Human Point-of-Regard Tracking using State Space and Modular Neural Network Models

Jason L. Mitchell, and Ravi Kothari*
Artificial Neural Systems Laboratory
Department of Electrical & Computer Engineering and Computer Science
University of Cincinnati
Cincinnati, OH 45221-0030
E-Mail: ravi.kothari@uc.edu

Abstract

The presence of saccadic and smooth movements in the eye makes modular neural networks composed of two experts, each individually responsible for saccadic and smooth movements of the eyes, well suited for the tracking of human point-of-regard. To establish a basis for comparison on our data, we also consider a scalar ARMA model and a (vector) state space model. The purpose of this analysis is to build a reasonable model of human eye motion to use in prediction of point-of-regard. The ability to predict the point-of-regard of a human subject has applications in eye-tracking for man-machine interfacing, vigilance detection, and as a tool in cognitive psychology.

I Introduction

The ability to track point-of-regard is a necessary step towards using the eyes as an alternate input modality for computer users in general and disabled users in particular [1]. In addition, accurate tracking of the point of regard might permit the realization of complex displays in which an image is displayed with a higher resolution around point-of-regard regions, and in the automated assessment of operator vigilance. For such a system to be of practical use, it must be unobtrusive and perform the necessary operations of detecting the locations of the eyes [2], determining the point-of-regard, and anticipating (forecasting) the point-of-regard. Over the past few years a significant amount of literature with clinical or physiological measurement inclinations have appeared on the analysis of eye tracking movements [3]-[5] (see also [6]). Since eye movements can be characterized by saccadic (fast) and smooth (slow) movements, it seems natural to use modular neural networks composed of two experts, each individually responsible for saccadic

*Author to whom correspondence should be addressed

0-7803-3550-3/96$5.00©1996

and smooth movements of the eyes. To establish a basis for comparison on our data, we also consider a scalar ARMA model and a (vector) state space model.

The data we consider is obtained as a set of $\{(x,y) : x \in \{0,1,\ldots,511\}, y \in \{0,1,\ldots,511\}\}$ locations of the point-of-regard of a radiographer examining a chest x-ray, 26.67×26.67 cms, from a distance of 55 cms. The point-of-regard of the left eye was sampled at 51.5 Hz, using an ASL4000SU head mounted optics system, a miniature video camera, and software which tracks the pupil center and the first Purkinjie reflection of an incident IR source. The head position was monitored using a Flock of Birds magnetic head tracker. The two signals were integrated to calculate point-of-regard. A 9-point square dot pattern was used to calibrate the display space to yield an accuracy of 0.25 deg—0.50 deg, as subtended at the observer's eye. The point-of-regard data for a full 30 seconds is shown in Figure 1.

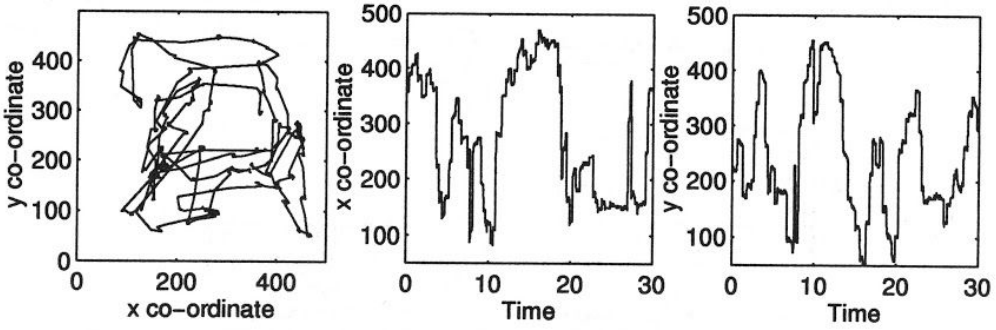

Figure 1: 30 seconds of point-of-regard data sampled at 51.5 Hz. The figure on the left shows the location (x,y), while the center and right panes show the x and y co-ordinates as functions of time

II Scalar ARMA Model

We begin the discussion of eye-movement prediction using simple ARMA modeling techniques [7],[8] to predict each component of the vector time series separately. It is clear, however, that there is some dynamic interdependence between the two components of the vector time series and it would be advantageous to use both when building a model. Thus a state space model of this vector-valued time indexed data is presented in the next section.

