Perceptive Agents and Systems in Virtual Reality

[Extended Abstract]

Demetri Terzopoulos
New York University, Media Research Lab
719 Broadway, 12th Floor
New York, NY 10003, USA
www.mrl.nyu.edu/∼dt

1. THE VISION: A REALITY EMULATOR

Imagine a computer-simulated world that approaches the realism and complexity of the real world. Inhabited by lifelike flora and fauna, this virtual world is populated by synthetic humans that look, move, and behave like natural humans. These autonomous agents are endowed with functional bodies and brains supporting motor control and locomotion, computational vision and audition, speech synthesis, natural language recognition, adaptive behavior, as well as cognitive skills, such as learning and reasoning. Such a *Reality Emulator* could be used in revolutionary ways and its impact across multiple scientific disciplines would be profound.

Offering unprecedented predictive power, the Emulator would enable experimentation with complex and/or dangerous scenarios that cannot be attempted safely or repeatedly in real life. For example, it could simulate the reactions of crowds in crisis situations, such as terrorist bombings, in order to assess likely consequences and plan appropriate emergency responses. A realistic simulation of a fire in a virtual building occupied by virtual humans could help assess the visibility and accessibility of exits and plan rescue and evacuation procedures as a function of the location and type of fire, smoke dispersal, ventilation, and so on. A less dramatic scenario would be the realistic simulation of automated CCD surveillance cameras directed in the lobbies and concourses of a busy train station or airport, in order to facilitate the development and testing of a distributed, vision-based human surveillance system capable of detecting suspicious behavior.

Although VR's holy grail may be a real-time, interactive synthetic reality of the sort portrayed in the motion picture *"The Matrix"*, by no means need our Emulator run in real time for it to be of immediate value in the sorts of applications described above. Rather, the fidelity and usefulness of such simulations is ultimately dependent on a detailed modeling of the abilities and limitations of human bodies, perception systems, and mental processes under real-world conditions. A non-real-time Reality Emulator of the sort described above is well within reach if one could deploy a web-based, distributed implementation infrastructure that would elicit the participation of a broad, multidisciplinary community of researchers.

VRST '03, Osaka, Japan

2. PROTOTYPE REALITY EMULATORS FOR COMPUTER VISION RESEARCH

Vision research today is both motivated and hampered by hardware technologies. The typical vision researcher is inspired by modern CCD cameras, pan-tilt ocular heads, frame-rate image processors, mobile robots, manipulators, controllers, etc. The hardware paraphernalia, however, can be expensive to acquire, a headache to interface and reconfigure, and a burden to maintain in good working order. It can also impose spuriously artificial constraints on the development and testing of perception algorithms.

In an effort to liberate a substantial segment of the computer vision research community from the "tyranny of hardware", we have proposed an alternative, software-based research methodology that relies only on commodity computing and graphics hardware [1, 2]. Our approach caters especially well to scientists who are motivated to understand and ultimately reverse engineer the powerful vision systems of humans and other animals.

Consistent with our aforementioned vision of the Reality Emulator, we advocate the use of realistic virtual environments populated by artificial humans and other animals as software laboratories within which researchers can develop and evaluate complex computer vision systems. Fig. 1 illustrates a prototype application of our approach to virtual humans [3], which has resulted in a biomimetic active vision system implemented within "DI-Guy", the commercially available graphical API of dismounted infantry (marketed by Boston Dynamics, Inc., [4] www.bdi.com). The figure shows one DI-Guy agent visually tracking another and autonomously navigating in pursuit. The observer soldier obtains its perceptual information by continually processing the image streams acquired by its foveated virtual eyes. The environmental modeling may be trivial in this prototype reality emulator, but the virtual world is inhabited by multiple autonomous virtual humans whose lifelike mobility presents a nontrivial challenge in designing robust sensorimotor control systems that are capable of identifying and pursuing moving targets.

3. ENABLING TECHNOLOGIES

Our paradigm for computer vision research is made possible by several enabling technologies:

- **Advanced artificial life** modeling of flora and fauna, especially human modeling, including biomechanical, perceptual, ethological, and cognitive component models [5];
- **Virtual reality** modeling of natural phenomena, such as terrain and weather, as well as man-made artifacts, including buildings, streets, bridges, tunnels, etc.;

