Computer Vision for Computer Interaction

William T. Freeman
Paul A. Beardsley
MERL, Mitsubishi Electric Research Lab

Hiroshi Kage
Ken-ichi Tanaka
Kazuo Kyuma
Mitsubishi Electric
System LSI Division

Craig D. Weissman
E.piphany Software

Introduction

Figure 1 shows a vision of the future from the 1939 World's Fair. The human-machine interface that was envisioned is wonderful. Both machines are equipped with cameras; the woman interacts with the machine using an intuitive gesture. That degree of naturalness is a goal today for researchers designing human-machine interfaces.

Vision-based Interactive Systems

It might seem that to achieve a natural interaction, an interface based on computer vision would require visual competence near the level of a human being, which is still beyond the state of the art. Fortunately, this is not the case. Interactive applications typically restrict the vision problem that needs to be solved. By clever system design, researchers can create the appearance of high level understanding with a system that is really solving a few low-level vision problems. For example, the television controlled by hand gestures (see Fast and Low-Cost Systems section) performs simply by identifying the location of a generic hand template in the image, without a fuller understanding of the activity of the human subject. A second advantage of vision for interactive applications comes because there is a human in the loop. Given immediate feedback, a user can adjust their motions to achieve the desired effect. The applications described in this paper, to varying degrees, all take advantage of these features of interactive vision applications.

Under the proper imaging conditions, one may only need to acquire binary images, which can be processed very quickly. Krueger showed in an early system that silhouette-based vision was sufficient for simple yet enjoyable games [10], while the San Francisco Exploratorium has long had an exhibit where the silhouette of participants controls a graphical display [13].

Some interactive systems focus just on the face or just on the hands of a subject. The popular 'Magic Morphin' Mirror' combined face detection technology with computer graphic image warpings to comically distort the faces of participants [2].

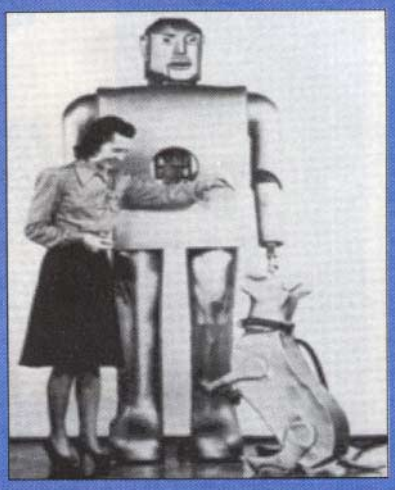

Figure 1: A vision from the past of the future: a natural, gesture-based interface, using camera-based input [4].

