There are still obstacles to achieving general, robust, high-performance computer vision systems. The last decade, however, has seen significant progress in vision technologies for human-computer interaction.

Computer Vision in the Interface

Visual information is clearly important as people converse and interact with one another. Through the modality of vision, we can instantly determine a number of salient facts and features about others, including their location, identity, approximate age, focus of attention, facial expression, posture, gestures, and general activity. These visual cues affect the content and flow of conversation, and they impart contextual information different from, but related to, speech—for example, a gesture or facial expression may be a key signal, or the direction of gaze may disambiguate the object referred to in speech as "this" or the direction "over there." In other words, vision and speech are co-expressive and complementary channels in human-human interaction [6]. Just as automatic speech recognition seeks to

Illustration by Sandra Dionisi

By Matthew Turk

build machines that perceive the verbal aspects of human communication, computer vision technology can be used to build machines that "look at people" and automatically perceive relevant visual information.

Computer vision[1] is the computing discipline that attempts to make computers "see" by processing images and/or video [2, 3]. By understanding the geometry and radiometry of image formation, properties of the sensor (camera), and properties of the physical world, it is possible (at least in some cases) to infer useful information about the world from imagery, such as the color of a swatch of fabric, the width of a printed circuit trace, the size of an obstacle in front of a mobile robot on Mars, the identity of a person's face in a surveillance system, the vegetation type of the ground below, or the location of a tumor in an MRI scan. Computer vision studies how such tasks can be performed robustly and efficiently. Originally seen as a subarea of artificial intelligence, computer vision has been an active area of research for almost 40 years.

THERE *has been growing interest in turning the camera around and using computer vision to "look at people," that is, to detect and recognize human faces, track heads, faces, hands, and bodies, analyze facial expression and body movement, and recognize gestures.*

Computer vision research has traditionally been motivated by a few main application areas, such as biological vision modeling, robot navigation and manipulation, surveillance, medical imaging, and various inspection, detection, and recognition tasks. In recent years, multimodal and perceptual interfaces [9] have emerged to motivate an increasingly large amount of research within the machine vision community. The general focus of these efforts is to integrate multiple perceptual modalities (such as computer vision, speech and sound processing, and haptic I/O) into the user interface. For computer vision technology in particular, the primary aim is to use vision as an effective input modality in human-computer interaction. Such video-based sensing is passive and non-intrusive, as it does not require contact with the user or any special-purpose devices; the sensor can also be used for videoconferencing and other imaging purposes. This technology has promising applications in vision-based interaction domains such

[1] Also known as machine vision, image understanding, or computational vision. The Computer Vision Homepage (www.cs.cmu.edu/~cil/vision.html) is a good starting point for investigating the field.

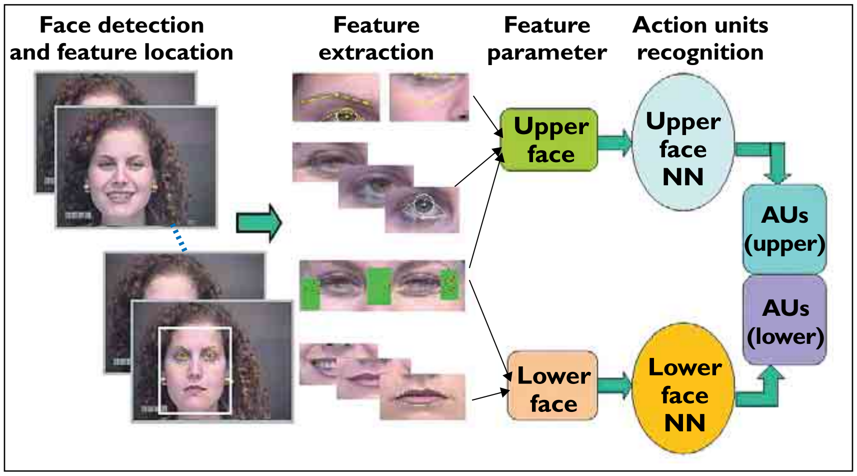

Figure 1. A feature-based action unit recognition system for facial expression analysis. (Reprinted with permission from [8].)

as games, biometrics, and accessibility, as well as general multimodal interfaces that combine visual information with other speech and language technologies, haptics, and user modeling, among others.

