Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Learning to Predict Intent from Gaze During Robotic Hand-Eye Coordination

Yosef Razin, Karen Feigh
School of Aerospace Engineering
Georgia Institute of Technology
270 Ferst Drive, NW Atlanta, GA 30332

Abstract

Effective human-aware robots should anticipate their user's intentions. During hand-eye coordination tasks, gaze often precedes hand motion and can serve as a powerful predictor for intent. However, cooperative tasks where a semi-autonomous robot serves as an extension of the human hand have rarely been studied in the context of hand-eye coordination. We hypothesize that accounting for anticipatory eye movements in addition to the movements of the robot will improve intent estimation. This research compares the application of various machine learning methods to intent prediction from gaze tracking data during robotic hand-eye coordination tasks. We found that with proper feature selection, accuracies exceeding 94% and AUC greater than 91% are achievable with several classification algorithms but that anticipatory gaze data did not improve intent prediction.

1 Introduction

In an increasingly autonomous world, while the physical and cognitive burdens of work will be offloaded to robots, there will necessarily be tasks that require human interaction with these ever-ubiquitous intelligent systems. While many of these interactions may remain simply physical, more intelligent systems will require cognitive interaction, such as activity recognition and intent prediction, to work safely and productively with humans.

In visuomotor coordination tasks, it has long been recognized that activity recognition is not limited to processing hand movements but can also be learned from eye behavior (Bednarik, Vrzakova, and Hradis 2012; Bondareva et al. 2013; Land, Mennie, and Rusted 1999; Gielen, Van den Heuvel, and Van Gisbergen 1984). It has also been shown that the eyes have a tendency to anticipate hand movements when a task's sequence and targets are being planned or when the target is moving (Mennie, Hayhoe, and Sullivan 2007; Thier and Ilg 2005; Ma-Wyatt, Stritzke, and Trommershäuser 2010). Thus, gaze data should be able to inform us of upcoming user intent. This anticipatory information could prove useful by giving us advance clues for earlier intent prediction. However, while past studies have applied machine learning techniques to classify intent from gaze data (Eivazi and Bednarik 2011; Bednarik, Vrzakova, and Hradis 2012; Bondareva et al. 2013), none have attempted to leverage anticipatory eye behavior to improve upon their predictive capabilities.

Therefore, this research addresses the questions:

1. Can gaze data be used to predict intent, as realized in the movements of the robotic grasper?
2. Does gaze data significantly improve such prediction over using "hand" movement data alone?
3. Which classification algorithm performs this intent prediction best?

2 Background

During hand-eye coordination tasks, a trained operator's gaze anticipates the target's position and tends to fixate on the goal (Sailer, Flanagan, and Johansson 2005; Johansson et al. 2001; Posner 1980). Fixations are associated with the task structure and have unique characteristics and routines depending on the function they serve (Land, Mennie, and Rusted 1999; Sailer, Flanagan, and Johansson 2005; Hayhoe 2000). In all cases, they tend to last for 200-400 ms and are characterized by their low velocities ($< 100^\circ/\text{s}$) (Salvucci and Goldberg 2000a). A key feature of task-oriented fixations is how far in advance they anticipate manipulations (Land, Mennie, and Rusted 1999; Pelz and Canosa 2001; Johansson et al. 2001; Hayhoe 2000). Depending on the level of cognitive processing involved, fixations may occur anywhere from 0.2-10 s before hand movement commences (Johansson et al. 2001; Land, Mennie, and Rusted 1999; Hayhoe and Ballard 2005). About 20% of fixations are composed of these look-ahead movements (Mennie, Hayhoe, and Sullivan 2007; Hayhoe et al. 2003; Pelz and Canosa 2001), and the majority of fixations help gather guiding information for hand movement planning (Mennie, Hayhoe, and Sullivan 2007).

In dynamic tasks another gaze type that exhibits anticipatory behavior is smooth pursuit. This behavior tracks a moving target to keep it centered on the fovea, taking 100 ms to lock on to the target using an open-loop strategy and then switching to closed-loop feedback once the target is acquired. At that point, the 100–ms lag gap is closed and sometimes even reversed, such that the gaze position leads the target's (Thier and Ilg 2005). Smooth pursuit tracking has been shown to improve control of a target's motion during

direct manipulation by both humans and monkeys. While the exact nature of this coupling in oculo-manual coordination is still a subject of debate (Vercher and Gauthier 1992; Reina and Schwartz 2003), the close coupling of smooth pursuits and target tracking as well as the anticipatory capabilities of smooth pursuits have been well established (Barnes and Marsden 2002; Mrotek and Soechting 2007).