Figure 2 shows the sample autocorrelation function (AF) of the data. One may observe that neither the time series of the x, nor the time series of the y component is stationary. A first difference operator, as in many cases, is sufficient to transform the non-stationary time series values into stationary time series values (see the AF and the sample partial autocorrelation function (PAF) in Figure 3).

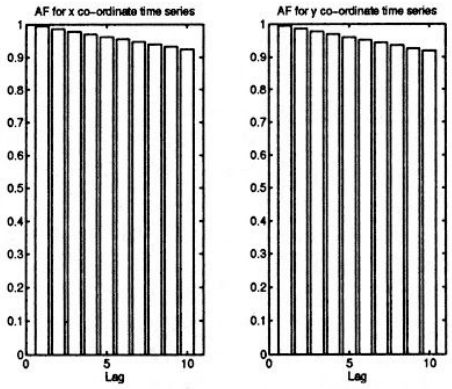

Figure 2: Autocorrelation function of the x co-ordinate (left), and the y co-ordinate time series (right)

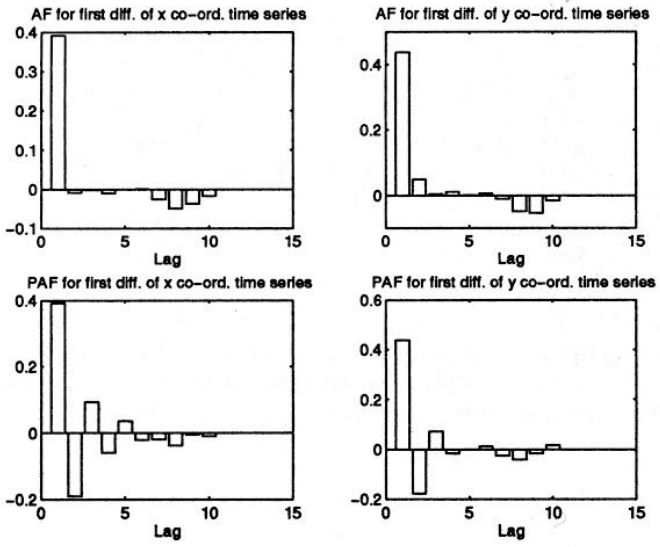

Figure 3: Autocorrelation function and partial autocorrelation function of the first difference of the x co-ordinate (left), and the first difference of the y co-ordinate time series (right)

Thus a generic ARMA model of order (p, q), as given below, can be used to track and predict the point-of-regard co-ordinates.

$$\phi_p(B)z_t = \delta + \theta_q(B)a_t \tag{1}$$

where, z_t are the stationary time series values, $\phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)$, and $\theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)$, δ is a constant term, a_t are 'random-shocks', and $B^k z_t = z_{t-k}$. In both the x and y cases, the AF displays a cut off after lag 1, and the PAF displays cut off after lag 1 or dies down fairly rapidly. Thus, we reduce the above ARMA model to a MA model of order 1, with $\delta = \mu_z$ — the true mean of the stationary time

series. Equation (1) above thus reduces to:

$$z_t = \mu_z + a_t - \theta_1 a_{t-1} \tag{2}$$

where $z_t = x_t - x_{t-1}$, and $z_t = y_t - y_{t-1}$ for the x and y components of the point-of-regard respectively.

The model parameter θ_1 was estimated using least squares with the first 25 seconds of the data, with the remaining 5 seconds used to determine the model's one-step-ahead prediction accuracy. The model parameter (θ_1) for the x-component was found to be 0.45211, while for the y-component it was 0.51557. Figure 4 shows the performance of the model of equation (2) for the x-component and the y-component respectively, and Figure 5 shows the difference between the actual time series values and the model output. Despite the simplicity of the model and despite ignoring the coupling between the x and y components of the series, one obtains a fairly accurate prediction of the point-of-regard. However, there remains a significant error at locations where large excursions of the point-of-regard occur. Statistics of the residuals in doing the one-step-ahead prediction are summarized in Table I.

