AFRL-AFOSR-VA-TR-2016-0222

Embodied Interactions in Human-Machine Decision Making for Situation Awareness Enhancement Systems

Juan Wachs
PURDUE UNIVERSITY

06/09/2016
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTA2
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
05-26-2016	Final	05/01/2013-04/30/2016

4. TITLE AND SUBTITLE

Embodied Interactions in Human-Machine Decision Making for Situation Awareness Enhancement Systems

5a. CONTRACT NUMBER

5b. GRANT NUMBER

FA9550-13-1-0141

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Wachs, Juan P.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Purdue University
155 S Grant Street
West Lafayette, IN 47907

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Office of Scientific Research
875 North Randolph Street
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The research objective of this proposal is to test the hypothesis that attention, action intention and physical performance (embodied interaction), such as gesturing, can enhance visual processing, leading to better solutions for spatial optimization problems. I will develop a framework to determine which body expressions best support complex decision making in human-computer mixed systems.
The technical approaches adopted in my research involve rigorous engineering-based methods for directly integrating focus of attention with embodied interfaces, and decision making techniques to assess the value of feedback. These methods include systematic characterization of gestures during complex problem solving.

15. SUBJECT TERMS

Embodied interaction, gestures, one-shot learning, TSP, Bayesian networks

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE		46	Juan P Wachs
					19b. TELEPHONE NUMBER *(Include area code)* 7654967380

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

Final Report for AFOSR YIP Program

Title: Embodied Interactions in Human-Machine Decision Making for Situation Awareness Enhancement Systems

AFOSR Award Number: FA9550-13-1-0141

Program Manager: Dr. James Lawton

Performance Dates: 01 MAY 2013– 30 APR 2016

Participants
PI: Juan Wachs, Purdue University (IE)

The three years of the YIP program was overall successful in achieving planned goals towards addressing scientific questions raised in the proposed work, as well as proposing new questions regarding fundamentals of modeling the operator's attention using network theory, during spatio-navigational complex problem solving. This summary provides an overview of the research questions addressed and progress achieved in the three years of this YIP program. Each of three Research Goals from the original proposal is discussed separately below. An Appendix report is attached, to which the reader is referred for specific details and information.

Research Goal 1: Determine physical actions to enhance decision-making problems

Task 1: Solving the TSP problem through embodied interaction

Fundamental Question(s) Addressed:

- Which physical activities occur during the solution of spatial navigation problems in time sensitive scenarios?

- How the quality of its solutions is affected/improved when using embodied interfaces compared to those solutions obtained using standard interfaces?

Key Achievements**:**

- Creation of a database with observations from 393 independent trials. Each trial included 5 to 8 commands describing a task related sequence of commands. From those, 193 trials were used to create a training dataset of 1200 observations, and the remaining 200 trials (1670 observations) were used for the creation of a testing set.

- Developed the computer based TSP problem version, capable of generating random instances with a fixed number of cities, and exponentially decreasing rewards. Performance metrics for TSP solution also were determined.

- Developed and implemented software to capture users' physical movements using a combination of Kinect cameras, Wii sensors, gloves, and DDR (dance on dance revolution). We leveraged primarily on open source software to detect and track human motions, and off-the shelf SDKs provided with the sensors purchased. In some other cases (e.g. DDR and Wii) we developed our own software.

Task 2: Motion, Gestures and Posture Recognition

Fundamental Question(s) Addressed:

- Is the proposed gesture recognition method agnostic of the classification technique used?
- Is the proposed gesture recognition method generalizable to different gesture datasets?

Key Achievements**:**

- Formulated a framework for generating a big-dataset of artificial "human like" gesture observations. The approach used is based on the conceptual features – what do we remember when we see a gesture and try to mimic it (we refer to these conceptual features as the "Gist of a Gesture"). Three different classifiers were implemented and compared as well as using two different datasets for training and testing: one is formed by customized gestures for image manipulation and the other is a publicly available dataset from a well-known gesture recognition competition.

Research Goal 2: Learn probabilistic models to assess user's focus of attention and intention

Task 1: Construct Models for User's Attention

Fundamental Question(s) Addressed:

- Can level of attention be assessed using non-disruptive, non-subjective models of attention?

Key Achievements**:**

- We developed in Year 1 the BAN model to help make inferences about the user's focus of attention under uncertainty. In Year 2, we looked into extending the Node Consensus Model (NCM). The NCM relied on finding consensus among candidate network solutions provided by humans and agents (an evolutionary approach). Now, we assigned different importance to each expert based on a gradient descent method so a cost function was minimized.
- To determine the right importance that each candidate solution must receive the gradient descent method attempts to minimize a cost function. In this year, we developed two cost

functions that represent the tradeoff between maximizing the evolutionary based network as opposed to maximizing the "knowledge" conveyed in the human based networks.

Task 2: Capturing Activity Cues through Sensors

Fundamental Question(s) Addressed:

- How to capture the BAN's evidence nodes to infer attention levels?

Key Achievements**:**

- We have done significant progress on this task in Year 1, and we will focus again on this task in Year 3. We purchased an EMOTIV EPOC headset to measure EEG signals and we are currently learning how it should be implemented in our framework.

Research Goal 3: Explore Effective and Efficient Feedback Techniques

Fundamental Question(s) Addressed:

- How can the tradeoffs between the interaction control and feedback modalities and their effect in attention be expressed through utility functions?

- Can the BANs developed in Research Goal 2 be applied to the CSA Explorer scenario to infer attention?

Key Achievements**:**

- We utilized a utility theory based approach to find the best combination of embodied interaction and feedback modality so the operator performance is maximized. We found empirical evidence that the combination of feet gestures with visual feedback provides the best task performance.
- We conducted experiments in cyber physical network scenarios comparing the multimodal embodied interface with traditional interface, and validated the BAN developed in Research Goal 2 using dual task approach (a common practice to measure the level of attention in phycology). Experimental results indicated the multimodal embodied interface outperformed the traditional interface. Also, statistical analysis indicated the consistency between BAN and dual task approach.

Challenges: We were interested in checking whether this utility functions and our findings will translate well to the CSA Explorer scenario. We encountered a major obstacle in trying to use the CSA explorer at the Rome facility Air Force Research Laboratory, Information Directorate, Rome NY, or alternatively getting a beta version to work from our campus. We were not able to have the people from AFRL help us none of these two options. We informed the program manager (Dr. Jamie Lawton) about this situation several times, and the last time was more than one year ago. Nothing was done to help us address this issue. Furthermore, due to this problem, we had to come-up with our own implementation of the CSA Explorer, which delayed the performance of the project by several months. In spite of that situation, we were not granted a non-cost extension to finish the project properly.

References

[1] S. Kita, I. van Gijn, and H. van der Hulst, "Movement phases in signs and co-speech gestures, and their transcription by human coders," in *Gesture and Sign Language in Human-Computer Interaction*, I. Wachsmuth and M. Fröhlich, Eds. Springer Berlin Heidelberg, 1998, pp. 23–35.

[2] P. Viviani and C. Terzuolo, "Trajectory determines movement dynamics," *Neuroscience*, vol. 7, no. 2, pp. 431–437, Feb. 1982.

[3] A. F. Bobick and A. D. Wilson, "A state-based approach to the representation and recognition of gesture," *Pattern Anal. Mach. Intell. IEEE Trans. On*, vol. 19, no. 12, pp. 1325–1337, 1997.

[4] M. Ahissar and S. Hochstein, "Task difficulty and the specificity of perceptual learning," *Nature*, vol. 387, no. 6631, pp. 401–406, May 1997.

[5] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante, "Multi-modal gesture recognition challenge 2013: Dataset and results," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 445–452.

[6] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992.

[7] M. G. Jacob and J. P. Wachs, "Context-based hand gesture recognition for the operating room," *Pattern Recognit. Lett.*, vol. 36, pp. 196–203, Jan. 2014

Accepted

Publications (2013-2016)

1. Y. T Li, J. P Wachs. "A Bayesian Approach to Determine Focus of Attention in Spatial and Time-Sensitive Decision Making Scenarios" in the *AAAI'14 Workshop on Cognitive Computing for Augmented Human Intelligence.*
2. Y. T Li, J. P Wachs. "Linking Attention to Physical Action in Complex Decision Making Problems" in *IEEE International on Systems, Man and Cybernetics, 2014*.
3. J. P Wachs. Designing Embodied and Virtual Agents for the Operating Room: Taking a Closer Look at Multimodal Medical Service Robots and Other Cyber-Physical Systems. Speech and Automata in Healthcare Voice-Controlled Medical and Surgical Robots Series: Speech Technology and Text Mining in Medicine and Healthcare. A. Neustein (Ed). De Gruyter, 2014; November 2014; ISBN: 978-1-61451-515-9.
4. M. E Cabrera, J. P Wachs, "Embodied Gesture Learning from One-Shot", In *The 25th IEEE International Symposium on Robot and Human Interactive Communication, 2016. RO-MAN 2016.* IEEE.
5. T. Zhang, B. S Duerstock, J. P Wachs. "A Computational Framework for Attention Inference Using a Bayesian Approach" presented in IEEE IROS 2015 Workshop, 8th International Symposium On Attention in Cognitive Systems (ISACS 2015)

Submitted (under review).

1. T. Zhang, Y. T Li, J. P Wachs. "The Effect of Embodied Interaction in Visual-Spatial Navigation" submitted to *ACM TiiS, Major Revision*.

Presentations

1. Y-T. Li, J. P Wachs, "A Bayesian Approach to Determine Focus of Attention in Spatial and Time-Sensitive Decision Making Scenarios," 2014 Graduate Student (and post-doctoral fellows) Research Award Competition, Purdue University, Feb. 12 2014.
2. Y-T. Li, J. P Wachs, "Embodied Interaction with Visualization and Spatial Navigation in Time-Sensitive Scenarios," 2014 Industrial Engineering Research Symposium, Purdue University, April 24 2014.
3. Y. T Li, J. P Wachs. "Linking Attention to Physical Action in Complex Decision Making Problems" Oral presentation at the IEEE International on Systems, Man and Cybernetics, 2014.
4. T. Zhang, B. S Duerstock, J. P Wachs. "A Computational Framework for Attention Inference Using a Bayesian Approach" Poster presented in IEEE IROS 2015 Workshop, 8th International Symposium On Attention in Cognitive Systems (ISACS 2015)

Personnel

Graduate Students

Yu-Ting Li
Ting Zhang
Maria Eugenia Cabrera

Professional Salaries

Prof Wachs – 100% one month summer salary, 2014 – 2015.

Prof Wachs: 50% January – April 2016

Appendix to Final Report for AFOSR YIP Program

Title: Embodied Interactions in Human-Machine Decision Making for Situation Awareness Enhancement Systems

AFOSR Award Number FA9550-13-1-0141

Program Manager: Dr. James Lawton

Performance Dates: 01 MAY 2013– 30 APR 2016

Participants
PI: Juan Wachs, Purdue University (IE)

Research Goal 1: Determine physical actions to enhance decision-making problems

Participants: (Task 1) Juan P Wachs (PI), Yu-Ting Li (graduate student).

Question(s) addressed: Which physical activities occurred during the solution of spatial navigation problems in time sensitive scenarios? How the quality of its solutions was affected/improved when using embodied interfaces compared to those solutions obtained using standard interfaces?

There were two major aspects of the project for Research Goal 1, addressed during year 1, 1) development of the experimental setting, the TSP navigation problem setup and metrics to assess its performance, 2) development of each of the embodied interfaces to interact with a graphical visualization of the TSP, and 3) we collected observations while users solved the TSP in the different conditions. These aspects are denoted in Figure 1, inside the dashed red rectangle.

Participants: (Task 2) Juan P Wachs (PI), Maria Eugenia Cabrera (graduate student).

During the first year we gathered data and characterize differences in spatial navigation strategies in a complex task, the Traveling Salesman Problem (TSP). For the second year, we developed a new theoretical framework to recognize movements performed during the solution of the TSP problem, based on single observations (one-shot learning). This year (year 3), we mainly focused on implementing the framework and test its generalization capabilities using different gesture vocabularies and different classification techniques. This is important since the developed framework is not dependent on the classification method used and it is currently generalizable to gestures performed with upper limbs with only one original instance coming from the operator.

Question(s) addressed: Is the proposed gesture recognition method (One Shot Learning) agnostic of the classification technique used and generalizable to different gesture datasets?

In Year 2 we focused our work in recognizing arbitrary gestures performed during the solution of the TSP. We formulated a framework for generating a big-dataset of artificial "human like" gesture observations. The strategy was to generate a dataset of realistic samples based on biomechanical features extracted from a single gesture sample. These features, called "the gist of a gesture", are considered to represent what humans remember when seeing a gesture and the cognitive process involved when trying to replicate it. By adding meaningful variability to these features, a large training dataset was created while preserving the fundamental structure of the original gesture. This large dataset was then used to train three different classification methods, namely Hidden Markov Model (HMM), Support Vector Machine (SVM) and Conditional Random Fields (CRF). Additionally, this procedure was executed for two different datasets: the publicly available ChaLearn Challenge Dataset from 2013 [ref] and a customized dataset for image manipulation.

Task 1: Solving the TSP problem through embodied interaction

We conducted experiments to test the effects of embodied interaction on task performance and its dependency on attentional levels. Institutional IRB permission was sought and obtained to conduct these experiments. 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.

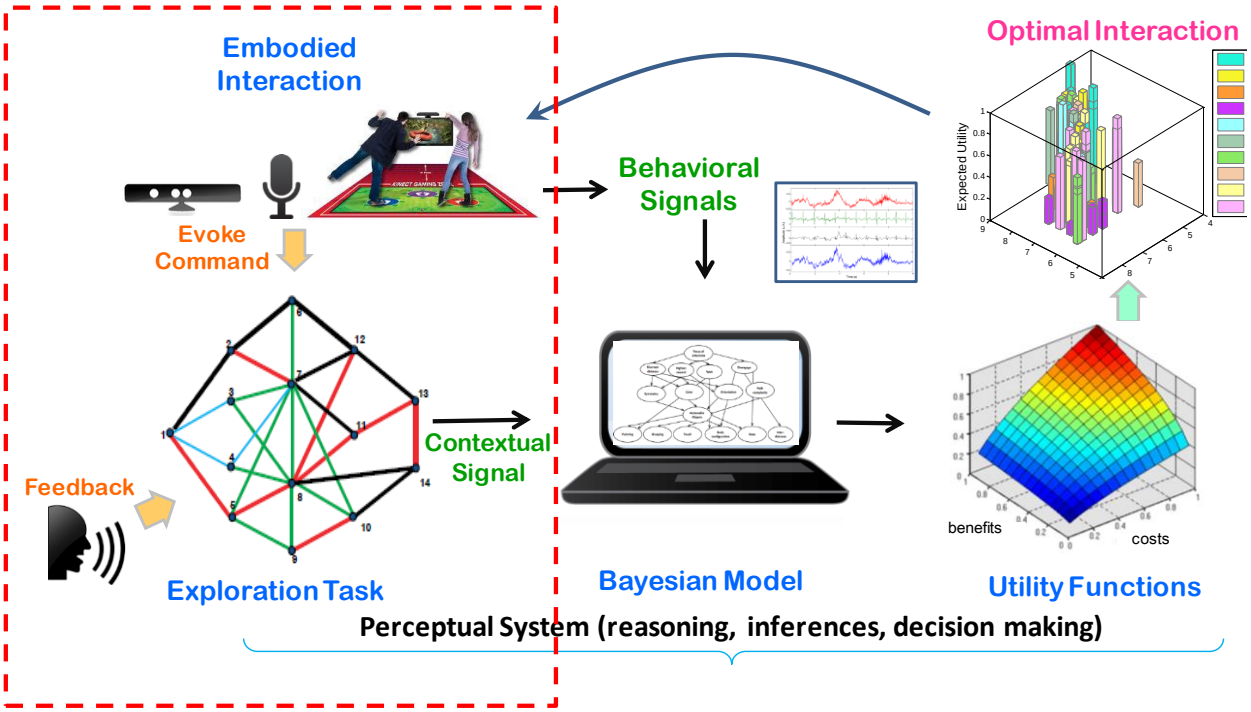


Figure 1. Research Objective 1 denoted under the dashed line

In those experiments we used the TSP layout following the Symmetric with Rewards setup [1], in which the distances between two cities are exactly the same in each direction; and there are prizes (rewards) assigned to the cities which value decays over time (see Fig. 2). We told each subject that their goal is to find a path such that minimizes total distance and it maximizes the

reward collected, subject to a limit on the total length of the path. The TSP was designed as a directed graph with a reward values exponentially decaying over time.