This pursuit of visual information about people has led to a number of research areas within computer vision focusing on modeling, recognizing, and interpreting human behavior. If delivered reliably and robustly, such vision technology can support a range of functionality in interactive systems by conveying relevant visual information about the user, such as identity, location, and movement, thus providing key contextual information. In order to fully support visual aspects of interaction, several tasks must be addressed:

- *Face detection and location:* How many people are in the scene and where are they?
- *Face recognition:* Who is it?
- *Head and face tracking:* Where is the user's head, and what is the specific position and orientation (pose) of the face?
- *Facial expression analysis:* Is the user smiling, laughing, frowning, speaking, sleepy?
- *Audiovisual speech recognition:* Using lip-reading and face-reading along with speech processing, what is the user saying?
- *Eye-gaze tracking:* Specifically where are the user's eyes looking?
- *Body tracking:* Where is the user's body and what is its articulation?
- *Hand tracking:* Where are the user's hands, in 2D or 3D? What are the specific hand configurations?
- *Gait recognition:* Whose style of walking/running is this?
- *Recognition of postures, gestures, and activity*: What is this person doing?

These tasks are very difficult. Starting with images from a video camera (or sometimes multiple cameras to get different views), this effort typically comprises at least 240 by 320 pixels (at 24 bits per pixel) delivered about 30 times per second. We seek to make sense of this barrage of data very quickly. Compare this with the problem of speech recognition, which starts with a one-dimensional, time-varying signal and tries to segment and classify into a relatively small number of known classes (phonemes or words). Computer vision is really a collection of subproblems, which may have little in common with each other, and which are all quite complex.

Vision-based Interface Tasks

Computer vision technology applied to the human-computer interface has some notable successes to date, and has shown promise in other areas. Face detection and face recognition have received the most attention and have seen the most progress. The first computer programs to recognize human faces appeared in the late 1960s and early 1970s, but it was not until the early 1990s that computers became fast enough to support these tasks in anything close to real time. The problem of face recognition has spawned a number of computational models, based on feature locations, face shape, face texture, and combinations thereof; these include principle component analysis, linear discriminant analysis, Gabor wavelet networks, and Active Appearance Models (AAMs). A number of companies, such as Identix, Viisage Technology, and Cognitec Systems, now develop and market face recognition technologies for access, security, and surveillance applications. These systems have been deployed in public locations such as airports and city squares, as well as in private, restricted access environments. For a comprehensive survey of face recognition research, see [12].

Face detection technology—to locate all faces in a scene, at various scales and orientations—has improved significantly in recent years with statistical learning approaches that run in real time. Head and face tracking works well in very constrained environments—for example, when markers are placed on the subject's face—but tracking face poses and facial feature positions in general environments is still a difficult problem. The same is true for facial expression analysis, which typically depends on accurate facial feature tracking as input. There have been several promising prototype systems that can recognize a limited range of facial features (see Figure 1 for example), but they are still very limited in performance and robustness.

Eye-gaze tracking has been commercially available for several years, generally for disabled computer users and for scientific experiments. These systems use active sensing, sending an infrared light source toward the user's eye to use as a reference direction, and they severely restrict the movement of the head. In their current form, they are not appropriate for general multimodal user interfaces.

In order to determine a person's location or to establish a reference coordinate frame for head and hand movement, it is useful to track bodies in the video stream. Early systems such as Pfinder [10] (as illustrated in Figure 2) produced a contour representation of the body's silhouette by keeping track of a static background model, and identified likely positions of the head and hands. More detailed and sophisticated articulated and dynamic body models are used by several researchers, though fitting image data to these models is complex and can be quite slow (see [4] for a recent survey of large-scale body motion technologies). Although motion capture systems are widely used in animation to capture precise body motion, these require the user to don special clothing or scores of sensors or markers, making this approach unsuitable for general-purpose multimodal interfaces.

Tracking hand positions in 2D and 3D is not difficult when the environment is controlled (for example, fixed lighting conditions, camera position, and background) and there is little or no occlusion of the hands; keying on skin color is the typical approach. However, in normal human behavior, the hands are often hidden (in pockets, behind the head) or temporarily occluded by the other arm or hand. In these cases, hand tracking is difficult and requires prediction based on human kinematics. A more difficult problem is tracking the complete hand articulation—the whole 29 degrees of freedom (DOF) defined by the physical hand structure (23 DOF above the wrist, and 6 DOF specifying position and orientation of the hand). Wu and Huang [11] provide a good review of hand tracking and hand gesture recognition.

