Predicting an Observer's Task using Multi-Fixation Pattern Analysis

Christopher Kanan[*]
California Institute of Technology

Nicholas A. Ray[†]
University of California San Diego

Dina N.F. Bseiso[‡]
University of California San Diego

Janet H. Hsiao[§]
University of Hong Kong

Garrison W. Cottrell[¶]
University of California San Diego

Abstract

Since Yarbus's seminal work in 1965, vision scientists have argued that people's eye movement patterns differ depending upon their task. This suggests that we may be able to infer a person's task (or mental state) from their eye movements alone. Recently, this was attempted by Greene et al. [2012] in a Yarbus-like replication study; however, they were unable to successfully predict the task given to their observer. We reanalyze their data, and show that by using more powerful algorithms it is possible to predict the observer's task. We also used our algorithms to infer the image being viewed by an observer and their identity. More generally, we show how off-the-shelf algorithms from machine learning can be used to make inferences from an observer's eye movements, using an approach we call Multi-Fixation Pattern Analysis (MFPA).

CR Categories: I.5.4 [Pattern Recognition]: Applications;

Keywords: Eye Tracking, Yarbus, Machine Learning

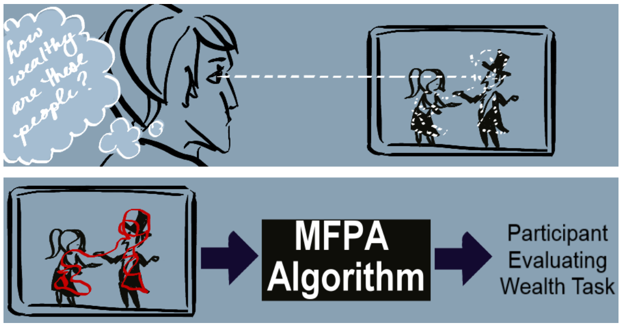
how wealthy are these people?
MFPA Algorithm
Participant Evaluating Wealth Task

Figure 1: *MFPA algorithms take a scan path's time-series features as input and use them to make inferences about a person solely from their eye movements. In this paper we infer the tasks given to subjects, the identities of the subjects, and the images viewed by the subjects.*

1 Introduction

Yarbus [1967] showed that an observers task can drastically alter their scan path (eye movement trajectory) when viewing a scene. While recording his subject's eye movements, Yarbus showed the subject a scene and gave various instructions, such as "estimate the material circumstances of the family in the picture" and "give the ages of the people in the picture." His subjects scan paths revealed considerable qualitative differences across the tasks. Future researchers confirmed Yarbus's general result in a variety of eye tracking experiments [Ballard et al. 1995; Castelhano et al. 2009; DeAngelus and Pelz 2009; Hagemann et al. 2010; Hayhoe et al. 2003; Kaakinen and Hyona 2010; Land et al. 1999; Tatler et al. 2011]. Since an observer's task influences their scan paths, it may be possible to infer from an observer's scan path what that observer is attempting to accomplish. We refer to algorithms that attempt to make inferences from eye tracking data as Multi-Fixation Pattern Analysis (MFPA).

In a study similar to Yarbus's [1967], Greene et al. [2012] attempted to use MFPA to infer their subject's tasks. In their experiment, the subjects' eye movements were recorded while photographs were shown to them. The subjects were asked to perform one of the following four tasks: (1) determine the decade in which the picture was taken, (2) memorize the picture, (3) determine the wealth of the people in the picture, or (4) determine how well the people in the picture know each other. Using both the visual and motor eye movement summary statistics from each trial as input, Greene et al's algorithm did not exceed chance at inferring the task given to their subjects. They concluded that "static scan paths alone do not appear to be adequate to infer complex mental states of an observer." We believe this conclusion deserves further scrutiny for two reasons.

The first reason is that Greene et al. were able to train their algorithm to successfully infer subject identity. Since this was possible, it suggests that subjects may have very individualized scan path patterns. For task-prediction, this means that it may be wise to train a task-prediction classifier for each subject individually using a within-subjects analysis. This was not explored in their study, and they trained their task-prediction classifier using data from all subjects. The second reason is that their approach turns each trial's time-series features into a single vector by calculating the trial's summary statistics, e.g., the number of fixations in the trial and the amount of dwell time on faces in the trial. This approach discards a large amount of data within each trial, including where the subject looked and the sequential pattern of eye movements that they made. Using a more sophisticated technique that preserves this information needs to be explored.