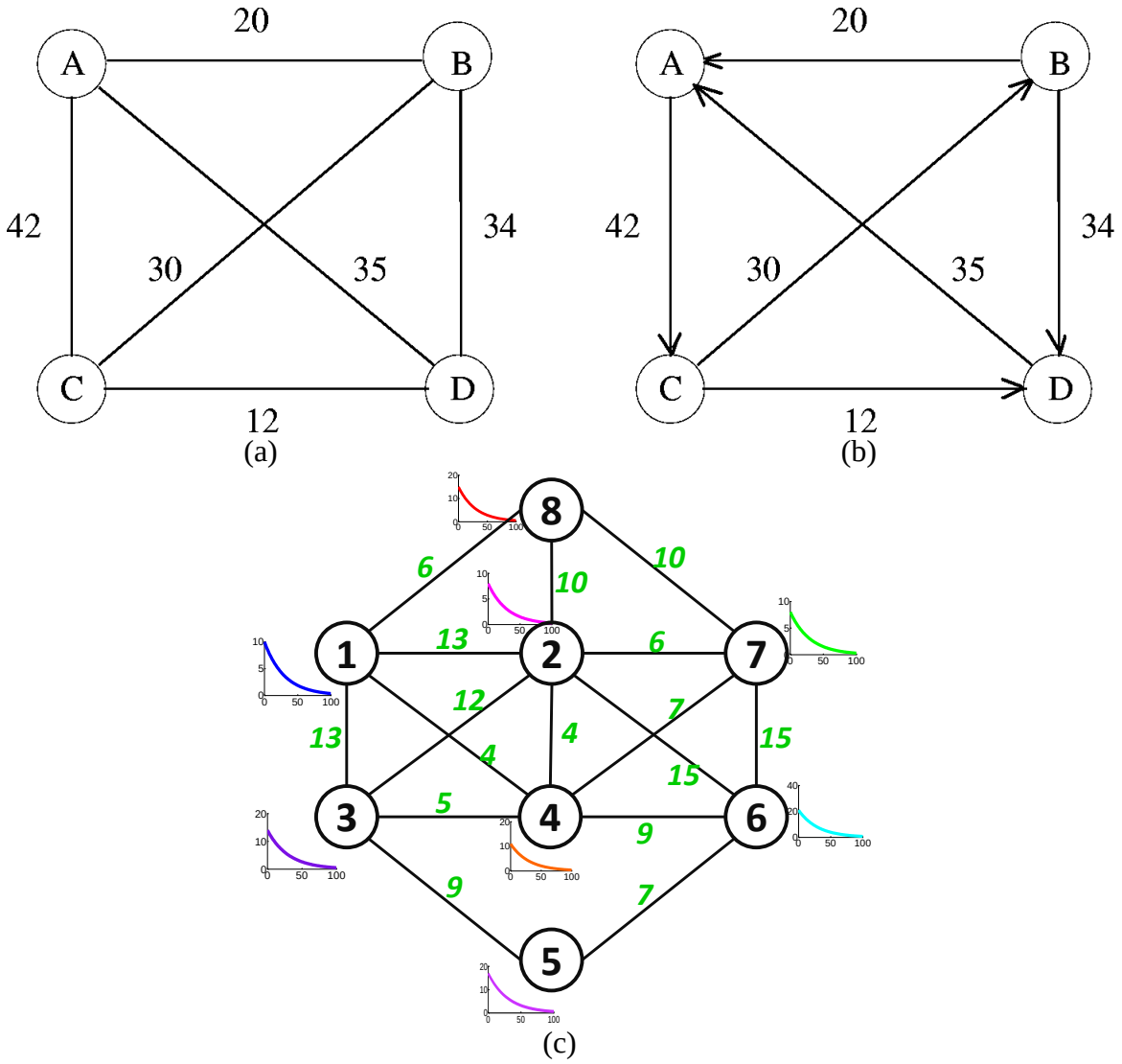


Figure 2. TSP graph (a) symmetric. The distance from A to B is the same as from B to A (20). TSP graph (b) is asymmetric. The distance from A to B is 42+30, and B to A is only 20. (c) . One instance of the 8-city TSP with reward at each city used in our experiments. The distance between two cities is marked as a text in green color. The exponential decay by each city expresses the change of rewards assigned to the city as a function of time.

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", since their domain knowledge for solving the TSP was as good as anyone else's knowledge. The five modalities adopted included gross gestures (using mainly the arms), fine gestures (using fingers configuration), speech, feet configuration (on dance pad controller), and body stance (using a Wii balance board) (see Fig. 3). Each city in each TSP was randomly assigned a reward which decreased exponentially over time. A sequence of those cities' reward decreasing over time is presented in Fig. 4. Each "active" cell in the 3x4 grid represents one of the cities. Warmer colors

indicate higher reward values while cooler colors represent lower rewards. As can be seen, all active cells become cool colors since rewards are reduced with increasing time. In the experiment, the reward value assigned to each node was displayed as a bar plot to avoid confusions (Fig. 6). The experimental apparatus consisted of a PC, a large 60" screen, a Kinect sensor, a microphone, a data glove, a dance pad, and the balance board. Those sensors were used to collect evidence including: torso and face orientations, hand gestures, utterance, body stance and elapsed time, which served as the raw observations (evidences) for the BANs.

Random instances of the TSP were assigned to the subjects in order to assess their performance. A trial is defined as a sequence of commands required to solve an instance of the TSP. Each command resulted in an observation vector which was stored for training and testing purposes.

Accomplishments: A total of 393 independent trials were collected. Each trial required 5 to 8 commands to complete the task. From those, 193 trials were used to create a training dataset of 1200 observations, and the remaining 200 trials resulted in 1670 observations which were used for the testing set. In addition, we developed the computer based TSP problem, capable of generating random instances with a fixed number of cities. We completed this task according to the original roadmap.

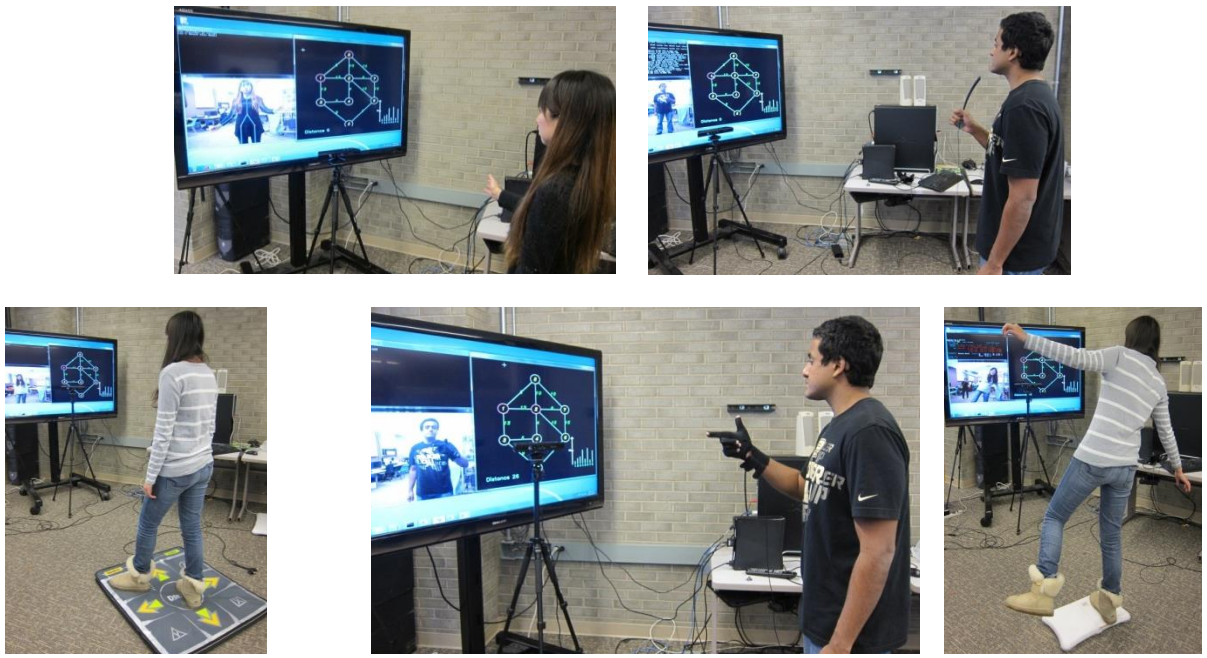


Figure 3. Five modalities used in the experiment. (a) gross gestures (Kinect) (b) speech (c) feet on dance pad (d) fine gestures (glove) (e) body stance on Wii balance board.

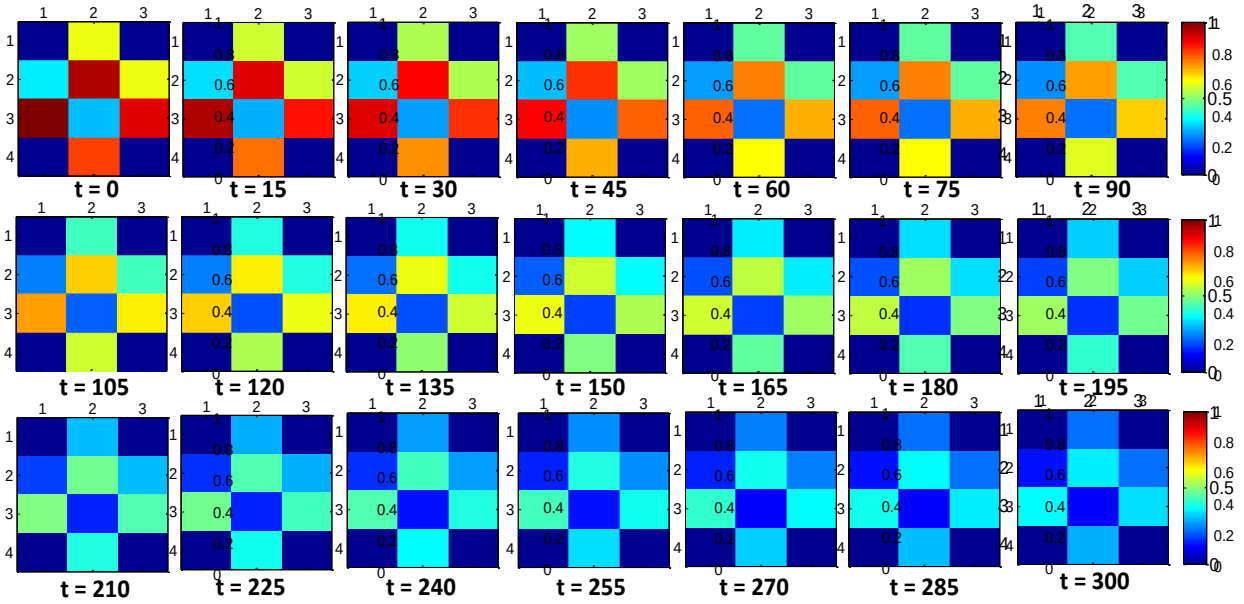


Figure 4. The exponentially decreasing reward of 8 cites over time

Discussion: We found that while the TSP interface computer based version included the main features of a characteristic TSP problem, namely, the distances, rewards and cities, the amount of variance between the instance was limited: only the distance and rewards values were changed. Ideally, we realize after the experiments that a better setting would include different TSP layouts other than the one presented in Figure 2. Furthermore, it would be useful to visualize the edges representing the distances between cities with a proportional length to the distance value associated with that edge. While this would be a better representation of real TSP problems, as found in the real world problems, it would constraint the amount of cities displayed on the screen (the screen has limited physical dimensions). We opted to present the lengths with the same physical size but to assure that all the nodes fit within the region of the screen and are large enough for the user to see the complete layout. Given these constraints, we concluded that the current setting is the most suitable for conducting the TSP experiments with human subjects.

Task 2: Motion, Gestures and Posture Recognition

In the experiment described in Task 1, the detection of commands representing the next city visited by the user depends on the scenario used to interact with the computer based TSP application. When mouse and keyboard commands were used, the recognition of these events was trivial. We used windows events manager to monitor and detect the use of those devices through messages and events. For embodied interaction there were two challenges. The first related to allowing the users adopt any arbitrary form of physical interaction desired, and being able to detect those forms of expressions (one shot learning). The second challenge is in the development of computer vision techniques for recognizing such events. These problems are a roadblock to collecting user interaction observations awhile solving the TSP. To circumvent this problem, we decided to tackle these challenges in two phases. The first phase tool place in Year 1, and consisted of using open source code and predefined hand and step gestures, speech commands, and stance

configurations (also referred as lexicons). The next year (Year 2) we will focus on the development of One Shot Learning techniques. Dividing the solution in two phases allow us to proceed with the research project (e.g. building the Bayesian networks) without further delays. Once the problem of One Shot Learning will be addresses, it will be integrated smoothly with the existing modules.

Therefore, for Phase 1, the Kinect and other sensors were used in conjunction with the SDK to obtain an easy method for tracking human motion. We considered the following interaction forms:

(a) Speech: Using spoken commands (e.g. "move left").

(b) Gross gesture movements: Using the arms and hands (e.g. waving the hand from the left to right for the command "left")

(c) Fine gesture configurations: Using static hand poses (e.g. thumbs-up for "up").

(d) Step gestures: Stepping over specific regions (e.g. jump right for "right").

(e) Body stance: Changing the body balance (e.g. bending forward for "select").

As an example, the Kinect sensor can deliver a stream of images wherein body parts are tracked using a "skeleton" model. This model can recover the 3D coordinates of every joint in the skeleton (see Fig. 5).

Discussion: Using predefined lexicons for each form of interaction modality, in order to navigate the computer version of the TSP was necessary to enable the modeling of the Bayesian networks in the next Research Objective. This also allowed us to compare the effectiveness of each type of interface with respect to our performance metrics (defined in Research Objective 3).
Nevertheless, without one shot learning algorithms we are still unable to tell what gestures (or other forms of expressions) are selected for the users for interacting with the TSP when they are not constrained by specific lexicons. This aspect of the research will be addressed in Year 2 of the project.

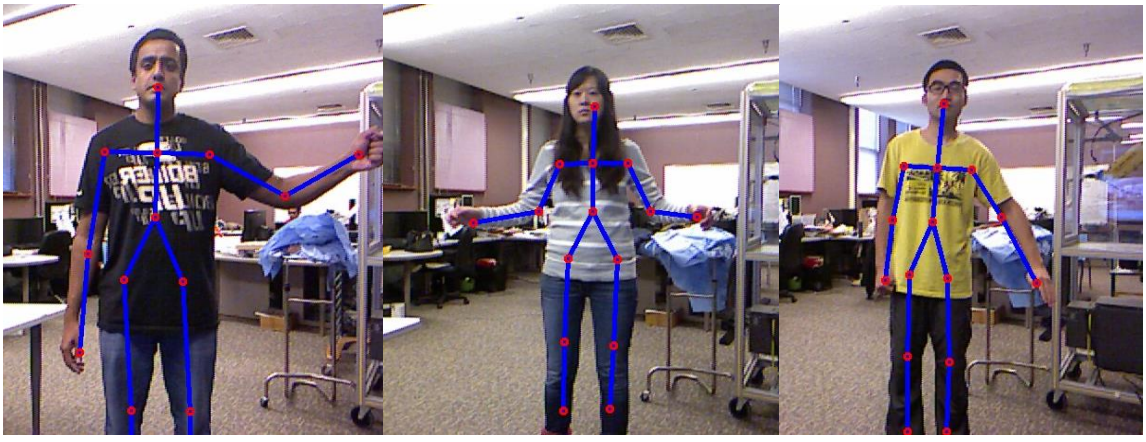


Figure 5. Kinect's skeleton of a user

For feedback, we considered two main forms of interaction: "speech" and "visual". These forms of interaction we used to acknowledge the command evoked, or simply to convey information about the TSP's quality of solution found. The feedback provided consisted of the overall travelled distance and current rewards. With this information, the subjects were better equipped to estimate

possible alternatives that would lead to shorter distances (better solutions) while accumulating the highest rewards. Visual feedback was presented by displaying the distance information on the screen, and the reward values for each city in real time (see Fig. 6). Reading back the distance information to the users through a text-to-speech program (Microsoft SAM) constituted the acoustic feedback delivered to the operator.

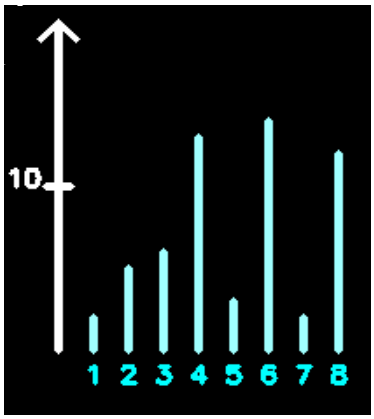


Figure 6. Rewards for each city in real-tine, presented as visual feedback to the operator

Accomplishments: We developed and integrated all the necessary software in C++ to detect and recognize each of the expressions described above (e.g. gestures). The software works in real-time and allows recording the sequence of commands evoked, together with the raw information used during the interaction (e.g. skeleton join angles, weight).

During this third year, efforts were dedicated to implement the theoretical framework developed during the second year. This approach focuses on the features used to represent the gestures, rather than the machine learning method to classify them. By understanding key features that human use to produce the gestures, more realistic artificial observations can be generated.

This framework was used to train three different classification algorithms, namely Hidden Markov Model (HMM), Support Vector Machine (SVM) and Conditional Random Field (CRF) in order to test the framework itself rather than the performance of a given classification paradigm, making the method agnostic to the classifier used. Additionally, two different gesture vocabularies were considered as both training and testing data. One of the datasets was customized for image manipulation whereas the other was selected as a subset of a publicly available dataset from a gesture recognition competition.

1. Overview of the implemented framework

The overview of the implemented framework is shown in Fig. 1. Initially the system requires a labeled example from a user. This is done in the following way: a gesture is performed by a user, which is detected and recorded using a Kinect sensor. Using the skeleton data provided by the sensor, core features are extracted, that we refer to as "*the gist of a gesture*", and used to recreate new realistic observations that resemble the one provided by the user. This process is repeated until a large dataset of observations is generated. Such data set constitutes the training set of an arbitrary classifier.

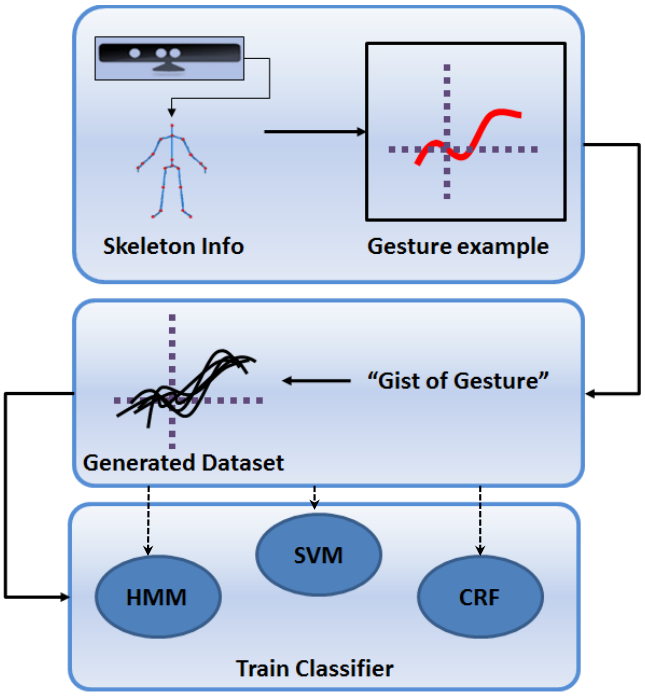


Figure 1. Overview of the System

2. *Extraction of "The Gist of a Gesture"*

In this section the method to produce artificially generated instances of a gesture that look realistic, as if they were generated by a human being, is presented. The goal is to have such a genuine representation that an arbitrary observer could not be able to tell which is which. In order to achieve this goal, it is necessary to leverage on bio-mechanical features reflecting the physical and dynamic limitations of humans during gesture production.

