UNIVERSITY OF MARIBOR
FACULTY OF TECHNICAL SCIENCES

MODERN MODES OF MAN - MACHINE COMMUNICATION

INTERNATIONAL WORKSHOP

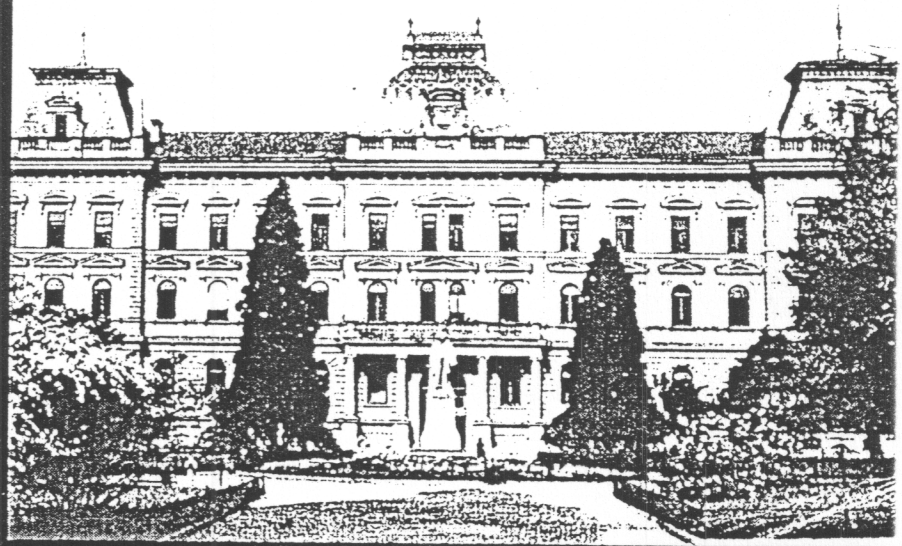
JUNE 6-10, 1994

MARIBOR, SLOVENIA

Edited by

Bogomir Horvat
Zdravko Kačič

PROCEEDINGS



MODERN MODES OF MAN-MACHINE COMMUNICATION

edited by
Bogomir Horvat
Zdravko Kačič

PROCEEDINGS

co-financed by TEMPUS and Ministry of Science and Technology of Republic of Slovenia

COMPUTER VISION TECHNIQUES IN MAN-MACHINE COMMUNICATIONS *

Franc Solina

INTRODUCTION

This article gives a brief introduction to computer vision and discusses some actual and future applications of computer vision techniques in man-machine communication. The field of computer vision started to evolve in the early 1970s. Since the hardware necessary for computer vision research and applications used to be expensive, computer vision techniques were predominantly limited to military, space, medical, and industrial domains. The cost of hardware and the introduction of personal computers changed that trend. Faster processors, cheaper high capacity memory devices, and local and global computer networks spawned a whole new generation of software, such as hypertext, multimedia, teleconferencing etc. Now, a new generation of personal computers is coming out of the factory already equipped with devices for capturing and producing images and sounds. The huge personal computing market that ten years ago changed or initiated completely new application areas, such as desktop publishing, spreadsheets, local area networks, and computer games, will now give a push for integration of computer vision techniques into man-machine interfaces.

The first section of this article gives a brief introduction to computer vision. The second section talks about the new evolving hardware and software platforms which enables simple adoption of computer vision techniques. In the third section, several existing and future applications of computer vision techniques in man-machine communication are described. The article ends with conclusions in the fourth section. Some of the references at the end are standard introductory texts to different aspects of computer vision. Other references are research articles and reports that brief on the latest developments in applying computer vision techniques to man-machine communication.

1 WHAT IS COMPUTER VISION?

Superficially are compute vision and computer graphics closely related — they both work with images. But, computer vision actually does just the opposite of computer graphics. Developers of computer graphics applications start with a set of models of the physical environment ranging from geometrical models for shape, color models, shading models, texture models etc., which are all employed in the process of producing still or animated images. Computer vision systems start on the opposite end of the above process and work backward. They start with an image or a sequence of images and try to recover some information from the images, taken with an input device, in order to identify a person, recognize an object, enable a robot to manipulate objects on the shop floor, drive a mobile robot, etc.

