2005

上海论坛文集

Collection of the Papers of Shanghai Forum 2005

IT卷

Economic Globalization and the Choice of Asia

复旦大学上海论坛组织委员会 编

复旦大学出版社

2005

上海论坛文集

Collection of the Papers of Shanghai Forum 2005

IT卷

Economic Globalization and the Choice of Asia

复旦大学上海论坛组织委员会 编

复旦大学出版社

编 者 说 明

一、《2005上海论坛文集》选入了2005年5月15日至17日召开的上海论坛大部分演讲稿和论文稿。这次论坛的部分演讲稿已经收入此前出版的《经济全球化与亚洲的选择》(复旦大学亚洲研究集刊第二辑，复旦大学出版社2006年1月版)，本文集不再重复收入。还有部分文稿或属于发言提纲，或属于内部讨论交流，虽然对于论坛的成功举办起了很大作用，也未收入本文集。

二、本文集根据各文稿的专业内容，分别编入能源卷、IT卷和金融卷。个别文稿的专业有交叉，我们酌情编入相应专业卷，可能与论坛交流时的专业划分不尽相同。

三、本文集各卷所收入文稿的先后排序，大体按照从宏观到微观、从全局到局部，中英文混排。由于所有文稿的具体题目均为作者自定，或专业性很强，或涉及多个方面，难以进行确定的分类，所以排序先后只是相对的，更不意味着水平的高下之分。

四、所有文稿均按照论坛举办时作者的来稿排印，部分文稿在论坛结束后经过作者的修订。我们对明显的文字和标点符号差错予以了订正。文稿作者身份介绍均按照来稿时的说明。

五、我们对所有文稿作者付出的学术努力表示由衷的敬意和感谢，并希望各位读者朋友提出宝贵的意见。

《2005上海论坛文集》编委会

2006年5月

目　录

Content

Cognitive Interfaces

Dana Ballard, Weilie Yi and Chen Yu*

Abstract: One of the principle difficulties in interacting with computers is the difficulty of developing programs. What ultimately we would like is to have computers have human-like interfaces so that we could ask them questions and direct them to do things using everyday language. Although we have made many significant steps towards developing human computer interfaces, we still have a long way to go before achieving the goal of anthropomorphic avatars. Nonetheless in the future, human-like avatars will have an everyday understanding of the world around us, have human-like graphic forms, and be capable of communicating with affect.

Achieving the goal of humanoid interfaces will require significant progress in understanding the fundamental structure of human cognition. The human cognitive architecture is nothing like conventional silicon computers. To compensate for neuronal circuitry that operates over millisecond time scales, the brain resorts to table -look-up, storing vast amounts of highly coded pre-computed responses. Such responses are continually updated with reinforcement mechanisms. Such an organization can loosely be compared to Google. That search engine spends vast amounts of effort souring the web and organizing its findings in way that allows fast answers to questions. In the same way that brain spends its everyday existence organizing its experience to achieve fast responses in the course of natural behaviors.

Perhaps the principal feature that the human brain uses in encoding information is its embodiment. The brain does not compute in isolation but is situated at the top of the extensive sensory-motor system that, it turn, is situated in the world. The world and body can each be seen as doing extensive computation on demand, leaving the brain with much less to do as a consequence. Although the importance of embodiment was first articulated in the 1940s, it is only recently that the tools have become available to fully appreciate its importance in cognitive function. We illustrate some of the rapid progress in this area with three examples from everyday activities using both real human subjects and model avatars.

How is flexible, goal-driven behavior organized? How do we respond to interrupts in the course of a task? What advantages do structured environments offer in answering these questions? We may be able to start to answer such questions owing to recent technological developments. One of the important advances that has allowed us to study this human cognitive organization is that of new instrumentation designed to track body state. Besides audio recording ability we now have lightweight eye trackers and

* Dana Ballard, Professor of Computer Science Department, University of Rochester; Weilie Yi, PhD. of Computer Science Department, University of Rochester; Chen Yu, graduate student of Computer Science Department, University of Rochester.

body position sensors. These allow us to record the allocation of the body's resources in the course of everyday tasks. We illustrate this capability with such observations by taken in the course of food preparation. Such recordings reveal the ways in which the brain organizes task components down to hundreds of milliseconds.