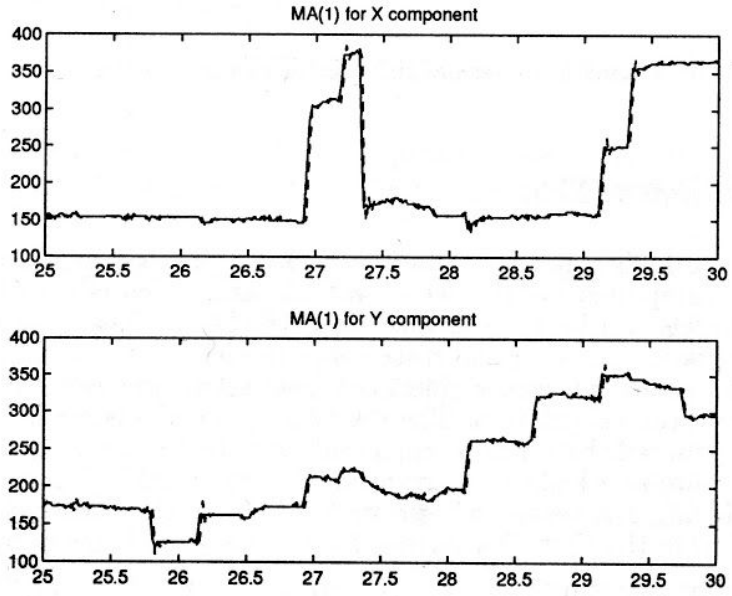

Figure 4: Performance of a first order moving average model at one-step-ahead prediction of the x-component (top), and the y-component (bottom) of the point-of-regard. The model output is shown by broken lines

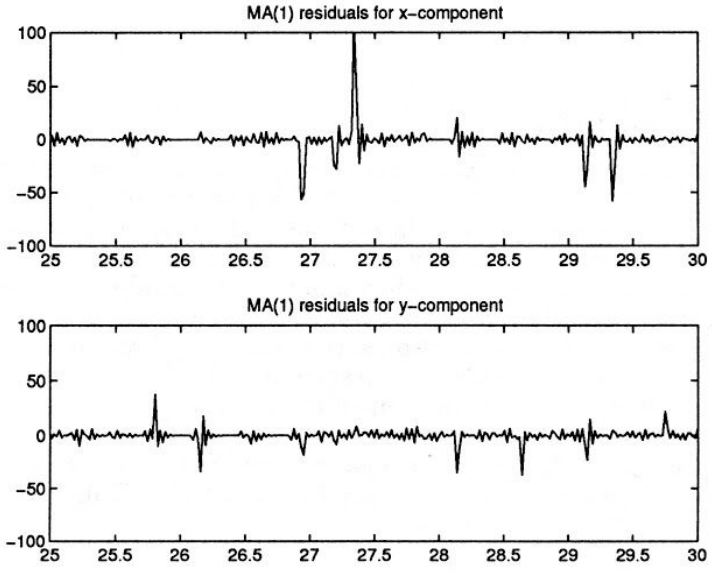

Figure 5: Difference between the actual time series values and the moving average model output

III State Space Model

A state space model is a natural choice to capture the dynamic interdependence between components of the vector-valued data. Theoretically however, state space models can be transformed into ARMA models systematically. The main motivation of using the state space model is thus to examine the benefits, if any, resulting from considering the interdependence between components of the vector-valued data. The state space model was generated using SAS [9] which proceeds by formulating a multivariate AR model whose order is chosen to minimize Akaike's Information Criterion (AIC) [10] and subsequently introducing MA terms to improve the fit [11]. The state space model thus obtained from the first 25 seconds of the time series is given by:

$$Z_{t+1} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.019 & 0.006 & -0.101 & -0.009 \\ 0.021 & -0.039 & -0.033 & 0.187 \end{bmatrix} Z_t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0.439 & -0.001 \\ 0.159 & 0.521 \end{bmatrix} a_t \tag{3}$$

where, $Z_t = [x_t \quad y_t \quad x_{t+1} \quad y_{t+1}]^T$. Figure 6 shows the performance of the model of equation (3) for the x-component and the y-component, and Figure 7 shows the difference between the actual time series values and the model output. As anticipated, developing a multivariate model to explicitly account for the coupling between the x and the y components has led to better performance. However, as in the case of the scalar ARMA model, there still remains a significant error at locations where large excursions of the point-of-regard occur. Statistics of the residuals in doing the one-step-ahead prediction are summarized in Table I.