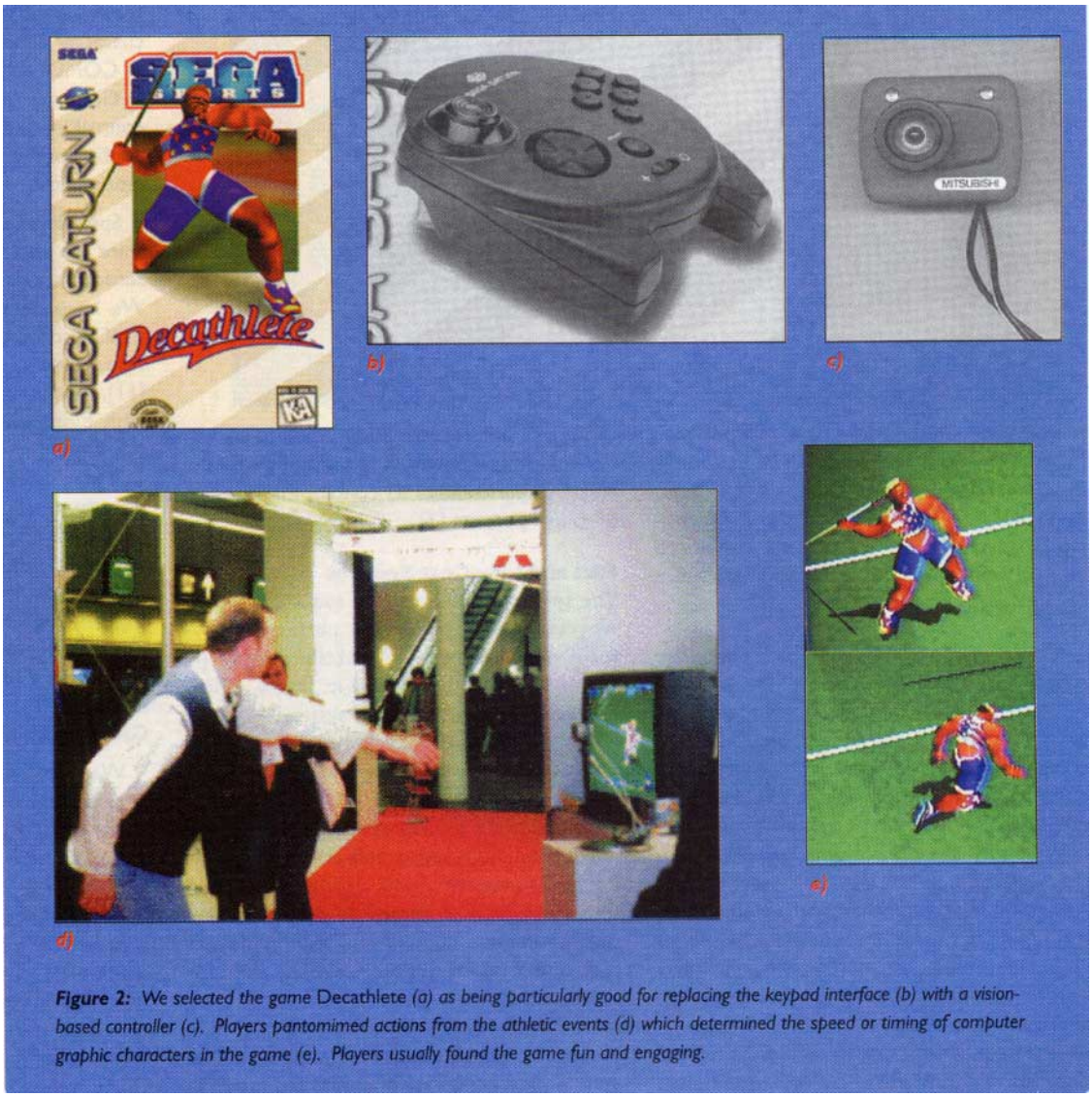

Figure 2: We selected the game Decathlete (a) as being particularly good for replacing the keypad interface (b) with a vision-based controller (c). Players pantomimed actions from the athletic events (d) which determined the speed or timing of computer graphic characters in the game (e). Players usually found the game fun and engaging.

