A Bayesian Approach to Determine Focus of Attention in Spatial and Time-Sensitive Decision Making Scenarios

Yu-Ting Li and **Juan P. Wachs**
School of Industrial Engineering
Purdue University
West Lafayette, IN 47906
{yutingli, jpwachs}@purdue.edu

Abstract

Complex decision making scenarios require maintaining high level of concentration and acquiring knowledge about the context of the task in hand. Focus of attention is not only affected by contextual factors but also by the way operators interact with the information. Conversely, determining optimal ways to interact with this information can augment operators' cognition. However, challenges exist for determining efficient mathematical frameworks and sound metrics to infer, reason and assess the level of attention during spatio-temporal complex problem solving in hybrid human-machine systems. This paper proposes a computational framework based on a Bayesian approach (BAN) to infer users' focus of attention based on physical expression generated from embodied interaction and further support decision-making in an unobtrusive manner. Experiments involving five interaction modalities (vision-based gesture interaction, glove-based gesture interaction, speech, feet, and body balance) were conducted to assess the proposed framework's feasibility including the likelihood of assessed attention from enhanced BAN and task performance. Results confirm that physical expressions have a determining effect in the quality of the solutions in spatio-navigational type of problems.

Introduction

There is currently a lack of fundamental theories and methods to analytically express the relationship between user physical interaction, attention and task performance. This is in spite of existing evidence in the cognitive psychology literature that these are tightly related (Bailey et al., 2001). Nevertheless, the multivariate nature of attention, makes its quantitative, objective and evidence based assessment be a hard challenge. To assess the effects and levels of users' attention, traditional tools rely on subjective metrics (e.g. SAGAT (Endsley 1998)). The existing limitation of such methods is that they are disruptive and therefore, add a confounding effect to the measured variable (Endsley, Pina, and Cummings 2009). Other approaches relying on physiological signatures (e.g. ocular movements (Poole and Ball 2006) and pulse) require the operator to stay seated so physical expressions such as hand movements do not interfere with the acquired signals. This opposes to the current trend in complex image analysis to use more the human bodies (Vogel and Balakrishnan 2005) to interact with spatio-navigational information, rather than passive analysis (users seated continuously in front of the computers).

This paper's main contribution is presenting a rigorous mathematically, biologically and psychologically-inspired method for assessing attention from disparate raw signals. These methods include a systematic characterization of operators' interaction during complex problem solving; probabilistic modeling of the links between attention and task performance; evolutionary inspired approaches for network generation. A key feature of our work is inferring users' focus of attention dynamically, in a non-intrusive fashion. This is done through the design of Bayesian Attentional Networks (BANs) along with its topology structure and parameters. This methodology is expected to have less confounding results compared to former studies. Our methodology addresses a key question in AI related to the design of cognitive experience interfaces: how to determine the optimal combination of control and feedback modalities to augment operators' cognition, enhance their performance, thus leading to better decision-making.

Figure 1 shows the system architecture of the BAN framework. It infers the user focus of attention based on the probability distribution of the query variable (attention – the dependent variable), given values from evidence variables (observations – the independent variables). To determine the probabilistic models for inferring users' focus of attention, a systematic approach is developed that integrates operator's knowledge and an automatic learning process. The enhanced BAN is further used to infer the probability of attention in different interaction scenarios.

