Integrating Perception, Cognition and Action for Digital Human Modeling

Daniel W. Carruth, Mark D. Thomas, Bryan Robbins, and Alex Morais

Center for Advanced Vehicular Systems
Mississippi State University
P.O. Box 5405, Mississippi State, MS 39762
{dwc2,mthomas,bryanr,amorais}@cavs.msstate.edu

Abstract. Computational cognitive models are used to validate psychological theories of cognition, to formally describe how complex tasks are performed, and to predict human performance on novel tasks. Most cognitive models have very limited models of how the body interacts with the environment. The present research examines a simple human-machine interaction task and a simple object manipulation task. A cognitive model is integrated with a human avatar within a virtual environment in order to model both tasks.

Keywords: Digital Human Modeling, Cognitive Models, Perception, Cognition, Action.

1 Introduction

Computational modeling and simulation such as CAD/CAE and FEA have become invaluable tools for manufacturing and design. Radical new designs can be virtually constructed and mathematical models of materials and physics can be used to accurately simulate the performance characteristics of these new designs without expending large amounts of time or money for prototyping and testing. Ergonomic tools exist that allow simulation of the human component for analysis. However, these tools currently require significant input from the user and can often generate unlikely results. For each alternative design for a single task, a designer must specify the postures (manually or through simulation) for each step of the task. If the task could be specified more generally and the digital human model could execute the task on multiple designs, analysis of designs could proceed much more quickly.

Cognitive modeling may provide a mechanism that would allow a generic specification of how to perform a task. In addition, cognitive modeling provides an analysis of the role of perceptual, cognitive, and motor factors in the usability of physical designs. The current research integrates a cognitive architecture with a simple, human avatar in a virtual environment and applies it to two basic tasks: human-machine interaction and object manipulation.

1.1 ACT-R Cognitive Architecture

ACT-R is a computational architecture that implements a formal specification of psychological theories of cognition, perception, and action [1]. Within the

V.G. Duffy (Ed.): Digital Human Modeling, HCII 2007, LNCS 4561, pp. 333–342, 2007.

architecture, individual models of human performance can be constructed to describe or to predict human behavior on many different tasks. ACT-R has been successfully used to investigate psychological phenomenon, to study human-computer interaction, and to develop intelligent tutoring systems.

The ACT-R cognitive architecture is a modular system with a production system as its core. The construction of a model within the architecture involves the specification of condition-action rules that represent the modeler's hypothesis of how the task is performed. The execution of the model involves matching the condition-action rules against the current state of the architecture's modules. Matching rules are then compared based on a utility function and the matching rule with the highest utility is executed updating the other modules. The functions of the other modules include goal-tracking, declarative memory, vision, audition, vocalization, and action. The architecture can be extended by modifying existing modules or creating entirely new modules.

2 Extending ACT-R for Upper-Body Task Modeling

In order to apply ACT-R to real-world tasks such as product assembly, vehicle maintenance, or human-machine interaction, the standard implementation of the cognitive architecture must be extended. The extensions in this research are limited to the perception and motor modules. The perception and motor modules contained in the standard implementation of ACT-R are largely based on the EPIC architecture developed by Meyer and Kieras [4].

2.1 Environment

The standard implementation of ACT-R supports interaction with user interfaces such as command line or GUI systems. This interaction is accomplished by generating information suitable for ACT-R's perception modules from the environment and translating ACT-R's action module's output into messages to the software environment. Since our goal is to model real-world tasks, our environment is a dynamic, virtual environment created in commercial off-the-shelf software. ACT-R is embodied by an animated human avatar within the virtual environment. Every 17 ms, the virtual environment analyzes the scene and generates a table of all of the visual and auditory features that the human avatar may be able to perceive. These tables are submitted to the appropriate modules in ACT-R where they are processed in the same way as the standard implementation processes information from user interfaces. Messages generated from our motor module are sent to the environment which activates stored animations, inverse kinematic simulations, and/or advanced models of human posture prediction.

2.2 Vision

The ACT-R vision module implements a feature-based theory of visual perception. The visual array is made up of features (i.e. color, size, shape, etc.) at particular locations in the array. In order to perceive an object in the environment, visual

attention must be shifted to a location in the visual array and the features at that location must be integrated into a representation of the object at that location.