*This work was supported in part by the Ministry of Science and Technology of the Republic of Slovenia (Project P2–1122).

The inverse problem of deriving information from images is much harder than just producing images from given models. There are at least two hard problems that make computer vision difficult:

1. It is not clear, what kind of models to select to adequately describe a given scene.
2. There is loss of information when the 3-D world is projected onto a 2-D image.

The first problem is addressed with a multitude of all kinds of models that serve for shape, texture, color, and motion, as well as control strategies (i.e. attention, focusing). There is probably not a single best general model for any of the above modalities and one of the difficulties of a computer vision system designer is to choose appropriate models.

The second problem has from a mathematical viewpoint no unique solution because the transformation from 3-D to 2-D is not reversible. For any given 2-D shape there exist many 3-D shapes that could project the same 2-D shape. However, human and animate vision in general is a witness that biological systems can and do solve this problem. Unfortunately, the human visual system, especially its higher level functions such as recognition or visual memory are not completely understood [12]. Higher level visual functions are tightly connected with general cognitive functions. Although computer vision methods do not try to emulate biological vision at least not on the implementation and algorithmic level, computer vision still offers an excellent test-bed for psychophysical research. For an introduction to vision from a cognitive science viewpoint see [14]. The problem of interpreting images seems to be tractable by introducing additional information or knowledge to the problem and by seeking among many possible solutions the most probable solution. Knowledge of image formation, scene structure, and contextual information can be provided by the proper selection of models used for image interpretation. The suitability of models in computer vision is hence judged not only on how accurate they can model the visual data but also on how robustly can those models be reconstructed from the image data.

Computer vision or machine vision research started with the very ambitious goal of endowing machines with the capabilities of human vision, influenced by the excitement of early artificial intelligence research that set off to build intelligent machines. Especially in low-level image processing, computer vision used the results of the then already well established field of pattern recognition [19]. But soon the lofty goal of emulating human vision led to the realization that vision is a hard problem and this diverted the research to devising individual modules that could cope with isolated problems (shape-from-X techniques, where X can stand for shading, stereo, texture etc.). Visual processing was supposed to proceed in stages, each achieving a further step towards the complete recovery of the 3-D information about the scene. A basic work in this tradition of bottom-up and top-down processing is by David Marr [16]. This research direction had a limited success. In order to solve the problems of such isolated modules, assumptions were made that unfortunately did not hold in general. Such modules had hence a limited applicability. Also, integration of such modules is not straightforward.

A newer direction in computer vision research are goal or task directed vision systems. Researchers realized that to cope with the complexity of vision one should not strive to reconstruct a given scene in its entirety and then use just some part of that information for a given purpose. A complete scene reconstruction would require enormous computational resources, too. To do real-time computer vision, which reacts to a changing environment, one should extract in a given instance only the information which is necessary for performing a given task [22]. This so called task-directed or active vision integrates in a closed loop image capture (controlling the gaze, aperture, focus), model recovery, and recognition, all

designed and tailored to a specific task [2,1]. Such systems can be much easier applied to different situations because they adapt much readily.

Computer vision systems must organize the data in terms of some primitives that would bridge the gap between low-level features (i.e. image pixels) and high-level symbolic structures useful for further higher level processing. This requires segmentation of the image into meaningful parts and recovering the metric properties of those parts (shape recovery). Recent research shows that those two processes are better solved in parallel [15, 23].

Computer vision uses as input devices a whole range of sensors. The most common devices are CCD TV cameras hooked to a frame-grabber. Normally, only intensity images (black & white) are used, although color images are getting introduced to specific applications (color images require three times the memory capacity of intensity images). In industrial settings, range images (obtained with laser scanners or structured light techniques) are often used [5].

Computer vision applications span today from industrial settings (assembly, quality control, robotics, reverse engineering), medical applications (3-D imaging with X-rays, PET, ultra-sound, etc.) to military (target recognition, autonomous navigation), and general security applications (identification). The major problem in introducing computer vision techniques into practice is the high initial cost. Even commercial computer vision systems are not a turn-key application. They require considerable adjusting and tuning to a particular setting. Introducing computer vision techniques to the personal computing market will lower the cost of applications and hopefully initiate some new and better methods.

