Combining Recurrent & Bayesian Models for Action Anticipation with Multiple Cues

Mariia Zimokha[1] **Lorenzo Jamone**[1,2] **Iran R. Roman**[1]
[1]Queen Mary University of London, School of EECS
[2]University College London, Computer Science
corresponding author: zimoham@gmail.com

Abstract

Human action prediction involves rapid sensory integration and deliberative reasoning. Inspired by human studies, we propose a dual-process model with Reservoir Computing (RC) for temporal processing and Bayesian Networks (BN) for uncertainty-aware probabilistic decisions. The RC integrates sensory cues while the BN processes the output RC states to refine predictions. We tested this integrated framework using simulated reaching tasks with cues such as gaze direction, hand movement, and hand shape. The results indicate that our combined system replicates key aspects of human behavior.

Introduction

Human action prediction represents a cognitive brain function enabling anticipation of others' movements during social interactions (Kilner et al., 2007; Ambrosini et al., 2015; Roman et al., 2019, 2023). When observing someone reaching for an object, the brain integrates cues like the other's gaze and hand trajectory (Kong et al., 2014; Schydlo et al., 2018). For human-robot interaction, replicating these predictive processes is crucial to develop systems that can anticipate human actions to enhance safety and efficiency (Park et al., 2024).

Computational models of behavior aid our understanding of underlying cognitive and neural mechanisms (O'Reilly et al., 2010). However, existing action prediction models face key limitations: (1) they fail to capture the human tendencies to dynamically rely on cues that evolve over time and may provide congruent/incongruent information (Groen et al., 2022) , and (2) they do not simulate computations observed in biological systems, such as recurrent signal propagation in neural circuits (Levi & Huk, 2020). These shortcomings make the integration of cues challenging; for example, during object reaching, one may look at one object but then reach for another (Ambrosini et al., 2015; Kong et al., 2014)

Our hypothesis is that the recurrent integration of multiple cues, combined with probabilistic graphical modeling will yield a plausible model for human action prediction. This hypothesis is grounded in Dual Process Theory, which distinguishes between Type 1 and 2 processes. Type 1 captures both rapid, heuristic-driven inference while Type 2 captures slower, reflective belief updating under uncertainty. In this paper, we present a model with recurrent neural processing to integrate temporal sensory cues over time, followed by Bayesian graphical model to perform structured probabilistic inference.

Methodology

Simulated Anticipation Task

We simulate a reaching task where an actor grasps a target object placed on a surface in front (left/right). While inspired by Ambrosini et al. (2015) video-based paradigm, we replace real videos for a five-dimensional and discrete time series with seven steps that correspond to an action of 600ms. The five dimensions are: (1) **Object Location** that represents the x-coordinate of the target object with a value of -1/1 (left/right). (2) **Object Size** represents the size of the target object with a value of 0.5 (small) or 1 (large). (3) **Hand Preshape** represents the hand configuration prior to grasping, with a value of 0.5 (small) or 1 (large). (4) **Gaze** direction is represented as -1 or 1 (left/right). (5) **Hand Trajectory** is a hand's x-coordinate that starts at zero and evolves via a directed walk with Gaussian noise to ends at -1/1. The walk models drift toward the target, capturing both goal-directed movement and variability observed in human reaching. Object Location, Object Size, Hand Preshape, and Gaze are variables that remain constant throughout a trial, while the Hand Trajectory evolves over time and, importantly, and can either align or misalign with object location.

Experimental conditions

We include four conditions of cue congruency. If the shape of the hand matches the size of the object, it is considered to be congruent; otherwise, it is incongruent. The gaze can be considered similarly. Therefore, trials are categorized in one of four conditions: *Gaze Congruent/Hand Preshape Congruent* (Gaze C/Hand C), *Gaze Congruent/Hand Preshape Incongruent* (Gaze C/Hand I), *Gaze Incongruent/Hand Preshape Congruent* (Gaze I/Hand C), and *Gaze Incongruent/Hand Preshape Incongruent* (Gaze I/Hand I). We assess the model's accuracy as a function of these conditions.