In this paper, we reanalyze Greene et al.'s data and conclusions. We found that their data does contain sufficient information to infer the tasks given to their subjects. To do this, we used two different MFPA algorithms. The first algorithm uses only summary statistics features to represent each trial, while the second algorithm can preserve time-series information within a trial. Further, we conduct a within-subjects analysis and show that this allows us to infer a subject's task using very little data.

[*] e-mail: ckanan@caltech.edu
[†] e-mail:niray@ucsd.edu
[‡] e-mail:dbseiso@ucsd.edu
[§] e-mail:jhsiao@hku.hk
[¶] e-mail:gary@ucsd.edu

Figure 2: *Four example images used in Greene et al.'s study.*

2 MFPA Algorithms

We compare two algorithms for MFPA. The first is similar to Greene et al.'s method in computing summary statistics; however, unlike Greene et al., we do not include any features that refer to the image - only the fixation locations over time. Summary statistics integrate over time, removing temporal information. The second algorithm makes explicit use of the temporal dynamics of the eye position recordings.

2.1 Summary Statistics Algorithm

The main difficulty in constructing a classifier for eye tracking data is that a variable number of fixations occur within a trial. This means that using an off-the-shelf classification algorithm requires turning this time-series data into a single vector. Greene et al. [2012] do this by generating a single vector of summary statistics per trial. We used a similar approach. For each trial's summary statistics features, we used 2-dimensional features consisting of the mean fixation duration and the number of fixations. Both of these motor features were used by Greene et al., but they also used mean saccade amplitude and three visual features: the amount of dwell time on faces, bodies, and objects. We normalized (z-scored) the training and testing data by subtracting the mean of the training data from the training and testing features and then dividing by the training standard deviation. Subsequently, a radial-basis function support vector machine (SVM) was used to classify these features. To train the classifier, we used the C-SVC algorithm from the LIBSVM toolbox [Chang and Lin 2011]. Four-fold cross-validation using the training data was used to tune the SVM cost parameter and the width of the radial-basis functions. Both were chosen from $2^{-8}, 2^{-7}, \dots 2^{8}$.

2.2 Fisher Kernel Learning Algorithm

Using summary statistics is one way to transform each trial's variable-length time-series features into a single vector, but it discards a great deal of information. Each trial's features are comprised of a variable number of 3-dimensional feature vectors (one vector per fixation), which contain the fixation's Cartesian screen coordinates and the duration of the fixation. Unlike using summary statistics, Fisher kernels are a way to preserve the temporal information within the time-series, while still condensing it into a single vector that can be used with a standard classifier [Jaakkola and Haussler 1998]. This is done by internally training a generative model, usually a hidden Markov model with Gaussian emissions, which can internally represent the temporal patterns. The parameters of the HMM, which include transition probabilities between hidden states and the probability of observing particular fixations given the internal state, will then reflect the sequential information in the data. The idea of a Fisher kernel is to compute how a new sequence of data would change the parameters of the model if the model were trained on them - that is, the parameter gradients of the generative model when given a novel time-series as input. This gradient is a fixed-length vector. Two time-series from the same category will likely change the model in the same way; the Fisher kernel computes a kind of inner product between them. However, Fisher kernel representations do not necessarily ensure that similar category data will lead to similar parameter changes, as it does not use the categories to compute the features. Fisher Kernel Learning (FKL) is an approach to rectify this limitation by using each time-seriess labels to improve the learned representation by explicitly attempting to keep the gradients within a category close together [van der Maaten 2011]. See van der Maaten [2011] for more details regarding Fisher kernel features and FKL.

We used FKL to turn each trial's time-series data into a single vector. To do this, we used van der Maaten's [2011] MATLAB software (available on his website), which uses a hidden Markov model with Gaussian emissions as the underlying generative model. We set the number of hidden states to 10, except for one of our within-subjects experiments where we set the number of hidden states to 5 instead. The number of hidden states alters the length of the FKL feature vectors, e.g., using 10 hidden states produces 140-dimensional vectors. Before training FKL, we normalized the 3-dimensional testing and training time-series features by subtracting the mean of the training features and then dividing by the training standard deviation across all trials.