Visual attention is shifted to locations in one of two ways: a voluntary shift to a known location or an involuntary shift to a significant change in the features of the visual array. In ACT-R's vision module, involuntary shifts are made when the model is not currently attending to a visual object and a significant change occurs in the visual array (i.e. a new feature appearing). Voluntary shifts generally require two steps. In the first step, a model requests a search for a specified set of features be performed on the visual array. This very efficient search makes available a location that matches the specified features. Then, the model can shift attention to the location to encode the location's features into a visual object. Encoded visual objects are then available to the production system for processing.

Extensions. The standard implementation of ACT-R does not specifically handle motion as a feature. The vision module was extended to better support dynamic, virtual environments with many moving objects. These extensions impacted three aspects of the vision module: search of the visual array, involuntary shifts of attention, and encoding of movement information. In order to support motion in visual search, motion magnitude and motion direction features can be used to specify a search of the visual array. A model can search the visual array for a fast object moving amongst slow objects by specifying a search for motion magnitude greater than the speed of the slow objects. Alternatively, a model can search for an object moving left amongst objects moving right. If a motion feature changes significantly, then visual attention may be involuntarily drawn to the location of the change. For example, if an object suddenly starts moving, visual attention will involuntarily shift towards it. Motion magnitude and motion direction is then encoded and made available to the production system as part of the visual object.

Spatial Information. In addition to motion, spatial information is needed for interactions with the virtual environment. The exact nature of human spatial encoding, representation, and manipulation is a matter of some debate [6] [7]. In our current model of spatial representation, egocentric spatial relationships can be encoded and stored in declarative memory based on the visual objects encoded by the vision module. An egocentric spatial relationship encodes the observer's bearing to and distance from an object in the visual array. Object-to-object relationships can be encoded based on two egocentric spatial relationships. These relationships can encode a view-independent, allocentric representation of the environment. The current models generate, store, and use both types of representations. Other spatial extensions to ACT-R are currently being investigated by researchers [2] [3].

2.3 Motor

The motor module of ACT-R models the initiation and programming of motor actions. The module also estimates the time required to execute motor actions. The standard ACT-R module contains movement styles primarily suited for human-computer interactions including keyboard presses and mouse movements. Our

extensions to the motor module include the addition of movement styles that drive animations for a human avatar within our virtual environment. The movements that our current model supports are walking, pushing buttons, reaching for and manipulating objects.

3 Empirical Data

In order to validate our extensions to the ACT-R vision and action modules and to test our integration with the virtual environment, human performance data was collected for multiple tasks including a human-machine interaction task and an object manipulation task.

3.1 Participants

Data was collected for 12 male participants (Mean age = 19.08 years) recruited from the Mississippi State University student population. Participants were screened for major health and vision problems. Two participants appeared to have below average visual acuity but were allowed to participate in the study.

Participants were asked to wear a Mobile Eye portable eye tracker manufactured by ASL. The Mobile Eye is a lightweight (76 grams) eye tracker that records (30 Hz) to a digital tape recorder worn on the waist. Computer software analyzes the recorded eye and scene videos for analysis. Three participants were dropped from data analysis because of technical difficulties with the eye tracking equipment. The whole-body motion of participants was also captured using an Animazoo Gypsy suit. The whole-body data is not currently relevant to the modeling effort.

3.2 The Vending Machine Task (Human-Machine Interaction)

Many tasks of interest to digital human modelers are industrial tasks that require some interaction between a human operator and a machine. In order to investigate a simple and common interaction between humans and machines, participants were asked to purchase a 20 oz. bottled drink from a vending machine. Each participant was given 10 dimes and directed to the location of the vending machine. At the vending machine, participants deposited the coins and selected a drink of their choice. The participants point of gaze and motions were recorded during the task.

As participants approach the vending machine, their focus is primarily on the coin slot. Similar to results reported by Pelz and Canosa [5], participants appear to look ahead at locations that will be necessary later in the operation of the machine. For example, during approach, participants would glance to the receptacle at the bottom of the machine where they would ultimately retrieve their drink. Also, while reaching from the coin slot to their hand for another dime, participants would often browse the button labels presumably to consider their drink selection. Throughout the task, eye movements (see Figure 1) primarily focused on the coin slot. Participants also verified their deposits by checking the display.