Saccades, which last under 100 ms (Salvucci and Goldberg 2000b) and achieve angular speeds well over 70°/s (Komogortsev and Karpov 2013), are generally used for basic visual input such as filtering, tracking peripheral movement, and updating visual memory. In hand-eye coordination tasks, saccades help to direct grasps to specific extensions or protrusions of the object to be manipulated (Johansson et al. 2001). Therefore, like fixations, saccades are used to plan future manipulations (Ma-Wyatt, Stritzke, and Trommershäuser 2010), or at least future fixations (Land 2006). However, saccades also occur at the onset of hand movement (Ma-Wyatt, Stritzke, and Trommershäuser 2010), as well as randomly during fixations, in order to reacquire the image on the fovea (Carpenter 1991).

Since visual feedback is generally too slow for dynamic tasks to be performed well, the predictive visual systems underlying fixations, saccades, and smooth pursuits are already naturally exploited by cognitive processes (Miall and Reckess 2002; Mrotek and Soechting 2007). Through them, humans learn to visually anticipate the effect of their actions (Crawford, Medendorp, and Marotta 2004; Johansson et al. 2001; Land and Hayhoe 2001; Barnes and Marsden 2002) using saccade-and-fixate and saccade-pursue strategies, supplemented with haptic feedback (Sobuh et al. 2014; Mrotek and Soechting 2007; Reina and Schwartz 2003).

Machine learning techniques have proven useful for intent estimation from gaze data. Steichen et al. (2014) and Bondareva et al. (2013) both found that logistic regression worked best, compared to Decision Trees, Support Vector Machine (SVM), Random Forest, and Naive Bayes in information–gathering tasks (53% and 78% accuracy, respectively), but results varied widely based on task granularity, likely due to the limited number of unique search strategies employed.

Eivazi et al. (2011) successfully used SVMs to classify high- and low- performers (73% accuracy) and their current task (95% accuracy). A further study using the same SVM found that predictions using fixation-saccade sequences were more successful than those using pupillary responses (Bednarik, Vrzakova, and Hradis 2012). However, these experiments only looked at binary classification problems with intentional gaze use. When eye movement is less directed, such as during natural hand-eye coordination tasks, expectations of gaze behavior are lower, as the task structure is less well known. Simola et al. (2008) found that a discriminative Hidden Markov Model (dHMM) predicted discrete processing states with 60% average accuracy, using fixation-saccade sequences during natural information search tasks. Again, predictive accuracy was found to vary widely between task types, with some tasks much easier to predict than others, due to task granularity.

These previous studies used various features related to either areas of interest (Eivazi and Bednarik 2011; Bondareva et al. 2013) or to specific gaze types, such as fixation duration, number of fixations, and mean saccade length (Simola, Salojärvi, and Kojo 2008; Steichen, Conati, and Carenini 2014). Furthermore, they have all solely focused on 2D interface interactions (Simola, Salojärvi, and Kojo 2008; Eivazi and Bednarik 2011; Bednarik, Vrzakova, and Hradis 2012; Steichen, Conati, and Carenini 2014; Bondareva et al. 2013) and not 3D tasks in physical work spaces. Therefore, further research into more effective features and learning in hand-eye coordination scenarios is strongly warranted.

3 Materials & Method

Experimental Setup

The data was collected from 7 participants (4 men, 3 women) with normal or corrected-to-normal vision, some of whom had taken part in previous gaze-tracking studies. All participants gave prior consent, and the experiment followed all regulations as required by the Georgia Institute of Technology's Central Institute Review Board.

Each participant played a modified version of the classic arcade-style "claw" game, where a grasper is moved in a small space to perform pick-and-place tasks in the presence of obstacles. Participants were instructed to (1) pick up a target object, (2) bring it to a pre-specified location, (3) hold it there until indicated, and then (4) drop the object off at the prize chute. Before the trial, a monitor to the side of the claw machine displayed the "hold" location. Some time after the "hold" was initiated, the monitor, through both visual and auditory cues, indicated to the participant when to stop holding and continue to take the object to the chute. In order to keep the participant's attention engaged, "holding" required actively pushing the target into the "hold" position.

The claw was controlled by three levers: two controlling one DoF each (left/right, forward/back) and a third which simultaneously lowered and opened (raised and closed) the grasper. The claw machine was limited to staying on for one minute, providing a natural cut-off for each trial.

After 3-5 practice runs to familiarize the participant with the experimental setup, each participant completed 10 trials, over which the target's orientation was randomly varied. The timing of each hold was randomized with a mean of 1.25 s ($\sigma^2 = 0.5$ s). The target's color, starting location, hold position, and the obstacles' positions were varied in a counterbalanced design, and the claw always started in the same position, over the goal, to ensure consistency.