We build on previous work, which highlights the major considerations regarding gestures. Gestures are defined as a concatenation of movement phases, with the following distinctions:

- Movement phases are separated by abrupt changes in orientation, and changes in speed [1].
- Phase segmentation is invariant with respect to the duration of movement [2].
- Gestures are bound by a sequential order but time invariant [3].
- Gestures performed by humans show spatial generalization and orientation specificity [4].

Given such considerations, the selected features are: the number and location of inflexion points for each hand's trajectory (an array of three-dimensional points), the type of curvature present between a pair of inflexion points (e.g. convex, straight, and concave), and the sequence of the movement described by the quadrant where each inflexion point is located with respect to the gesturer's shoulder, considering the y-z plane as above and below the shoulder, and closer to or further away from the body centroid (this is an anthropometric feature).

For each observation in a given vocabulary, the proposed feature representation is depicted in Fig. 2 and Fig. 3. The depicted gesture vocabulary in Fig. 2 is a subset of the 10 selected gestures out of the 20 available in the dataset corresponding to the 2013 ChaLearn Challenge [5]. The choice of dataset between the ChaLearn Challenges has to do with the availability of skeleton information in the 2013 and the absence of it in the 2011 dataset.

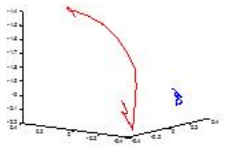
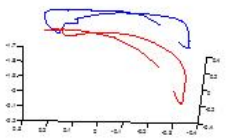
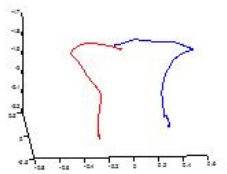
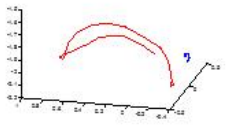
Gesture	Kinect Image	Hand Trajectory	Feature Representation
"Vieniqui" G2			3 Inflexion points Quadrant Sequence: III,II,III Curvature: +, +
"Chevuoi" G6			5 Inflexion Points Quadrant Sequence: III,II,IV,II,III Curvature: +,+,+,+
"Basta" G13			4 Inflexion Points Quadrant Sequence: III,IV,III,III Curvature:+,-,+
"Tantotempo" G17			5 Inflexion Points Quadrant Sequence: IV,II,II,II,III Curvature: +,+,+,+

Figure 2. Examples of gestures with the respective feature representation

Fig. 3 shows some of the gestures in the image manipulation dataset. Both datasets include *iconic* and *metaphorical* gestures. In the case of the dataset from ChaLearn 2013, they represent Italian cultural/ anthropological signs. The types of gestures are extracted from McNeil's classification taxonomy [6].

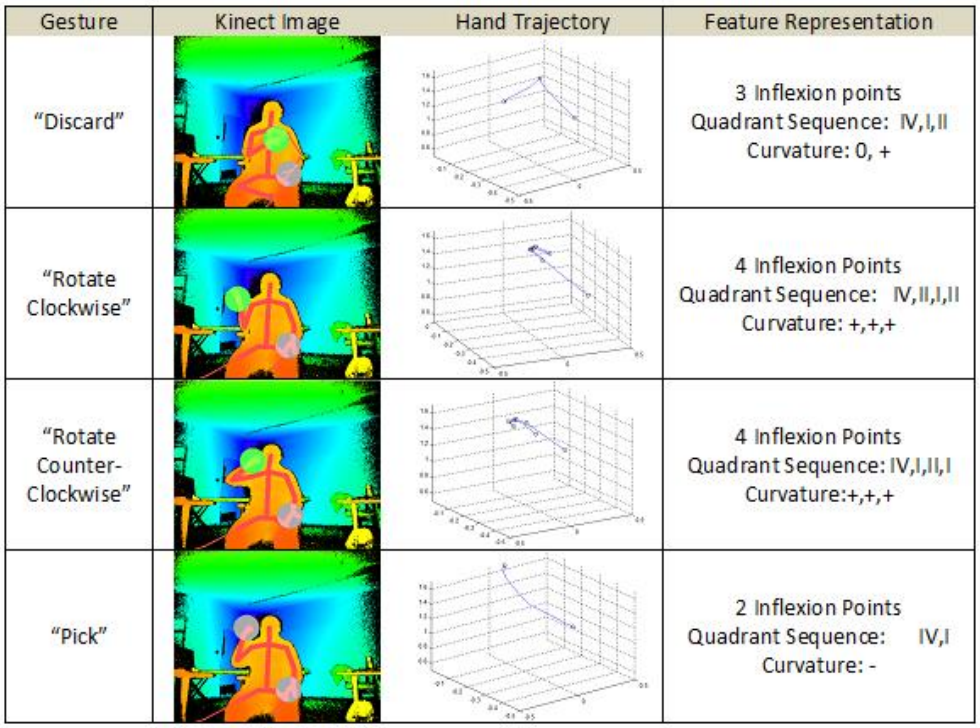


Figure 3. Examples of gestures with the respective feature representation from the customized dataset for image manipulation

The described feature representation is used as input to generate new observations with the objective of populating a training set to train a state-of-the-art classifier.

3. Artificial Observation Generation

Features obtained from a user-generated example are used to recreate observations that maintain the core motion aspects of the gesture, while adding variability to improve recognition of future instances.

A visual representation of the process is presented in Fig. 4. The original trajectory is decomposed using "the gist of a gesture" to the set of representing features previously mentioned.

The stored location for each inflexion point is used as the mean value for mixture of Gaussians, while the quadrant information relative to the user's shoulder is used to estimate the variance. Considering a vector of dimension *d* (in this case 3), a Gaussian Mixture Model (GMM) is fitted using the parameters shown in (1), (2) and (3).

$$(x; \mu_k, \sigma_k, \pi_k) = \sum_{k=1}^{m} \pi_k p_k(x), \quad \pi_k \geq 0, \sum_{k=1}^{m} \pi_k = 1, \tag{1}$$

$$p_k(x) = \frac{1}{(2\pi)^{d/2}\, {\sigma_k}^{1/2}} \exp\{-\frac{1}{2}(x-\mu_k)^T {\sigma_k}^{-1}(x-\mu_k)\} \tag{2}$$

$$\Theta = \{(\mu_k, \sigma_k, \pi_k): \mu_k \in \mathbb{R}^d, \sigma_k = {\sigma_k}^T > 0, \sigma_k \in \mathbb{R}^{dxd}, \pi_k \geq 0, \sum_{k=1}^{m} \pi_k = 1\} \tag{3}$$

Where *m* represents the three mixtures in the model, p_k is the normal distribution density with mean μ_k using the location of each inflexion point and a covariance matrix σ_k which is positive semi definite; π_k is the weight of the k^{th} mixture, and all the weights are equal adding up to one.

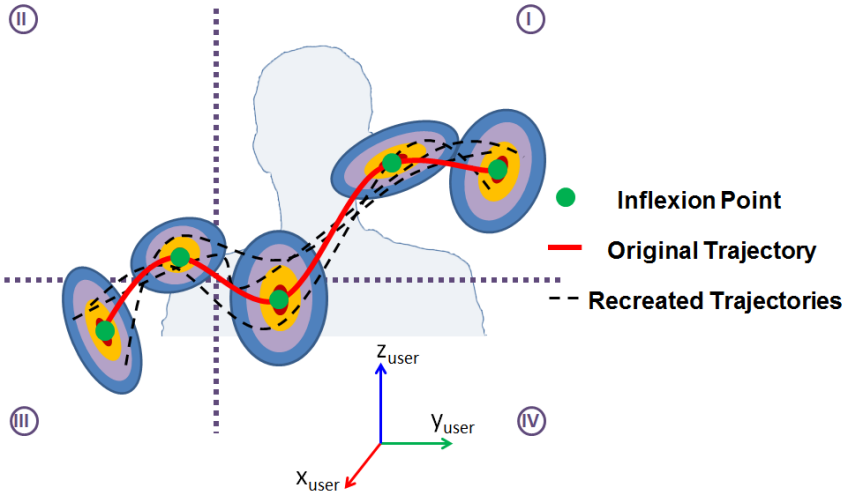


Figure 4. Visual Representation of how features corresponding to the "Gist of the Gesture" are used to create new observations.

To estimate the variance, each point in the sample trajectory is assigned to a quadrant with respect to the user's shoulder, using the reference frame shown in Fig. 4. The next step uses the points from each quadrant as samples to estimate variance of each quadrant as shown in (4), which in turn adjust the parameters for the generated GMM for each inflexion point.

$$\sigma_k = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(P_i - \mu_k)^2 \ , P_i \in q(x_i)_k \in \mathbb{R}^3 \ \& \ k = 1,2,3,4 \qquad (4)$$

With different sets of inflexion points, generated using GMM, and the curvature information related to the original gesture trajectory, artificial trajectories are generated. The points p_i and p_{i+1} are used along with curvature c_i to generate smooth segments, for all i in the set of inflexion points. The basic algorithm is outlined next, considering that the input for the algorithm is a single labeled example, in the form of a trajectory $\mathbb{X}$ consisting of A points defined in (5), acquired using the skeleton information from a Kinect.

$$\mathbb{X} = \{(x_1, y_1, z_1), (x_2, y_2, z_2), \dots, (x_k, y_k, z_k), \dots, (x_A, y_A, z_A)\} \qquad (5)$$

Algorithm 1: Generate artificial observations from one sample

Input:
$\mathbb{X}$ – 3D hand trajectory
$\boldsymbol{x_s} = (x_s, y_s, z_s)$ – 3D position of the shoulder
N – Number of artificial trajectories to generate

$\boldsymbol{x_{IP}} \leftarrow \boldsymbol{x_i} \in \mathbb{X}, \left.\frac{dx^2}{d^2t}\right|_{x=x_i} = 0, i = 1, \dots M$ *// Inflexion Points (IP)*
$I_j = \{x \mid x \in (x_i, x_{i+1})\}, j = 1, \dots M-1$ *// Interval between IP*
for M - 1 iterations **do**
  $c_j = sign\left(\frac{dx^2}{d^2t}\right), x \in I_j$ *// Convexity for interval I_j*
end for
if $y_i > y_s$, $z_i > z_s$
  $q(\boldsymbol{x_i}) = 1$ *// Determine quadrant location based on $\boldsymbol{x_s}$*
else if $y_i < y_s$, $z_i > z_s$
  $q(\boldsymbol{x_i}) = 2$
else if $y_i < y_s$, $z_i < z_s$

11 $q(\boldsymbol{x_i}) = 3$
12 **else** // $y_i > y_s$, $z_i < z_s$
13 $q(\boldsymbol{x_i}) = 4$
14 **end if**
15 $\sigma_k = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(p_i - \mu_k)^2$, $p_i \in |q(x_i)|_k \in \mathbb{R}^3$ // *Variance estimation*
16 $\Gamma_i = \Sigma \sim N(\boldsymbol{x_i}, \sigma_k)$, $i = 1, \dots M$, $k = 1, \dots 4$ // *Generate GMM*
17 **for** *N* iterations **do**
18 $\boldsymbol{x^*_{IP}} \leftarrow \boldsymbol{x_i^*} \in \Gamma_i$, $i = 1, \dots M$ // *Sample* Γ_i *to obtain* $\boldsymbol{x_i^*}$*new IP*
19 **for** *M* - 1 iterations **do**
20 $a_l \leftarrow \cup\, arc(\boldsymbol{x_i}^*, \boldsymbol{x}^*_{i+1}, c_i)$ // *Smoothly connect new IP*
21 **end for**
22 **end for**

Output:
$\boldsymbol{a} = \{a_1, a_2, \dots a_N\}$ – Set of artificial trajectories

Fig. 5 and Fig. 6 show generated trajectories for some of the gestures included in each of the datasets used to test the proposed framework. The thicker lines represent the original trajectory gathered from the user, while the fine lines show different artificial trajectories for the right and left arm, respectively.

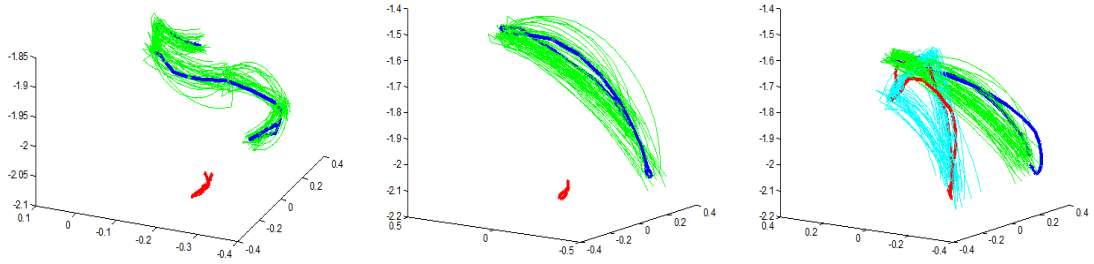


Figure 5. Recreated trajectories for some of the gestures from the ChaLearn 2013 Challenge. The thicker lines represent the original trajectory whereas the finer ones were artificially generated.

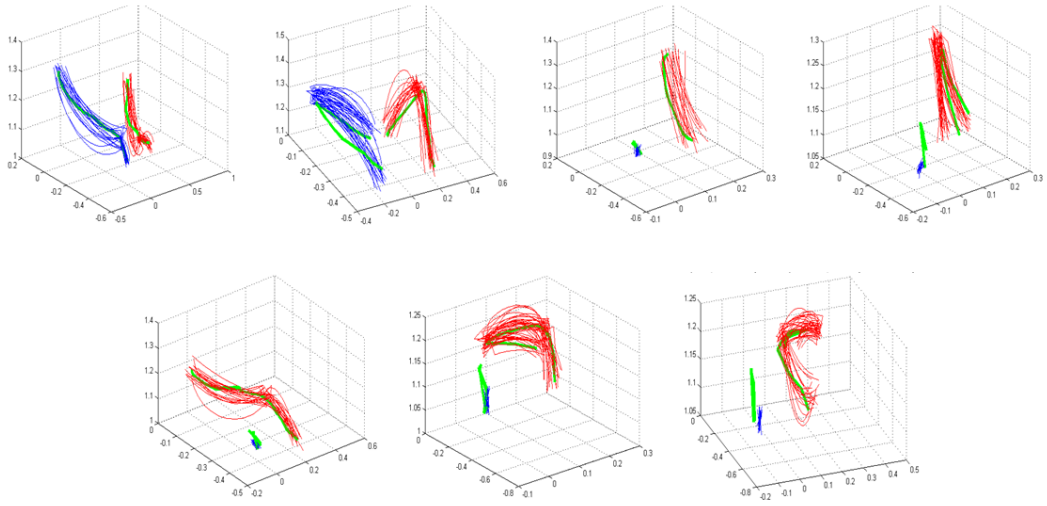


Figure 6. Recreated trajectories for each gesture. From upper left to lower right: Zoom In, Zoom Out, Pick, Drop, Discard, Rotate CCW, Rotate CW

The next step is to use the resulting artificial observations, as inputs to a classifier in order to train and posteriorly classify future instances of the same gesture.

4. *Classification*

The classification strategy is based on a one-vs-all scheme. The approach presented is independent to any specific classifier. It is recommended to use the one that delivers best performance. As a mode of an illustrative example, three different classifiers were implemented, namely HMM, SVM and CRF to test the previously described method. Nevertheless, note that the approach presented is agnostic to the selected classifier.

In the case of HMM, each HMM is comprised by five states in a left-to-right configuration and trained using the Baum-Welch algorithm, which has been previously shown to generate promising results in hand gesture recognition [7]. An observation is classified based on the specific HMM chain that explains better that observation. That is, determining which of the trained HMMs outputs has the highest probability for a state sequence $\vec{z}$ given a new observation $\vec{x}$ and its intrinsic parameters A and B to assign the corresponding label to the new sample. The formulation is as follows in (7):

$$\arg\max_i \{ \log(P_i(\vec{z}\,|\vec{x}; A_i, B_i)\}, i = 1, \dots, N \qquad (7)$$

For the SVM, each was trained using the RBF kernel function. In the case of CRF, the training examples were encoded using BIO, to determine the beginning (B), inside (I), and outside (O) of a gesture.

The training dataset includes 500 artificially generated examples. The testing dataset is comprised by 100 examples extracted from the dataset for Italian gesture recognition from ChaLearn 2013 [5]. Thus, one-shot learning is accomplished by training the classifier on artificially generated instances. Said procedure enables the recognition of future instances of each gesture in the dataset.

5. *Experiments and Results*

The experimental section of this work has two sections: one where the implemented framework was trained using the customized gestures for image manipulation, and tested against 30 observations coming from a set of users who performed the gestures in the dataset; and another section in which the trained classifiers are tested against 100 testing samples of each gesture from the development stage of the ChaLearn Gesture Challenge for 2013. In both cases, the classifiers' performance was compared using ROC curves and confusion matrices as the metrics.

It is important to specify that even when the ChaLearn 2013 dataset was used, the challenge itself was not part of the experiment. Each gesture used for testing had a label assigning it to one of the 10 gestures selected for this particular experiment. No spotting technique was applied. This results in a testing dataset of 1000 gestures.

In order to obtain the ROC interaction curves for each classifier, a free parameter was selected in each to vary and obtain different values for hit rate and false alarm. In the case of HMM and CRF, given that their configuration is intrinsically related to probabilities, the parameter was assigned as the ratio between the highest and the second highest probability obtained from each classifier. In the case of the SVM, the selected parameter was the scaling factor in the Gaussian radial basis function kernel. These parameters were varied and used as threshold with three different values and the curves were completed with the two extremes: (0,0) and (1,1).

The same parameter was used 4 times, dividing the datasets and reshuffling in groups of 25. Fig. 7 and Fig. 8 show the means in the obtained ROC curves for the three classifiers for each dataset respectively.