Coherence and Target Determination

Gaze, hand, and position are random variables, each independently coherent or incoherent with object position with probability 0.5; that is, each coherence variable is an independent Bernoulli random variable with success probability of 0.5. A coherence strength factor (0.25) determines the influence of each coherent cue on increasing the probability that the final hand position will match object location. The probability of final hand position matching object location is therefore

$$p = \frac{1}{2} + \frac{1}{4} \times (\text{gaze\_coherent} + \text{handshape\_coherent}), \quad (1)$$

and it is clipped to the range [0.1, 0.9]. This factor reflects the human tendency to rely on coherent cues for predicting actions and aligns with empirical findings showing prediction accuracy is highest with both cues congruent, intermediate with one cue, and lowest when incongruent (Ambrosini et al., 2015). In the simulated dataset the trials are balanced for the variable of target action (left/right, 50/50 split).

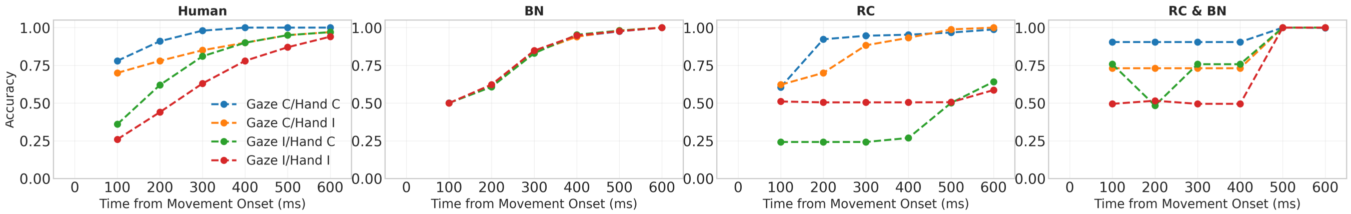

Figure 1: Accuracy in predicting if a person will reach for an object placed on the left or right. The task uses 600 ms trials with cues that are either congruent—gaze directed to the object (Gaze C) and hand shape matching the object (Hand C)—or incongruent—gaze directed away (Gaze I) and mismatched hand shape (Hand I). The left panel shows human data by Ambrosini et al. (2015), and subsequent panels show a Bayesian Network (BN) model, the Reservoir Computing (RC) model, and the combined RC + Bayesian Network (BN) model.

Models

Reservoir Computing (RC): An RNN architecture with fixed random recurrent connections that project inputs into a high-dimensional space, enabling temporal integration without requiring weight optimization (Lukosevicius & Jaeger, 2009; Maass et al., 2002). The input is the five data dimensions (object position, object size, gaze, hand preshape, and hand trajectory). The RC has a 200-neuron reservoir size and recurrence scaled by $\rho = 0.9$. ρ is a crucial hyperparameter to govern dynamics, memory capacity, and stability. The output layer is trained via logistic regression to map reservoir states to binary predictions (left/right grasp) using gradient descent and binary cross-entropy loss.

The recurrent component emulates continuous evidence accumulation in the dorsolateral prefrontal cortex (PFC), where attractor dynamics maintain task-relevant information over time (Jensen et al., 2024; Bullmore & Sporns, 2009). This implementation enables fast, intuitive Type 1 processing by continuously integrating sensory inputs to form rapid predictions (Kahneman, 2011).

RC and Bayesian Network (BN): A temporally-aware BN that processes the states of an RC model. For each timestamp, the BN selects the top seven informative features using mutual information and creates a histogram of these to carry out its probabilistic prediction (left/right grasp).

The BN mirrors inference mechanisms found in cortical and subcortical loops, using structured graphical models to represent uncertainty and update beliefs over time (Koller & Friedman, 2009; Deng et al., 2016), thus implementing slower, deliberative (Type 2) processing (Kahneman, 2011).

Results

Figure 1 shows accuracy by humans in the task, as reported by (Ambrosini et al., 2015). When gaze and hand are congruent with object location, human performance starts around 80% and reaches 100% at 400ms. The initial prediction of the RC model is around 60% and it quickly rises approaching 100%. The RC&BN model starts at 90% reaches 100% around 500ms. When the gaze is congruent but the hand is incongruent, human performance starts at approximately 70% and increases steadily approaching 100%. The RC model begins with a lower initial accuracy (around 62%) but also steadily climbs to 100%. The RC&BN model starts at around 75% and after 400ms, its performance jumps to 100%. When the gaze is incongruent but the hand preshape is congruent, human performance starts much lower, around 36%, but progressively improves to near-perfect performance by the end. The RC model demonstrated the worst initial performance, starting around 24% for an extended period and ending at about 64% accuracy. The RC & BN model shows high initial performance (≈87%), fluctuates, and eventually reaches 100% accuracy. Finally, when gaze and hand preshape are incongruent, human observers start at a very low accuracy and gradually improve over the trial to reach near perfect accuracy by the end. The RC model is unable to overcome the incongruence cues, sitting around 50% accuracy (i.e. chance), until the late stages when it reaches ≈60%. The RC & BN stars near chance and it reaches perfect accuracy at the end after a sudden leap at the 400ms mark.

