MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Computer vision for computer interaction

William T. Freeman, P. A. Beardsley, H. Kage, K. Tanaka, K. Kyuma, C. D. Weissman

TR99-36 October 1999

Abstract

It might seem that 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, the interactive application often constrains the vision problem to be solved, allowing fast and simple vision algorithms to be used. This paper gives a brief survey of existing vision-based interactive systems. These systems typically use a number of basic vision algorithms. We describe the basic algorithms used by some systems we have built at MERL: vision-based computer games, a television set controlled by hand gestures, and 3-D head tracking.

SIGGRAPH Computer Graphics magazine, November 1999

MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

Computer vision for computer interaction

W. T. Freeman*, P. A. Beardsley*,
H. Kage†, K. Tanaka†, K. Kyuma†, C. D. Weissman*‡

TR-99-36 October 1999

Abstract

It might seem that 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, the interactive application often constrains the vision problem to be solved, allowing fast and simple vision algorithms to be used.

This paper gives a brief survey of existing vision-based interactive systems. These systems typically use a number of basic vision algorithms. We describe the basic algorithms used by some systems we have built at MERL: vision-based computer games, a television set controlled by hand gestures, and 3-D head tracking.

To appear in: SIGGRAPH Computer Graphics magazine, November, 1999.

*MERL, a Mitsubishi Electric Research Lab, 201 Broadway, Cambridge, MA 02139. *freeman@merl.com*

†Mitsubishi Electric, Advanced Technology R&D Center, 8-1-1, Tsukaguchi-Honmachi, Amagasaki City, Hyogo 661, Japan

‡present address: E.piphany Software, San Mateo, CA

Computer vision for computer interaction

W. T. Freeman*, P. A. Beardsley*,
H. Kage†, K. Tanaka†, K. Kyuma†, C. D. Weissman*‡

1 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.

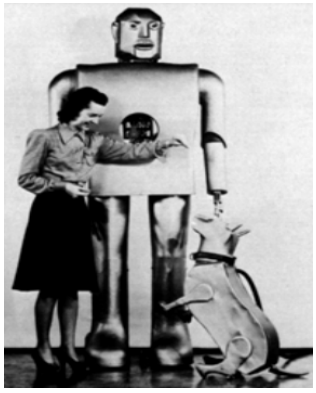


Figure 1: A vision from the past of the future: a natural, gesture-based interface, using camera-based input. Reprinted from [15].

2 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 which is really solving a few low-level vision problems. For example, the television controlled by hand gestures (Sect. 3) 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 his or her 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 [9], while the San Francisco Exploratorium has long had an exhibit where the silhouette of participants controls a graphical display [12].

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]. Segen and collaborators have built on work in hand gesture recognition [13] to make interactive games and fly-bys using hand gesture input [14]. Wilson used 3-D 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 [11] 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 [6, 7]. A system by Sony observed players making different fighting gestures and translated those into a computer game [8]. A collaboration of several research groups allowed the dance of participants to control the motions of virtual puppets [3].

*MERL, a Mitsubishi Electric Research Lab, 201 Broadway, Cambridge, MA 02139. *freeman@merl.com*

†Mitsubishi Electric, Advanced Technology R&D Center, 8-1-1, Tsukaguchi-Honmachi, Amagasaki City, Hyogo 661, Japan

‡present address: E.piphany Software, San Mateo, CA

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, USA 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 [4].

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 which can compete in cost or speed.

We have made prototypes which 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 made 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 it out 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 [10], 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 [4]. Figure 3 shows a schematic diagram of the artificial retina chip, a photograph of it, and a commercial product which uses the chip, the Nintendo GameBoy Camera. Over 1 million of these have been sold.

We also made a gesture-based television remote control, again designing the system to make the vision task simple [5]. 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, Fig. 4.

Finally, Figure 5 shows 3-D 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 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.

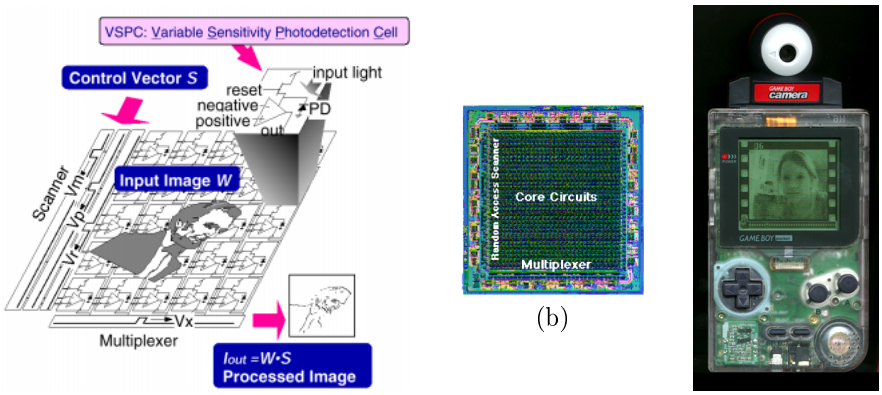
4 The present and the future

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, 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. Interdisciplinary approaches, combining human studies as well as computer vision, will contribute. If interface-builders can match the ease of use shown in Fig. 1, the prediction of that photograph should come true in at least one aspect: vision-based interfaces should become ubiquitous.