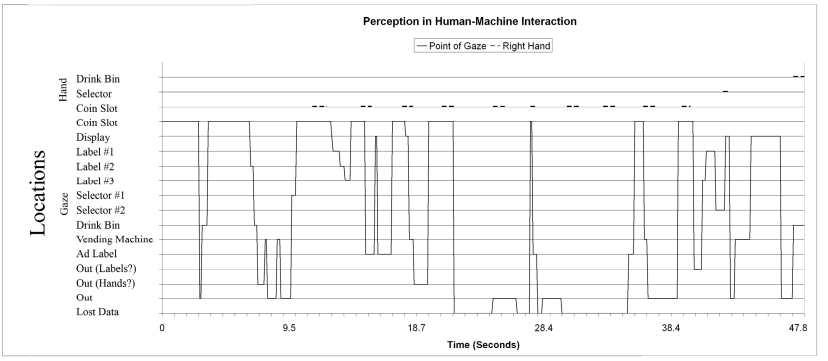

Fig. 1. Example eye and right hand movements

Participants had almost no errors during their interaction with the vending machine. One participant dropped a dime while depositing a coin in the coin slot. During one participant's trial, the vending machine did not accept one of the dimes. The participant used the coin return and successfully completed the task on the second attempt. In the human-machine interaction task, there are two types of errors: human errors and machine errors. In this task, human error could be a physical error (i.e. dropping a dime), a decision making error (i.e. choosing the wrong drink), or a visual encoding error (i.e. not noticing that a drink was available). The human operator also has to respond to machine errors including deposit errors (i.e. a coin getting stuck), out of stock errors, and unresponsive buttons.

3.3 The Block Modeling Task (Object Manipulation)

Many real-world tasks also involve the assembly and/or disassembly of parts. Participants were shown three models constructed from 9 of 14 children's blocks. Figure 2 shows the three models (M1, M2, and M3) and lists the 14 blocks available for constructing the models. Participants were asked to construct a copy of the example models by selecting the correct blocks from the 14 available blocks and placing them in the correct positions.

All participants performed the task quickly and successfully with few errors. Initially, participants assess the overall shape of the example model. Following this first assessment of the model, participants appear to employ one of two strategies. In the single-block strategy, participants select a single piece from the example model, search for a matching piece amongst the available blocks, and then orient and place the block in the appropriate position relative to the already placed blocks.

In the pattern strategy, participants appeared to identify patterns of blocks that make up the example model. The participants would then find all of the blocks making up the pattern and construct the pattern without referencing the example until the pattern was completed. For example, in *M1*, the top three blocks of the model (2 small blocks and the large triangle) could be encoded together and constructed at the beginning of the construction attempt. Once the pattern is completed, it might be set

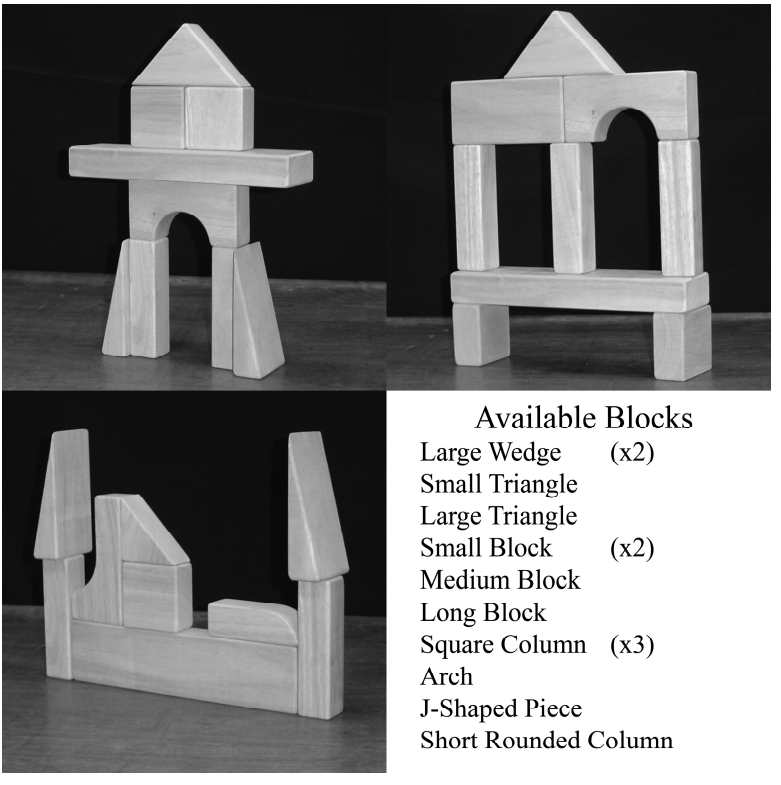

Fig. 2. Three models (*M1* in the top left, *M2* in the top right, and *M3* in the bottom left) constructed from children's blocks. Each model consists of 9 blocks drawn from a total of 14 blocks (listed in the bottom right).