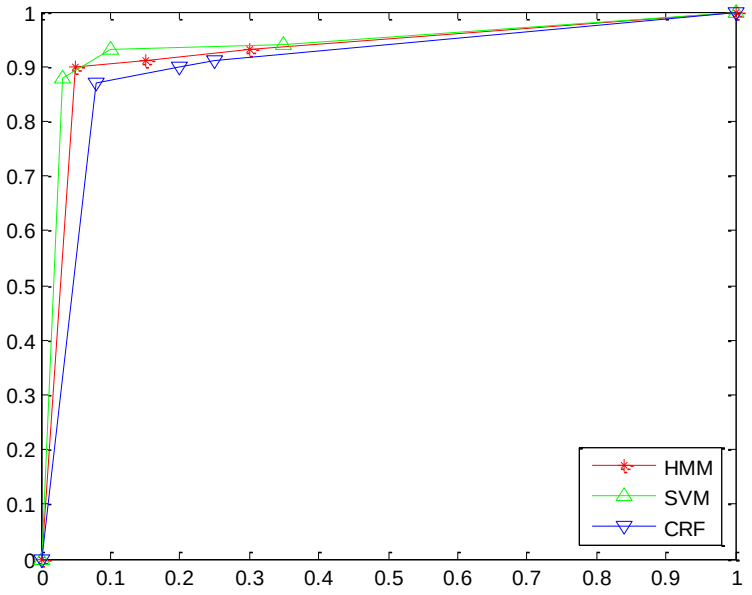


Figure 7. ROC curves for all three implemented classifiers using the image manipulation dataset.

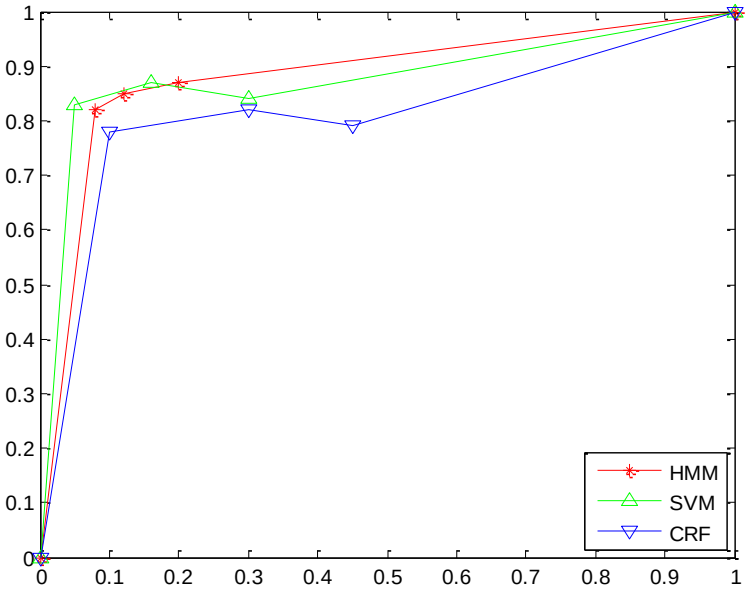


Figure 8. ROC curves for all three implemented classifiers using ChaLearn 2013 data

By visual inspection, all three classifiers show rather similar performance around 90% recognition rate for the image manipulation dataset and 80% for the ChaLearn data. However, HMM and SVM show higher hit rates than CRF. The highest hit rate was obtained with SVM in both datasets at 92% and 87% respectively, followed by HMM with 91% and 85%.

Confusion matrices were also used to determine the classifiers' performance on the ChaLearn dataset and are shown in the following figures, Fig. 9, Fig. 10 and Fig. 11 respectively for HMM (85.9%), SVM (85.2%) and CRF (81%). Previous results with the same dataset are mentioned in

[5], where the winning teams using the test data achieved scores of 87.24%, 84.61% and 83.19% respectively.

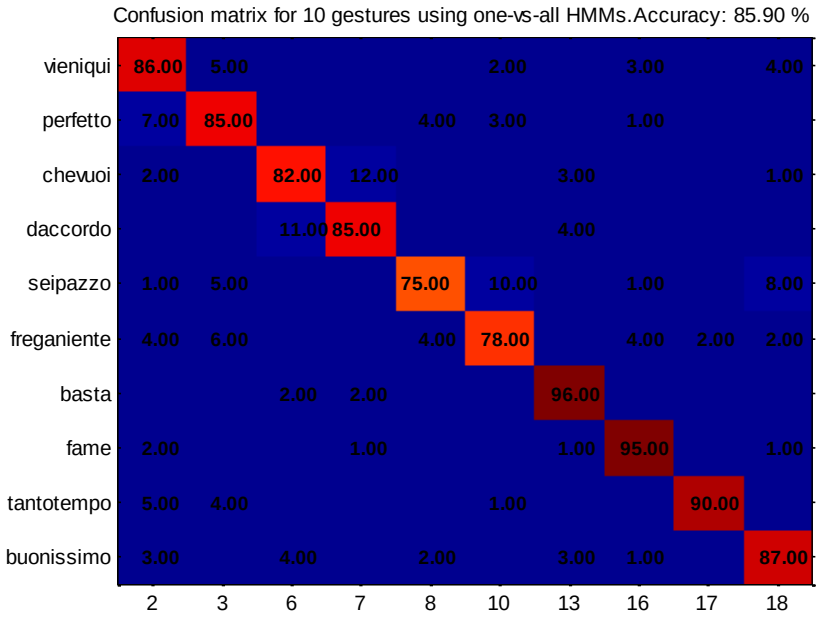


Figure 9. Confusion Matrix for 10 gestures from ChaLearn 2013 using HMM

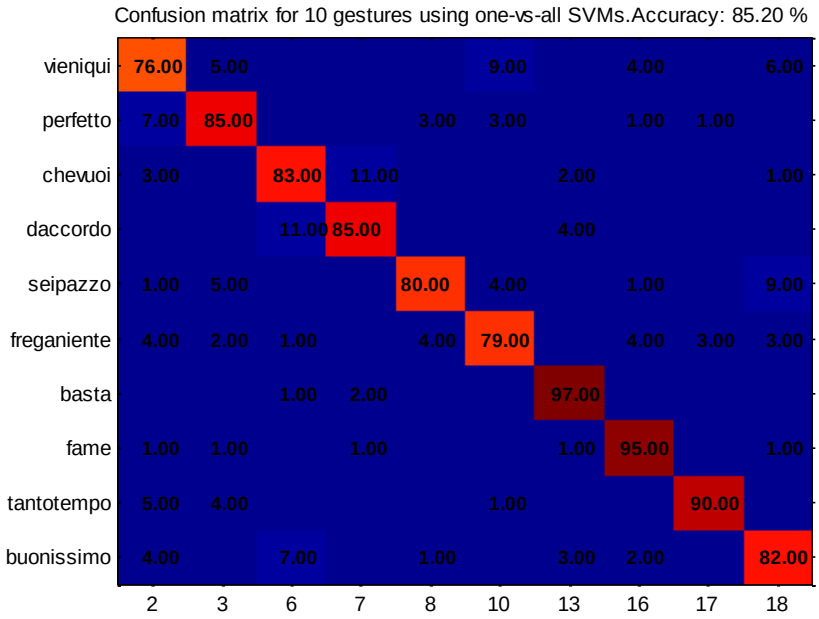


Figure 10. Confusion Matrix for 10 gestures from ChaLearn 2013 using SVM

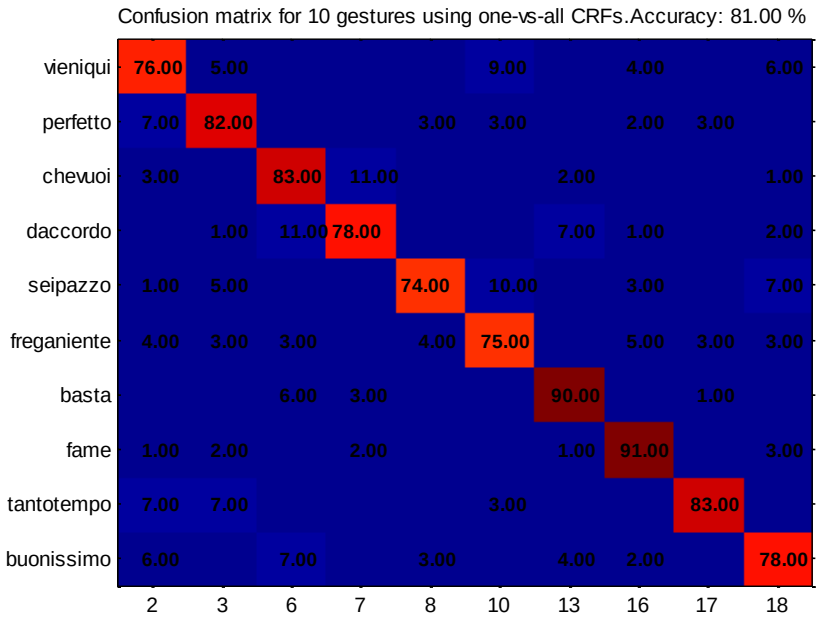


Figure 11. Confusion Matrix for 10 gestures from ChaLearn 2013 using CRF

6. Discussion

The One Shot Learning approach taken in this project, which is Phase 2 of this Task, differs from the common ones where the learning and classification aspects are given more relevance than the data inference process. The implemented framework leverages human biomechanics and motion information, to use as context and to extract "the gist of the gesture". This reduction to a minimum expression which encapsulates the commonality in the gesture itself, allows us to generate artificial observations. Such observations incorporate variability into the classification process.

The "gist of the gesture" framework is not dependent on the classification method used, shown by implementing, testing and comparing three different methods: HMM, SVM and CRF. All methods show similar performances not only one customized gesture vocabulary, but in a public dataset originated from the ChaLearn gesture recognition competition. The reported performance adds to the validity and relevance of the framework.

References

[1] S. Kita, I. van Gijn, and H. van der Hulst, "Movement phases in signs and co-speech gestures, and their transcription by human coders," in *Gesture and Sign Language in Human-Computer Interaction*, I. Wachsmuth and M. Fröhlich, Eds. Springer Berlin Heidelberg, 1998, pp. 23–35.

[2] P. Viviani and C. Terzuolo, "Trajectory determines movement dynamics," *Neuroscience*, vol. 7, no. 2, pp. 431–437, Feb. 1982.

[3] A. F. Bobick and A. D. Wilson, "A state-based approach to the representation and recognition of gesture," *Pattern Anal. Mach. Intell. IEEE Trans. On*, vol. 19, no. 12, pp. 1325–1337, 1997.

[4] M. Ahissar and S. Hochstein, "Task difficulty and the specificity of perceptual learning," *Nature*, vol. 387, no. 6631, pp. 401–406, May 1997.
[5] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante, "Multi-modal gesture recognition challenge 2013: Dataset and results," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 445–452.
[6] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992.
[7] M. G. Jacob and J. P. Wachs, "Context-based hand gesture recognition for the operating room," *Pattern Recognit. Lett.*, vol. 36, pp. 196–203, Jan. 2014

Research Goal 2: Learn probabilistic models to assess user's focus of attention and intention

Participants: Juan P Wachs (PI), Ting Zhang (graduate student).

In this research goal we investigated the hypothesis that the modalities of interaction selected in Research Goal 1 while solving the different versions of the TSP indicate the user's focus of attention. Our central assumption was that meaningful forms of interaction (in terms of effectively solving the TSP problem) release the user from thinking about how to map intentions to commands, and instead allow her to allocate most attentional resources on how to solve better the problem from the computational stand point.

Question(s) addressed: Can level of attention be assessed using non-disruptive, non-subjective models of attention?

During Year 2 of the project, we focused most of our efforts in the optimization of the probabilistic models, built and learnt in Year 1. Ten Bayesian Attention Networks (BANs) were built in Year 1, which were named as ten candidate BANs. Five of the networks are built by human experts, with the other five learnt based on evolutionary approach. The networks built by human experts inherited experts' knowledge (which was indirectly shaped by the shape of their bodies) and indicated their preferences for simple networks, while the networks learnt through evolutionary approach represented the observations from experimental data. In Year 2, to improve the performance of BANs generated from the evolutionary approach, the experimental data was optimally divided, and five new BANs were learnt.

To integrate the findings from both experts' knowledge and experimental observations, a Node Consensus Model (NCM) was applied in Year 1 to produce an integrated BAN from the ten candidate BANs, with equally distributed importance for each candidate. However, it is not realistic for each BAN representing the same importance since different experts show distinct levels of knowledge and the networks generated by stochastic evolutionary approach showed different levels of fitness to the experimental data. Therefore, the previously applied NCM was extended and integrated with a neighborhood search algorithm to find the optimal BAN that fulfills not only the expectations of human experts, but also the fitness to experimental data. Two score functions were studied and applied in the neighborhood search algorithm to evaluate the

performance of integrated BANs. The contribution of each candidate BAN to form the optimal integrated BAN was also obtained from the neighborhood search process.

The main accomplishments of Year 2 are: (a) the development of new BANs learnt from evolutionary approach using optimally divided experimental data (b) the optimization of BANs by extending NCM and integrated with a neighborhood search algorithm. So far the main scenario for our study is the TSP. We plan in Year 3 implementing and validating our approach with the cyber-physical system (given that we are given access to the CSA Explorer) . The aspects studied in Year 2 are denoted in Figure 8, inside the dashed red rectangle.

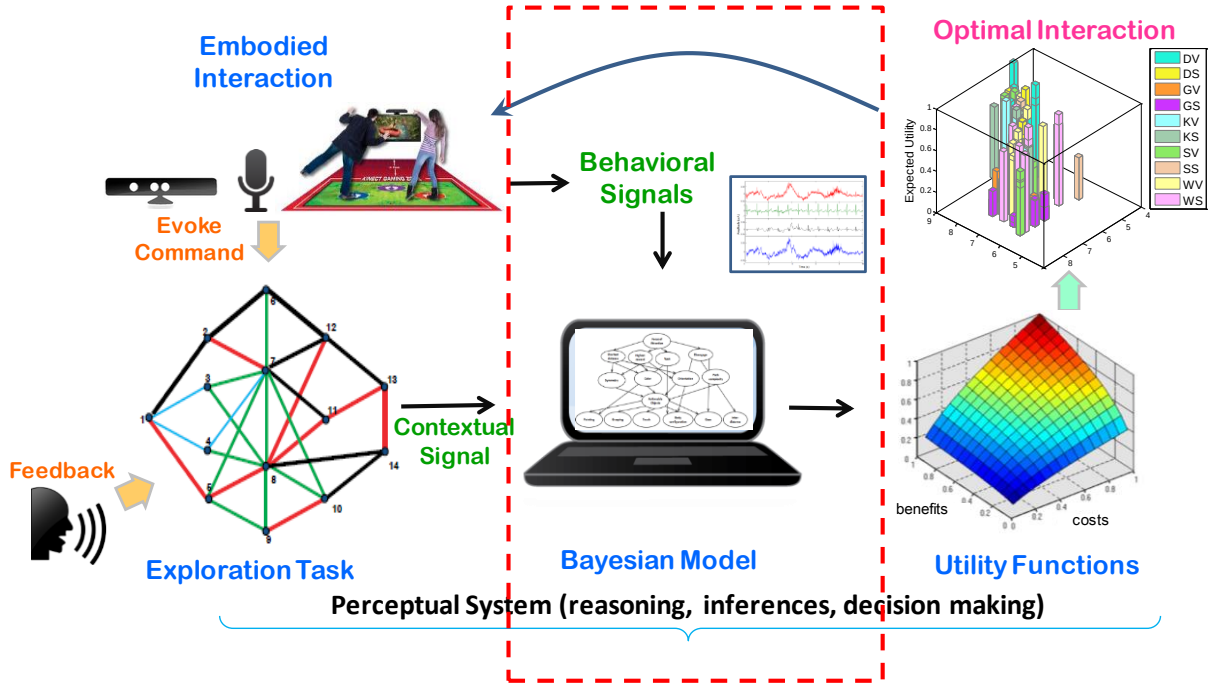


Figure 8. Research Objective 2 denoted under the dashed line

Task 1: Construct Models for User's Attention

During Year 2, we modified and extended the procedure to construct models representing user's attention. Figure 9 shows the system architecture of the Bayesian Attentional Network (BAN) framework. The parts within the dashed red rectangle were updated and improved.

First, the experimental data was divided in a different way that more data can be used to train the BANs making our results more statistically significant. Five new BANs were trained and learnt following this new procedure. Second, as in Year 1, a systematic approach was developed to integrate human-operator's knowledge and agents' solutions (solutions from evolutionary approach). However, different from previously applied NCM, in Year 2, each candidate BAN was assigned a factor, representing its importance or contribution to the integrated BAN. Instead of directly considering the integrated BAN from NCM as the representative network, the optimal BAN was determined by applying a neighborhood search algorithm integrated with the extended NCM. Different score functions were developed to evaluate the performance of BANs produced in the searching process. These procedures are explained in details in the following sections.