Experiments with humans can reveal components of intelligent behavior but to go further we will need models with extensive predictive behavior. We need to achieve this goal for two reasons. In the first place we need a unified theory of human cognition for its own sake. In the second place we will need substantial components of such a theory to achieve the goal of intelligent avatars. There have been huge advances in simulation that allow one to simulate the physical extent of humanoid graphical figures. Such figures can provide a platform for studying the essential links between brain and behavior by providing complicated environments with which to carry out complex natural tasks. These platforms can contain the ability to model physical collisions at a useful level of detail so that the consequences of manipulation in the world can be explored.

An essential feature of social communication is language. Recently researchers have come to recognize that the crucial problem of language is grounding. How are the brain's symbols tied to structure in the world? Evidence is accruing that everyday tasks serve as a grounding for human language. Language learning has been initially modeled as a statistical encoding problem but new evidence suggests that the body plays an essential role. Such evidence shows that the body's movements during the course of a task can be seen as a language-body language-that may provide a necessary scaffold for the development of spoken language. Studies using Mandarin as a target language show that adult human learners are far more effective if they have access to the intentions of a teacher as revealed through eye head and hand movements.

Humanoid platforms allow us to address many central issues. For example, how do humans handle multiple concurrent tasks? Simulations show that there can be extremely complex interplay between the management of the physical resources of the body, predominantly the hand and eyes, and the computational resources of the body as reflected in neural representations of the task. Avatar simulations allow us to address these issues for the first time and generate theories that can be compared with human data.

1. Introduction

The role of the body in shaping intelligent behavior is vastly under-appreciated[9,24]. The body provides the brain with enormous amounts of economical special-purpose computation that may allow the brain's programs to be expressed as simple collections of programs. Extensive evidence exists that such programs may be expressed as Markov decision processes (MDPs). MDPs are an attractive model as they can both capture the uncertainty of the natural world and be programmed by a secondary reward mechanism[34]. Thus at this point one can test the hypothesis that very large sets of reward-based Markov Decision Processes (VLS-MDPs) can account for significant segments of the behaviors found in natural tasks. The principal way this testing can be done is to make use of the extensive humanoid modeling advances that have recently come available. Such software allows several new opportunities.

- Advances in simulation speed allow the rapid prototyping of cognitive architectures. Such architectures must learn to be predictive and learning algorithms require extensive iterations that simulate human development.
- The simulation of real-world physical constraints can achieve useful levels of realism. While no software can capture all the complexity of the natural world, recent systems are sufficiently complex so as to be able to approximate the behavior in the real world at useful levels.
- It is now possible to simulate natural computation provided by the body. Peripheral gaze control and muscle systems drastically reduce the bandwidth of the brain's more central systems that need to control and analyze the interaction between humans and their environment.

No software system can perfectly simulate all the complexities of the world experienced by a standard robot. However robots have to operate in real-time and therefore suffer a tremendous disadvantage in the development of on-line learning algorithms that adapt based on a lifetime of experience. In contrast, integrated software systems that model physical collisions have reached the point where they can experience both the advantages of real-world complexities and have the speed of software simulation. We have spent considerable effort integrating a software platform that allows a humanoid figure to perambulate in virtual environments. Our system has realistic models of human visual system and collision detection that allows it to have realistic interactions with other bodies.

2. Background

A long time research program goal has been set up to develop a functional, embodied cognitive architecture. This is a complete architecture at a given level of abstraction that is capable of directing human like behaviors in real time. However, so far this research goal has not been achieved. Extensive research in psychology has focused on components of intelligent behavior, but these have been studied in comparative isolation, with the result that it is not possible at this time to combine the compartmentalized knowledge gleaned into a working system description. Two examples illustrate this state of affairs. We know a lot about working memory but have not been able to use this concept in a complete system that can account for its role in ongoing behaviors. Extensive research has been done on attention, but the current notion of attention per se as a general resource is still too simplistic to capture the complexity of even simple behaviors in natual environments.