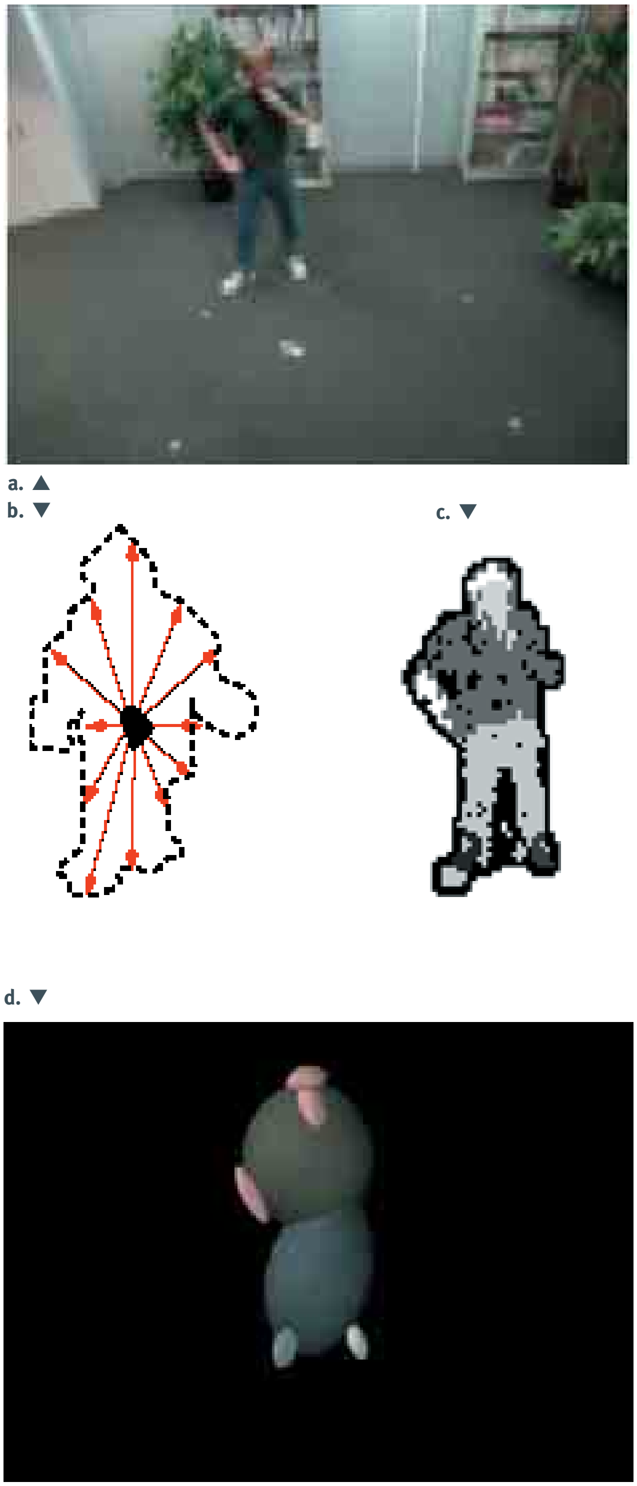

Figure 2. The MIT Pfinder system [10] for body tracking: video input, computed silhouette, segmentation, a 2D-representation of the blob statistics. (Reprinted with permission.)

Locating, identifying, and tracking the human body and its constituent parts is only the first step for the purposes of interaction; recognizing behavior is also required. The behaviors of interest may be structured, isolated gestures (as in a system for signaling at a distance), continuous natural human gesture, or behaviors defined over a range of time scales (for example, leaving the room, or eating lunch at one's desk). Gesture recognition may be implemented as a straightforward pattern recognition problem, attempting to match a certain temporal sequence of body parameters, or may be a probabilistic system that reasons about statistically defined gesture models. The system must distinguish between unintentional human movements, movement for the purpose of manipulating objects, and those gestures used (consciously or not) for communication. The relationship between language and gesture is quite complex [6], and automating general-purpose, context-independent gesture recognition is a very long-term goal.

THERE *has been significant progress toward building real-time, robust vision techniques, helped partly by advances in hardware performance driven by Moore's Law.*

Although simple state-space models have worked in some cases, statistical models are generally used to model and recognize temporal gestures. Due to their success in the field of speech recognition, Hidden Markov Models (HMMs) have been used extensively to model and recognize gestures. An early example is a system to recognize a limited amount of American Sign Language developed by Starner and Pentland [7]. There have been several variations of the basic HMM approach, seeking better matches with the wider variety of features and models in vision. Because many gestures include several components, such as hand-motion trajectories and postures, the temporal signal is more complex than in the case of speech recognition. Bayesian networks have also shown promise for gesture recognition.