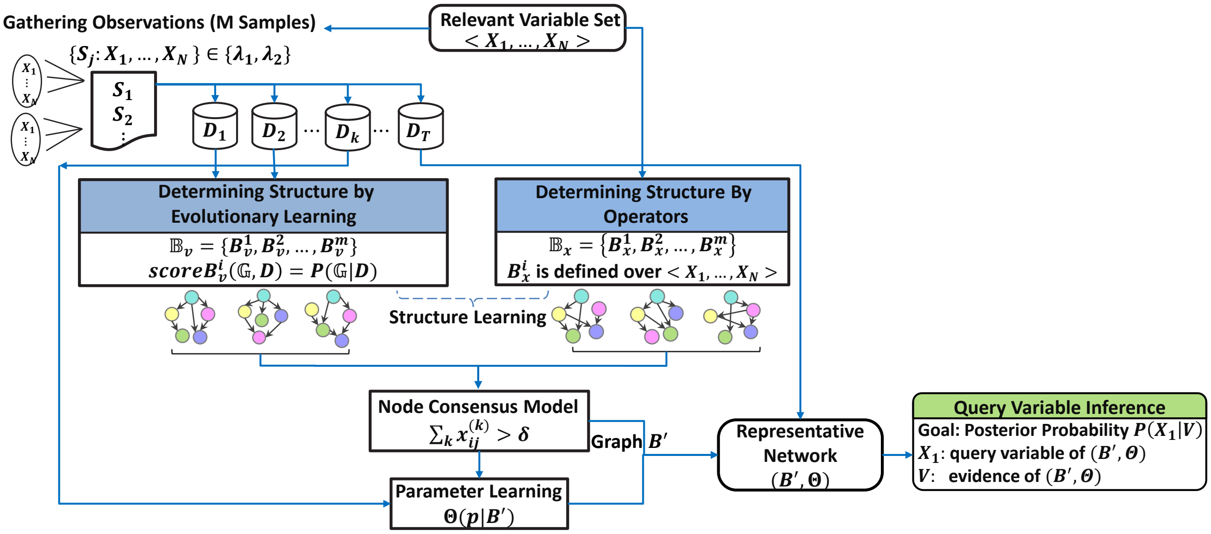

Gathering Observations (M Samples)
Relevant Variable Set
$< X_1, \dots, X_N >$
$\{S_j : X_1, \dots, X_N\} \in \{\lambda_1, \lambda_2\}$
X_1
X_N
X_1
X_N
S_1
S_2
D_1
D_2
D_k
D_T
Determining Structure by Evolutionary Learning
$\mathbb{B}_v = \{B_v^1, B_v^2, \dots, B_v^m\}$
$scoreB_v^i(\mathbb{G}, D) = P(\mathbb{G}|D)$
Determining Structure By Operators
$\mathbb{B}_x = \{B_x^1, B_x^2, \dots, B_x^m\}$
B_x^i is defined over $< X_1, \dots, X_N >$
Structure Learning
Node Consensus Model
$\sum_k x_{ij}^{(k)} > \delta$
Graph B'
Parameter Learning
$\Theta(p|B')$
Representative Network
(B', Θ)
Query Variable Inference
Goal: Posterior Probability $P(X_1|V)$
X_1: query variable of (B', Θ)
V: evidence of (B', Θ)

Figure 1: System Architecture of constructing the representative BAN

Contextual Framework

In this paper, the spatial navigation decision making problem used is the traveling salesman problem (TSP), where a salesman must visit *N* different cities using the shortest path without visiting any city more than once. Studies (Bureš, Burešová, and Nerad 1992; Tenbrink and Wiener, 2009) indicate that people (and animals) can find near-optimal solutions to computer generated versions of the TSP using perceptual information, however there is a large variability on the strategies adopted by each individual. This is the reason why it is of paramount importance to investigate how humans solve this problem and what factors affect their solution. For example, it was also found that symmetry of the city layout and other aesthetic factors have an effect on the optimality of the solutions given by each individual (MacGregor et al. 2004; Vickers et al. 2006). In this paper the TSP layout will follow the Symmetric with Rewards setup (Blum et al. 2003), in which the distances between two cities are exactly the same in each direction; and there are prizes (rewards π_v) assigned to the cities v (see Fig. 2). The goal is to find a path such that minimizes the total distance and it maximizes the reward collected subject to the selections of the cities.

Bayesian Attention Network (BAN)

The Bayesian network represents the operators' attentional levels while solving spatial navigation and decision making problems in time sensitive scenarios. Such a Bayesian model can capture cognitive key processes which are characteristic to strategies issued by the operators to solve decision-making problems and their effect on