There have been several system-building efforts, but they have all had their drawbacks. Newell[25] embarked upon an ambitious attempt to model human problem solving in a production-system composed of a large library of pattern triggered rules. A similar line of research was initiated by Anderson[1]. Both these systems are faithful to the time constants observed in problem solving, but the level of abstraction chosen does not lend themselves to modeling the details of visuo-motor processing, or the many issues associated with embodiment.

Another compositional approach is that of symbolic planning. Its principal assumption is that the agent has a central repository of symbolic knowledge. In this paradigm, the purpose of perception is to translate sensory information into symbolic form. Actions are selected that result in symbolic transformations that bring the agent closer to goal states. The "sense-plan-act" approach is typified in the robotics community by early work on Shakey the robot[26], and in the cognitive science community by the theories of David Marr[23]. In principle, the symbolic planning approach is very attractive, since it suggests that sensation, cognition and action can be studied independently, but in practice each step of the process turns out to be difficult to characterize in isolation.

The difficulties with the symbolic planning approach led to a focus on systems composed of behavioral primitives such as those championed by Brooks[6]. His alternate approach is to describe whole visuo-motor behaviors that each have very specific goals. Behavior-based control involves a different approach to composition than planning-based architectures: the goal is to combine and sequence simple micro-behaviors to solve arbitrarily complex problems. Brooks' own *subsumption* architecture worked by organizing behaviors into fixed hierarchies, where higher level behaviors influenced lower level behaviors by over-writing their inputs. Subsumption works well for trophic, low-level tasks, but generally fails to scale to handle more complex problems[13]. The upshot is that the best approach to attaining behavioral composition is an active area of research[10,7].

Brooks also has stressed the need to work with real robots, but real robots have the disadvantage that the basic learning algorithms underlying intelligence have developmental roots and require sufficient computation to simulate liftimes. Partly for this reason, the use of graphic humanoid figures in modeling was started in several laboratories[20]. Most recently Terzoupolos has made several important contributions in specifying architectures for such figures[35,16]. Schaal and Mataric have focused on body primitives in terms of learning movements[17,18] but so far have skirted the crucial issues of cognitive architecture. So far we are the only laboratory focusing on the graphic figures' embodiment exclusively to its relationship to cognition and in particular the use of gaze.

We emphasize gaze as it is a central component of the human cognitive architecture. Yarbus was the first to call attention to the way that seeing is inextricably linked to the observers' cognitive goals. The relationship between eye movements and cognitive processes has been studied extensively since that time, and recently a number of separate lines of inquiry have coalesced to give an increasingly coherent understanding of the elaborate and intricate role of eye movements in cognitive function. Although under some circumstances task processing can be covert, in our experience the visual environment can almost always be arranged so that overt gaze is used. We have argued that the problem of covert processing is simply one of timing. Only in cases where the internal computations that can use gaze are faster than the time to change gaze are overt gaze changes not seen[28].

Three principal complementary advances have had a direct bearing on understanding the role of gaze. The first is the description of the role of eye movements in executing everyday visually guided behaviors[27,21,14,19]. The major findings of these studies have been the ubiquitous role of the task in such movements, and the importance of learning where and when to fixate. The second advance has been the recognition of the role of internal reward in guiding eye and body movements[16-20]. This has been revealed especially in neurophysiological studies. The third important advance has been the theoretical developments in the field of reinforcement learning, together with tremendous advances in graphic simulation[21-23]. Together, these developments have allowed the simulation of reward-based systems that incorporate realistic models of eye movements over extended time scales.

Our work differs from other computer graphic approaches in that 1) we specifically build reward-based models that predict human performance and 2) we test these predictions with real human subjects in real world and virtual reality environments. Our laboratory environment incorporates several recent simulation advances that we have integrated into a single system.

3. Laboratory

Our laboratory consists of one Silicon Graphics Onyx Reality Engine and three Dell PCs as well as one Apple G5. The Dell PCs have rendering cards that enable them to exceed in speed the Silicon Graphics machine, but the latter machine has essential software for running a Sensible Phantom dual force feedback device.

In addition the laboratory has two eye trackers, one mounted in one of two V-8 Head-mounted displays. Six dof position measurements are available via a magnetic -based Fastrack system as well as via a Highball six dof tracker that uses ceiling mounted infrared LED emitters.