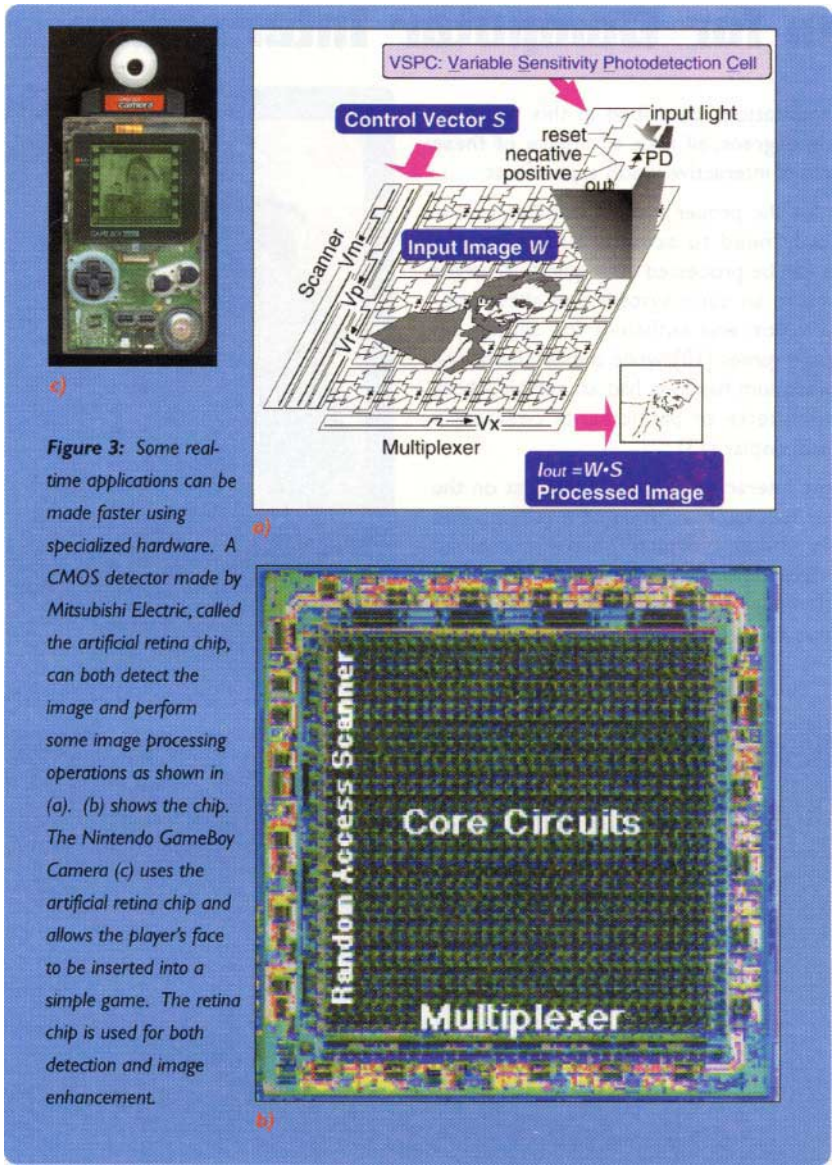

Figure 3: *Some real-time applications can be made faster using specialized hardware. A CMOS detector made by Mitsubishi Electric, called the artificial retina chip, can both detect the image and perform some image processing operations as shown in (a). (b) shows the chip. The Nintendo GameBoy Camera (c) uses the artificial retina chip and allows the player's face to be inserted into a simple game. The retina chip is used for both detection and image enhancement.*

Segen and collaborators have built on work in hand gesture recognition [14] to make interactive games and fly-bys using hand gesture input [15]. Wilson used 3D hand positions derived from color and stereo to control the flapping of a virtual seagull in a flight graphics system [16].

Other interfaces attempt to identify the rough 3D pose and motion of a subject. The MIT Media Lab made the ALIVE interactive environment [2] and successors [17], one component of which used vision to estimate body pose and location. This information was combined with artificial agent methods to create a virtual world of synthetic characters that respond to a person's gestures. The Advanced Telecommunication Research Institute (ATR) in Japan has developed a variety of vision mediated graphical systems, including virtual kabuki and the resynthesis of human motions observed from multiple cameras [7, 8]. A system by Sony observed players making different fighting gestures and translated those into a computer game [9]. A collaboration of several research groups allowed the dance of participants to control the motions of virtual puppets [3].

Fast and Low-Cost Systems

The systems above typically require powerful workstations for real-time performance. A focus of our work at Mitsubishi Electric (in Cambridge, MA, U.S.A. and in Osaka, Japan) has been low-cost, real-time systems. We have built prototypes of vision-controlled computer games and televisions with gesture-based remote control [5].

The existing interfaces for these systems impose daunting speed and cost constraints for any computer vision algorithm designed to replace them. A game pad or a television remote control costs a few tens of dollars and responds in milliseconds. The components of a vision-based interface covering the same functionality as those interfaces include a camera, digitizer and a computer. The system must acquire and analyze the image in little more time than it takes to press a button on a keypad interface. It may seem impossible to design a vision-based system that can compete in cost or speed.

We have made prototypes that address the speed and cost constraints by exploiting the restrictions to the visual interpretations imposed by the interactive applications. For example, at some moment in a computer game, it may be expected that the player is running in place. The task of the vision algorithm may then be simply to determine how *fast* the player is running, assuming they are running, a relatively easy vision problem. Such application constraints allow one to use simple and fast algorithms and inexpensive hardware.