The participant's gaze was recorded in real-time by a gaze-tracker, and their arm, head, and eye movements were wholly unconstrained. Furthermore, an egocentric scene camera mounted to the top of the gaze-tracker recorded the visual workspace data from the user's perspective.

Apparatus & Calibration

For gaze-tracking, a glasses-style head-mounted monocular (right eye) tracker with a scene camera that employs IR reflection was used (MobileEyeXG; Applied Sciences Laboratories, Bedford MA). The gaze-tracker recorded at 30 Hz with a latency of 100 ms and an accuracy of $0.5 - 1^\circ$. The

eye-tracking equipment can be seen in Figure 1. The MobileEyeXG software package was used to export the relative gaze position in the scene plane (2D, pixels) and egocentric video of the scene (RGB, 480x640 pixels).

Figure 1: The gaze-tracking system.

The calibration procedure followed that outlined by the gaze-tracker supplier with few variations (MobileEye 2013; 2014a; 2014b). A calibration scene with nine carefully-spaced points was placed parallel to the visual plane, at the back panel of the claw machine. The gaze-tracker was then calibrated by having the participants look at the nine points sequentially with the experimenter manually marking the intended point of gaze in the scene plane with the supplied software. After calibration, the participant's distance from the calibration plane was measured in order to calculate visual angle.

Feature Extraction

State Features Two sets of basic features were first extracted from the raw data: gaze and claw location. The former was a direct output of the EyeHead module of ASL Results Plus, which pre-processed the gaze data via the calibration. Since both the claw and gaze positions on the scene plane were recorded from an unstable platform, the participant's moving head, their relative position from a single fixed coordinate system was estimated using video stabilization and manual frame alignment. This was accomplished using a point feature mapping constructed by fast corner detection and then extracting the Fast Retina Keypoint (FREAK) descriptor at each point. The Hamming distance was then used as the matching cost between these descriptors. To refine correspondences between points, a variant of the RANSAC algorithm, M-estimator Sampling Consensus (MSAC), calculated the inlier correspondences and got the homogeneous geometric transform matrix between the collections of points in each frame. Instead of calculating this transform between each successive frame, the joint transform was taken between each frame and the initial frame, as the initial frame always captured the entire scene, including the calibration plane, from the most straight-on perspective due to its proximity to calibration. Once the current and initial frames were aligned, the transforms were applied to the raw gaze position data for each frame.

Post-stabilization, the claw's position in the scene plane was recorded manually every 10 frames, due to its significantly slower velocity and absence of internal sensors. Intermediate positions were then interpolated and its trajectory smoothed with a 5-point moving average filter. The gaze and claw data had to also be temporally synced. This was accomplished by manually identifying the location of the gaze point in the scene plane at the start of the trial and finding its matching frame number in the gaze-tracker data.

Finally, the scene data across all trials was aligned by marking the points at the upper left and lower right of the game apparatus' viewing window. Then an appropriate homogeneous transform was calculated and applied to both the claw and eye data. This last step enabled calculation of consistent gaze data statistics across trials in order to extract gaze type and remove outliers.

Due to the physical constraints of the oculomotor system, gaze velocity during normal usage has a Gaussian distribution and maximal velocity. Thus, outliers greater than 3σ were attributed to faulty sensor readings and blinks, and removed from the feature set (Komogortsev and Karpov 2013; Larsson et al. 2015). This exclusion of outliers is further justified by noting that they account for less than 2% of the entire training set (see Results below), thus proving statistically insignificant with regard to the results from the fully trained models. Outliers were also excluded from the claw data, for which jumps were rarer and mostly attributable to quick movements of the head, where the participant turned away from the task and the eye lost contact with the claw's workspace.

In addition to the positional data, the eye's and claw's velocity components were extracted as features.

Gaze Classification Based on previous gaze-tracking work involving eye-hand coordination tasks, gaze behavior is likely to be a strong predictor of imminent tasks (Hayhoe 2000; Mennie, Hayhoe, and Sullivan 2007). Most gaze classifiers just differentiate between two states, fixations and saccades (Salvucci and Goldberg 2000b; Duchowski 2007; Munn, Stefano, and Pelz 2008; Olsen 2012). However, as our task structure contains dynamic, continuous movements of the claw, smooth pursuit is expected to weakly dominate other gaze behaviors. Thus, a ternary classifier that identifies smooth pursuits, as well as fixations and saccades, was implemented (Larsson et al. 2015; Gyllensten 2014).

Before gaze behavior classification, a series of preprocessing steps was carried out. First the raw gaze data was transformed into visual angle v° as

$$v^\circ = 2 \arctan\left(\frac{p - p_c}{2d}\right) \quad (1)$$

where p is the gaze position, p_c is the position of the center of the calibration plane, and d is the participant's distance from the calibration plane. Then blinks and one-sample spikes were detected and removed per Larsson et al. (2013), using local extrema detection around missing data samples for blinks (blink window = 0.5 s) and a median filter (length 3) with a minimum activation threshold of 0.3°.