The centerpiece of our modeling platform consists of four integrated systems 1) A basic environmental model coded either in Performer or Open GL, 2) Boston Dynamics DI-Guy software for humanoid figures that have a repertoire of basic motion commands, 3) The Vortex collision simulator that allows collision detection and force modeling of the dynamic interactions between complex bodies, and 4) virtual reality immersion support software that allow human subjects to participate in ambulatory or hand-eye settings. While no simulation system can be expected to model reality perfectly, our system comes sufficiently close. An earlier version for driving developed entirely in VR was able to detect traffic lights and stop signs in urban settings rendered via a hand held camera on busy streets. In addition, the simulation systems allows us to model sensory and motor aspects that cannot be captured directly with current technology. For example we can model the foveation of the human eye as well as the degrees of freedom and muscle properties of the human system. Current robotic technology cannot come close on the latter owing to an inability to duplicate skeletal and muscle strength and weight properties.

The DI-Guy software package from Boston Dynamics has an extensive set of subroutine calls that direct the motion of the humanoid as well as facial expression and eye gaze. Motion is created by libraries of human motions that are captured and associated with subroutine calls. Special features include interpolation between different motions and scaling the complexity of the figure with distance for rendering economies that allow many figures to be rendered at the same time.

DI-Guy vision and motor control is provided by software that we have created. For simple visual routines, we use Silicon Graphics, but for more expensive processing such as that required by multiscale filters, we use a DataCube FPGA that can compute the necessary convolutions in real-time on modest sized (200 $\times$ 128 pixel) images. The humanoid has binocular imaging capability. All gaze control algorithms have to be programmed. So far we have programmed saccades, vergence and a vestibulo-occular reflex. For learning trials with simple vision algorithms the system can run at ten times real time.

DI-Guy motor control is provided by a standard library but also by a spring muscle model package that we have developed. Our software uses virtual forces generated by the Vortex software package. This is a beautiful program that computes the effects of collisions between objects. Collisions between simple objects can be easily calculated in real-time. To use this package, we wrap Vortex objects around the DI-Guy, such as cylinders for limbs, calculate the collision results and use them to drive the graphic figure.

Humans can be immersed directly into the humanoid graphic environment via a V-8 head mounted binocular display. A very low latency image update (<50 ms) avoids the swimming associated with longer latency displays. The six degrees of freedom of the head coordinate system needed for image rendering are provided by a Hi-Ball which is based on ceiling mounted LED emitters. The environment allows humans to ambulate over a substantial area. Motion over long linear tracks can be made possible by mapping circular body paths onto linear visual paths. Thus to experience a visual straight line subjects have to walk in a circle. This techniques is very effective: subjects walk 4 to 5 labs of the 20 x 20 laboratory without noticing[22]. In addition the VR environment allows for hand-eye tasks to be conducted while sitting at a bench. Our Sensible Phantom force feedback device allows subjects to execute two-fingered grasps on virtual objects over a volume of one-cubic meter.

4. Demonstrations

We have done extensive work in measuring the course of eye, head, and hand movements while subjects are engaged in natural tasks. We have studied human performance in driving, walking and hand-eye tasks in virtual environments, and sandwich making and story book reading in the real world. In these tasks all the body's movements are important but easily, and the most important is that of eye gaze. The moment-by-moment fixations of the eye during the course of behavior are an extraordinarily reliable indicator of ongoing mental processing. These studies have revealed a plethora of data about the human organization of mental processing in natural tasks.

(1) In driving, eye gaze is very sensitive to the driver's task[31]. Subjects fixate stop signs 15% of the time when instructed to follow a car but 45% when instructed to follow a car and obey traffic signs. In addition, when approaching a corner while following, subjects will switch back and forth between the lead car and the intersection. Our interpretation of this result is that the subjects have both the following process and an intersection process active and are using gaze to update each.

(2) In a virtual reality (VR) driving intervention where a no-parking sign is changed to a STOP sign for one second, subjects may subsequently stop, but only if they have fixated the sign during that interval. We interpret this to imply that the information as to the identity of the sign is obtained by a visual routine[38] running at that moment. We have been able to model this data with a virtual car that uses solely

visual routine based vision.