Whitened principal component analysis (PCA) was used to reduce the dimensionality of the FKL features. Each trial's FKL features were classified using a radial-basis function SVM. Using the training data, the SVM cost parameter, the width of the radial-basis functions, and the number of principal components were tuned using 4-fold cross validation. The cost parameter and radial-basis function width were both chosen from $2^{-8}, 2^{-7}, \dots 2^{8}$. The number of principal components was chosen from $1, 5, 10, 20, 30, \dots, D$, where D is the dimensionality of the FKL features.

3 Results

3.1 Dataset

In their study, Greene et al. [2012] gathered eye movement data in three different experiments. We analyzed the data from their Experiment 3[1]. In this experiment, 20 grayscale photographs from Time Life magazine were shown to 16 subjects for 60 seconds each. Four of these images are shown in Figure 2. Each subject viewed

[1] The data from experiments 1 and 2 was not made available to us.

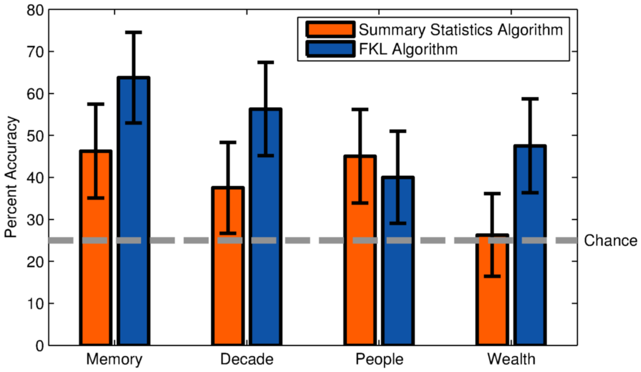

Figure 3: *Mean accuracy and 95% confidence intervals for each of the four tasks in the first within-subjects evaluation protocol. See Section 3.1. for more information about the tasks.*

the images in the same order. The experiment was divided into four blocks of five images each, with a Latin square design used to assign the block order to the participants. In the four blocks participants were either told to (1) memorize the picture [Memory], (2) determine the decade in which the picture was taken [Decade], (3) determine how well the people in the picture know each other [People], or (4) determine the wealth of the people in the picture [Wealth].

3.2 Task Prediction using Greene et al.'s Protocol

In each of Greene et al.'s [2012] cross-validation trials, they trained their classifiers using data from all participants and then tested on a single hold-out trial, so classifiers were trained and evaluated 320 times. In their experiment, prediction of task was at chance (25.9% correct, 95% CI = 21–31%, $p = 0.70$; chance = 25%).

In our experiment, the summary statistics algorithm did not significantly differ from chance (26.3% correct, 95% CI = 21.4–31.1%, $p = 0.61$), which is consistent with Greene et al.'s result; however, the FKL algorithm did perform above chance (33.1% correct, 95% CI = 27.9–38.3%). This suggests that using summary statistics alone discards a great deal of discriminative data for predicting an observer's task, which is preserved using FKL.

3.3 Within-Subject Task Prediction

In each of Greene et al.'s [2012] cross-validation trials, they trained their classifiers using data from all participants and then tested on a single hold-out trial. In this experiment, we explored training and evaluating classifiers using data from each subject individually. Because there are only 20 trials per subject, with 5 trials per task, we did this using two different evaluation protocols to cope with the small sample size. The small sample size is particularly problematic for FKL, because it has many more parameters than the summary statistics algorithm.

In the first, we used leave-one-out cross-validation using only the data from each participant individually, i.e., we trained each classifier using 19 trials, then tested on the one hold-out trial, and repeated this procedure 20 times per subject. Using an unbalanced number of training examples per class with a very small training set can impair predictive accuracy. To cope with this, for both algorithms we set the SVM cost parameter to be variable per class, with it being multiplied by $w_k = \left(n_k \left(\sum_{k'} n_{k'}^{-1}\right)\right)^{-1}$, where n_k is the number of training instances for category k. Note that this was only done in this experiment. Using this approach, both algorithms achieved above chance accuracy, with the summary statistics algorithm achieving 38.8% correct (95% CI = 33.4–44.1%; chance = 25%) and the FKL algorithm achieving 52.9% accuracy (95% CI = 46.4–57.4%). We show the performance per task in Figure 3.