Classification of saccades was performed via a simple velocity threshold $\tau_v = 80^\circ$/s per Komogortsev et al. (2010). Saccade and blink locations were then used to segment the gaze data into windows for fixation/smooth pursuit classification. Four criteria were used for this purpose: dispersion, directional consistency, positional displacement, and spatial

Table 1: Parameters for gaze processing. Sources are listed for parameter definitions and starting points for parameter selection, with some differences due to sampling rate. Parameters were tuned experimentally for best performance.

Parameter	Symbol	Value	Source
Max. Window Size	w_{mx}	166 ms	(Gyllensten 2014)
Velocity Threshold	τ_v	$80^\circ/s$	(Komogortsev et al. 2010)
Min. Fixation Duration	t_{mn}	67 ms	(Larsson et al. 2015)
Parameter thresholds	η_D	0.5	(Larsson et al. 2015)
	η_{CD}	0.6	
	η_{PD}	0.3	
	η_{Fix}	2.6	
	η_{Smp}	2.3	
	ϕ	$\pi/4$	

range. A complete discussion of these criteria and their calculations can be found in Larsson et al. (2015) and Gyllensten (2014), while our parameters for these calculations are listed in Table 1.

Ground Truth Data Labelling Three classes of intention are of interest: *'move', 'ready', 'hold'*. We approximated these intentions based on when the claw was moving and not moving (*'move'/'hold'*) and about to move (*'ready'*). The latter was defined as 210 ms before movement commenced, which is around the mean reaction time for hand movement during hand-eye coordination (Gielen, Van den Heuvel, and Van Gisbergen 1984). Movement was defined with velocity ($\geq$ 6.7px/ms) and dispersion ($\geq$ 80px) thresholds, both determined experimentally. A Hampel filter was applied to claw movement to remove one-sample spikes, reducing noise.

Feature Offsetting To test if intent classification improved by considering eye movements that precede hand movement, feature sets were re-calculated with the claw data offset by an increasing number of frames. The combined features form something akin to a Hankel matrix:

$$\begin{bmatrix} [c_0\ g_0] & [c_1\ g_0] & \cdots & [c_q\ g_0] \\ [c_1\ g_1] & [c_2\ g_1] & \cdots & [c_{q+1}\ g_1] \\ [c_2\ g_2] & [c_3\ g_2] & \cdots & [c_{q+2}\ g_2] \\ \vdots & \vdots & \ddots & \vdots \\ [c_{n-q}\ g_{n-q}] & [c_{n-q+1}\ g_{n-q}] & \cdots & [c_n\ g_{n-q}] \end{bmatrix} \quad (2)$$

such that the claw features at a given time c_x are offset by q in time with the gaze features g_{x-q}.

Learning Algorithms

Eight classifiers, including all five which had previously been used in intent estimation from gaze data (see Background) and three other popular techniques, were compared against each other in combination with varying feature sets. The algorithms used were Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), k-Nearest Neighbors (KNN), Naive Bayes (NB), C4.5 Decision Tree (C4.5), multinomial logistic regression, Support Vector Machine (SVM), and Hidden Markov Models (HMM). The first five were implemented using MATLAB's Statistics and Machine Learning Toolbox, logistic regression from Chen (2016), SVM from LibSVM, and HMM from PMTK3 (Dunham and Murphy 2010). The list of classifiers is not meant to be exhaustive, just demonstrative of how various feature sets perform and whether anticipatory gaze data proves useful in intent prediction. The SVM and logistic regression were implemented with one-against-all coding, and the SVM parameters were chosen by implementing a 5x5 grid-style parameter selection with 5-fold validation. The HMM used a mixed Gaussian model with 3 nodes, which was determined to work best experimentally. All algorithms were externally tested with 5-fold validation, with 20% of the data unseen in training being tested. Averages of accuracy and area under the curve (AUC) from the Receiver Operating Characteristic scores were calculated for all of the classifiers.

4 Results

Out of 70 trials, 7 trials from 6 participants were deemed unsuitable for feature extraction due to the amount of head movement, image quality, and poor calibration. The remaining 63 trials yielded 16,373 gaze samples. While samples with all gaze types, including unclassified ones, were used for learning, 1.8% of the samples had unclassifiable gaze types due to blinks, reflections, and dropped frames.

The class labels were imbalanced such that *'ready'/'move'/'hold'* were respectively represented by 8.1%, 77.0%, and 14.9% of the samples. Thus, when comparing classifiers, metrics that take into account precision and sensitivity, such as AUC, should be considered and not just accuracy.