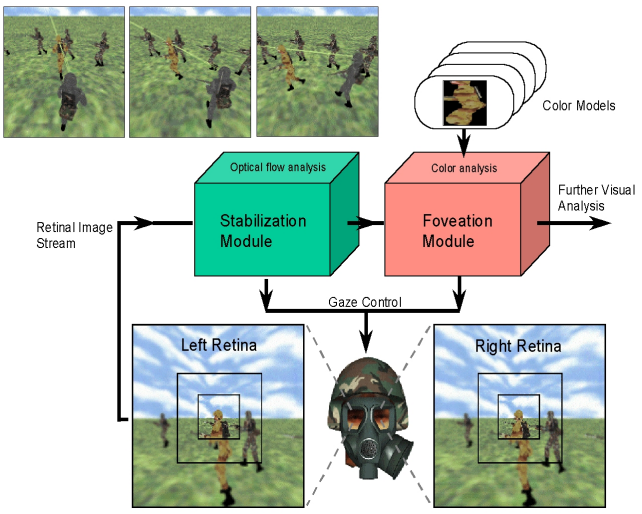

Figure 1: *A prototype, biomimetic, active vision system for virtual humans. The three images at the upper left show a vision-enabled "DI-Guy" virtual soldier visually tracking a fellow DI-Guy and autonomously navigating in pursuit (the added lines indicate the gaze direction of the observer's eyes). The diagram below illustrates the active vision system at work in the observer. The binocular retinal image stream acquired by the observer's foveated eyes is stabilized and the gaze is actively controlled based on a color (mental) model of the moving target.*

- **Photorealistic graphical rendering** using computer graphics software and performance commodity graphics pipelines (e.g., ATI, NVIDEA).

Given their remarkable advances, which have been fueled by the interactive computer game and movie special effects industries, the convergence of these technologies now beckons the computer vision community.

A useful application of the above technologies has been in active vision research. The late psychologist, J.J. Gibson [6] stresses the importance of modeling the active observer situated in the dynamic environment. Versions of this so called active vision paradigm were introduced in mainstream computer vision by Bajcsy [7], Ballard [8], and Aloimonos *et al.* [9], and the approach was further established by many others (see, e.g., [10, 11]). Mobile robots (see, e.g., [12]) have served as the primary testbeds for designing and experimenting with active vision systems. Undeniably, however, efforts to equip real-time mobile robots with general-purpose active vision systems have been hampered by the inadequacies of the hardware and the on-board computers. First, readily available hardware systems are woefully inadequate models of natural animals—animals do not have CCD chips for eyes, electric motors for muscles, and wheels for legs. Recent humanoid/animaloid robots attempt to rectify some of these deficiencies, with partial success.[1] Second, mobile robots typically lack the compute power necessary to achieve real-time response within a fully dynamic world while running active vision algorithms of much sophistication. We believe that virtual animals and humans, such as those depicted in Fig. 1, offer a more convenient and more biomimetic alternative for active vision research.

[1]Honda Motor Co. Ltd. `world.honda.com/ASIMO/`; Sony Co. Ltd. `www.aibo.com`.

4. ADVANTAGES OF THE APPROACH

Our emulation approach leads to rich environments for vision research that may be implemented entirely in software on inexpensive, universally accessible, commodity PCs, thereby circumventing the need for specialized vision hardware. Our approach offers additional crucial advantages:

- **Exact ground truth data.** The quantitative photometric, geometric, and dynamic information that is needed to model and render the virtual world is available explicitly. Generally, the autonomous agents must glean their visual information from egocentrically acquired retinal image streams, but the readily available ground truth data can be extremely useful in assaying the effective accuracy of the vision algorithms or modules under development.
- **Ease of experimentation.** By implementing vision systems in a virtual world, not only does one have exact ground truth data, but one also has the opportunity to run controlled, repeatable experiments. In theory, this will accelerate progress in computer vision research, by eliminating the cumbersome aspects of experimenting in the real world, and permitting many more iterations of the scientific method.
- **Advanced vision systems in dynamic offline scenarios.** The time clock in the virtual world can be retarded arbitrarily relative to real (wall-clock) time, so that each agent can perform an arbitrary amount of computation per unit virtual clock time. This enables us to equip our virtual agents with elaborate collections of vision algorithms that cannot possibly be run in real time on current hardware. Within the virtual world, however, we can evaluate these experimental vision systems "offline" in complex, fully dynamic scenarios.

4.1 The Suitability of Synthetic Imagery

Some vision researchers have questioned the suitability of synthetic imagery. Such criticism boiled over in the 1990s in reaction to the excessive use of simple, synthetic imagery during the 1970s and 1980s; e.g., the prototypical Lambertian sphere used to develop and test shape from shading algorithms. While it is widely acknowledged that the computer graphics industry has made dramatic progress in the realistic simulation and rendering of the real world, vision researchers generally remain skeptical about the suitability of synthetic imagery in computer vision research. For some reason, however, they seem less skeptical of a different form of oversimplification: Developing vision algorithms using real imagery, but of artificially simplistic real scenes: e.g., the so called "blocks worlds" scenarios.

Both types of simplification approximate reality in order to facilitate the development and/or testing of vision algorithms and systems. The criticism of synthetic images still holds, although more so in the context of low-level vision than intermediate or high-level vision, which are more concerned with the application of models, higher level knowledge, and system integration than with imaging. Furthermore, the best (global illumination) renderings of virtual scenes are barely distinguishable from real images, and their algorithmic analysis poses a challenge commensurate with that posed by real images. Furthermore, both real and synthetic images can be made arbitrarily more challenging through corruption with noise, distortions, and other artifacts associated with real-world sensors and associated hardware.