We constructed a vision-based version of the Sega game, *Decathlete*, illustrated in Figure 2. The player pantomimes various events of the decathlon. Knowing which event is being played, simple computations can determine the timing and speed parameters needed to make the graphical character move in a similar way to the pantomiming player. This results in natural control of rather complex character actions. We demonstrated the game at COMDEX '96 in the U.S. and at CeBIT '97 in Germany. Novice users had fun right away, controlling the running, jumping or throwing of the computer character by acting out the motions themselves.

Specialized detection and processing hardware can also reduce costs. Low-cost, CMOS sensors are finding many vision applications. We have designed a low-power, low-cost CMOS sensor with the additional feature of some on-chip, parallel image computations [11], named the Artificial Retina (by analogy with biological retinas which also combine the functions of detection and processing). The chip's computations include edge-detection, filtering, cropping and projection. Some of the computer game applications involve the computation of image moments, which can be calculated particularly quickly using the on-chip image projections [5]. Figure 3 shows a schematic diagram of the artificial retina chip, a photograph of it and a commercial product that uses the chip, the Nintendo GameBoy Camera.

We also made a gesture-based television remote control, again designing the system to make the vision task simple [6]. The only visual task required is the detection and tracking of an open hand, a relatively distinct feature and easy to track. When the television is turned off, a camera scans the room for the appearance of the open hand gesture. When someone makes that gesture, the television set turns on. A hand icon appears in a graphical menu of television controls. The hand on the screen tracks the viewer's hand,

allowing the viewer to use his or her hand like a mouse, adjusting the television set controls of the graphical overlay (see Figure 4).

Finally, Figure 5 shows 3D head tracking. The visual task of head tracking allows for a template-based approach, described in the caption. This could be used for a variety of interactive applications, such as a graphical avatar in a videoconferencing application, or to adjust a graphical display appropriately for the viewer's head position. In addition to the entertainment uses described above, vision interfaces have applications for safety. Such tracking may be used in automobile applications to detect that a driver is drowsy or inattentive.

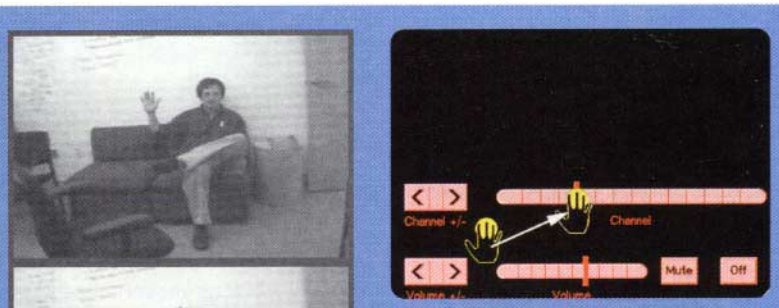

Figure 4: Adjusting a television set by hand signals. To get the attention of the television, the viewer raises his hand, palm toward the camera. When that gesture is detected, the television turns on, and a graphical overlay appears. A yellow hand tracks the position of the viewer's hand, and allows the viewer to use his hand like a mouse, selecting the desired menu item from the graphical interface. Normalized correlation with pre-stored hand templates performs the tracking.