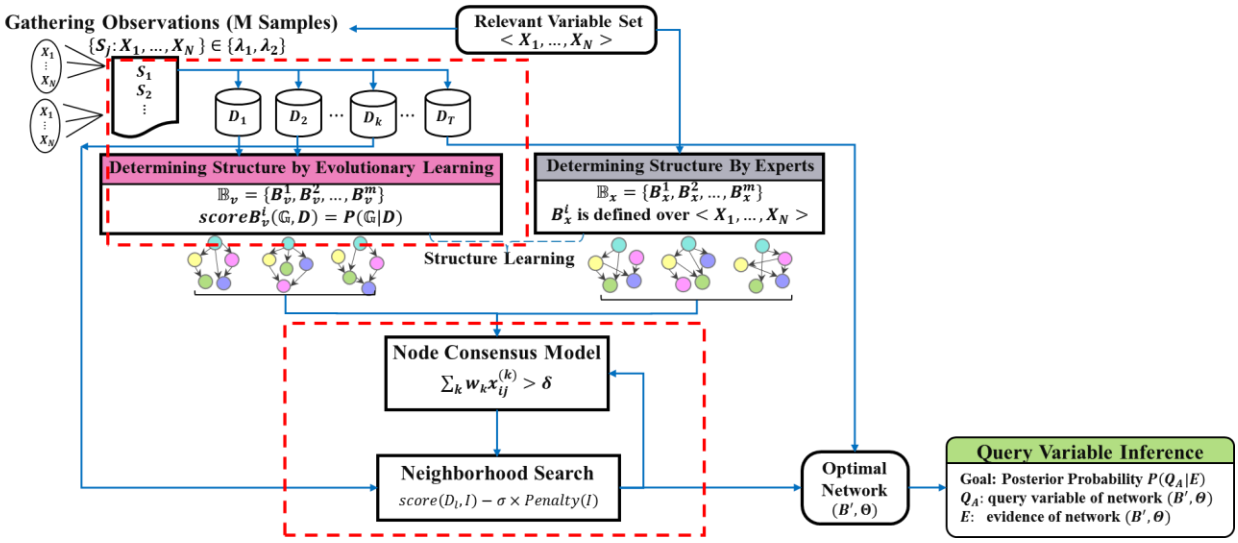
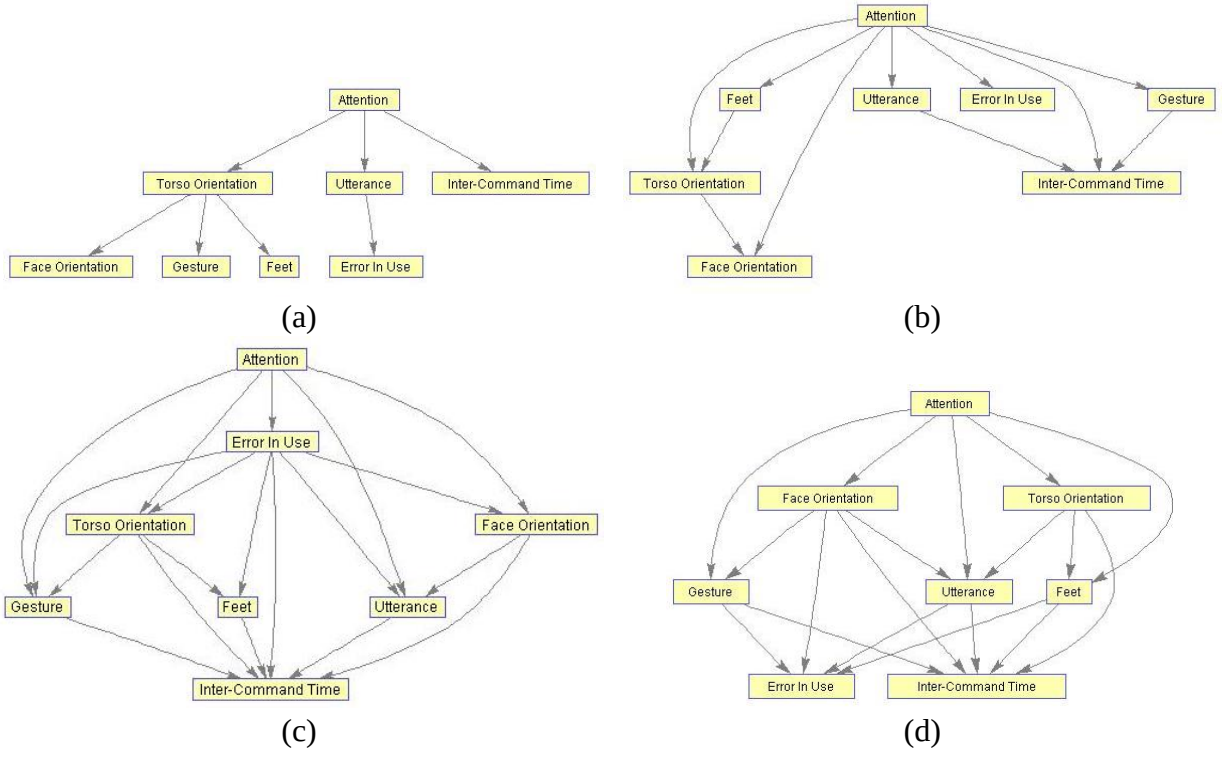
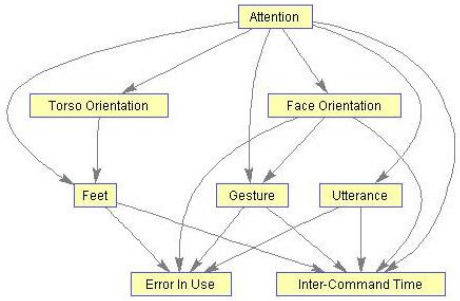


Figure 9. Architecture of the representative BAN construction process

A. Determining the BAN Structure through Operators' Knowledge

We completed this aspect of the project in Year 1. The five BANs obtained from human-operators are shown in Figure 10.





(e)

Figure 10. Five BANs built by human-operators.

B. Determining the BAN Structure through Evolutionary Learning

An evolutionary-based modeling approach was adopted to construct the Bayesian network in Year 1. In Year 2, to improve the performance of BANs generated from this approach, the experimental data used was divided in a different way. Figure 11 illustrates how the experimental data was divided. Five different BANs were learnt following the evolutionary modeling process. To learn these BANs, the experimental data was divided into two parts: the training data used for learning, and the testing data used for validation. Using the k-fold paradigm, the training data was then randomly divided into five parts (k=5), and each BAN was learnt using four out of the five parts of the training data. Since the amount of training data should be at least larger than the amount of testing data to produce representative BANs, four-fifth amount of the training data should be larger than the amount of testing data.

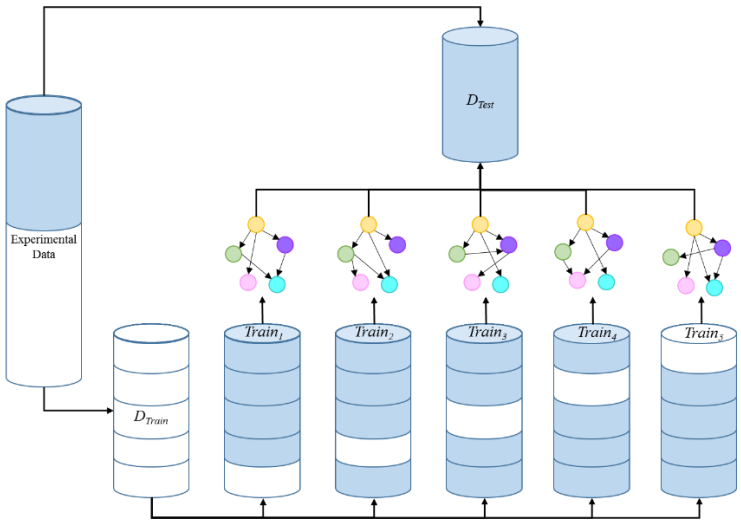


Figure 11. Experimental data division for BAN training and testing.

The five new BANs trained following the data division rule described above are shown below in Figure 12.

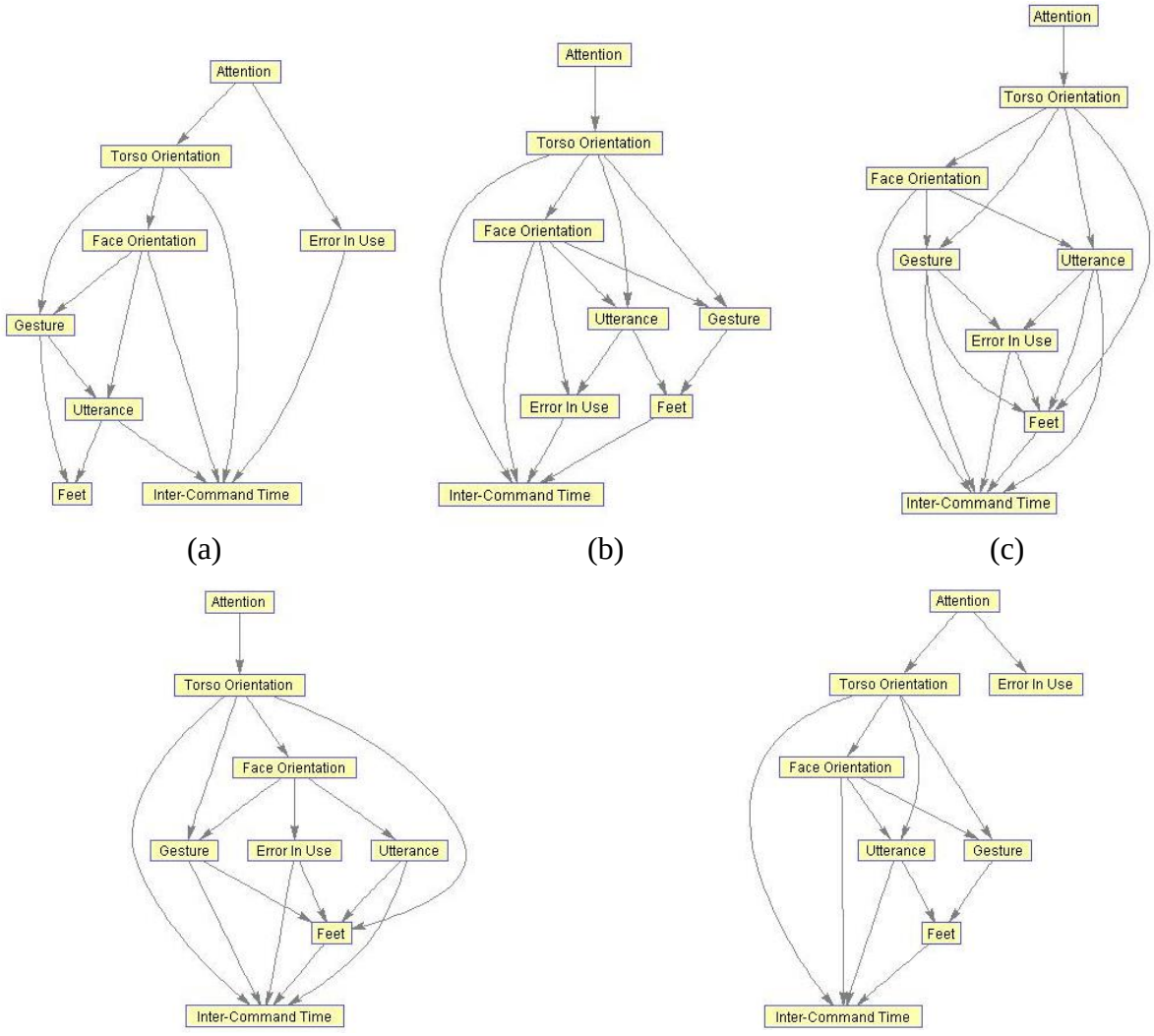


Figure 12. Five BANs learnt by evolutionary approach.

C. Node Consensus Model with Relative Importance

The previously applied Node Consensus Model (NCM) seeks the majority of votes among all candidate BANs and weights all candidates equally. However, this is not realistic since human operators show various level of expertise (same with virtual agents) and expressed by different levels of the BANs's fitness to experimental data. Therefore, in Year 2, we assigned a weight for each candidate BAN representing expertise and seek the largest agreement when the weights applied.

Assume there are K BANs in the candidate set, and each candidate can be represented as an adjacency matrix A_k with each element in the matrix denoted by x_{ij}, where i , $j \in \{1 \cdots N\}$. N is the number of nodes in the network. An entry "1" assigned to x_{ij} means that nodes i is connected to node j, and "0" is assigned otherwise. Figure 13(a) is an example of adjacency matrix for a model with 4 nodes. To represent different levels of knowledge and fitness to data, there is an importance factor, w_i, assigned to each candidate BAN. The importance factors are normalized, so that all weights add up to 1. The importance factors are multiplied with each element in the adjacency

matrix, thus generated a weighted matrix for each candidate (see Figure 13 (b)). These *K* weighted matrices are then added up by performing additions with the elements in the same position of all matrices, to form an integrated matrix, *I* (shown in Figure 13(c)). The consensus rule is at last performed to obtain the representative model. A consensus value, *c*, is determined empirically, so that if the element in the integrated matrix I_{ij} is larger than *c*, the element should be 1, which means node *i* is connected to node *j*, and the element should be 0, otherwise. An example of seeking consensus with weights assigned to each candidate is shown in Figure 13 and also summarized in Algorithm 2.

Algorithm 2: Node Consensus Method with relative importance

Input:
A_k – matrices representing a set of graphs, each with order *N*
w_k – importance factors for each candidate BAN
K – the number of candidate BANs

for all *i,j<=N* **do** // *given i,j as the source and destination indices of nodes x*
  $I(i,j) \leftarrow 0$
end for
for *k<=K* **do**
  for all *i,j<=N* **do** *// apply importance factor to each element in the matrix*
    $A_k(i,j) \leftarrow w_k \times A_k(i,j)$
  end for
  for all *i,j<=N* **do** *// add elements of the same position in all candidates*
    $I(i,j) \leftarrow I(i,j) + A_k(i,j)$
  end for
end for
for all *i,j<=N* **do** *// apply consensus rule*
  if *I(i,j)>=0.5* **then** *// majority is more than 50% agreement*
    $I(i,j) \leftarrow 1$
  end if
end for
$\mathcal{G} \leftarrow$ Mat2Dag(*I*) // *convert the adjacency matrix to the directed graph*

$\mathcal{G}$:= enhanced graph with majority consensus
Output: enhanced graph $\mathcal{G}$ with adjacency matrix *I*

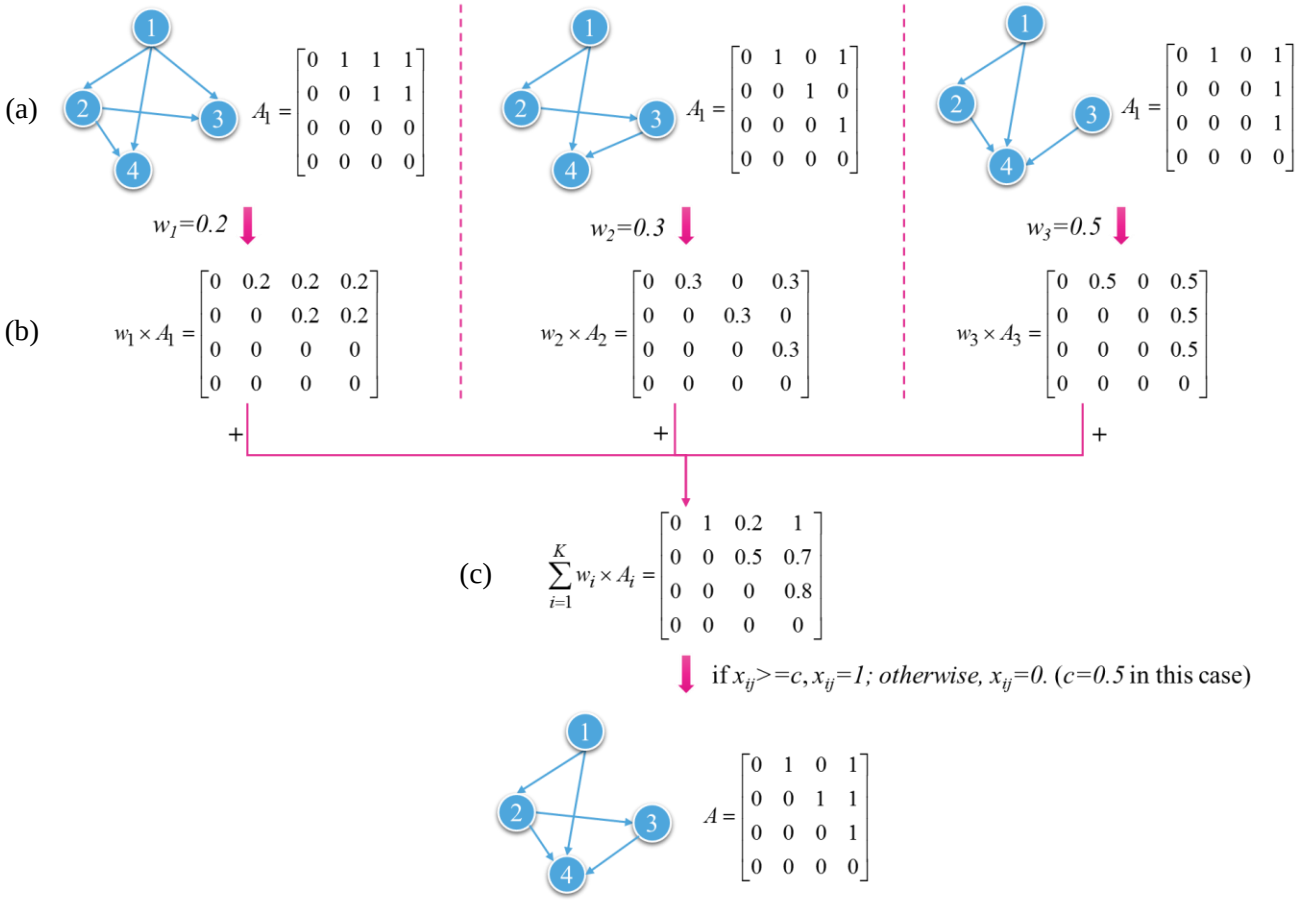


Figure 13. Example of NCM with relative importance.

D. Determining the Relative Importance

To further optimize the Node Consensus Model (NCM) obtained from Algorithm 2, and investigate the contribution of each candidate model to the final optimized one, a Neighborhood Search Algorithm (Wachs et al., 2005) was applied in this study as well. Neighborhood search algorithm is aimed at finding the solution that optimize a score function by searching the solution space's neighborhood repeatedly. In this study, the optimal BAN is considered as a structure that not only can explain the empirical observations but also can reflects the intellectual insights of human operators. Two score functions (explained in the next section) were studied and tested in this research to evaluate a BAN's performance considering both criterion. Therefore, two sets of neighborhood search processes were conducted to seek the optimal BAN using two different score functions (results shown in part *F. Experiments and Results*).

Regardless of different score functions, the neighborhood search follows the same procedure. First, each candidate BAN is randomly assigned an initial importance factor w_k and all factors add up to 1, where K is the number of candidate BANs.

$$\sum_{k=1}^{K} w_k = 1 \tag{5}$$

The NCM with relative importance algorithm is then applied to produce an integrated model. Score function $f_S(I)$ is applied next, to evaluate the performance of this model, represented with a score P. The score P is considered as the temporary maximum score P_{max} in the searching process, and I is considered as the optimal model I_{best} found so far. Assume the searching process lasts M iterations, in each iteration, we change the importance factor for all candidate BANs one by one. Every time after we change the importance factor for one candidate BAN, we generate the integrated model with the modified importance factors. The score for this model is then computed using $f_S(I)$ and compared with P_{max}. If the current P is greater than P_{max}, then P_{max} is updated with P and the optimal model I_{best} is updated with current I. By the end of each iteration, the best configuration of importance factors is selected and used as the initial importance factors for the next iteration. After several iterations, we can observe that P_{max} is not increased and the BAN survived after M iterations will be the optimal representative BAN. The number of iterations, M, is determined empirically. This process is also summarized in Algorithm 3.