(a) (b) (c) (d) (e)

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



(a) (b) (c)

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, (a). (b) shows the chip. The Nintendo GameBoy Camera 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.

References

[1] P. Beardsley. Pose estimation of the human head by modelling with an ellipsoid. In *Intl. Conf. on Automatic Face and Gesture Recognition*, pages 160–165, Nara, Japan, 1998. IEEE Computer Society.

[2] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 601–609, Santa Barbara, CA, 1998.

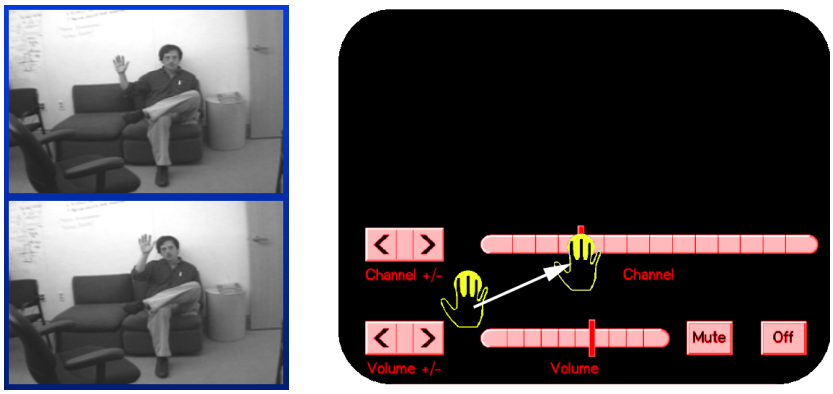
[3] K. Ebihara, J. Kurumisawa, T. Sakaguchi, J. Ohya, L. S. Davis, T. Horprasert, R. I. Haritaoglu, A. Pentland, and C. Wren. Shall we dance? *SIGGRAPH Conference abstracts and applications*, page 124, 1998. Enhanced Realities.

[4] W. T. Freeman, 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.

[5] W. T. Freeman and C. Weissman. Television control by hand gestures. In M. Bichsel, editor, *Intl. Workshop on automatic face- and gesture-recognition*, pages 179–183, Zurich, Switzerland, 1995. Dept. of Computer Science, University of Zurich, CH-8057.

[6] H. Ishii, K. Mochizuki, and F. Kishino. A human motion image synthesizing by model based recognition from stereo images. In *IMAGINA*, 1993.

[7] S. Iwasawa, J. Ohya, K. Takahashi, T. Sakaguchi, S. Kawato, K. Ebihara, and S. Morishima. Real-time, 3d estimation of human body postures from



viewer television overlay

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.

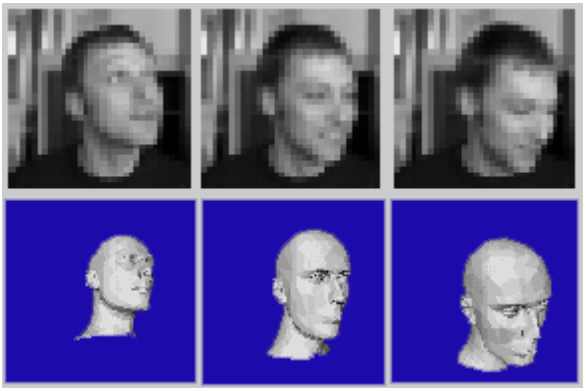


Figure 5: The scheme for head tracking shown here is based on template matching [1]. At initialisation time, a frontal image of the subject is registered to a generic 3D model of the human head, and synthetically generated templates showing the appearance of the subject for a range of head poses are 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).

trinocular images. In *ICCV'99 Workshop on Modelling People*, Corfu, Greece, 1999.

[8] S. Kobayashi, Y. Qiao, and A. Chugh. Optical gesture recognition system. *SIGGRAPH Visual Proceedings*, page 117, 1997.

[9] M. Krueger. *Artificial Reality*. Addison-Wesley, 1983.

[10] K. Kyuma, E. Lange, J. Ohta, A. Hermanns, B. Banish, and M. Oita. Artificial retinas–fast, versatile image processors. *Nature*, 372(197), 1994.

[11] P. Maes, 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.

[12] San Francisco Exploratorium, 1999. www.exploratorium.edu.

[13] J. Segen. Gest: A learning computer vision system that recognizes gestures. In *Machine Learning IV*, pages 621–634. Morgan Kauffman, 1994. edited by Michalski et. al.

[14] J. Segen and S. Kumar. Gesture VR: gesture interface to spatial reality. *SIGGRAPH Conference abstracts and applications*, page 130, 1998. Digital Pavilions.

[15] Elektro and sparko. Westinghouse Historical Collection. 1939 New York World's Fair.

[16] A. Wilson. Seagull, 1996. SIGGRAPH '96 Digital Bayou. http://www-white.media.mit.edu/vismod/demos/smartspaces/smartspaces.html.

[17] C. R. Wren, 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.