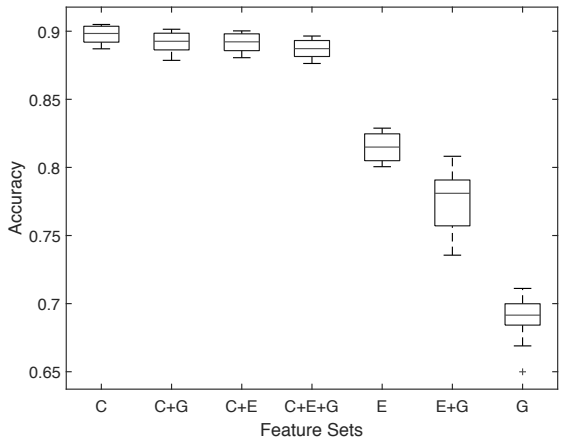

Figure 2: Feature set performance averaged over all classifiers and time offsets. Feature sets are a combination of claw (C), eye (E), and gaze type (G) features.

Feature Set Performance

Figure 2 compares performance between claw (C), eye (E), and gaze (G) type features and their combinations. As expected, claw features are best for predicting intention (89% accuracy, 0.82 AUC), since intention was attributed based on claw behavior. While not as effective, eye position data still performed fairly well on average ($\approx$ 80% accuracy, 0.63 AUC), while gaze type alone was the weakest predictor (69% accuracy, 0.51 AUC).

Another trend to note is that the combined feature sets tended to have weaker performance (*e.g.* C vs. C+E, E vs E+G) suggesting that perhaps despite 5-fold cross-validation, overfitting occurred due to inclusion of too many features. This trend does not hold for E vs. C+E, where the claw features clearly improve upon the accuracy of E alone.

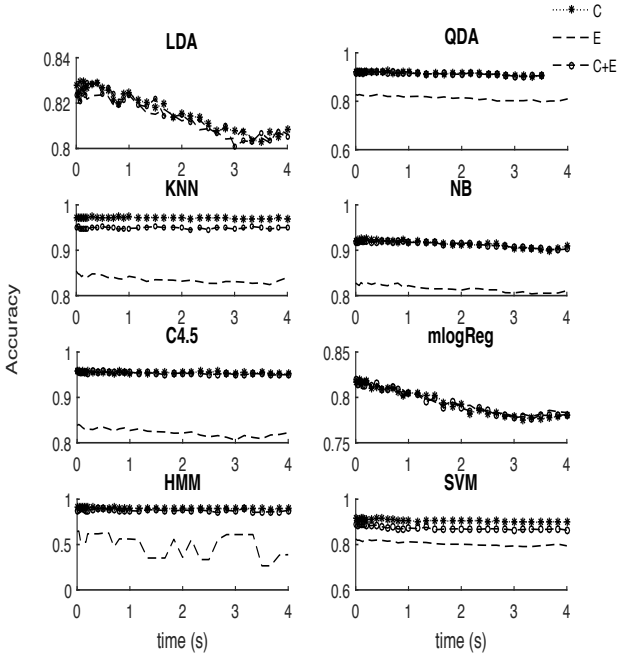

Figure 3: Feature set performance over time for each classifier over claw features (C), eye features (E), and their combination (E+C). mlogReg stands for multinomial logistic regression.

Classifier Performance Over Time

Figure 3 depicts classifier accuracy with increasing prediction horizon. Feature sets that included the claw (C and C+E) always outperformed eye features alone, with the exceptions of LDA and logistic regression. On average (Figure 4), time had a small negative effect on accuracy (-0.62%/s) but KNN on gaze type alone actually yielded a small positive effect of 0.31%/s. Conversely, the classifier/feature combination with the largest negative effect was HMM with eye and gaze type features, with an effect of -4.5%/s. While accuracy generally decreased smoothly against prediction time, this was not the case for the HMM. The HMM results display that training

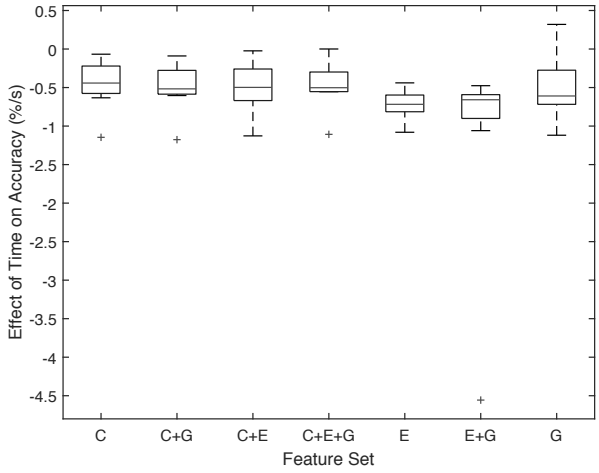

Figure 4: Effect of offset time on accuracy averaged over all classifiers on claw features (C), eye features (E), and their combination (E+C).

tended toward one of two end-states, such that the results appear to oscillate between them. Note, however, that independent HMM's were trained at each time slice. One group of HMM's reached a "high performing" end-state and the other group, a "low performing" one; thus, careful attention should be paid to model validation if this HMM is to be used again.