aside until other construction was completed or placed immediately. When placed, participants were observed moving the pattern as a whole (i.e. lifting the three pieces at once) and disassembling and reassembling the pattern in parts.

Typical errors included orientation errors (i.e. in *M3* attempting to place the wedges facing in rather than facing out) or size errors (i.e. in *M1* or *M2* attempting to use the small triangle rather than the large triangle).

4 Modeling the Empirical Data

Based on the results of the study of human participants, a model of task performance was constructed for the two tasks within the ACT-R architecture using our extensions.

4.1 The Vending Machine Task (Human-Machine Interaction)

The vending machine task requires participants to perform a common interaction with a machine. In order to model this interaction, a virtual environment must be

constructed with the appropriate objects and a cognitive model must be developed within the selected architecture.

The Environment. The virtual environment was developed within Virtools, commercial off-the-shelf software designed for rapid prototyping of dynamic, 3D environments. Models of the objects relevant to the task were initially created in CATIA then imported into the Virtools environment. The Virtools environment communicates with the ACT-R cognitive architecture via a TCP/IP network connection.

Due to the level of interaction, the environment model of the vending machine requires considerable detail. Each of the components that a participant may interact with was modeled as a distinct object making up the whole of the machine. Each of the modeled objects is tagged with the symbolic information that can be visually encoded. For example, the coin slot, selector, and vending machine body are tagged with their VALUE slot equal to "SLOT", "BUTTON", and "MACHINE." The labels next to the selectors are tagged with their VALUE slot equal to the drink name represented by the label (i.e. "Water," "Lemonade," etc.).

The Model. Once the environment is modeled and imported into the virtual environment, a cognitive model for the specific task can be created. It should be possible to create a single model that would perform both of the current tasks and many other tasks but, for the current research, separate models have been created for simplicity.

Based on the human performance data, four basic tasks were identified: deposit coins, determine a desired drink, select the desired drink, and retrieve the drink. The first two tasks are somewhat independent and are often interleaved during actual human performance. Selecting the desired drink cannot be done until the coins have all been deposited and the desired drink has been identified. Likewise, the drink cannot be retrieved from the drink bin until a selection is made.

4.2 The Model Assembly Task (Object Manipulation)

The object manipulation task requires participants to encode the shape, size, orientation, and spatial relationship of the blocks that make up the model. Participants must then use the stored information to construct a matching model from available parts. In order to model this task, the virtual environment must contain all of the parts that make up the block models.

The Environment. The object manipulation task required 15 virtual object prototypes: the 14 blocks used in the models and the workbench. All of the object prototypes were constructed using CATIA and imported into the Virtools environment. Two instances of the 14 blocks were created in the environment: the blocks available to the digital human (*source pile*) and the blocks used in the example model.

As in the vending machine environment, each object was tagged with symbolic information that could be encoded visually. In this environment, there are two classes of objects. The workbench is encoded using the basic visual-object tag and only basic

information is available to the model. This includes location, size (in the visual array), distance, color, motion, etc. The blocks all share a block tag derived from the visual-object tag with additional parameters including: shape, size, and orientation.

When the model attends to one of the blocks, the block's shape, size, and orientation are encoded. For some parts of the model assembly, the model must encode the stability of the block. The model accomplishes this by monitoring the motion values of the basic visual-object tag. If a support block or a top block is moving, the model can perceive the possible instability and attempt to adjust the blocks. This is important for ensuring that bottom blocks successfully support top blocks. The model also encodes spatial relationships between the blocks. This encoding is critical for determining the correct position to place blocks.

The Model. The model for object manipulation is more complicated than the model for human-machine interaction. The object manipulation task includes encoding of block information and spatial relationships between blocks as well as additional movement styles to orient and place the blocks. However, participants tend to perform the block task in a clear, serial fashion without the interleaving of two sub-tasks as seen in the vending machine task.