We believe that for the purposes of vision system research and development, the potential benefits of the use of modern, high-quality synthetic images of dynamic virtual worlds far outweigh their disadvantages.

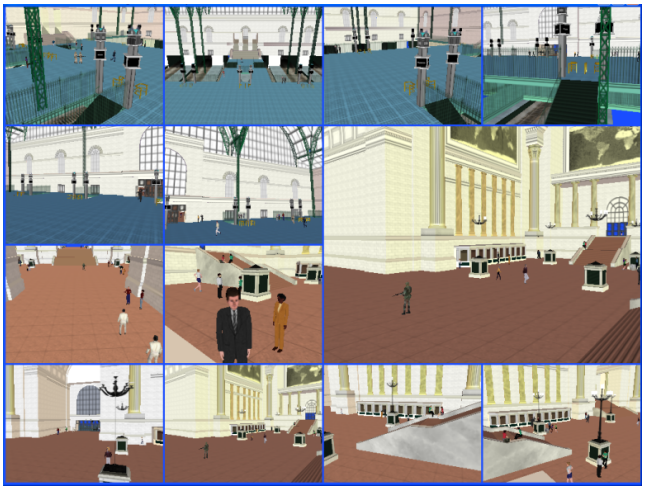

Figure 2: *A prototype, virtual surveillance system. Multiple virtual cameras monitor a virtual train station populated by autonomous DI-Guy pedestrians. The sensors, which include fixed wide-field and steerable telephoto cameras, are controlled by a computer vision system that monitors human activity in the station through automated visual analysis of the incoming video streams acquired by the virtual cameras.*

5. A VIRTUAL SURVEILLANCE SYSTEM

Recent events have redoubled interest in the application of non-obtrusive computer vision technologies to person identification and tracking, with potentially significant military, national security, and commercial implications. We have recently been working within our reality emulator paradigm in the service of such applications.

As Fig. 2 illustrates, we have been developing software that accurately models an array of CCD video surveillance cameras, including both wide field of view static cameras and steerable, variable focal length active cameras. These virtual cameras image the environment within a virtual train station, including virtual pedestrians moving about in that space. Given our prior experience, the DI-Guy software from Boston Dynamics, Inc., has been our API of choice for emulating the appearances and actions of virtual humans. The cameras monitor portals into the train station, the main lobby space of the building, and the various concourses leading to the train platforms.

The most challenging technical aspect of this research has been to combine a suite of active vision algorithms, including appropriate algorithms based on color, motion, and shape analysis, to identify the presence of virtual humans from the incoming stream of video imagery and visually track these autonomous agents as they move around the building, passing control from camera to camera. Color and motion serve as primary visual cues for tracking. A formidable task is to determine if any of the simulated pedestrians may depict terrorists who harbor hostile intentions towards the building and its occupants. A more readily attainable goal would be to determine, through multicamera visual analysis, if simulated pedestrians that are approaching the building are carrying firearms. Suspicious individuals can then be subjected to greater scrutiny, including analysis by face recognition algorithms, using close-up virtual cameras monitoring portals.

6. ACKNOWLEDGMENTS

Tamer Rabie made important contributions to the Animat Vision paradigm. My thanks to Mauricio Plaza-Villegas for his ongoing work on the virtual surveillance system. This research is funded by the Defense Advanced Research Projects Agency (DARPA).

7. REFERENCES

[1] D. Terzopoulos and T.F. Rabie. Animat vision. In *Proc. Fifth International Conference on Computer Vision*, pages 801–808, Cambridge, MA, June 1995. IEEE Computer Society Press.

[2] D. Terzopoulos and T.F. Rabie. Animat vision: Active vision in artificial animals. *Videre: Journal of Computer Vision Research*, 1(1):2–19, September 1997.

[3] T. Rabie and D. Terzopoulos. Active perception in virtual humans. *Vision Interface 2000 (VI 2000)*, Montreal, Canada, May, 2000, 16–22.

[4] J. Koechling, A. Crane, and M. Raibert. Applications of realistic human entities using DI-Guy. In *Proc. of Spring Simulation Interoperability Workshop*, Orlando, Florida, 1998.

[5] Terzopoulos, D. Artificial life for computer graphics. Communications of the ACM 42(8):32–42, 1999.

[6] J. J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, MA, 1979.

[7] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):996–1005, 1988.

[8] D. Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.

[9] Y. Aloimonos, A. Bandyopadhyay, and I. Weiss. Active Vision. *Int. J. Computer Vision*, 1:333–356, 1987.

[10] A. Blake and A. Yuille, editors. *Active Vision*. MIT Press, Cambridge, MA, 1992.

[11] M.J. Swain and M.A. Stricker. Promising directions in active vision. *Inter. J. Computer Vision*, 11(2):109–126, 1993.

[12] R. C. Arkin. *Behavioral Robotics*. MIT Press, 1998.