Overall Classifier Performance

Overall classifier performance is given in Table 2 and shows that the C4.5 decision tree algorithm yielded the highest average accuracy and AUC (90%, 0.86) over all feature sets and time offsets, while KNN had the single maximum accuracy (97%, 0.97 AUC) averaged over time for a specific feature set. C4.5, KNN, HMM, QDA, and NB all performed very well, and their results compare favorably to previous work; however, these results stand in contrast to those works in which SVM and logistic regression were found to perform best (Eivazi and Bednarik 2011; Simola, Salojärvi, and Kojo 2008; Steichen, Conati, and Carenini 2014; Bondareva et al. 2013). In our research they ranked as the two worst performers among those tested. While they had fairly high accuracy, their AUC scores were significantly weaker than the other classifiers' scores. Overall, the classifier/feature combinations used in this research yielded higher average accuracies than nearly all previous studies.

5 Discussion

Effectiveness of Gaze Data for Prediction

In regard to the first part of the research question, we can now answer that while not as effective as using the claw features, eye features including location, visual angle, and velocity all yield respectable classification accuracies of 75-85% (0.63 AUC), while gaze type alone performed weakly

Table 2: Classifier performance ranked by $\overline{\text{AUC}}$, that is AUC averaged over all classes and features. $\overline{\text{Acc}}$ is the classifier's accuracy at $\overline{\text{AUC}}$. Acc* is the maximum accuracy obtained by each classifier on any given feature set, and AUC* is the AUC at Acc*. mlogReg is multinomial logistic regression.

Method	$\overline{\text{Acc}}$	$\overline{\text{AUC}}$	Acc*	AUC*
C4.5	0.900	0.861	0.959	0.947
KNN	0.879	0.811	0.974	0.970
HMM	0.845	0.806	0.918	0.942
QDA	0.881	0.796	0.926	0.985
NB	0.881	0.781	0.925	0.912
LDA	0.817	0.604	0.831	0.565
SVM	0.852	0.590	0.917	0.589
mlogReg	0.798	0.576	0.825	0.546

(69%, 0.51 AUC). Thus, eye data can yield fairly accurate classification of intent when the end-effector features are unavailable during task performance.

Improvement of Classification With Incorporation of Gaze Data

When gaze data was no longer considered alone but in combination with the claw data, it was not found to significantly improve overall accuracy. Time in general had a small negative interaction effect with all feature sets, and including eye location or gaze type did not seem to mitigate this effect, with the exception of gaze type when classified with KNN. In general, KNN accuracy was affected the least by lag time, falling only -0.1%/s on average. Perhaps this is due to KNN's dependence on local feature structure and the local, if not offset, correlation between hand and eye position.

Overall, we had expected to find that including gaze data would improve classification because of the anticipatory nature of many gaze behaviors, which does not seem to be the case here. One possible explanation is the dominance of the smooth pursuit behavior (58%) among the gaze types. While smooth pursuit can be anticipatory, it tends to slightly lag or stick very close to the tracked object, and only occasionally overtake that object (Thier and Ilg 2005). Another factor that must be taken into account is that only 20% of fixations, which themselves accounted for 34% of all gaze types, are used to look ahead at relevant targets and goals (Mennie, Hayhoe, and Sullivan 2007). While saccades are often also anticipatory, they only accounted for 6% of the observed gaze behavior. Many saccades were likely also lost due to the 30–Hz frame rate of the camera, given that saccades tend to last less than 100 ms.

A final explanation is that eye behavior changes with task-specific experience and expertise. Sobuh at al. (2014) showed that novices learning to operate an upper body prosthesis tended to have more fixations, more erratic gaze behavior, and more transitions between areas of interest. Experts, on the other hand, tended to have a more systemic and simpler fixation trajectory. Furthermore, novices tend to oscillate their fixations between the targets and the manipulator, while experts focus on the targets. While participants in our study were given the chance to get comfortable with operating the machine, they may have fixated more on their hands as opposed to the robotic manipulator to make sure they were placed correctly on the control levers. These possibilities require further investigation before the usefulness of gaze data in furthering intent prediction is ruled out.