Algorithm 3: Neighborhood Search

Input:

A_k – matrices representing a set of graphs, each with order N

K – the number of candidate BANs

```
1   w_k ← random(K)
2   I_best ← NCM(A_k,w_k,K) // see Algorithm 2
3   P_max ← f_S(I_best)
4   for M iterations do
5       w_k ← w_best // update importance factors
6       for each matrix in A_k do
7       // 1. add step to one importance factor
8           w_t ← normalize(w_k+δ)
9           I ← NCM(A_k,w_t,K)
10          P ← f_S(I)
11          if P>P_max
12              I_best ← I; P_max ← P; w_best ← w_t
13          end if
14      // 2. subtract step to one importance factor
15          w_t ← normalize(w_k-δ)
16          I ← NCM(A_k,w_t,K)
17          P ← f_S(I)
18          if P>P_max
19              I_best ← I; P_max ← P; w_best ← w_t
20          end if
21      end for
22  end for
23  G ← Mat2Dag(I_best) // convert the adjacency matrix to the directed graph
```

$\mathcal{G}$ ≔ enhanced graph with majority consensus by neighborhood search

Output:

enhanced graph $\mathcal{G}$ with adjacency matrix I_{best}

$\sum_1^M \delta_{ij}^{(k)}$ – total variance of each matrix

E. Comparing score functions

In the neighborhood search process discussed above, a score function is used to evaluate the performance of an integrated BAN. To evaluate a BAN's performance based on their marginal contribution of a human operator and a virtual agent, two different score functions were studied in Year 2. We represent the score functions in a general form shown in the following equation

$$f_S(\mathrm{D}\,|\,\mathrm{I}) = score(D_l, I) - \sigma \times Penalty(I) \tag{6}$$

score(D_l,I) is the score function used for evolutionary approach (see the equation below), which represents a BAN's level of fitness to the experimental data. The penalty function *Penalty(I)* represents human operator's effect on the score. σ is a normalization factor that balances the effect between scores from evolutionary approach and penalty functions.

$$score(D_l, I) = P(D_l|\mathrm{I}) = \sum_{i}^{2^M} P(d_i \,|\, \mathrm{I}) \tag{7}$$

Where

$$P(d_i|I) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(N_{ij})}{\Gamma(N_{ij} + M_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(a_{ijk} + s_{ijk})}{\Gamma(a_{ijk})} \tag{8}$$

.

The two equations above were explained in Year 1.

In Year 2, two different penalty functions were studied, while the score function used to reflect how well the observation data was the same. One penalty function considers the complexity of a Bayesian network, while the other penalty function focuses on the contribution of the human operators represented by the model.

1. Complexity of a Bayesian network

When we are using the penalty function to evaluate the complexity of a Bayesian network, we prefer simple networks over complex networks. And the complexity of a Bayesian network is defined by the concept of Minimum Description Length (MDL) (Rissanen, 1983). MDL measures the minimum number of bits required to encode a network. The penalty function is defined in the equation below:

$$Penalty(I) = \frac{\log N * p_i}{2} \tag{9}$$

where N is the number of data entries, and p_i is the number of parameters for node X_i, which is the number of configurations of node X_i's parents. Complex networks requires longer encodings using this penalty function.

2. Physical Shape and Appearance of the Human Body

The BAN constructed in this project contains several observation nodes that represent the effect of the operator in the knowledge generation process. Since, our assumption is that the shape of the human body shapes the thought process, we include a shape mode of the body into the penalty function. It includes Torso Orientation, Face Orientation, Hands and Feet. This is summarized in Figure 14. From Figure 14, we can observe that Torso is connected to Head, Hands and Feet physically. Therefore, we penalize a model if any of these three connections are missing. The penalty function is defined in the equation below,

$$Penalty(I) = \frac{\log N}{2} * \sum\left(1 - I_{ij}\right) \tag{10}$$

where i is the index for the node of Torso and j is the index for the nodes of Face, Hands and Feet.

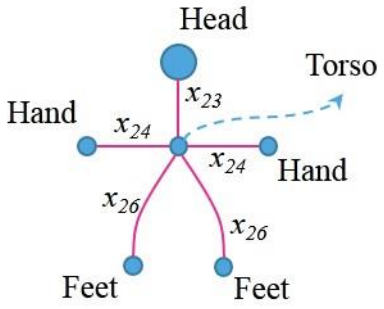
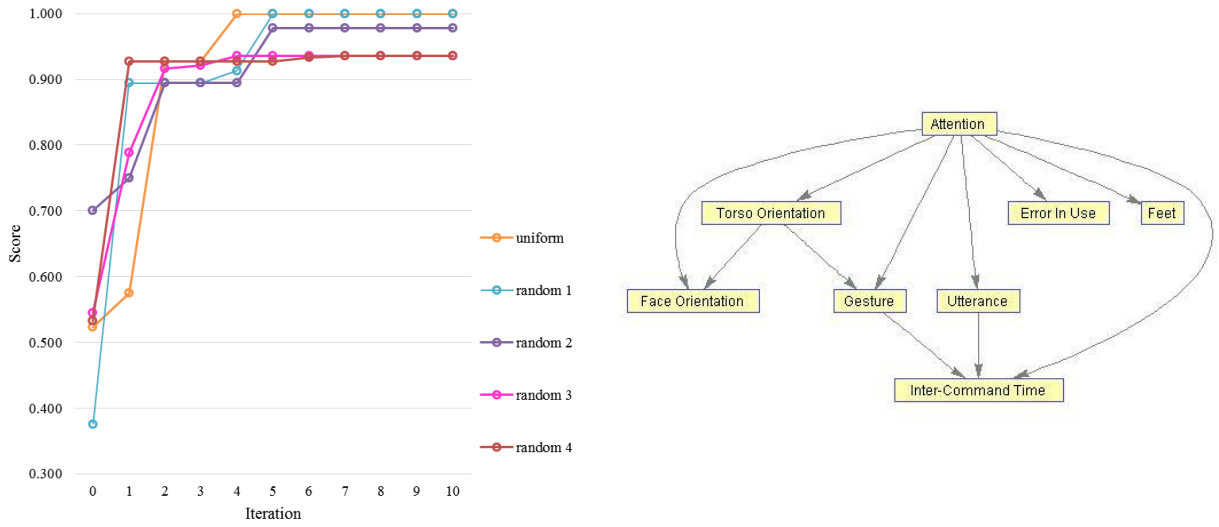


Figure 14. Physical connections between body parts.

F. Experiments and Results

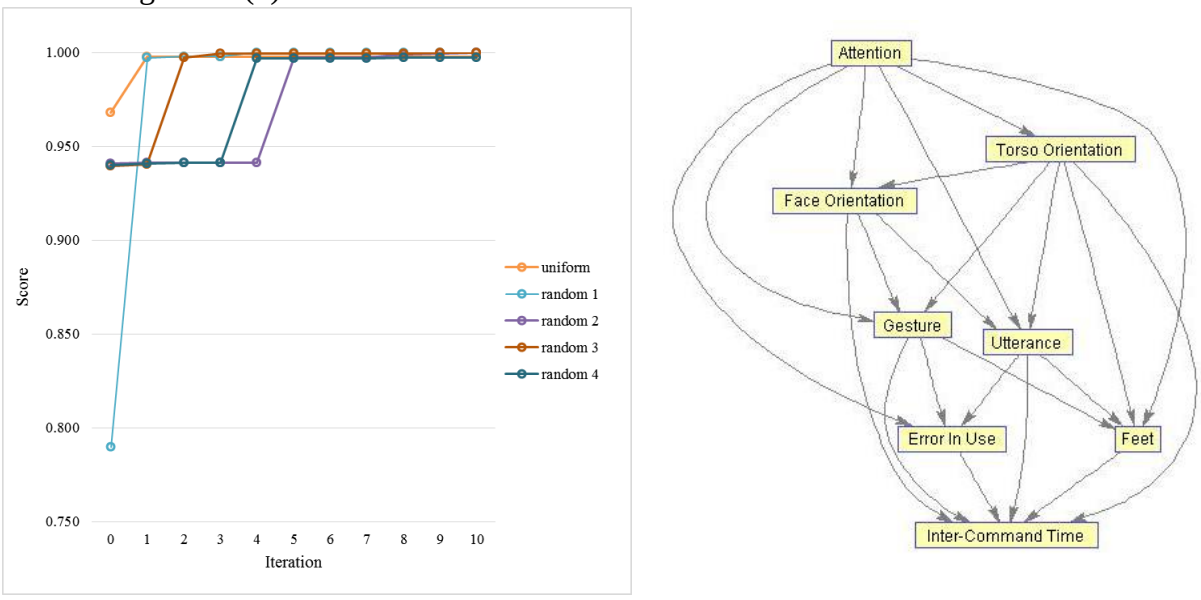
We conducted two sets of experiments using the two different score functions for the neighborhood search process to find the representative BAN. 20 rounds of searching processes are conducted with different initial importance factors. Each searching process ended at the tenth iteration. Figure 15 (a) shows an example of 5 rounds for the searching process using score function with complexity penalty and the optimal BAN obtained from this searching process is shown in Figure 15 (b).



(a) (b)

Figure 15. Neighborhood search process using score function with complexity penalty and its result BAN.

Figure 16 (a) shows an example of 5 rounds for the searching process using score function with physical appearance based penalty and the optimal BAN obtained from this searching process is shown in Figure 16 (b).



(a) (b)

Figure 12. Neighborhood search process using score function with physical appearance based penalty and its result BAN.

Comparing these two sets of searching process, we measured the number of rounds that reached the maximum score and the average number of iterations taken to reach the maximal. From these two measurements, we found there are more rounds of searching process reaching the maximum score using the score function with physical appearance based penalty; however, it always took more iterations as well to converge (an average of 7 iterations) to reach the maximum. The

searching process with score function of complexity penalty reached the maximum by the end of 4.5 iterations on the average.

Task 2: Capturing Activity Cues through Sensors

To infer the user's focus of attention, we need to obtain the evidence nodes in Figure 10. We obtained the evidence nodes from observations of the user while solving a complex, time-sensitive decision making problem. In Year 1, the TSP was used as the main case scenario. The evidence nodes were the discrete measurements (sensors' outputs). This information was calculated directly from the sensors, which included: head orientation (degrees), torso orientation (degrees), gestures (true/false), step gestures (true/false), utterances, and contextual information such as inter-command time, and user errors. The contextual information was directly obtained from the task performance. The embodied related signals were obtained in a similar fashion as that explained in Research Objective 1, Task 2. This means that the physical expressions were pre-determined and the detectors trained in advance. In Year 2, we will adopt the One Shot Learning paradigm previously described. In addition we will apply offline algorithms to extract further information that could indicate attention. These algorithms will be applied to the existing datasets that we have already collected in Year 1 (Research Objective 1, Task 1).

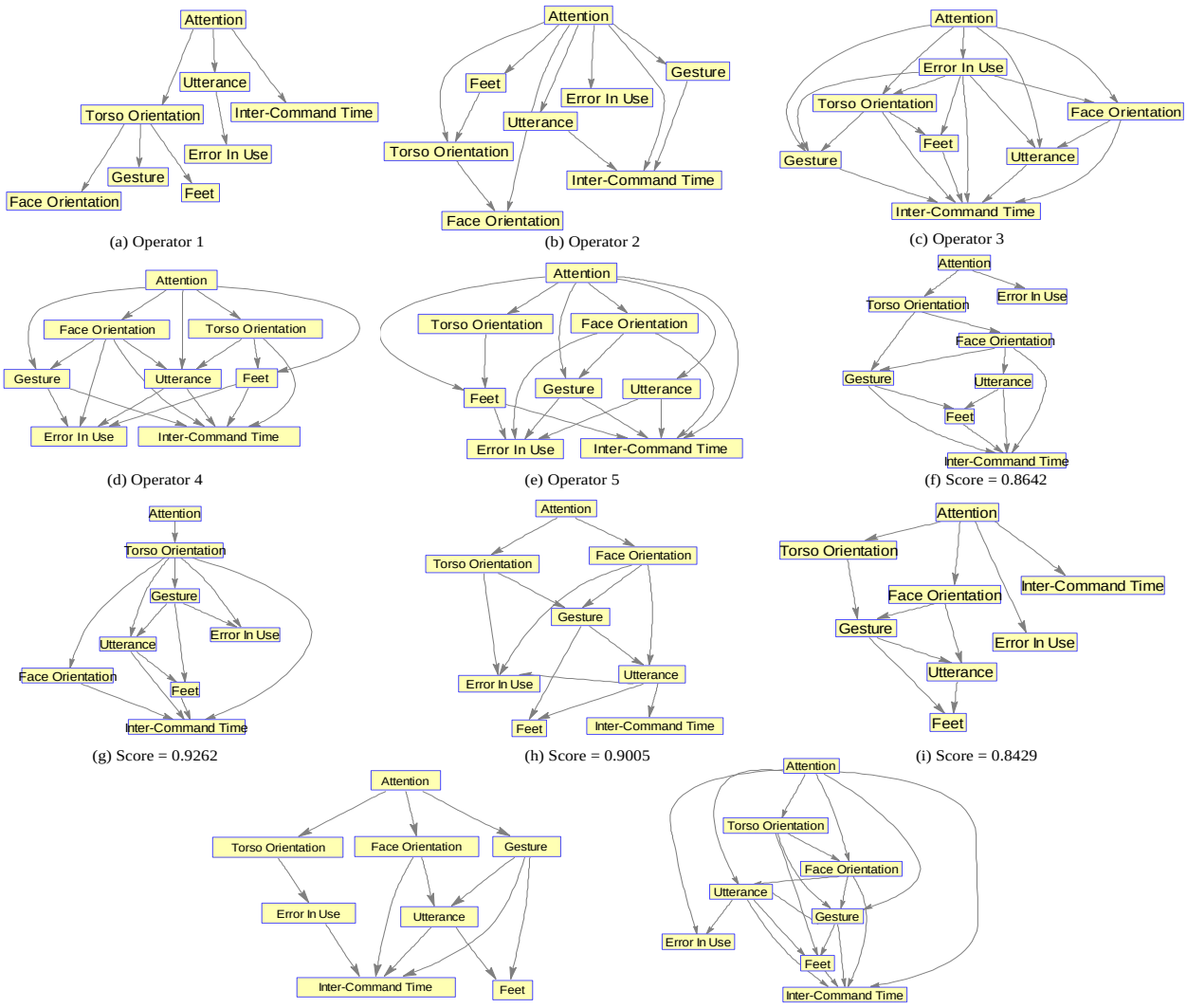


Figure 13. Bayesian Attentional Network's structure obtained by (a) – (e) Operator based, (f) – (j) Evolutionary learning (k) NCM method

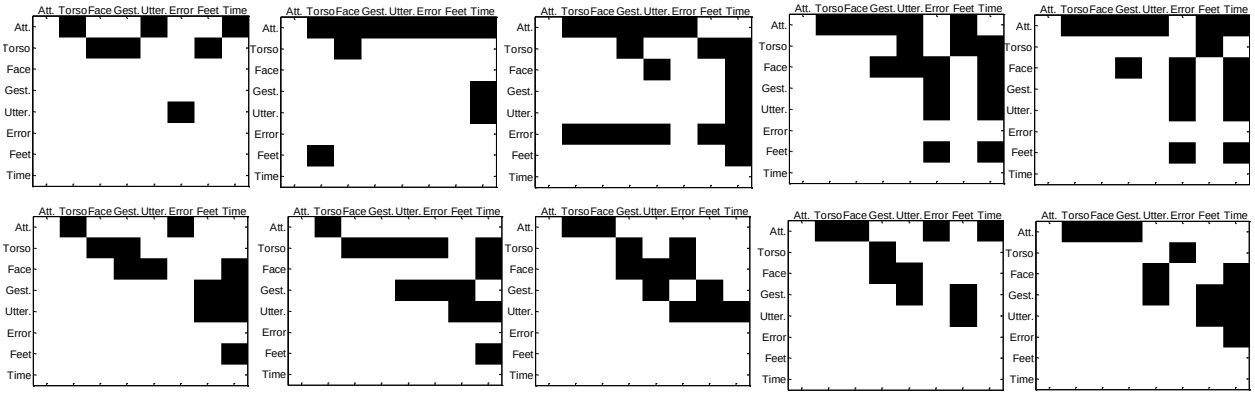


Figure 14. The adjacency matrix of BANs obtained by (a) – (e) Operator based, (f) – (j) Evolutionary Learning

Discussion: We created a new method to generate BANs by combining networks obtained through experts and generated using an evolutionary approach, this method is referred as the Node Consensus Model (NCM). The concept guiding this idea is that using the consensus about the best results given by experts, and the best results found computationally, their combination should be good as well (in a way getting the best from both worlds). One limitation with this method is that the probabilities associated with each edge in the network cannot be obtained, since NCM combines only the topologies of the networks. Without the probability associated with each edge, the previous observations need to used and the EM method is required to find these probabilities. Future work (Year 2) involves finding alternative ways to obtain the probabilities directly from the NCM method.

Another interesting aspect that we realized is that we assumed that each network has an equal vote that is as important for determining the appearance of the final network. In the next step, Year 2, we plan to experiment adjusting different "weights" to the networks based on some performance metrics. For the experts' based networks, this weight could reflect the level of expertise or familiarity with the problem being solved and/or the interface. For the genetic based networks, the weight could be proportional to their corresponding score (the objective function associated with each network).

Accomplishments: Given the constructed Bayesian networks (defined in Section III), a discrete probability distribution of the focus of attention is computed by updating the values of evidence nodes and considering the conditional dependencies of all evidences (based on the observations). We developed the concept of Bayesian Attention Networks (BAN) and found two ways to generate the BANs. One using subjective opinions of experts (or operators) and the other method is using an evolutionary approach. From the pool of solutions using each of the methods, we seek for consensus among these candidates and pick a solution that was agreed by the majority of the candidates. We accomplish this goal using new method called Node Consensus Model (NCM).