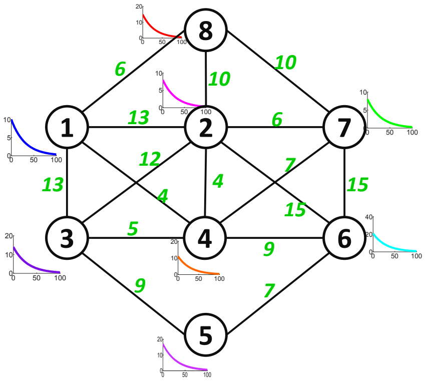


Figure 2: A 8-city TSP. The distance between two cities is marked as a text in green color .The exponential decay function, $\gamma(t_v) = e^{-t_v/T_m}$ expresses the change of rewards as visiting city v at time t_v (T_m is the maximum time allotted for visiting). Total reward function is given by $\sum_v^N \pi_v \gamma(t_v)$

attention. The representative Bayesian network, describing an operator's attentional behavior, is obtained by (1) selecting an operator highly familiar with the task in hand (Korb and Nicholson 2003) (e.g. radiologist, intelligence analysts, air traffic controllers), or by (2) adopting a genetic programming paradigm whereas the network evolves automatically as a result of genetic operations towards an incumbent solution. The structure of the BAN is defined as an assignment over N variables $< X_1, X_2, \dots, X_N >$, each of which takes a binary value in finite domain $\{0,1\}$. The description of a BAN $\mathbb{B}$ consists of the directed acyclic graph $\mathbb{G}$ which includes directed edges between variables and associated parameters vectors Θ that specify the associated conditional dependencies. In this paper, the variables include observations of the user while solving a spatial decision-making problems using embodied interaction. Let us define a variable X_i (i=1,..,k,..,N) such that its value $\lambda_i = f(X_i)$ is a Boolean. Also let X_1 be the query variable (focus of attention).

Attention can be discretized in states $\{0,1\}$ (representing "High attention", and "Low attention", respectively). We use sensors to collect raw instances S about the users' physical behavior (body movements), and contextual information (e.g. task completion time) during the experiment. Those raw instances are transformed into the states' value of the variables, $f(S) \rightarrow X_2, X_3, \cdots, X_N$ (see Table 1).

Table 1
Definition of Discrete States for Each Variable

Variable	Description	States
X_1	Focus of Attention	{High Attention, Low Attention}
X_2	Torso Orientation	Detection of frontal torso {True, False}
X_3	Face Orientation	Detection of frontal face {True, False}
X_4	Hand Gesture	{Evoked, Not evoked}
X_5	Utterance	{Present, Not present}
X_6	Feet in location	{Yes, No}
X_7	Inter-command Elapsed Time (t)	$\{\lvert t-\mu\rvert \leq \sigma, \lvert t-\mu\rvert > \sigma\}$ [a]
X_8	Error in Use	{Wrong command delivered, Correct command delivered}

[a] μ: mean of the inter-command elapsed time of all observation; σ: standard deviation of the inter-command elapsed time

Determining the BAN Structure through Operators' Knowledge

In the operator-centered based modeling, each of the networks is elicited by operators who have domain knowledge, considering the systems' requirement and user centric preferences. The consideration of having operators be responsible for the design of the BAN is rooted in the fact they have experience not only with effective problem solving in the given domain, but they are highly familiar with the interaction process itself. The procedure used by the operators for building the networks is described in the Algorithm 1.

Determining the BAN Structure through Evolutionary Learning

Evolutionary-based modeling was used to construct the Bayesian network in order to obtain several candidate BANs. This method is based on the concept of Genetic Programming (GP) where the dependencies between nodes are inducted following GP's operations. Thus, to build a number of BANs through evolutionary learning, the observations collected during the experiment were used to construct the datasets $(D_1, D_2, \ldots, D_T)$. Each dataset D_i is constituted by a number of feature vectors $\boldsymbol{\Psi} \in \mathbb{R}^M$, in other words, $D_l \in \mathbb{R}^{M,N-1}$, where M is the number of observations assigned to D_i. An observation is defined as a feature vector $\Psi = \{\lambda_2, \ldots, \lambda_N\}$ where the binary value $\lambda_k = f(X_k)$ corresponds to the X_k evidence variable computed from the operator's evoked command. The feature vector only contains the variables whose states are observable, and therefore λ_1 is not included (since it is inferred).