6 Conclusion

This research investigated the potential of gaze data to improve intent prediction in visuomotor coordination tasks. In doing so, it yielded the first comparison of classification algorithms for intent prediction during hand-eye coordination in a 3D workspace, finding that the C4.5 decision tree, KNN, and HMM algorithms worked best, with accuracies above 94%. It found that gaze features alone could yield reasonable predictions (80% accuracy) and that accuracy for nearly all classifier and feature combinations decreases slightly with lengthening horizon. Adding gaze features to claw features was not found to significantly improve prediction accuracy, despite theoretical expectations. Future work should focus on explaining the mismatch between the science of hand-eye coordination and the findings presented here. By working toward a more complete understanding of anticipatory gaze–tracking, not only may intent prediction be improved but also an important contribution can be made to cognitive human-robot interaction.

7 Acknowledgements

We would like to thank Srishti Gupta for helping to extract feature data. Funding for this project was provided by the National Science Foundation through an NSF NRI-Small Project grant NSF IIS-1317718

References

Barnes, G., and Marsden, J. 2002. Anticipatory Control of Hand and Eye Movements in Humans During Oculo–Manual Tracking. *The Journal of Physiology* 539(1):317–330.

Bednarik, R.; Vrzakova, H.; and Hradis, M. 2012. What do You Want to do Next: A Novel Approach for Intent Prediction in Gaze–Based Interaction. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, 83–90. ACM.

Bondareva, D.; Conati, C.; Feyzi-Behnagh, R.; Harley, J. M.; Azevedo, R.; and Bouchet, F. 2013. Inferring Learning from Gaze Data During Interaction with an Environment to Support Self-Regulated Learning. In *International Conference on Artificial Intelligence in Education*, 229–238. Springer.

Carpenter, R. H. 1991. *Eye Movements*, volume 8. Boca Raton, Flor.: CRC Press LLC, 1st edition.

Chen, M. 2016. Logistic Regression for Classification (software). www.mathworks.com/matlabcentral/fileexchange/55863-logistic-regression-for-classification/. Last accessed: 11-22-2016.

Crawford, J. D.; Medendorp, W. P.; and Marotta, J. J. 2004. Spatial Transformations for Eye–Hand Coordination. *Journal of Neurophysiology* 92(1):10–9.

Duchowski, A. 2007. *Eye Tracking Methodology: Theory and Practice*. Springer.

Dunham, M., and Murphy, K. 2010. Probabilistic Modeling Toolkit for Matlab/Octave (software). github.com/probml/pmtk3. Last accessed: 11-22-2016.

Eivazi, S., and Bednarik, R. 2011. Predicting Problem–Solving Behavior and Performance Levels from Visual Attention Data. In *Proc. Workshop on Eye Gaze in Intelligent Human Machine Interaction at IUI*, 9–16.

Gielen, C.; Van den Heuvel, P.; and Van Gisbergen, J. 1984. Coordination of Fast Eye and Arm Movements in a Tracking Task. *Experimental Brain Research* 56(1):154–161.

Gyllensten, O. C. 2014. Evaluation of Classification Algorithms for Smooth Pursuit Eye Movements: Evaluating Current Algorithms for Smooth Pursuit Detection on Tobii Eye Trackers. Master's thesis, NADA, KTH, Stockholm, Sweden.

Hayhoe, M., and Ballard, D. 2005. Eye Movements in Natural Behavior. *Trends in Cognitive Sciences* 9(4):188–94.

Hayhoe, M. M.; Shrivastava, A.; Mruczek, R.; and Pelz, J. B. 2003. Visual Memory and Motor Planning in a Natural Task. *Journal of Vision* 3(1):6.

Hayhoe, M. 2000. Vision using Routines: A Functional Account of Vision. *Visual Cognition* 7(1-3):43–64.

Johansson, R. S.; Westling, G.; Bäckström, a.; and Flanagan, J. R. 2001. Eye-Hand Coordination in Object Manipulation. *The Journal of Neuroscience* 21(17):6917–32.

Komogortsev, O. V., and Karpov, A. 2013. Automated Classification and Scoring of Smooth Pursuit Eye Movements in the Presence of Fixations and Saccades. *Behavior Research Methods* 45(1):203–215.

Komogortsev, O. V.; Gobert, D. V.; Jayarathna, S.; Koh, D. H.; and Gowda, S. M. 2010. Standardization of Automated Analyses of Oculomotor Fixation and Saccadic Behaviors. *IEEE Transactions on Biomedical Engineering* 57(11):2635–2645.

Land, M. F., and Hayhoe, M. 2001. In What Ways do Eye Movements Contribute to Everyday Activities? *Vision Research* 41(25):3559–3565.

Land, M.; Mennie, N.; and Rusted, J. 1999. The Roles of Vision and Eye Movements in the Control of Activities of Daily Living. *Perception* 28(11):1311–1328.