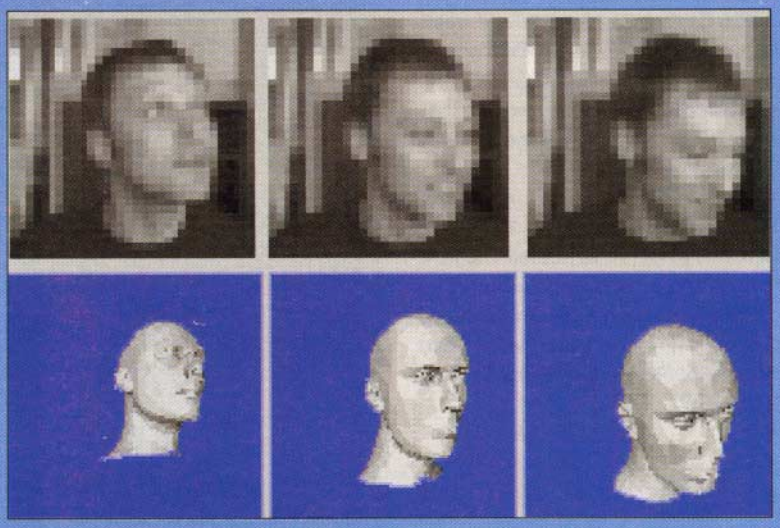

Figure 5: The scheme for head tracking shown here is based on template matching [1]. At initialization time, a frontal image of the subject is registered to a generic 3D model of the human head, and synthetically generated templates showing the appearance of the subject for a range of head poses are computed and stored. Subsequent head motion is determined by matching incoming images of the subject to the templates. The system achieves frame-rate performance by operating on subsampled images (32x32 resolution as shown in the figure). The approach is robust to non-rigid motion of the face (eyes closing, mouth opening etc).

The Present and the Future

Computer analysis of images of people is an active research area. Specialized conferences, such as the International Conference on Automatic Face and Gesture Recognition and the Workshop on Perceptual User Interfaces (PUI), present the state of the art. Relevant papers also appear in the major computer vision conferences: Computer Vision and Pattern Recognition (CVPR) and the International Conference on Computer Vision (ICCV).

Systems are now beginning to move beyond the research community, and to become viable commercial products. The Me2Cam, due in the Fall of 1999 from Intel and Mattel, will allow children to pop or become trapped by bubbles on the computer screen, depending on their movements. As the field progresses and the sophistication and reliability of the vision algorithms increases, applications should proliferate. Inter-disciplinary approaches, combining human studies as well as computer vision, will contribute. If interface-builders can match the ease of use shown in Figure 1, the prediction of that photograph should come true in at least one aspect: vision-based interfaces should become ubiquitous.

References

1. Beardsley, P. Pose estimation of the human head by modelling with an ellipsoid, Intl. Conf. on Automatic Face and Gesture Recognition, IEEE Computer Society, pp. 160-165, Nara, Japan, 1998.
2. Darrell, T., G. Gordon, M. Harville and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection, *Proc. IEEE Computer Vision and Pattern Recognition*, Santa Barbara, CA, 1998, pp. 601-609.
3. Ebihara, K., J. Kurumisawa, T. Sakaguchi, J. Ohya, L.S. Davis, T. Horprasert, R. I. Haritaoglu, A. Pentland and C. Wren. Shall we dance? *SIGGRAPH Conference abstracts and applications*, page 124, 1998. Enhanced Realities.
4. Elektro and sparko, Westinghouse Historical Collection, 1939 New York World's Fair, as printed in *Yesterday's Tomorrows* by J. J. Corn, Johns Hopkins University Press, 1996.
5. Freeman, W. T., D. B. Anderson, P. A. Beardsley, C. N. Dodge, M. Roth, C. D. Weissman, W. S. Yerazunis, H. Kage, K. Kyuma, Y. Miyake and K. Tanaka. "Computer vision for interactive computer graphics," *IEEE Computer Graphics and Applications*, 18(3):42-53, May-June 1998.
6. Freeman, W. T. and C. Weissman. Television control by hand gestures. In M. Bichsel, editor, Intl. Workshop on automatic face- and gesture-recognition, pp. 179-183, Zurich, Switzerland, 1995. Dept. of Computer Science, University of Zurich, CH-8057.
7. Ishii, H., K. Mochizuki and F. Kishino. A human motion image synthesizing by model based recognition from stereo images, *IMAGINA*, 1993.
8. Iwasawa, S., J. Ohya, K. Takahashi, T. Sakaguchi, S. Kawato, K. Ebihara and S. Morishima. Real-time, 3d estimation of human body postures from trinocular images, *ICCV'99 Workshop on Modelling*