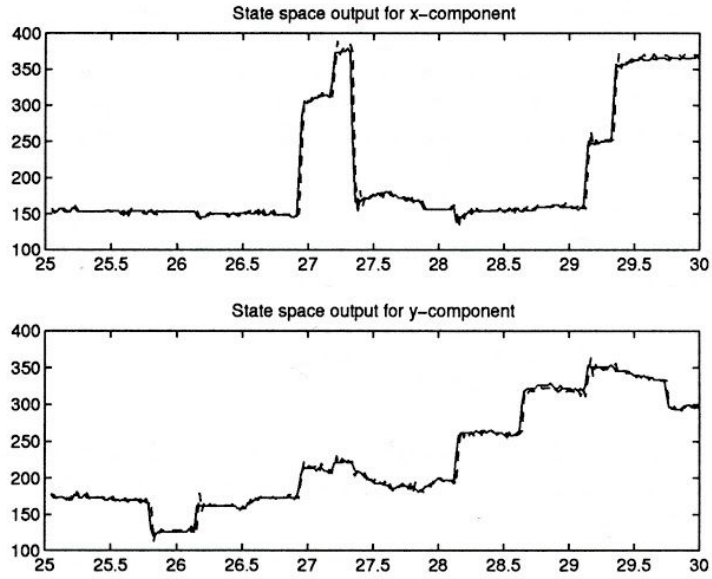

Figure 6: Performance of the state space model at one-step-ahead prediction of the x-component (top), and y-component (bottom) of the point-of-regard. The model output is shown by broken lines

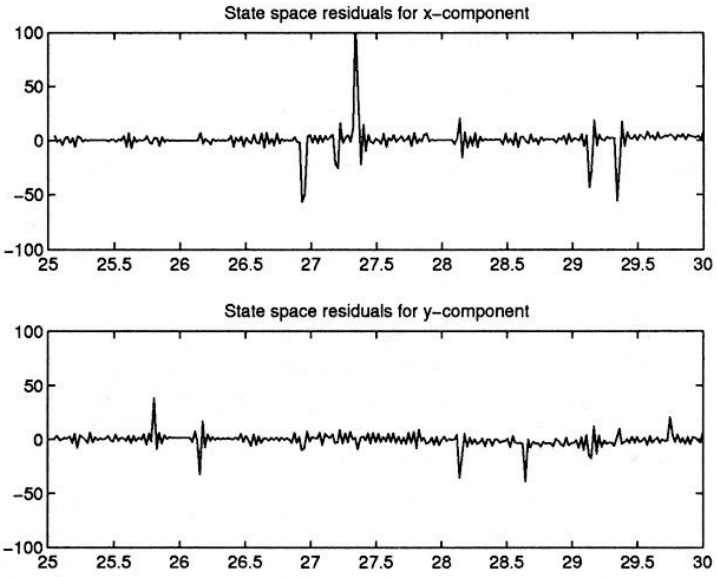

Figure 7: Difference between the actual time series values and the state space model output