Land, M. F. 2006. Eye Movements and the Control of Actions in Everyday Life. *Progress in Retinal and Eye Research* 25(3):296–324.

Larsson, L.; Nyström, M.; Andersson, R.; and Stridh, M. 2015. Detection of Fixations and Smooth Pursuit Movements in High-Speed Eye-Tracking Data. *Biomedical Signal Processing and Control* 18:145–152.

Larsson, L.; Nyström, M.; and Stridh, M. 2013. Detection of Saccades and Postsaccadic Oscillations in the Presence of Smooth Pursuit. *IEEE Transactions on Biomedical Engineering* 60(9):2484–2493.

Ma-Wyatt, A.; Stritzke, M.; and Trommershäuser, J. 2010. Eye-hand Coordination While Pointing Rapidly Under Risk. *Experimental Brain Research* 203(1):131–45.

Mennie, N.; Hayhoe, M.; and Sullivan, B. 2007. Look-Ahead Fixations: Anticipatory Eye Movements in Natural Tasks. *Experimental Brain Research* 179(3):427–442.

Miall, R., and Reckess, G. Z. 2002. The Cerebellum and the Timing of Coordinated Eye and Hand Tracking. *Brain and Cognition* 48:212–226.

MobileEye, A. 2013. Eye Tracker Systems Manual: MobileEyeXG. Technical Report Manual Version 1.5, Applied Science Laboratories, Bedford, MA.

MobileEye, A. 2014a. MobileEye EyeHead Integration Manual. Technical Report Manual Version 1.05, Applied Science Laboratories, Bedford, MA.

MobileEye, A. 2014b. Technical Note 5094: Using Polhemus FASTRAK as Head Tracker with MobileEye EyeHead Integration. Technical Report Manual Version 1.5, Applied Science Laboratories, Bedford, MA.

Mrotek, L. A., and Soechting, J. F. 2007. Target Interception: Hand–Eye Coordination and Strategies. *The Journal of Neuroscience* 27(27):7297–7309.

Munn, S. M.; Stefano, L.; and Pelz, J. B. 2008. Fixation–Identification in Dynamic Scenes: Comparing an Automated Algorithm to Manual Coding. In *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization*, 33–42. Los Angeles, Calif.: ACM.

Olsen, A. 2012. The Tobii I-V Fixation Filter. Technical report, Tobii Technology.

Pelz, J. B., and Canosa, R. 2001. Oculomotor Behavior and Perceptual Strategies in Complex Tasks. *Vision Research* 41(25):3587–3596.

Posner, M. I. 1980. Orienting of Attention. *Quarterly Journal of Experimental Psychology* 32(1):3–25.

Reina, G. A., and Schwartz, A. B. 2003. Eye–Hand Coupling During Closed-Loop Drawing: Evidence of Shared Motor Planning? *Human Movement Science* 22(2):137–152.

Sailer, U.; Flanagan, J. R.; and Johansson, R. S. 2005. Eye-hand Coordination During Learning of a Novel Visuomotor Task. *The Journal of Neuroscience* 25(39):8833–8842.

Salvucci, D. D., and Goldberg, J. H. 2000a. Identifying Fixations and Saccades in Eye–Tracking Protocols. In *Proceedings of the 2000 Symposium on Eye tracking research & Applications*, 71–78. Palm Beach Gardens, Flor.: ACM.

Salvucci, D. D., and Goldberg, J. H. 2000b. Identifying Fixations and Saccades in Eye-Tracking Protocols. In *Eye Tracking Research & Applications Symposium*, 71–78.

Simola, J.; Salojärvi, J.; and Kojo, I. 2008. Using Hidden Markov Model to Uncover Processing States from Eye Movements in Information Search Tasks. *Cognitive Systems Research* 9(4):237–251.

Sobuh, M. M. D.; Kenney, L. P. J.; Galpin, A. J.; Thies, S. B.; McLaughlin, J.; Kulkarni, J.; and Kyberd, P. 2014. Visuomotor Behaviours when using a Myoelectric Prosthesis. *Journal of Neuroengineering and Rehabilitation* 11(1):72.

Steichen, B.; Conati, C.; and Carenini, G. 2014. Inferring Visualization Task Properties, User Performance, and User Cognitive Abilities from Eye Gaze Data. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 4(2):11.

Thier, P., and Ilg, U. J. 2005. The Neural Basis of Smooth–Pursuit Eye Movements. *Current Opinion in Neurobiology* 15(6):645–652.

Vercher, J.-L., and Gauthier, G. 1992. Oculo–Manual Coordination Control: Ocular and Manual Tracking of Visual Targets with Delayed Visual Feedback of the Hand Motion. *Experimental Brain Research* 90(3):599–609.