We include a BN baseline showing performance starting at chance and reaching 99% accuracy across all conditions, reflecting higher uncertainty at trial onset due to independent processing at each time step. The RC variants address this simplification by projecting data to high dimensions and maintaining a memory state, similar to human working memory.

It is important to note that in our dataset, simulated trajectories approximate key features of human reaching but rely on idealized assumptions about motor variability. While conceptually inspired by Ambrosini et al. (2015), their study reports only endpoint statistics (whether the hand reached the same side as the object) and does not provide temporal hand movement data. Since the original videos are unavailable for tracking, validating our simulated trajectories remains an important direction for future empirical work using motion capture or similar methods.

Conclusion

We studied action anticipation when cues may congruently or incongruently inform target actions. Humans resolve incongruent cues by accumulating evidence over time, achieving perfect performance. Two of the three model variants that we studied, the RC and RC+BN models, also carry out this task by accumulating evidence over time, similar to how humans do it. However, we observed that the RC model struggles to resolve the ambiguity when cues are incongruent. In contrast, the RC & BN model performed better due to its feature-selective strategy, even qualitatively surpassing human performance. Overall, these results highlight mechanisms behind effective action prediction and suggest benefits of feature-selection strategies in computational models facing ambiguity.

References

Ambrosini, E., et al. (2015). The eye in hand: predicting others' behavior by integrating multiple sources of information. *Journal of Neurophysiology*, *113*(7), 2271–2279. doi: 10.1152/jn.00464.2014

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3), 186–198. doi: 10.1038/nrn2575

Deng, Z., et al. (2016). Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4772–4781).

Groen, I. I. A., et al. (2022, Oct). Temporal dynamics of neural responses in human visual cortex. *The Journal of Neuroscience*, *42*(40), 7562–7580. Retrieved from https://doi.org/10.1523/JNEUROSCI.1812-21.2022 doi: 10.1523/JNEUROSCI.1812-21.2022

Jensen, K. T., et al. (2024). A recurrent network model of planning explains hippocampal replay and human behavior. *Nature Neuroscience*, *27*, 1340–1348. Retrieved from https://doi.org/10.1038/s41593-024-01675-7 doi: 10.1038/s41593-024-01675-7

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kilner, J. M., et al. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing*, *8*, 159–166.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.

Kong, Y., et al. (2014). A discriminative model with multiple temporal scales. In *European conference on computer vision* (pp. 596–611).

Levi, A. J., & Huk, A. C. (2020). Interpreting temporal dynamics during sensory decision-making. *Current opinion in physiology*, *16*, 27–32.

Lukosevicius, M., & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, *3*(3), 127–149. doi: 10.1016/j.cosrev.2009.03.005

Maass, W., et al. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Computation*, *14*(11), 2531–2560. doi: 10.1162/089976602760407955

O'Reilly, R. C., et al. (2010). Computational models of cognitive control. *Current Opinion in Neurobiology*, *20*(2), 257–261. doi: 10.1016/j.conb.2010.01.008

Park, J., et al. (2024). Towards safe human-robot interaction: A pilot study on a deep learning-assisted workspace monitoring system. *2024 IEEE 24th International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*, 1197–1202. doi: 10.1109/QRS-C63300.2024.00157

Roman, I. R., et al. (2019). Delayed feedback embedded in perception-action coordination cycles results in anticipation behavior during synchronized rhythmic action: A dynamical systems approach. *PLoS computational biology*, *15*(10), e1007371.

Roman, I. R., et al. (2023). Hebbian learning with elasticity explains how the spontaneous motor tempo affects music performance synchronization. *PLOS Computational Biology*, *19*(6), e1011154.

Schydlo, et al. (2018). Anticipation in human-robot cooperation: A recurrent neural network approach for multiple action sequences prediction. In *2018 ieee international conference on robotics and automation (icra)* (p. 5909-5914). doi: 10.1109/ICRA.2018.8460924