Discussion: We created a new method to optimize the BANs built and learnt through experts and an evolutionary approach; this method is referred as a hybrid method of Neighborhood search process integrated with the Node Consensus Model (NCM). The concept guiding this idea is that each candidate BAN can contribute differently to produce an integrated BAN and the contribution for each candidate is unknown. Therefore, by performing a search around the neighborhood of a possible configuration for contributions, we can gradually find a better solution. Future work (Year 3) involves testing alternative ways to generating the BANs based on induction rules.

Another interesting aspect is that score functions applied in the searching process defined the property of result models. For instance, with a score function seeking simple networks, the best structure found after the searching process would be a model with relative simple structures. While with a score function seeking the maximizing the shape of human appearance, the best model found after the searching process would be a network that optimally describes the physical connections between body parts.

Accomplishments: We accomplished research goal 2 by developing a hybrid approach for probability model optimization. This hybrid approach applies a neighborhood search algorithm integrated with an extended Node Consensus Model (NCM). Candidate models are treated differently in the extended NCM.

Research Goal 3: Explore Effective and Efficient Feedback Techniques

Participants: Juan P Wachs (PI), Yu-Ting Li, PhD (graduated) and Ting Zhang, (grad student)

In this research objective we explored effective and efficient feedback techniques considering cognitive limitations (attentional resources) together with physical limitations (encumbered vs. unencumbered sensors) through a utility function. We developed a decision-making based approach to guide feedback delivery based on the user's focus of attention and the task being performed.

Question(s) addressed:

- How can the tradeoffs between the interaction control modality and feedback modality and their effect in attention be expressed through utility functions?
- Can the BANs developed in Research Goal 2 be applied to the CSA Explorer scenario to infer attention?

Task 1: Determine the effective and efficient combination of interaction control and feedback modality.

During this past year one of the main focused of our work was to discover the relationship between interaction forms, expected utility and performance, several metrics of task performance. To design the relation functions between benefits/costs and the task performance, four performance metrics were used. Let B_i and C_i be defined as the design benefit and cost associated with performance metric i (i=1,,..,4): Recognition, Time, Quality of solution and Operators' satisfaction.

The net value of performance metric i can be computed as the difference of benefits and costs, $B_i - C_i$. The utility obtained by measuring the performance metric i is expressed as a linear function of both B_i and C_i:

$$U_i(B_i - C_i) = (B_i - C_i)/P_{i,max} \quad (11)$$

This difference is normalized by dividing it by $P_{i,max}$, which is the maximum level of performance metric i. Thus, the expected utility function $U(I_k, F_j)$ of using interaction modality I_k and feedback modality F_j is given by:

$$U(I_k, F_j) = \sum_i \omega_i U_i(B_i - C_i) \quad (12)$$

where ω_i is the weighting factor assigned to performance metric i.

The goal is to find the feedback and the interaction modality (considering operators' level of attention and task performance) that yields the highest utility (Eq. 13). The level of attention is represented as a discrete probability distribution (right factor of Eq. 13). Thus, the higher the likelihood of high attentional level, the higher is the utility obtained with these performance

measures. Thus, by multiplying the probability of high attentional level by the expected utility (Eq. (12)), the optimal interaction and feedback modalities can be determined (Eq. 13). The most suitable modalities are those that maximize the expected utility function, considering the probability distribution of attention level, given the observed evidence $\boldsymbol{e}$ and performance metrics:

$$\underset{I_k,F_j}{\operatorname{argmax}}\, U(I_k,F_j)p(X_1=\theta_1|\boldsymbol{e}) \tag{13}$$

Substitution from Eq. 2 gives the following:

$$\underset{I_k,F_j}{\operatorname{argmax}} \sum_i \omega_i U_i(B_i-C_i)\, p(X_1=\theta_1|\boldsymbol{e}) \tag{14}$$

where the probability is inferred by the representative BAN with N variables $< X_1, X_2, \dots, X_N >$, each of which takes a binary value in finite domain $\{\theta_1 = 1, \theta_2 = 0\}$.

The utility of interaction used by the subjects was computed using Eqs. (11 –14), based on the testing data. The testing data consisted of 10 scenarios (5 interaction modalities and 2 feedback modalities) with 20 samples for each scenario. Post hoc power analysis, for a significance level of 5%, is over .99 for that sample size (effect size (η^2>0.36)). The testing data was randomly assigned to subsets. We used letter acronyms to represent different types of modality: *D* for feet movement as interaction modality on dance pad; *G* for gesture with glove; *K* for gesture recognized by Kinect; *S* for speech; *W* for body stance measured by Wii Balance Board. A similar procedure was followed for the two feedback modalities: *V* for visual; *S* for speech. Likewise, acronyms with the first letters of modality and feedback denote, respectively, the modality/feedback condition (e.g. "DS" means feet on dance pad as control, and speech as feedback modalities). To show that the best scenario (or alternatively the worst) is significant, one-way ANOVA (Analysis of variance) was conducted on each independent trial. This cell means model (one-way ANOVA) is a suitable approach for analysis since we want to analyze a two-factor treatment structure in terms of a one-way structure where the treatments are all the combinations of the two factors [57]. Results of one-way ANOVA (F(9,190)=58.75, p=3.44e-50<.05) revealed statistical differences between group means. Repeated Measure Analysis was conducted and it was found that there are no significant changes in the interaction's utility over repeated trials (p > .05). Fig. 1, shows the boxplot of expected utility for each trial in the 10 different scenarios. At the top of each box are the first and third quartiles, while the band inside the box represents the median. The ends of the whiskers of the boxes represent the minimum and maximum of the utilities.

To identify the specific groups that differ, post-hoc test was conducted again. All pair-wise comparisons using Dunnett's test, show that there are 5 modalities (GV, GS, KV, SV, SS) whose expected utilities are significantly lower than modality DV. Then, it was observed that using gestures from the data glove as the interaction modality was significantly worse than using the remaining four interaction modalities.

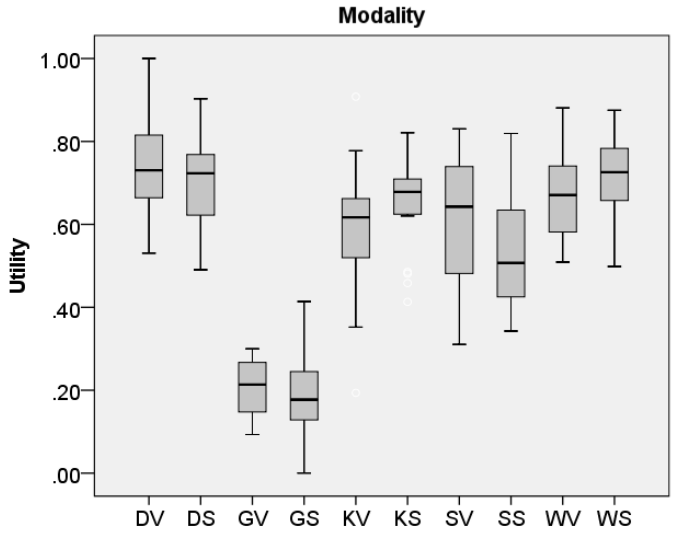


Figure 10. Boxplot of 10 interaction scenarios

A. The authors also tested the hypothesis that the "Quality of solution" (exceeded distance) and "Inter-command elapsed time" can indicate differences between combinations of modalities and feedback with statistical significance. This implies that using BAN's utility metric (Eq. 3) is not necessary to identify significant differences between the modalities. The ANOVA test presented below shows the F-ratio and statistical significance, when each of these metrics was used to identify differences between combinations of modalities and feedback.

Table 3. Results of F-statistics for different dependent variables

Dependent variable	**F statistics**	**Effect Size (η^2)**	**C.I. for η^2**
BAN's utility [Eq. 3]	F(9,190) = 58.75, p<0.001	0.7356	[0.6719, 0.7627]
Quality of Solution	F(9,190) = 0.96, n.s.	0.0433	[0, 0.0518]
Inter-command Elapsed Time	F(9,190) = 6.75, p<0.001	0.2424	[0.1267, 0.2922]
Error in Use	F(9,190) = 1.58, n.s.	0.0697	[0, 0.0907]

It can be seen from Table III that when "Quality of Solution" or "Error in Use" was used as the dependent variable, the significance level was above 0.05, and therefore the comparison between groups has no statistical significance. Thus, the variable "Quality of Solution" or "Error in Use" cannot be used alone.

While the results of ANOVA for "Inter-command elapsed time" show significant difference, the F-ratio and effect size of the of BAN's utility is much larger than the "Inter-command elapsed time" value (F-ratio: 58.75>6.95; η^2: 0.7356>0.2424).

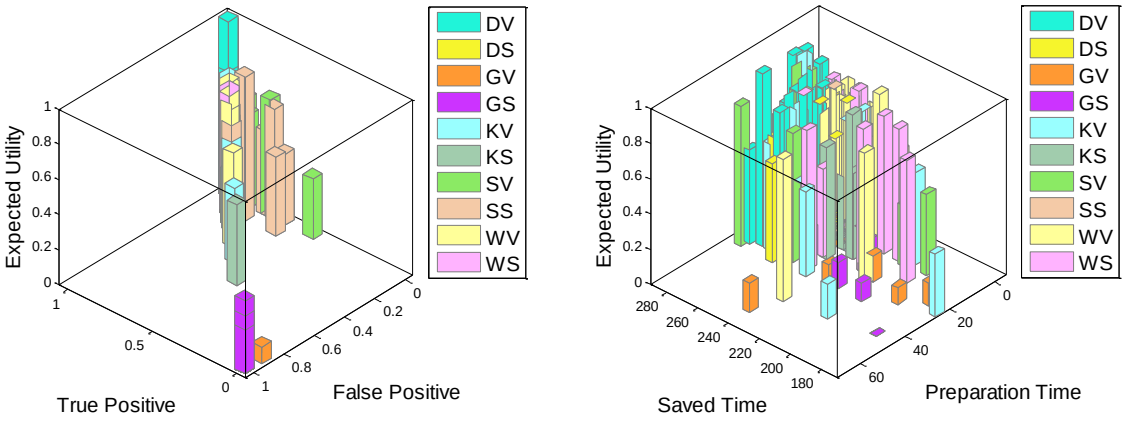
When the BAN's utility (Eq. 16) was not used, and instead the "exceeded distance" or the "error in use" was used as the dependent variable, it was not possible to tell which interaction modality and feedback combination was better than others. This is because there are complex dependencies between physical and contextual variables, and attention, which are not well represented when only the variables "quality of solution", "inter-command elapsed time" and

"error in use" are used. Again, this reinforces the embodied cognition thesis that body shapes cognition.

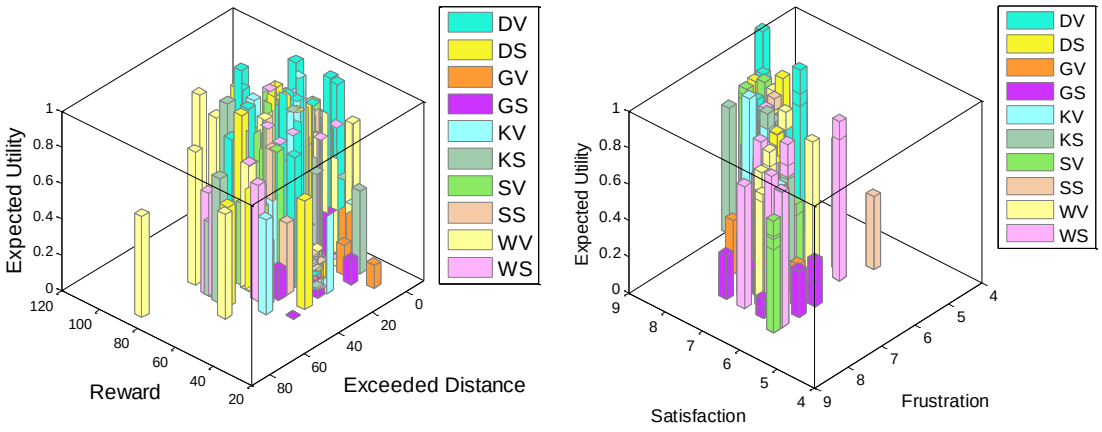
To discover the relationship between interaction forms, expected utility and performance, several metrics of task performance, including recognition rate of interaction, total task completion time, preparation time, and solution quality were measured during the experiment and further compared. In addition, a satisfaction survey was administered after task completion. The relationship between the utility of interaction and performance metrics was determined through post-experimental data analysis (see Fig. 10). The expected utility versus recognition rate is shown in Fig. 11a.

The modality DV received the highest average true positive rate and the lowest average false positive rate. The plots of expected utility versus performance metric B_2, saved time and C_2 and preparation time are shown in Fig. 14. It can be seen that instances obtained from modality DV resulted in higher expected utility, as well as shorter completion time and preparation time.

The bar plots of expected utility versus B_3, reward and C_3, and exceeded distance (the difference between optimal and incurred distances) are presented in Fig. 11. Even in these plots, the modality DV delivered a better solution (higher reward and shorter exceeded distance). Again in Fig. 11, it can be seen that modality DV received a higher operator satisfaction score. Further, better task performance is associated with higher attentional level and interaction utility.



(a) (b)



(c) (d)

Figure 11. Expected utility vs. Operators' performance. The performance is represented by (from top left to bottom right) (a) Recognition; (b) Time; (c) Quality of Solution; (d) Operator's experience.

Accomplishments: We found an analytical formulation for the utility of interaction between different combinations of control and feedback modalities. The results show that the performance achieved by using feet on the dance pad controller was better, in terms of expected utility for each of the four metrics (recognition, time, quality of solution, and satisfaction), than that achieved by using fine gestures (recognized through a data glove) for control and speech as feedback.

The dance pad led to the highest expected utility with respect to accuracy (100% true positive and 0% false positive) without affecting the operator's level of attention on the visualization surface. Besides, by adopting the solution found using the proposed approach (dance and visual), we can achieve high accuracy, save task completion time, and mitigate operators' frustration. Participants were satisfied with the dance modality (turquoise-color bars in Figure 11) explained by the fact that feet movements can convey the navigational physical experience more realistically than the hand movements do. Feet movements resemble "walking" during exploration.

Challenges: We were interested in checking whether this utility functions and our findings will translate well to the CSA Explorer scenario. We encounter a major obstacle in trying to use the CSA explorer at the Rome facility Air Force Research Laboratory, Information Directorate, Rome NY, or alternatively getting a beta version to work from our campus. We were not able to have the people from AFRL help us none of these two options. We informed the program manager about this situation.

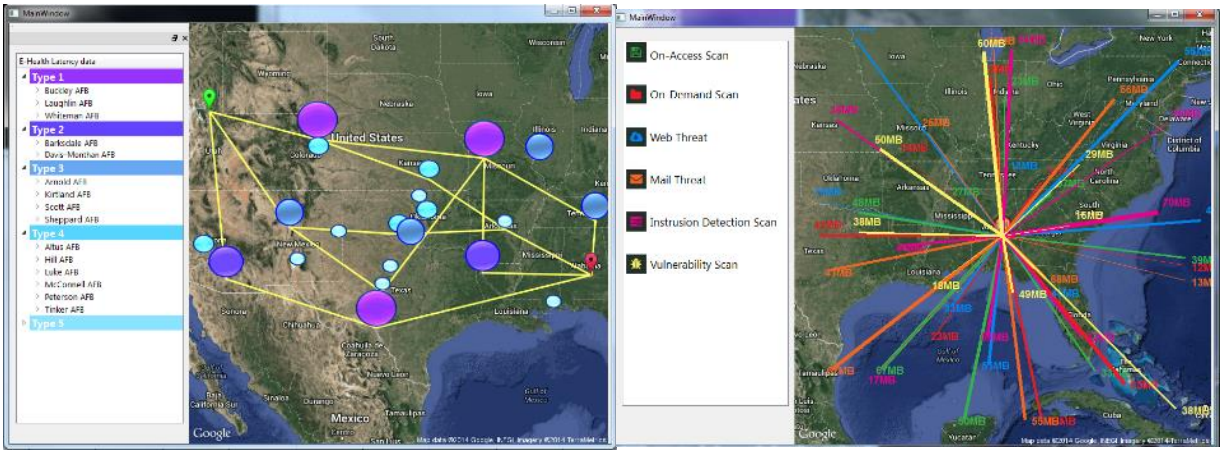
Addressing these questions will be the main purpose of Year 3, therefore currently there is no accomplishments to report on this Research Goal.

Task 2: Cyber physical network case study.

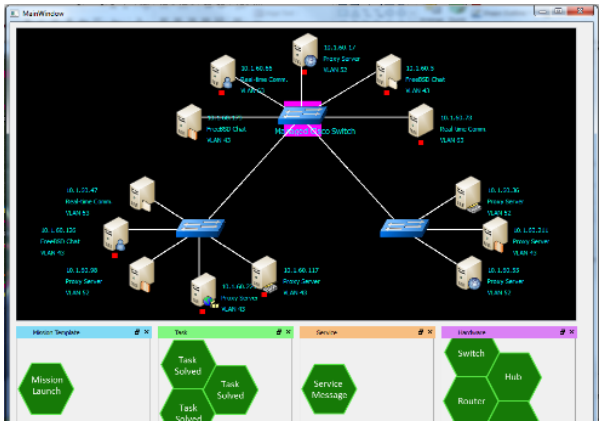
During Year 3, we evaluated the effect of a multimodal embodied interaction interface on user's level of attention in a case study in cyber physical network scenarios. Also, the proposed interface was compared with a traditional interaction interface (keyboard and mouse). The BAN structure obtained in Research Goal 2 is also validated using a dual task approached. Detailed explanation is presented in the following sections.

A. Task

A series of cyber-physical network problems was designed for this case study, including three tasks: (1) The goal for this task is to transfer a data packet from the origin base (marked as green) to the destination (marked as red) in the bases network through a path with less congested bases (shown in Fig. 19(a)); The task is based on a map of United Air Force Bases located around the central United States. In the map, a node represents a base, with the relative size representing the congestion level of bases' intra-network. (2) The goal of the second task is to identify the direction with largest data transmission in the network under 6 different scenarios (shown in different colors in Fig. 19(b)), including the network showing the amount of data transmission of on-access scan, on-demand scan, web threat, mail threat, instruction detection scan, and vulnerability scan. The amount of data transmission is represented through the thickness of lines. (3) The goal involves accessing the machines under threat (marked in red in Fig. 19(c)). Operators are required to traverse in the network and access all the machines under threat.