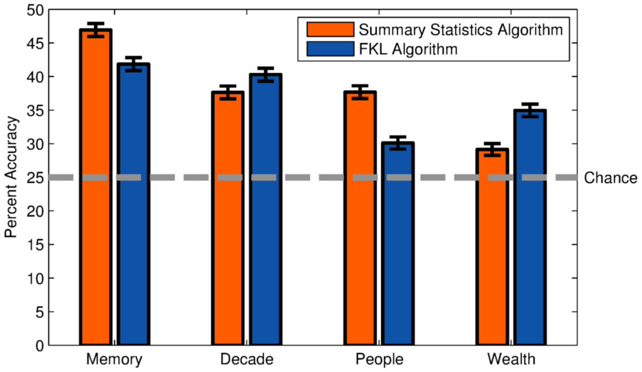

Figure 4: *Mean accuracy and 95% confidence intervals for each of the four tasks in the second within-subjects evaluation protocol. These results used 125 times more data than the first within-subjects protocol, which produced tighter confidence intervals. See Section 3.1. for more information about the tasks.*

In the second protocol, for each subject we train each classifier using 4 trials per category (16 trials total), and then test on the remaining 4 trials (1 per category). This was done for all 625 possible combinations of train and test data for each of the 16 subjects. Because FKL is relatively slow, for this experiment we used five hidden states instead of ten. Again, both algorithms achieved above chance accuracy, with the summary statistics algorithm achieving 37.9% accuracy (95% CI = 37.4–38.3%; chance = 25%) and the FKL algorithm achieving 36.8% accuracy (95% CI = 36.4–37.3%). The accuracy per task is shown in Figure 4. FKL does not perform as well as in the first within-subjects protocol. One potential reason why is that it is trained using five states instead of ten. When we use only 5 states in the first protocol, FKL's accuracy was reduced to 34.1% (95% CI = 28.8–39.3%). Other potential reasons why there may be a difference in performance between the two protocols are that (1) FKL is trained with slightly less data, so it may be over-fitting to a greater degree; and (2) adjusting the cost parameter in the first protocol may have inadvertently biased the model's predictions toward the correct answer.

3.4 Participant Identity Prediction

We also trained our algorithms to predict subject identity, using the same leave-one-out cross-validation protocol used by Greene et al. [2012], as described in Section 3.2. They achieved 42.8% correct (95% CI = 37 – 48%; chance = 6.3%), using a combination of motor and visual features. Using only motor features, the summary statistics algorithm achieved 31.3% accuracy (95% CI = 26.1–36.4%) and the FKL algorithm achieved 52.5% accuracy (95% CI = 47.0–58.0%). The likely reason why their summary statistics features outperformed ours is that they incorporated visual features. This suggests that if visual features were incorporated into the time-series data input into FKL, then its performance would markedly improve at identity prediction.

3.5 Image Prediction

Greene et al. [2012] also used their algorithm to predict which image was being viewed by a subject, using the same evaluation procedure as described in Section 3.2. They achieved 54.4% correct

(95% CI = 48–60%; chance = 5%). We used an evaluation procedure identical to theirs. Our summary statistics algorithm did not perform significantly better than chance (4.1% correct, 95% CI = 2.2–6.0%). In additional to the summary statistics features we used, Greene et al. also used visual features (the amount of dwell time on faces, bodies, and objects). Our results suggest that these visual features were key to predicting which image was being viewed in their study. However, the FKL algorithm did perform significantly above chance (41.6% correct, 95% CI = 36.1–47.0%), indicating that by incorporating information from the sequence of eye movements it is still possible to perform the inference, albeit with less accuracy compared to incorporating visual information as well.

3.6 Discussion & Future Work

We showed that there is sufficient information in Greene et al.'s [2012] data to infer the task given to their subjects using only motor information, i.e., fixation duration and the Cartesian locations of the fixations. While we confirmed their result that using summary statistics is not sufficient to infer the the subject's task when trained on data from all subjects, we found that we could make this inference using FKL. Moreover, we found that using a within-subjects analysis would allow us to infer the task given to the subject using either algorithm. Because our within-subjects analysis used very little training data but nevertheless achieved above-chance accuracy, this result suggests that individuals have distinct eye movement patterns when performing the same task, and thus one person's eye movement strategy may not generalize to other people.