(3) Additional evidence for the visual routines hypothesis comes from a VR block task. Subjects are twice as likely to report changes to a feature of an object when it is part of the task agenda than when it is not, even though they are often looking at the block when it changes. We interpret this to imply that subjects must be actively testing for the color feature to notice its change[37].

(4) In sandwich making in the natural world, the limitations of the body forces subtasks to be done sequentially. Nonetheless during the course of a task, subjects engaged in the hand coordination for one subtask, such as picking up a coke cap, will frequently look ahead to an item, such as a jar lid, that will be used to for the subsequent subtask[15]. We interpret this as evidence for two or more internal processes being simultaneously active.

(5) In our studies of infant language learning, we used a model of an adult learning a second language. By simultaneously recording eye, head and hand movements and phonemes while a reader was reading a story book we were able to recognize nouns and verbs in an unsegmented phoneme stream[8] (This research won the Marr prize at Cognitive Science 2004). To do this we used hidden Markov models (HMMs) to recognize eye fixations and familiar body movements.

4.1 Cognitive Architecture Overview

We have designed a cognitive architecture model based on reinforcement learning for the arbitration of multiple concurrent behaviors that share gaze as the same perceptual resource[33]. This model uses the uncertainty about the current sate estimates together with the impending costs for not updating these estimates to decide, which behavioral program obtains control over vision in order to update the respective state information. This model was implemented for a walking environment, where the goals are to stay on a walkway, avoid obstacles, and pickup litter. The agent learns the optimal strategy of switching between the three behaviors' access to the agent's eyes so as to maximize the expected reward over a trial. The corresponding sate-action mappings reflect the distributions of objects along the path and the respective rewards.

The model of behavior-based control centers around primitives that termed *micro-behaviors*. A microbehavior is a complete sensory/motor routine that incorporates mechanisms for measuring the environment and acting on it to achieve specific goals. For example a collision avoidance micro-behavior would have the exclusive goal of steering the agent to avoid collisions with objects in the environment. A micro-behavior has the property that it cannot be usefully split into smaller subunits. The micro-behavior control architecture follows work on behavior based control (e. g. [10,7,16]) that allows the agent to address changing goals and environmental conditions by dynamically activating a small set of appropriate behaviors. Each micro-behavior is triggered by a template that has a pattern of internal and environmental conditions. The pattern-directed activation of micro-behaviors provides a flexibility not found in other embodied architectures e. g. subsumption.

One of the principal findings of these simulations is that the information from fixations supports behavior in different ways. In this respect is very helpful to think of the visual computations that need to be done as hierarchically organized, a viewpoint that has not been as obvious from the attention/working-memory perspective. Our representative hierarchy that has three levels Behavior, Arbitration and Context. The issues that arise for vision are very different at each of the levels in this hierarchy.

(1) *Behavior.* At the level of individual behaviors, the gaze point enables the computation of state information necessary for meeting behavioral goals. (For example, is there an obstacle in the path?) Almost invariably, the visual computation needed in a task context is vastly simpler than that required by general-purpose vision and, as a consequence, can be done very quickly[38,29]. Psychophysical evidence for such context-specific visual computations within a fixation was described above, and context-specific neural activity in visual cortex provides a neural basis for such state computations.

(2) *Arbitration.* At the arbitration or resource allocation level, the principal issue for eye movement control is that the center of gaze (or attentional capacity) is not easily shared and instead must be allocated sequentially to different locations (For example, different fixation locations are needed for obstacle avoidance and controlling heading direction). Since the set of active behaviors must share perceptual and physical resources, there must be some mechanism to arbitrate when they make conflicting demands. Task arbitration, particularly dual-task performance, has been extensively studied but not from the perspective of gaze allocation. One arbitration model, described below[20], shows how gaze allocations may be selected to minimize the risk of losing the reward associated with a given behavior.

(3) *Context.* The context level maintains an appropriate set of active behaviors from a much larger library of possible behaviors, given the current goals and environmental conditions (For example, when reaching an intersection, a walker must detect the new context and choose from a different set of behaviors, such as stopping at the curb). Thus the function of vision at this level is to achieve an appropriate balance between agenda-driven and environmentally driven visual processing demands. This issue has not received much attention in the cognitive literature but has been extensively studied in robotic models.