Algorithm 1: Constructing BAN through Operators

```
Input: A set of relevant variables $< X_1, X_2, \ldots, X_N >$ that describe the problem domain
  Step 1. Start by placing the children nodes of the network (raw evidence) at the lower level arranged in the same level
  Step 2. Add the highest node of the network, Attention, in the top level.
  Step 3. Assign a variable $X_i$ with its description to each node
  Step 4. Add nodes in between the lowest level and the highest level, exhibiting a cause-effect relation, from the bottom to the top.
    Step 4.1 For each node added, determine its connection between node $X_i$ and the set of nodes already in the network.
    Step 4.2 If a cycle exists, remove the last node.
  Step 5. Return to Step 4 until all the nodes have been placed and all variables are assigned to nodes
```

In the evolutionary-based modeling, first, an initial population was generated randomly. Then, selected individuals were used to generate a new generation. This was done through genetic operators: crossover and mutation. Assume that a BAN with graph $\mathbb{G}$ consists of N nodes, where v_i indicates the i-th node. An arc $x_{ij} = (v_i, v_j)$ equals to 1 if it is directed from v_i to v_j, whereas 0 if it is not directed. The directed acyclic graph was represented as a bit string (Larrañaga et al. 1996), $x_{12}x_{13} \ldots x_{2k} \ldots x_{N-1,N}$. The individuals remaining (each individual is a single Bayesian network) are those which outperform the antecedents in terms of a given performance metric. The fitness (the performance metric) of the individual is assessed using a scoring measure (1), which is the probability of observing the dataset D_l by an individual in each population (Friedman 1997):

$$score(D_l, \mathbb{G}_H) = P(D|\mathbb{G}_H) = \sum_i^{2^M} P(d_i|\mathbb{G}_H) \qquad (1)$$

$$P(d_i|G_H) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(N_{ij})}{\Gamma(N_{ij} + M_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(a_{ijk} + s_{ijk})}{\Gamma(a_{ijk})} \qquad (2)$$

where $\mathbb{G}_H = (V \cup H, E)$ are the disjoint sets of observable variables ($V = \{X_2, \ldots, X_8\}$) and the latent variable (level of attention) is $H = \{X_1\}$, with edges E (between pairs of variables). In Eq. (1), the computation of the scoring metric takes exponential time in terms of M. To tackle this problem, an efficient calculation (Neapolitan 2004) was carried out consisting of computing $P(d_i|\mathbb{G}_H)$ of repetitive observations in the dataset only once, and then, multiplying the derived probability by the number of its occurrences. This process does not affect their statistical effect on the latent variable. A number of observation tables can be generated by concatenating the original table D_l with a new column $c_i \in \mathbb{R}^{M,1}$ each time. More formally, $B_i = D_l \cup c_i$,

$i = 1 \dots 2^M$, $B_i \in \mathbb{R}^{M,N}$. The overall procedure of evolutionary-based modeling for building the networks is described in the algorithm below:

Algorithm 2: Constructing BAN through Evolutionary Approach

```
Input:
        Table $\boldsymbol{D_i}$ – binary values of observable variables
        $M$ – number of iterations; $i$ – iteration index; $\epsilon -$ threshold
Initialization: generate a set of feasible $\mathbb{G}_H{}^c$ solutions randomly
while score($\boldsymbol{D_l}, \mathbb{G}_H^{(i)^*}$) - score($\boldsymbol{D_l}, \mathbb{G}_H^{(i-1)^*}$) $\geq \epsilon$ do
    $\mathbb{G}_H^{(i)} \leftarrow crossOver(\mathbb{G}_H^{(i)})$
    $\mathbb{G}_H^{(i)} \leftarrow mutation(\mathbb{G}_H^{(i)}, p_m)$ // $p_m$ as mutation probability
    $\mathbb{G}_H^{(i)^*} \leftarrow eliteSelection(\mathbb{G}_H^{(i)})$
    if $any(\mathbb{G}_H^{(i)})$ is infeasible then
        update $\mathbb{G}_H^{(i)}$ //replace a infeasible solution by a new random one
    end if
    increment $i$
end while
Output: Incumbent DAG $\mathbb{G}_H^{(m)^*}$
```

Node Consensus Model (NCM)