While our algorithms were able to infer the observer's task, this is somewhat surprising because the subjects were not experts at any of the tasks given by Greene et al. [2012], e.g., prior to the experiment most of them probably lacked experience with judging the decade of a photo. We would expect that using tasks that people are experts at, such as recognizing facial expressions, would elicit scan paths with more stereotyped patterns per task. This would likely improve classification accuracy significantly. Using faces as stimuli also may be algorithmically beneficial for FKL. FKL does not compensate for stimuli in which discriminative information could appear in random positions (although it did work well in our experiments). Its performance would likely be improved if discriminative information only appeared in particular locations, which would be the case with aligned frontal face images. We are currently exploring this in a subsequent study.

The two feature construction algorithms we used turn a time-series into a single vector, which was then used as input to a separate classifier. Instead of this two-step process, an alternative would be to train a single model that can draw inferences from time-series directly, such as a Hidden-State Conditional Random Field [Truyen et al. 2008]. Using a single model could lead to improvements in predictive accuracy.

Our algorithms are general in design, and have many potential applications. One promising area of study is the identification of neurophysiological diseases. Tseng et al. [2012] pioneered this approach, showing that their algorithm could discern whether their subjects had Attention Deficit Hyperactivity Disorder, Parkinson's Disease, Fetal Alcohol Syndrome, or were disease free by combining visual saliency and eye movement data. Our results suggest that using motor activity alone may be sufficient to make these inferences. Further development of MFPA techniques could yield a diagnostic clinical tool that is both low-cost and high throughput.

Acknowledgements

C.K. was affilitated with UC San Diego when this work was completed. This project was supported in part by NSF REU Site grant SMA-1005256 as well as NSF grant SBE 0542013 and SMA 1041755 to the Temporal Dynamics of Learning Center.

References

BALLARD, D. H., HAYHOE, M. M., AND PELZ, J. 1995. Memory representations in natural tasks. *Journal of Cognitive Neuroscience 7*, 1, 66–80.

CASTELHANO, M. S., MACK, M. L., AND HENDERSON, J. M. 2009. Viewing task influences eye movement control during active scene perception. *Journal of Vision 9*, 3, 6.

CHANG, C., AND LIN, C. 2011. A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2*, 3, 1–27.

DEANGELUS, M., AND PELZ, J. B. 2009. Top-down control of eye movements: Yarbus revisited. *Visual Cognition 17*, 6-7, 790–811.

GREENE, M. R., LIU, T., AND WOLFE, J. M. 2012. Reconsidering Yarbus: A failure to predict observer's task from eye movement patterns. *Vision Research 62*, 1–8.

HAGEMANN, N., SCHORER, J., CANAL-BRULAND, R., LOTZ, S., AND STRAUSS, B. 2010. Visual perception in fencing: Do the eye movements of fencers represent their information pickup? *Attention, Perception, and Psychophysics 72*, 8, 2204–2214.

HAYHOE, M. M., SHRIVASTAVA, A., MRUCZEK, R., AND PELZ, J. 2003. Visual memory and motor planning in a natural task. *Journal of Vision 3*, 49–63.

JAAKKOLA, T., AND HAUSSLER, D. 1998. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems (NIPS-1998)*, 487–493.

KAAKINEN, J. K., AND HYONA, J. 2010. Task effects on eye movements during reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition 36*, 6, 1561–1566.

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.

TATLER, B. W., BADDELEY, R. J., AND VINCENT, B. T. 2011. The long and the short of it: spatial statistics at fixation vary with saccade amplitude and task. *Vision Research 46*, 1857–1862.

TRUYEN, T., PHUNG, D., BUI, H., AND VENKATESH, S. 2008. Hierarchical semi-markov conditional random fields for recursive sequential data. In *Advances in Neural Information Processing Systems (NIPS-2008)*.

TSENG, P., CAMERON, I., PARI, G., REYNOLDS, J., MUNOZ, D., AND ITTI, L. 2012. High-throughput classification of clinical populations from natural viewing eye movements. *Journal of Neurology 260*, 1, 275–284.

VAN DER MAATEN, L. 2011. Learning discriminative fisher kernels. In *Proc. 28th International Conference on Machine Learning (ICML-2011)*.

YARBUS, A. 1967. *Eye Movements and Vision*. New York: Plenum Press.