For a more detailed introduction to computer vision we refer the readers to the following reference books: [3,13,16,24].

2 BETTER HARDWARE AND SOFTWARE PLATFORMS

The two up till now separate worlds of personal computers and workstation are slowly merging. New generations of RISC microprocessors bridge the performance gap and standardization of operating systems takes care of software compatibility across different hardware platforms [8]. The so called client-server architecture, multimedia [4,11] and interactive video communications are the newest trends in modern business information systems. Images can convey much more information in a short time and people have always used pictorial representation of information. Cheaper and faster hardware enables now to display, manipulate, produce or capture pictorial information on computers. One of the first signs of this trend were visually oriented operating systems, built on their own (Macintosh OS Finder) or on top of existing operating systems (Microsoft Windows on MS-DOS, X-windows on Unix). Computer manufacturers are offering so called multimedia-ready systems with built-in stereo speakers and microphones, as well as digital video cameras. Such systems, connected to high speed computer networks can offer audiovisual contact with colleagues across the world, access to time-variant information, and three-dimensional representations of computer-aided design data. The currently most popular multimedia standard on the *Internet* is the *World Wide Web – WWW*. It enables the interplay of text, sound, still and video imagery across the whole Internet.

Computers equipped with special input-output devices can create so called *virtual environments* or *virtual realities* [10]. Data gloves or whole data suits equipped with sensors that record their position can enable users to directly manipulate with virtual objects which they can observe through stereoscopic head-mounted displays.

Since pictorial data used in teleconferencing, multimedia and virtual-reality systems

are not only synthetic images, produced by computer graphic techniques, but also images of real objects, people and environments, computer vision techniques are used to process and extract information from images.

3 COMPUTER VISION IN USER INTERFACES

In this section we talk about some present or future applications of computer vision techniques in man-machine communication. Since a specified speed of processing, storage capacities etc. are now fairly easy to meet, software engineering is focusing now more and more on designing proper user-interfaces [20]. A good user interface enables faster learning of new software applications and effective work with them. A new research field, called *user modeling*, drawing researchers from several disciplines, such as artificial intelligence, linguistics, psychology, education and man-machine interface design is forming [7]. Since people use imagery in solving problems, remembering, and thinking in general, all kinds of pictorial information are playing an important role in designing user-interfaces.

3.1 Intelligent Image Coding

Image compression is normally not concerned with the actual physical structure of coded images. The main objective is to compress the data as much as possible with the least possible image degradation. A much higher compression rate could be achieved by using computer vision techniques in the following way:

1. Recover models of the scene,
2. Transmit only the parameters of the models,
3. Reconstruct the image of a scene from the models.

Such intelligent image compression or coding would require much more computational power at the first stage but is well suited for applications where the scene must be understood also in terms of its physical structure, i.e. in teleoperation.

In virtual-reality applications the images that the user sees must be constantly adapted to his apparent motion or manipulation. Those images are in present systems usually synthetic images derived by computer graphics techniques from synthetic 3-D models. To use models of real scenes, objects, and people in virtual-reality systems will require computer vision techniques that will built adequate models of those real entities and in this way enable to produce again images of those entities from different viewpoints. Hand-coding of such entities on a grand scale, even with the help of powerful computer graphics tools, would be too time consuming and costly. A similar problem on a smaller scale is reverse-engineering, where the goal is to automatically derive a CAD model of an object from images of that object [21,15,23].

3.2 Gesture recognition

Gesture recognition (with hands or with the whole body) as part of user interface can be used in situations when the user cannot use a keyboard, a mouse, or some other common input device. Some situations require only eye tracking—recognizing where the user is looking in a given moment. With the help of such interfaces the user can interact with the computer in a similar way as with other standard pointing devices. Such special user interfaces have been built not only for handicapped people but also to capture some specific motions such as dancing, conducting or recognizing the sign language that deaf

people use for communication [25]. Recognition of complex hand or body motions can be done via glove-based input devices or whole body input devices used in virtual-reality research.