IV Modular Neural Network Model

Modular neural networks consist of a group of networks, called *local experts*, and a *gating* network which is responsible for mediating the output of the local experts to form the final output of the system. The local expert networks and the gating networks are simultaneously trained [12],[13]. Assuming the output of the k^{th} local expert to follow a probability density function given by:

$$p(y_k) = e^{-0.5||d-y_k||^2} \tag{4}$$

The output of the entire network can then be expressed by a mixture probability density function:

$$p(y) = \sum_{k=1}^{M} g_k e^{-0.5||d-y_k||^2} \tag{5}$$

where, M is the number of local experts. One can thus perform maximum likelihood estimation i.e. for a given set of outputs $Y = \{y^{(t)} : t = 1, 2, \ldots, T\}$, we maximize:

$$\begin{aligned} J &= \ln\left[\prod_{t=1}^{T} p(y^{(t)})\right] \\ &= \sum_{t=1}^{T} \ln\left[\sum_{k=1}^{M} g_k e^{-0.5||d-y_k||^2}\right] \\ &= \sum_{t=1}^{T} J^{(t)} \end{aligned} \tag{6}$$

The output of the gating network is based on the softmax activation function:

$$g_k = \frac{e^{u_k}}{\sum_{i=1}^{K} e^{u_i}} \tag{7}$$

where u_k is the weighted sum received by a gating neuron.

Replacing g_k from (7), into (6), we obtain a cost function in terms of the weights of the experts (y_k is a function of each local expert weights), and the gating network (g_k is a function of the weights of the gating network). Using gradient ascent, one can thus maximize (6) [12],[13].

Figures 8 and 9 show the performance of the modular neural network model, consisting of two experts, each a multi-layered network with a single hidden layer of 10 hidden neurons, 4 inputs (the location co-ordinates at the current and previous time instants), and two outputs (the co-ordinates at the next time instant). We note the reduced one-step-ahead prediction error at locations where large excursions of the point-of-regard occur, while maintaining a maximum error of a few pixels during the smooth eye movements.

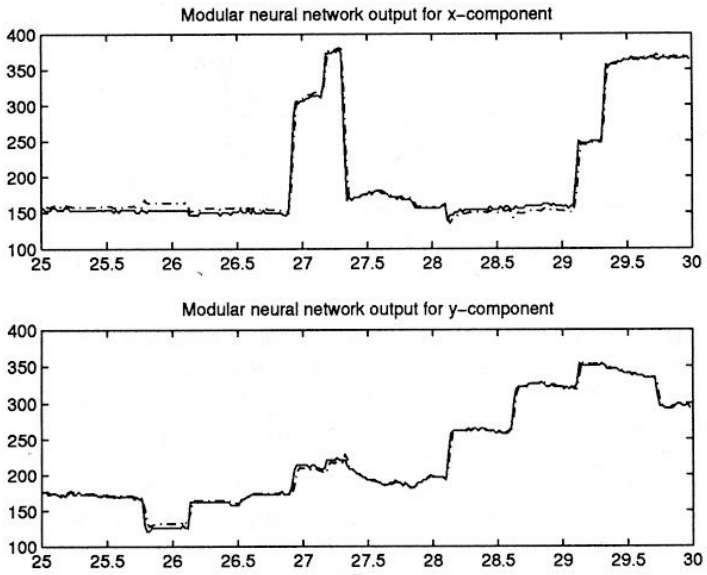

Figure 8: Performance of the modular neural network model at one-step-ahead prediction of the x-component (top), and y-component (bottom) of the point-of-regard. The model output is shown by broken lines

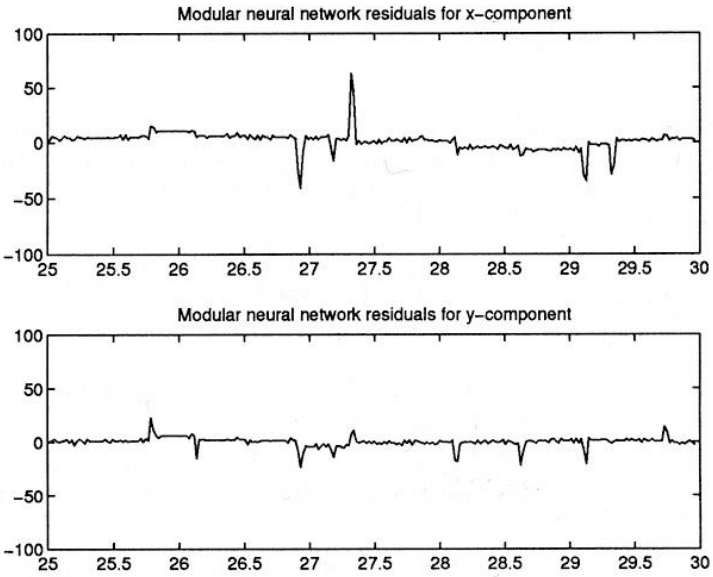

Figure 9: Difference between the actual time series values and the modular neural network model output