The representative graph structure is obtained from candidate BANs previously found using operator-based modeling and the evolutionary approach. The procedure used, coined *Node Consensus Model (NCM)*, consists of iteratively deriving an agreed graph among most of the candidates. The NCM attempts to find a BAN with consensus among the majority of the candidate BANs. The enhanced network is derived iteratively by examining the existence (and popularity) of edges among each BAN candidates. Assume there are K BANs in the candidate set, and for each, an adjacency matrix $\boldsymbol{A_k}$with each element x_{ij}, where $i, j \in \{1 \dots N\}$, is constructed to represent the network. This means that an entry "1" assigned to x_{ij} means that nodes i and j are connected, an "0" otherwise. The representative BAN starts from an initial empty graph in which nodes are not connected ($\boldsymbol{A_k}$ with all entities equal to 0). Let us hypothesize that there is an edge from two nodes v_i and v_j . Then, we ask how many of the remaining graphs agree with this hypothesis. Thus the existence of an edge is decided by iteratively examining the consensus among the remaining graphs. The edge reaches a consensus if and only if the number of graphs which have the same connectivity exceed some threshold. For example, the value of 10 at entry (i, j) indicates that 10 BANs agreed that there is a link (cause-effect) between node i and node j. Figure 3 shows the resulting adjacency matrix of the optimal BAN. Each entry in each adjacency matrix of a BAN included only 0-1 values, and thus the total values for entry (i, j) can be at most 10. For example, the top left value indicates that 10 BANs agreed that there is a link (cause-effect) between attention and torso orientation. This process is summarized in Algorithm 3.

$$\mathcal{A} = \begin{bmatrix} 0 & 10 & 7 & 5 & 5 & 4 & 3 & 4 \\ 0 & 0 & 4 & 6 & 2 & 3 & 4 & 3 \\ 0 & 0 & 0 & 5 & 6 & 3 & 0 & 6 \\ 0 & 0 & 0 & 0 & 4 & 3 & 5 & 6 \\ 0 & 0 & 0 & 0 & 0 & 4 & 5 & 8 \\ 0 & 1 & 1 & 1 & 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 0 & 0 & 2 & 0 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 3: The adjacency matrix of the representative BAN for the 10 candidate BANs in Figure 4.