Instead of using wired gloves, computer vision techniques can be used to analyze the silhouette and gestures of a person. A novel environment for studying situated vision and behavior was built recently in MIT Media Lab which allows wireless full-body interaction between a human participant and a graphical world inhabited by autonomous agents [9]. An image of the participant is projected onto a large screen in front of the participant. The participant can act on the agents purely by his motions. The agents inhabiting the world are modeled as autonomous behaving entities which have their own sensors and goals and which can interpret the actions of the human participant and react to them in real-time.

3.3 Preserving eye-contact in teleconferencing

Teleconferencing requires in principle only the transmittal of images in real time. Images of teleconferencing parties are usually displayed in a window on a computer monitor. Technically, it is in practice not possible to position the camera and the monitor in the same optical axis. This causes the loss of eye-contact for both parties in such discourse which can be very annoying and dissatisfying. To preserve eye contact a user would need to look simultaneously at the monitor and into the camera. A group of researchers [18] is trying to solve this problem by using two or more cameras placed on top and bottom of the monitor, making a three dimensional reconstruction of the scene by solving the stereo correspondence problem, and constructing an image seen from a desired point of view. This would result in a virtual camera placed behind the monitor screen.

3.4 Lip reading

Researchers are trying to improve the performance of speech recognition systems by incorporating visual information about the corresponding lip movements [6]. This can be especially beneficial in the presence of noise and crosstalk when the performance of acoustic only systems severely degrades. The crucial part of the lip reading system is extracting the visual features of the speech from the image of the speaker's face. The high dimensional sampled images must be represented by appropriate models that convey only the relevant information. Researchers [6] use active contour models to recover and track the shape of lips during continuous speech.

3.5 Classification, verification, and recognition of users

Identifying a person seems straightforward—people do it every day in business and social encounters. But modern society sometimes requires a more reliable identification for security or prevention reasons [17]. Passwords used for access to computing resources, for automated teller machines, credit cards and so on are not too reliable. Computer vision techniques offer some possibilities to classify, verify or recognize a person from physiological (face, hand, eye, fingerprint) or even behavioral characteristics (signature, movements). Classification, verification, and recognition are different but interrelated tasks;

Classification has the goal of identifying an individual as a member of a class or category. A class or category can be defined by sex, age or some other outer features that are important in a specific situation.

Verification has the goal of identifying an individual by comparing the user's claim to an identity (using a card or code) with some of the user's characteristics.

Recognition has the goal of identifying an individual by comparing his characteristics to many such characteristics stored in a database.

Classification of users in man-machine communications may help in adjusting the user interface to a particular user group. Such context dependent user-interface might offer to children or elderly a more explanatory way of communication.

Verification is a very common problem, needed in many situations of modern life, not only in controlling access to computer interfaces, but also to secure areas. Face recognition by computer is not a simple task. Even humans, who are well tuned for that task, have severe problems if an image of a face is turned upside down. Automatic face recognition is difficult because people change hair styles, makeup, even their temper is reflected on their faces. Verification using the texture of the iris of the eye by automatic means seems more reliable because the striations, furrows, and other detail of the iris is at least as unique for a person as his fingerprints are [17]. The problem with this technique of identification is the rather intrusive way of taking images—people do not want to put their eyes as close to the device as necessary.

Recognition is normally based on exhaustive search through databases. There are many situations when one would want to identify a person, not only in security situations but also for statistical and other monitoring reasons.

3.6 Searching in Pictorial Databases

Pictorial collections in national libraries and other large collections are getting digitized. Large picture databases are being offered on CD-roms or on-line on computer networks. Searching through pictorial databases is now possible mainly by key word indexing. Key words, however, may miss a certain aspect of an image that a user might be interested in. Searching by content, using computer vision techniques, will enhance present search techniques or make the key word indexing process at least semi-automatic.

4 CONCLUSION

Computer vision techniques used to be applied only in specialized areas such as in military, medicine and industry. The rapid development of better, faster and cheaper hardware and software computing platforms is bringing computer vision techniques also to the mass computing market. A new generation of personal computers is already factory equipped with microphones and speakers, cameras and capabilities to display high-quality images in video-rate. User interface design will be affected by this new technology enabling visual user identification, visual interfacing by tracking eye or body motion of the user. Like the explosive growth of the personal computer software market a decade ago which affected the whole software industry, the new emerging multimedia computing environment distributed on local and global networks will initiate a completely new generation of software applications.