V Discussion and Conclusion

Statistics for the three models considered in this paper are shown below:

Model	*x-component*		*y-component*	
	μ	σ	μ	σ
ARMA	4.4835	10.3077	3.4192	5.2552
State space	4.7748	10.3830	3.9552	5.1708
Mod NN	5.6708	6.6523	2.8154	3.7025

Table I: Statistics of the one-step-ahead prediction errors for the different models

Our results show that the modular neural network model performs better than both the scalar MA and the vector state space model, specially when the point-of-regard undergoes rapid excursions. Additionally, these results must be seen in the context of the effort involved in creating the state space and the modular neural network model. While, the selection of the state space model order was carefully selected using AIC, the model order for the modular neural network was chosen arbitrarily. In addition, the evolution of the state space model from an initial AR to an ARMA model is a complex optimization process. Additional research with more complex experts in the modular neural networks might lead to a model of choice in tracking the point-of-regard.

References

[1] T. E. Hutchinson, K. P. White, W. N. Martin, K. C. Reichert, and L. A. Frey, "Human-Computer Interaction Using Eye-Gaze Input," **IEEE Transactions on Systems, Man and Cybernetics**, vol. 19, no. 6, pp. 1527-1534, November 1989.

[2] R. Kothari, and J. L. Mitchell, "Detection of Eye Locations in Unconstrained Visual Images," **Proc. IEEE International Conference on Image Processing**, Lausanne (Switzerland), September 1996, (to appear).

[3] M. Bach, D. Bouis, and B. Fisher, "An Accurate and Linear Occulometer," **Journal Neuro-sciences**, vol. 9, pp. 9-14, 1983.

[4] A. Bahill, M. J. Iandolo, B. T. Troost, "Smooth Pursuit Eye Movements in Response to Unpredictable Target Waveforms," **Vision Research**, vol. 20, pp. 923-931, 1980.

[5] B. Sauter, B. J. Martin, N. Di Renzo, and C. Vomscheid, "Analysis of Eye Tracking Movements Using Innovations Generated By a Kalman Filter," **Medical & Biological Engineering & Computing**, pp. 63-69, January 1991.

[6] A. S. Willsky, and H. L. Jones, "A Generalized Likelihood Ratio Approach to State Estimation in Linear Systems Subject to Abrupt

Changes," **Proc. IEEE Conference on Decision and Control**, Phoenix, Arizona, pp. 846-853, 1974.

[7] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, **Time Series Analysis: Forecasting and Control**, 3rd Ed., NJ: Prentice-Hall, 1994.

[8] P. J. Brockwell, and R. A. Davis, **Time Series: Theory and Methods**, 2nd Ed., NY: Springer-Verlag, 1991.

[9] J. C. Brocklebank, and D. A. Dickey, **SAS System for Forecasting Time Series**, SAS Institute, 1986.

[10] H. Akaike, "Canonical Correlation Analysis of Time Series and the Use of an Information Criterion," in **Systems Identification: Advances and Case Studies**, R. K. Mehra and D. F. Lainiotis (eds.), NY: Academic Press, 1976, pp. 27-96.

[11] M. Aoki, **State Space Modeling of Time Series**, 2nd Ed., NY: Springer-Verlag, 1990.

[12] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive Mixture of Local Experts," **Neural Computation**, vol. 3, pp. 79-87, 1991.

[13] M. I. Jordan, and R. A. Jacobs, "Hierarchies of Adaptive Experts," **Advances in Neural Information Processing Systems**, vol. 4, pp. 985-992, 1992.

Acknowledgment: The raw eye-movement data used in this paper was provided by the Pendergrass Laboratory in the Department of Radiology at the Hospital of the University of Pennsylvania.