(a) (b)



(c)

Figure 13. Three tasks of cyber physical network problem

B. Apparatus

The experimental apparatus included a PC and a large 60" screen to deliver feedback, a Kinect sensor, a microphone, a data glove, a dance pad, and a balance board, which were used as interaction modalities. A keyboard and mouse were used for the control experiment.

C. Participants

15 participants were recruited, including 8 males and 7 females, from 20 to 30 years old. Institutional Review Board (IRB) permission was obtained to conduct these experiments.

D. Variables

The two interfaces were the independent variables, while error rate (indicating user's performance) was considered as the dependent variable. The error here refer to the errors when users failed to give commands as they intended. For instance, if the user wanted to go to the node on the upper right corner, however they evoked the wrong action, and went towards the left instead.

E. Procedure

Participants were asked to solve the three spatial navigational tasks using a multimodal embodied interface and a traditional interface. By using the multimodal interface, participants were required to either use hand gestures (which were recognized through a Kinect camera or by a cyber-glove), stepping movements (on a dancing pad with arrows indicating 8 directions), or body stance (through the change of center of gravity on a Wii balance board). Each participant completed 10 trials of each interface.

A secondary task was administered to the participants while they were solving the primary task (one of the three tasks mentioned earlier), to serve as the baseline for assessing the user's level of attention. The 1-back task [Baumann et al. 2007] was applied as the secondary task in this study. Two different T-like visual stimulus were presented each briefly to the participants. The sequence of the two different stimulus were generated randomly. The task for the participants was to identify whether the current stimulus is the same as the first one presented. The correct responses are recorded as hit. The hit rate is used as an attention level estimate for the primary task. If the primary task requires high level of attention from users, then the hit rate of the secondary task will decrease.

F. Results

To test whether the results (level of attention) found through the BAN approach are consistent with those found using the secondary task approach, equivalence tests were conducted (Table 7). For

instance, refer to the first row: a criterion of 0.10 indicates that the measured probability of attention from the two approaches is different at a value of 0.10. This hypothesis is not rejected which means the difference between these two approach can be as large as 0.10. Therefore, by looking at the last row, conclusion can be made that the difference between the BAN and Dual Task approach is less than 0.12, which indicates the similarity of these two approaches.

Table 7. Statistical summary of equivalence tests for attention measure

Metric	Criterion	Dissimilarity	p-value
Level of attention (BAN vs. Dual Task)	0.10	Not rejected	0.113
	0.11	Not rejected	0.063
	0.12	Rejected	0.032*

To compare the effect of embodied interaction in multimodal interface with traditional interaction, the error rate of task completion was measured. Fig. 20 shows the error rate of both interfaces. It can be observed that the embodied interaction has significant lower error rate than the traditional interaction (4.37% < 7.82%). Embodied interaction has shown advantages over traditional interfaces. Users are required to switch their visual focus of attention from the display to the keyboard when using the traditional keyboard interface, which increase error rates (due to the distraction). This does not occurred with embodied interfaces, and thus the errors were fewer.

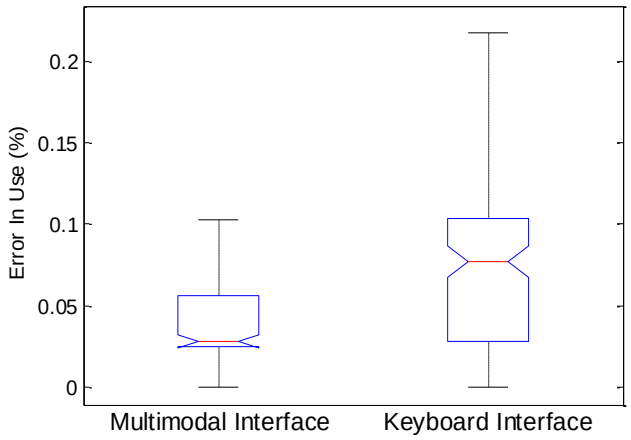


Figure 15. Comparison of error rate between two interfaces

Discussion: In the case study, embodied interaction interfaces was compared with traditional interface (i.e., keyboard). Despite of the fact that users are more familiar with traditional interfaces, the use of the keyboard resulted in higher errors (7.82%), while the embodied interface resulted in an error rate of 4.37%. This finding confirms the claims that traditional interfaces have limitations for spatial navigational tasks [Lee et al. 2012; Schöning et al. 2009] not found in more physical forms of interaction. The findings also indicated a dissonance between the user's intent and its execution via traditional interfaces, not present in embodied interfaces. As Foglia and Wilson [Foglia and Wilson 2013] argued, spatial movements, like front, back, up and down, arise from and are articulated by particular body shape as well as the way people navigate the surrounding space using their bodies.

The representative BAN was also evaluated and compared with dual task approach. Equivalence test was carried out indicating the difference between these two approaches is less than 0.12. Further power analyses have been conducted to validate the effectiveness of the selected criteria. The power analysis indicated a power of 0.959 with 149 trials and an effect size of 0.12. The high power validated the BAN approach is consistent with the secondary task approach at a difference level of 0.12. Future work would include other physiological assessment techniques such as eye tracking and brain control interfaces (EEG) to assess attention and compare to the presented method based on BAN.

Accomplishments: we completed Research Goal 3 by conducting a case study in cyber physical network scenarios. In this case study, we evaluated the multimodal embodied interaction interface by comparing with traditional interface, and validated the BAN structure obtained in Research Goal 2 using dual task approach.

Task 3: Investigating the effect of different gesture lexicons on the level of attention.

In Task 2, we evaluated the effect of a multimodal embodied interaction interface using one predefined gesture lexicons. To further understand and analyze the effect of embodied interaction on attention, we want to develop a multimodal embodied interaction interface with multiple sets of gesture lexicons determined by users, and compare the effects of these lexicons on the level of attention using EEG headsets. Detailed procedures are explained in the following sections.

A. Task Description.

To compare the effect of different gesture lexicons, we want to develop the embodied interface under a scenario of assembly task. Users will be asked to control a robotic arm using the embodied interface to finish the task, Tower of Hanoi. In this study, the Tower of Hanoi will have three pieces of blocks, sizes ranging from small, medium to large. Also, there are three stacks, named as left, middle and right. At first, the three blocks are stacked in the left stack, with the large one on bottom and small one on top. Users are required to move the three blocks from left stack to the right stack, with the large block on bottom and small one on top by moving one block at one time. In each step, users can only place a smaller block on the top of a larger block, and also can only pick the block on the top of a stack. For instance, users can put a small or medium block on a large block. However, they can't put a middle or large block on a small block.

B. Construction of gesture lexicons.

To finish the task, the users can use 14 commands that are divided into two groups. One group is low level commands, including “Go Left”, “Go Right”, “Go Up”, “Go Down”, “Go Forward”, “Go Backward”, “Gripper Open” and “Gripper Close”. The other group is high level commands, including “Pick from the left stack”, “Drop to the left stack”, “Pick from the middle stack”, “Drop to the middle stack”, “Pick from the right stack” and “Drop to the right stack”.

To build gesture lexicons for these 14 commands, 101 gestures were recorded, including hand gestures, arm movements, feet movements, leg movements, body instances, head movements, and shoulder movements. These 101 gestures were displayed as short videos in a webpage and users can play them individually. 30 subjects were recruited to build their own gesture lexicons by selecting one gesture for one command, from these 101 gestures. Therefore, we obtained 30 gesture lexicons for this task.

From the 30 gesture lexicons, we selected eight of them to compare with. The popularity of each lexicon was computed, considering the popularity for each assignment between gestures and commands, using equation 1.

$$\Phi(L) = \log \prod_{i=1}^{14} \rho(i, j) \tag{1}$$

where i is the index for command, and j is the index for selected gesture. $\rho(i,j)$ is the normalized number of selected assignments from gesture j to command i. The histogram of popularity for these lexicons are presented in Figure 9.

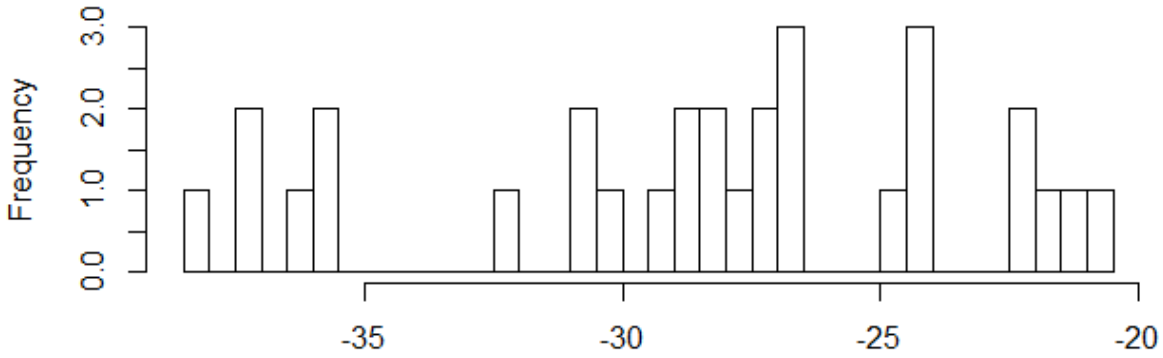


Figure 16. Histogram of popularity

It can be observed from the histogram, that there are approximately five divisions for popularity, including the divisions with high, middle and low popularity, and the divisions between high and middle popularity, and between middle and low popularity. We selected 2 lexicons from each of the divisions with high, middle and low popularity, and 1 lexicons from each of the rest two divisions. Thus, in total, 8 lexicons were selected, including the lexicon ranked number 1 (Φ =-*20.55*), number 5 (Φ =-*22.20*), number 8 (Φ =-*24.37*), number 12 (Φ =-*26.65*), number 18 (Φ =-*28.66*), number 24 (Φ =-*32.33*), number 26 (Φ =-*35.77*) and number 30 (Φ =-*38.08*). These 8 gesture lexicons will be evaluated by participants using these lexicons to complete the Tower of Hanoi task.

Discussion: Among the 30 lexicons obtained from 30 participants, we found nobody have built exactly the same lexicon. But there is overlapping between different lexicons, especially for low level commands. For instance, most participants selected “Open Hand” for the command “Gripper Open”, and “Hand to Fist” for the command “Gripper Close”. For low level commands, gestures

are more intuitive and can be mapped directly using hands, arms or feet. However, for high level commands, people made a variety of selections.

Accomplishments: In this task, we obtained 30 gesture lexicons from 30 participants and made a selection of 8 using popularity measurement. The experiments comparing these 8 lexicons haven't been finished yet and unfortunately had to be interrupted due to program manager (Dr. Jamie Lawton) rejecting our request for a non-cost extension.

Challenges: We were interested in checking whether this utility functions and our findings will translate well to the CSA Explorer scenario. We encountered a major obstacle in trying to use the CSA explorer at the Rome facility Air Force Research Laboratory, Information Directorate, Rome NY, or alternatively getting a beta version to work from our campus. We were not able to have the people from AFRL help us none of these two options. We informed the program manager (Dr. Jamie Lawton) about this situation several times, and the last time was more than one year ago. Nothing was done to help us address this issue. Furthermore, due to this problem, we had to come-up with our own implementation of the CSA Explorer, which delayed the performance of the project by several months. In spite of that situation, we were not granted a non-cost extension.

References

1. Cobos, S., Ferre, M., Uran, S., Ortego, J., & Pena, C. (2008, September). Efficient human hand kinematics for manipulation tasks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008. IROS 2008*. (pp. 2246-2251). IEEE.
2. De Boor, C. (1972). On calculating with B-splines. *Journal of Approximation Theory*, 6(1), 50-62.
3. Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. Applied Linear Statistical Models, vol. Fifth. 1996, p. 1408.
4. Plagenhoef, S., Evans, F. G., & Abdelnour, T. (1983). Anatomical data for analyzing human motion. *Research quarterly for exercise and sport*, 54(2), 169-178.
5. Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. The Annals of statistics, 416-431.
6. Wan, J., Ruan, Q., Li, W., & Deng, S. (2013). One-shot learning gesture recognition from RGB-D data using bag of features. *The Journal of Machine Learning Research*, *14*(1), 2549-2582.
7. Wu, D., Zhu, F., & Shao, L. (2012, June). One shot learning gesture recognition from rgbd images. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012* (pp. 7-12). IEEE.
8. Wachs, J., Stern, H., Edan, Y (2005) Cluster Labeling and Parameter Estimation for the Automated Setup of a Hand-Gesture Recognition System. IEEE Trans. on Systems, Man and Cybernetics. Part A. vol. 35, no. 6, Nov. 2005, pp. 932-944.

1.

1. Report Type

Final Report

Primary Contact E-mail

Contact email if there is a problem with the report.

jpwachs@purdue.edu

Primary Contact Phone Number

Contact phone number if there is a problem with the report

765-4967380

Organization / Institution name

Purdue University

Grant/Contract Title

The full title of the funded effort.

EMBODIED INTERACTION IN HUMAN-MACHINE DECISION MAKING FOR SITUATION AWARENESS ENHANCEMENT SYSTEMS

Grant/Contract Number

AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-13-1-0141

Principal Investigator Name

The full name of the principal investigator on the grant or contract.

Juan Wachs

Program Manager

The AFOSR Program Manager currently assigned to the award

James Lawton

Reporting Period Start Date

05/01/2013

Reporting Period End Date

04/30/2016

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. During Year 2 of the project, we focused our effort in extending the probabilistic model developed in Year 1, that we coined Bayesian Attention Networks (BANs). The objective of the BAN is to help make inferences about the user's focus of attention under uncertainty. The BAN method was extended by enhancing the Node Consensus Model approach. This method consists of building a Bayesian network combining candidate solutions provided by an evolutionary learning approach and candidate networks provided by experts. So far, all the candidate solutions were equally important and had the same effect (weight) in determining the appearance and parameters of the best network during the NCM method. This is not realistic since some experts are more familiar with the task in hand than others (the experts that we refer can be both human-operators and agents (solutions from the evolutionary approach). Therefore, our first contribution to the

project in Year 2 we assigned different importance to each expert based on a gradient descent method so a cost function was minimized. We experimented with cost functions that highlight the advantages provided by the evolutionary approach and the agent based approach (this is the second contribution).
In Year 3 we focused on two areas: (a) developing the equations, methods and algorithms involved in the One Shot Learning approach discussed earlier; (b) extending the inference methods of attention to include secondary task performance; (c) created our own version of the CSE Explorer for cyber-physical navigation.
A different aspect of the project involves characterizing strategies used to solve the TSP based on recognized gestures. Nevertheless, if the gestures were not observed before, how can they be recognized? This was the focused of Year 3. We solve this using a "one shot learning" method that takes a single observation of a gesture and recognizes in the future similar instances of that gesture. As opposed to traditional learning where hundreds/thousands of samples are required to recognize an action, we started developing a framework where a single observation of such gesture will be used to recognize it onward. Our one shot learning approach does not focus on the learning mechanism, as commonly done, but on the process of generating artificial "human like" gesture observations instead. For the observation to be realistic, the process in which they are generated must capture key features that humans use to produce those. In our case, we focus on the conceptual features – what do we remember when we see a gesture and try to mimic it (we refer to these conceptual features as the "Gist of a Gesture"), and bio-mechanical features which deal with the physical and ergonomic limitations that constrain the way that a gesture is produced (this resulted in a conference paper accepted to IEEE ROMAN 2016).
The last contribution deals with the development and validation of benefit and cost functions to test different combinations of modalities and feedback channels and their total cost encumbered on their use. We relied on utility theory to evaluate the trade-off relationships between task performance and users' utility when feedback is provided. We utilized this framework to find the best combination of embodied interaction and feedback modality so the operator performance is maximized.

Distribution Statement

This is block 12 on the SF298 form.

Distribution A - Approved for Public Release

Explanation for Distribution Statement

If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

SF298 Form

Please attach your SF298 form. A blank SF298 can be found here. Please do not password protect or secure the PDF The maximum file size for an SF298 is 50MB.

sf298.pdf

Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF . The maximum file size for the Report Document is 50MB.

Annual Report AY3_JPW.pdf

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.

Archival Publications (published) during reporting period:

1. Y. T Li, J. P Wachs. "A Bayesian Approach to Determine Focus of Attention in Spatial and Time-Sensitive Decision Making Scenarios" in the AAAI'14 Workshop on Cognitive Computing for Augmented Human Intelligence.
2. Y. T Li, J. P Wachs. "Linking Attention to Physical Action in Complex Decision Making Problems" in IEEE International on Systems, Man and Cybernetics, 2014.
3. J. P Wachs. Designing Embodied and Virtual Agents for the Operating Room: Taking a Closer Look at Multimodal Medical Service Robots and Other Cyber-Physical Systems. Speech and Automata in Healthcare Voice-Controlled Medical and Surgical Robots Series: Speech Technology and Text Mining in Medicine and Healthcare. A. Neustein (Ed). De Gruyter, 2014; November 2014; ISBN: 978-1-61451-515-9.
4. M. E Cabrera, J. P Wachs, "Embodied Gesture Learning from One-Shot", In The 25th IEEE International Symposium on Robot and Human Interactive Communication, 2016. RO-MAN 2016. IEEE.

5. T. Zhang, B. S Duerstock, J. P Wachs. "A Computational Framework for Attention Inference Using a Bayesian Approach" presented in IEEE IROS 2015 Workshop, 8th International Symposium On Attention in Cognitive Systems (ISACS 2015)

18. New discoveries, inventions, or patent disclosures:

Do you have any discoveries, inventions, or patent disclosures to report for this period?

Please describe and include any notable dates

Do you plan to pursue a claim for personal or organizational intellectual property?

Changes in research objectives (if any):

None

Change in AFOSR Program Manager, if any:

Yes. The program manager was changed from Jay Myung to James Lawton

Extensions granted or milestones slipped, if any:

No extensions were granted.

AFOSR LRIR Number

LRIR Title

Reporting Period

Laboratory Task Manager

Program Officer

Research Objectives

Technical Summary

Funding Summary by Cost Category (by FY, $K)

	Starting FY	FY+1	FY+2
Salary			
Equipment/Facilities			
Supplies			
Total			

Report Document

Report Document - Text Analysis

Report Document - Text Analysis

Appendix Documents

2. Thank You

E-mail user

May 26, 2016 11:28:57 Success: Email Sent to: jpwachs@purdue.edu