Based on human performance data, 6 components of the task were identified: model assessment, target selection, search for a block, orientation of a block, placement of a block, and target re-assessment. The model begins the task by finding the example and examining its makeup.

Participants appeared to use one of two strategies for constructing a copy of the example. Some participants worked from the bottom up and selected a single target block. Other participants appeared to identify patterns, then select blocks to build the pattern. Once a pattern was built, participants moved on to the next piece or pattern. In the initial assessment, the actor may identify pattern components or single components. After the assessment, the actor selects one of the identified components as the target. If a pattern is available, the actor may select the pattern even if it is not the base. If a pattern is constructed, the construction of the pattern is completed before moving to another part of the example model. For example, if the 2-blocks-1-triangle pattern at the top of M1 is selected, the actor will construct the pattern without referring back to the example model. Once the pattern is completed, it is treated as a single object to be added to the rest of the actor's construction. If no pattern was identified, the model will select a single base object.

The visual features (i.e. shape, size, etc.) of the target block are encoded to guide the actor's search of the source pile. The actor shifts attention to the source pile and searches for a block matching its internal representation of the target block. As attention shifts to possible matches in the source pile, the features of the currently attended block are compared to the stored features of the target block. Once a target block was found, participants would sometimes refer back to the target block. Multiple interpretations of this reference are possible. Participants could be verifying their selection. They could have failed to encode orientation or position when the target was selected. This target re-assessment can occur after almost every step in the construction process when the actor needs more information. In the pattern strategy,

participants may perform this re-assessment less as they may build an entire pattern before looking back to the main model.

Placing a block requires orienting the model correctly (and potentially re-assessing the target) and placing the model correctly (and potentially re-assessing the target). During placement, features such as motion are visually monitored to ensure the stability of the construction.

The model continues in this pattern until the example model has been successfully copied.

5 Discussion

The current research applies an initial integration of cognitive modeling and digital human modeling to two basic tasks. While the tasks currently being modeled are relatively simple, the fundamental cognitive processes may be applied to more complex tasks. For example, the perception and motor components of the operation of complex machinery are not significantly different from those used to accomplish the vending machine task. The operator encodes the features of the machine including display information and operators (buttons/levers/etc.). The operator performs a task based on if-then rules stored in procedural memory. The operator executes motions to act on the machinery. In the assembly task, the cognitive and perceptual components are similar. The motor capabilities in assembly tasks are particularly important and the use of a more effective digital human model than a simple animated avatar is required.

By extending ACT-R's sensory and action systems and integrating the architecture with a virtual environment, the possibilities for cognitive modeling may be greatly expanded. Future extensions may be made to include advanced computational models of human posture and motion. Once validated, these integrated digital human models will be powerful tools allowing designers to investigate both the cognitive and the ergonomic aspects of workspaces and products at the earliest stages of design.

Acknowledgments. This research was performed at the Center for Advanced Vehicular Systems at Mississippi State University.

References

1. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, D., Lebiere, C., Qin, Y.: An Integrated Theory of the Mind. Psychological Review 111, 1036–1060 (2004)
2. Harrison, A.M., Schunn, C.D.: ACT-R/S: Look Ma, no cognitive map! In: Proceedings of the Fifth International Conference on Cognitive Modeling, Bamberg Germany, 2003, pp. 129–134 (2003)
3. Johnson, T.R., Wang, H., Zhang, J.: An ACT-R model of human object-location memory. In: Proceedings of the 25th Annual Meeting of the Cognitive Science Society, Boston, MA, 2003, p. 1361 (2003)

4. Meyer, D.E., Kieras, D.E.: A Computational Theory of Executive Control Processes and Human Multiple-Task Performance: Part 1. Basic Mechanisms. Psychological Review 104, 3–65 (1997)
5. Pelz, J.B., Canosa, R.: Oculomotor Behavior and Perceptual Strategies. Vision Research 41, 3587–3596 (2001)
6. Shelton, A.L., McNamara, T.P.: Systems of spatial reference in human memory. Cognitive Psychology 43, 274–301 (2001)
7. Tversky, B.: Structures of mental spaces: How people think about space. Environment and Behavior 35, 66–80 (2003)