., Edwards, M., Pizlo, Z., Erkelens, C. J., & Collewijn, H. (1995). The function of visual search and memory in sequential looking tasks. *Vision Research*, 35(23-24), 3401–3422.

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, 3(1), 6.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194.

Hendrickson, A. E., & Yuodelis, C. (1984). The morphological development of the human fovea. *Ophthalmology*, *91*(6), 603–612.

Land, M. F., & Furneaux, S. (1997). The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *352*(1358), 1231–1239.

Patla, A. E., & Vickers, J. N. (1997). Where and when do we look as we approach and step over an obstacle in the travel path? *Neuroreport*, *8*(17), 3661–3665.

Pelz, J. B., & Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research*, *41*(25-26), 3587–3596.

Schütz, A. C., Braun, D. I., & Gegenfurtner, K. R. (2011). Eye movements and perception: a selective review. *Journal of Vision*, *11*(5).

Sprague, N., Ballard, D., & Robinson, A. (2007). Modelling embodied visual behaviors. *ACM Transactions on Applied Perception*, *4*(2), 11

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: reinterpreting salience. *Journal of Vision*, *11*(5), 5.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.