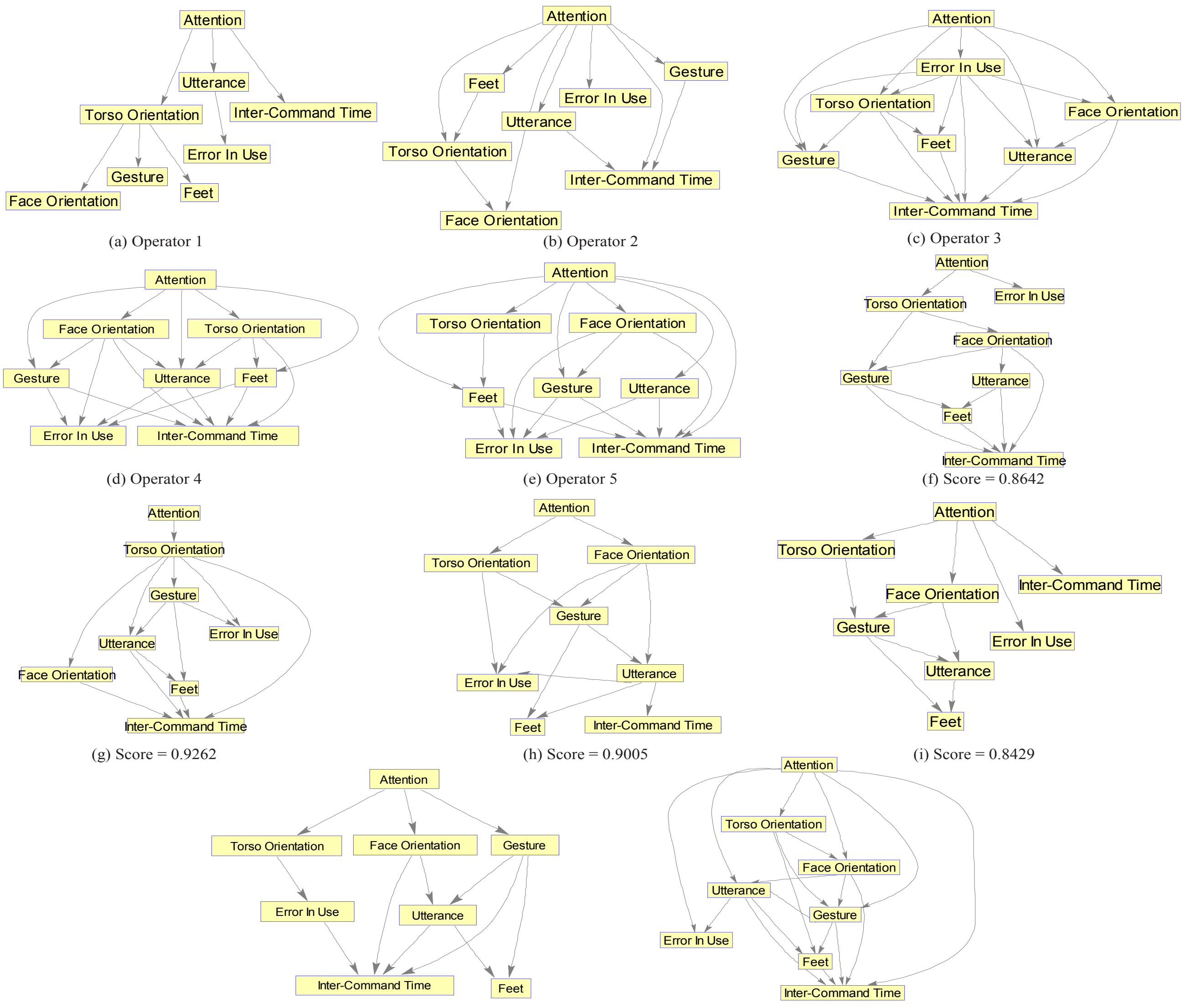
Algorithm 3: Node Consensus Method

```
Input:
        $\boldsymbol{A_k}$ matrices representing a set of $k$ graphs each with order $N$
        $K$ – the number of iterations performed
for all $i, j \leq N$ do // given i,j as the source and destination indices of
nodes $x$
    $nCon \leftarrow \sum_k^K x_{ij}^{(k)}$
    if $nCon > K/2$ then // majority is more than 50% agreement
        $\boldsymbol{\mathcal{A}(i,j)} \leftarrow nCon$
    end if
end for
$\mathcal{G} \leftarrow$ Mat2Dag($\mathcal{A}$) // convert the adjacency matrix to the directed
graph
$\mathcal{G} \coloneqq$ optimal graph with majority consensus
Output: optimal graph $\mathcal{G}$ with adjacency matrix $\boldsymbol{\mathcal{A}} = [x_{ij}]$
```

Experimental Results

Experiments were conducted to assess the validity of the framework. Twenty graduate and undergraduate students were recruited, including 13 males and 7 females, all 20 to 30 years old. The users were given instances of the TSP problem to solve. Each user was given 20 different TSPs to solve in 4 different scenarios (5 TSPs in each scenario). In each scenario, the subject used a different interaction and feedback modality, which was randomly assigned in advance. Each user acted as an "operator". The five modalities adopted included gross gestures (recognized by Kinect), fine gesture (finger configurations recognized through a data glove), speech, feet configuration (on dance pad controller), and body stance (using a Wii balance board). Those sensors were used to collect evidence including: torso and face orientations, hand gesture, utterance, body stance and elapsed time, which served as the raw observations (evidences).

The instances of the TSP problems presented included the layout of cities, labeled edges representing the distance between cities and the reward assigned to each city in a bar graph. As the subject travels to the next city using one of the aforementioned interaction modalities, feedback is displayed or read back to the subject through a text-to-speech program (Microsoft SAM). The feedback information provided consisted of the overall travelled distance. With this information, the subjects were better equipped to estimate possible alternatives that would lead to shorter distances.