Computer vision can only benefit from this trend, not only by opening up new application areas but also by a wider interest in basic computer vision research.

REFERENCES

[1] John (Yiannis) Aloimonos. *Integration of Visual Modules.* Academic Press, Boston, 1989.

[2] R. Bajcsy. Active Perception. *Proceedings of the IEEE.* Vol. 76, No. 8, pages 996–1005, August 1988.

[3] D. H. Ballard and C. M. Brown. *Computer Vision.* Prentice-Hall, Englewood Cliffs, NJ, 1982.

[4] Andreja Balon and Franc Solina. Multimedijska tehnologija. *Elektrotehniški vestnik,* 60(1):59–72, 1993.

[5] P. J. Besl. *Surfaces in Range Image Understanding.* Springer-Verlag, New York, 1988.

[6] C. Bregler and Y. Konig. "Eigenlips" for Robust Speech Recognition. Technical report, TR-94-002, International Computer Science Institute, Berkeley, CA, January 1994. (To be presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Adelaide, Australia, 1994).

[7] R. Cohen, B. Kass, C. Paris, W. Wahlster. *Third International Workshop on User Modeling (UM'92),* Dagstuhl-Seminar-Report, No. 44, Schloss Dagstuhl, Germany, 1992.

[8] R. Comerford. PCs and Workstations. *IEEE Spectrum,* Vol. 31, No. 1, pages 35–37, January 1994.

[9] T. Darrell, P. Maes, B. Blumberg, A. P. Pentland. *A Novel Environment for Situated Vision and behavior.* IAPR/IEEE Workshop on Visual Behaviours, Seattle, WA, June 1994.

[10] S. R. Ellis. What are virtual environments? *IEEE Computer Graphics and Applications.* Vol. 14, No. 1, pages 17–22, January 1994.

[11] J. Encarnação, J. Foley. *Multimedia – System Architectures and Applications.* Dagstuhl-Seminar-Report, No. 51, Schloss Dagstuhl, Germany, 1992.

[12] M. J. Farah. *Visual Agnosia, Disorders of Object Recognition and What They Tell Us about Normal Vision.* MIT Press, Cambridge, MA, 1990.

[13] B. K. P. Horn. *Robot Vision.* MIT Press, Cambridge, 199?.

[14] P. H. Johnson-Laird. *The Computer and The Mind, An Introduction to Cognitive Science.* Harvard University Press, Cambridge, MA, 1988.

[15] A. Leonardis, F. Solina, and A. Macerl. A Direct Recovery of Superquadrics in Range Images Using Recover-and-Select Paradigm, *Electrotechnical Review,* Vol. 60, No. 4, pages 240–250, 1993.

[16] D. Marr. *Vision.* W. H. Freeman, San Francisco, 1982.

[17] B. Miller. Vital signs of identity. *IEEE Spectrum,* Vol. 31, No. 2, pages 22-30, February 1994.

[18] M. Ott, J. P. Lewis, I. Cox. Teleconferencing Eye Contact Using a Virtual Camera. Technical Report, C&C Research Laboratories, NEC USA, Princeton, NJ 08540, USA.

[19] A. Rosenfeld and A. C. Kak. *Digital Picture Processing I, II.* Academic Press, Orlando, FL, 1982.

[20] B. Shneiderman. *Designing the User Interface, Strategies for Effective Human-Computer Interaction.* Addison-Wesley, Reading, MA, 1992.

[21] F. Solina and R. Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-12(2):131–147, 1990.

[22] F. Solina and A. Leonardis. Selective scene modeling. In *Proceedings of the 11th International Conference on Pattern Recognition*, pages A:87–90, The Hague, The Netherlands, September 1992. IAPR, IEEE Computer Society Press.

[23] F. Solina, A. Leonardis, and A. Macerl. A direct part-level segmentation of range images using volumetric models. In *Proceedings of the IEEE International Conference on Robotics and Automation*, San Diego, CA, May 1994.

[24] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis and Machine Vision.* Chapman & Hall, London, 1993.

[25] D. J. Sturman and D. Zeltzer. A survey of glove-based input. *IEEE Computer Graphics and Applications.* Vol. 14, No. 1, pages 30–39, January 1994.