People, Corfu, Greece, 1999.
9. Kobayashi, S., Y. Qiao, and A. Chugh. Optical gesture recognition system, *SIGGRAPH Visual Proceedings*, page 117, 1997.
10. Krueger, M. *Artificial Reality*, Addison-Wesley, 1983.
11. Kyuma, K., E. Lange, J. Ohta, A. Hermanns, B. Banish and M. Oita. "Artificial retinas-fast, versatile image processors," *Nature*, 372(197), 1994.
12. Maes, P., T. Darrell, B. Blumberg and A. Pentland. The alive system: Wireless, full-body interaction with autonomous agents, *ACM Multimedia Systems*, 1996, Special Issue on Multimedia and Multisensory Virtual Worlds.
13. San Francisco Exploratorium, 1999, **www.exploratorium.edu.**
14. Segen, J. "Gest: A learning computer vision system that recognizes gestures," *Machine Learning IV*, pp. 621-634, Morgan Kauffman, 1994, edited by Michalski et. al.
15. Segen, J. and S. Kumar. "Gesture VR: gesture interface to spatial reality," *SIGGRAPH Conference abstracts and applications*, page 130, 1998. Digital Pavilions.
16. Wilson, A. Seagull, 1996, SIGGRAPH 96 Digital Bayou, **http://www-white.media.mit.edu/vismod/demos/smartspaces/smartspaces.html.**
17. Wren, C. R., F. Sparacino, A. J. Azarbayejani, T. J. Darrell, T. E. Starner, A. Kotani, C. M. Chao, M. Hlavac, K. B. Russell and A. P. Pentland. "Perceptive spaces for performance and entertainment: untethered interaction using computer vision and audition," *Applied Artificial Intelligence*, 11(4), June 1997.

About the Authors

William T. Freeman is a Senior Research Scientist at MERL, a Mitsubishi Electric Research Lab in Cambridge, MA, where he studies Bayesian models of perception and interactive applications of computer vision. As part of his doctoral work at the Massachusetts Institute of Technology (1992), he developed "steerable filters," a class of oriented filters useful in image processing and computer vision. In 1997 he received the outstanding paper prize at the Conference on Computer Vision and Pattern Recognition for work on applying bilinear models to separate "style and content."

Paul A. Beardsley is a Research Scientist at MERL in Cambridge, MA, working in computer vision. His research interests include 3D reconstruction and development of partial 3D representations for image-based rendering. He received a Ph.D. in computer vision from the University of Oxford in 1992. His postdoctoral work was on the recovery of 3D structures from "uncalibrated" image sequences, with a particular application to robot navigation.

Hiroshi Kage is a Researcher of the Business Promotion Project of Artificial Retinas, the System LSI Division of Mitsubishi Electric in Sagamihara, Japan. He has been engaged in developing various machine vision algorithms for artificial retina chips. He received his B.E. and M.E. degrees in information science from Kyoto University, Japan in 1988 and 1990, respectively.

Ken-ichi Tanaka is a Group Manager of the Business Promotion Project of Artificial Retinas, the System LSI Division of Mitsubishi Electric in Sagamihara, Japan. His research interests include electric propulsion systems for spacecraft and neural networks. He received his B.E. and M.E. degrees in aeronautical engineering from Kyoto University, Japan in 1979 and 1981, respectively.

Kazuo Kyuma is the Project Manager of the Business Promotion Project of Artificial Retinas, the System LSI Division of Mitsubishi Electric in Sagamihara, Japan. He is also a professor at the Graduate School of Science and Technology, Kobe University, Japan, and a lecturer at Osaka University and the Tokyo Institute of Technology. His research interests cover optoelectronics, advanced LSI systems and neurocomputing. He received B.S., M.S., and Ph.D. degrees in electronic engineering from the Tokyo Institute of Technology, Japan, in 1972, 1974 and 1977, respectively.

Craig D. Weissman is Director, Software Engineering at E.piphany, Inc. in San Mateo, CA, a leading provider of customer-centric analytic solutions that help businesses personalize their interactions with customers by connecting and analyzing data from inside and outside the enterprise. He graduated in 1995 from Harvard University with an B.A./M.S. in applied math and computer science.

William T. Freeman
Paul A. Beardsley
MERL, Mitsubishi Electric Research Lab
201 Broadway
Cambridge, MA 02139
Email: **freeman@merl.com**

Hiroshi Kage
Ken-ichi Tanaka
Kazuo Kyuma
Mitsubishi Electric
System LSI Division
1-1-5-7, Miyashimo
Sagamihara City, KA
229-1195 Japan

Craig D. Weissman
E.piphany Software
San Mateo, CA