(a) Operator 1

(b) Operator 2

(c) Operator 3

(d) Operator 4

(e) Operator 5

(f) Score = 0.8642

(g) Score = 0.9262

(h) Score = 0.9005

(i) Score = 0.8429

(g) Score = 1.0000

(k) Combined

Figure 4: Bayesian Attentional Network's structure obtained by (a) – (e) Operator based, (f) – (j) Evolutionary learning (k) NCM method

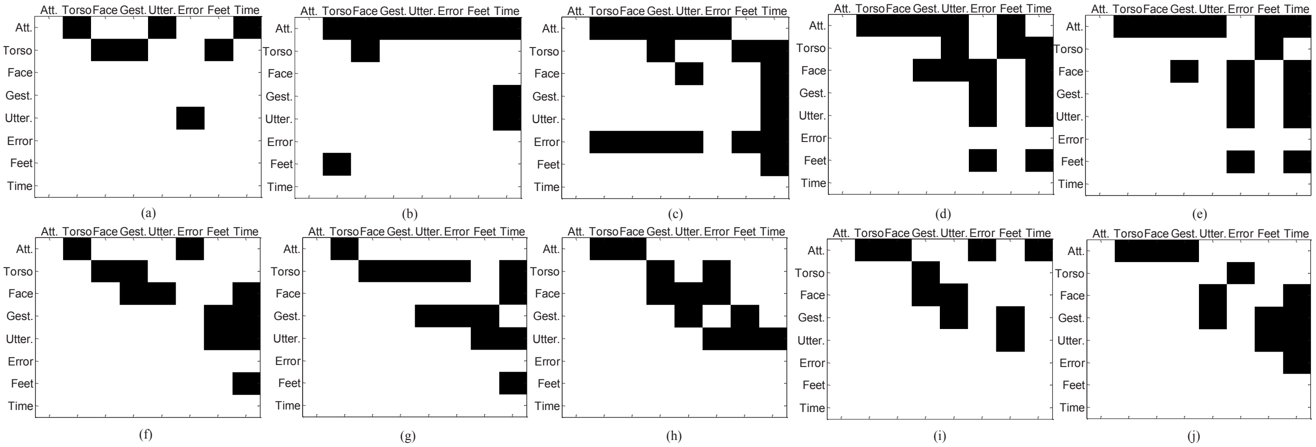


Figure 5: The adjacency matrix of BANs obtained by (a) – (e) Operator based, (f) – (j) Evolutionary Learning

Five topologies were acquired using the evolutionary BANs approach from 100 observations in each dataset. Additionally another 5 BANs were obtained by operators. The parameters (conditional probability distribution for each node) that quantify relationships between connected nodes were computed using the Expectation-Maximization (EM) algorithm (Friedman 1997). Figure 4 (a) – (j) shows the example BANs elicited by 5 operators and learned through the evolutionary process, respectively. The adjacency matrix for each BAN represents the connection between nodes in Figure 5 (a) – (j). The representative BAN determined by NCM method is shown in Figure 4 (k), and its adjacency matrix is shown in Figure 3. Figure 6 shows the evolutionary learning process of five BANs in each generation, obtained through Algorithm 2. The figure shows the best scores among the populations in each generation. From the figure can be learnt that after 170 generations, the solution increased significantly (25.08% at most, and 9.77% at least) from their initial values.

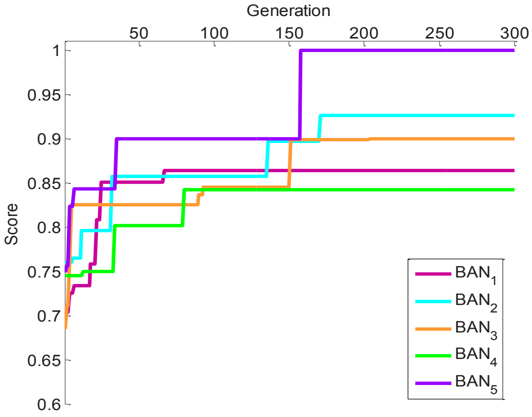


Figure 6: Convergence characteristics of 5 evolutionary

Additionally, the inferred probability of attention at the state of “high focus of attention” $P(X_1|V)$ in 10 difference scenarios is presented in Table 2. The highest value of probability of attention occurred when using step gestures as the input modality, and speech as feedback. Also the second best combination was step gesture and visual feedback. In order to show the optimal scenario (or alternatively the worst) to be significant, the ANOVA (Analysis of variance) is conducted on each independent trial. Results of one-way ANOVA (F(9,190)=96.16, p< .05) indicated that there are statistically differences between group means.

Table 2
Inferred probability of attention at the state of “high focus of attention” in different scenarios

	Step	Glove	Kinect	Speech	Wii
Visual	**0.5967**	0.4821	0.4690	0.3438	0.5747
Speech	**0.6079**	0.4871	0.4663	0.3290	0.5559

Discussion

In this paper we applied inference and reasoning to assess the level of operators’ attention using BANs. The main experimental result consists of a network automatically created based on consensus between the candidate solutions. This network was obtained through the NCM method proposed. It explains why the focus of attention not only affects the physical action but also the task performance (elapsed time, and operator error). Moreover, the torso orientation determined largely the direction where the user was facing and her feet movement. The gesture and utterance were determined significantly by the orientation of the users’ face (which in turn is a proxy of focus of attention). The elapsed time varied among users depending on the time taken to evoke the gestures or utterances. Through the use of cause-effect networks, five types of interaction modalities and two feedback modalities were cross-compared through a set of experiments. The results show that using step gestures on the dance pad controller lead to higher focus of attention than using other three interfaces (fine gestures recognized through a data glove, gross gesture recognized by Kinect, and speech) for control.

Conclusion

Bayesian attentional networks (BANs) are a structure describing the cause-effect relationship between operator’s focus of attention, physical action and decision-making in a spatio-temporal complex and time-sensitive problem. The proposed framework considers both operators’ knowledge and a biologically inspired method to compute the BAN with the highest performance metric. This BAN was obtained through an innovative method called Node Consensus Method (NCM). This method automatically creates a representative BANs based on the consensus level among the candidate solutions. Results showed that using step gestures allowed operators to solve spatial navigational problem while keeping high level of attention. Future work involves incorporating feedback information and testing this approach with a larger dimensional decision making problem (tools for visualizing cyber-operations). In addition, multiple modalities of user command and feedback will be adopted for operator-machine interaction.

Acknowledgment

This work is supported in part by the US Air Force Office of Scientific Research (AFOSR) Young Investigator Research Program (YIP) award No. FA 9550-13-1-0141 to the Juan P. Wachs.

References

Bailey, B. P.; Konstan, J. A.; and Carlis, J. V. 2001. The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In *Proceedings of INTERACT* (1): 593-601.

Blum, A.; Chawla, S.; Karger, D. R.; Lane, T.; Meyerson, A.; and Minkoff, M. 2003. Approximation algorithms for orienteering and discounted-reward tsp. In *Proceedings of 44th Annual IEEE Symposium on Foundations of Computer Science*, 46-55.

Bureš, J.; Burešová, O.; and Nerad, L. 1992. Can rats solve a simple version of the traveling salesman problem? *Behavioural Brain Research*, 52(2): 133-142

Endsley, M. R. 1988. Situation awareness global assessment technique (SAGAT). *In Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, 789-795.

Friedman, N. 1997. Learning belief networks in the presence of missing values and hidden variables. In *Proceedings of Fourteenth ICML*, 125-133. San Francisco, CA: Morgan Kaufmann

Korb, K. B.; and Nicholson, A. E. 2003. *Bayesian Artificial Intelligence*. CRC press.

Larrañaga, P.; Poza, M.; Yurramendi, Y.; Murga, R. H.; and Kuijpers, C. M. H. 1996. Structure learning of Bayesian networks by genetic algorithms: A performance analysis of control parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18(9): 912-926.

Poole, A.; and Ball, L. J. 2006. Eye tracking in HCI and usability research. *Encyclopedia of human computer interaction*, 211-219.

Tenbrink, T.; and Wiener, J. 2009. The verbalization of multiple strategies in a variant of the traveling salesperson problem. *Cognitive Processing* 10(2):143-161.

Vogel, D.; and Balakrishnan, R. 2005. Distant freehand pointing and clicking on very large, high resolution displays. In *Proceedings of the 18th annual ACM symposium on User interface software and technology*, 33-42. ACM.