Progress in Vision-based Interface Technology

Despite several successful niche applications, computer vision has yet to see widespread commercial use, even after decades of research. Several trends seem to indicate this may be about to change. Moore's Law improvements in hardware, advances in camera technology, a rapid increase in digital video installation, and the availability of software tools (such as Intel's

OpenCV library; www.intel.com/research/mrl/research/opencv) have led to vision systems that are small, flexible, and affordable. In recent years, the U.S. government has funded face recognition evaluation programs: the original Face Recognition Technology (FERET) Program from 1993 to 1997, and more recently the Face Recognition Vendor Tests (FRVT) of 2000 and 2002. These programs have provided performance measures for assessing the capabilities of both research and commercial face recognition systems. FRVT 2002 [5] thoroughly tested 10 commercial systems, collecting performance statistics on a very large dataset: 121,589 face images of 37,437 individuals, characterizing performance along several dimensions (indoor vs. outdoor, male vs. female, younger vs. older, time since original image registration of the subject). Figure 3 shows results of face verification of the best systems for five categories of frontal face images.

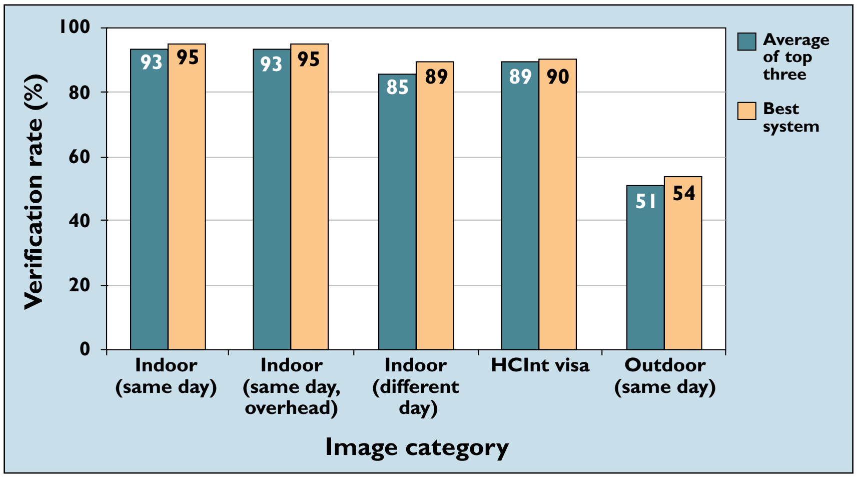

Figure 3. Results from the 2002 Face Recognition Vendor Test: Verification performance for five categories of frontal facial images. Performance is reported for the best system and average of the top three systems in each category. The verification rate is reported at a false accept rate of 1%. (Reprinted with permission from [5].)

In recent years, DARPA funded large projects devoted to recognizing people at a distance and video surveillance. The ongoing Human Identification at a Distance (HumanID) Program will pursue multimodal fusion techniques, including gait recognition, to identify people at long range (25–150 feet). The Video Surveillance and Monitoring (VSAM) Program sought to develop systems to recognize activity of interest for future surveillance applications. The National Science Foundation has awarded several Information Technology Research (ITR) grants in areas related to vision-based interface technology. Industry research labs at companies such as Microsoft, IBM, and Intel have made significant efforts to develop technology in these areas, as have companies in industries such as personal robotics and entertainment.

The biometrics market has increased dramatically in recent years, with many companies providing face recognition (and usually face detection and face tracking) technologies, including 3D approaches (for example, Geometrix, A4Vision, and 3DBiometrics; see the article by Jain and Ross in this issue for a more detailed description of biometrics involving computer vision and other modalities). Several research groups and companies have developed face tracking technologies, especially for use in computer graphics markets (games and special effects).

A nice example of simple vision technology used effectively in an interactive environment was the KidsRoom project (www-white.media.mit.edu/ vismod/ demos/kidsroom) at the MIT Media Lab [1]. The KidsRoom provided an interactive, narrative play space for children. Using computer vision to recognize users' locations and their actions helped deliver a compelling interactive experience for the participants. There have been dozens of other compelling prototype systems developed at universities and research labs, several of which are in the initial stages of being brought to market.

Technical Challenges

Aside from face recognition technologies geared for the biometrics market, there are few mature computer vision products or technologies to support interaction with users. There are, however, a large and growing number of research projects and prototypes of such systems. In order to move from the lab to the real world, a few basic issues must be addressed:

- *Robustness.* Most vision technologies are brittle and lack robustness; small changes in lighting or camera position may cause them to fail. They need to work under a wider variety of conditions and gracefully and quickly recover from failures.
- *Speed.* For most computer vision technologies, there is a practical trade-off between doing something thoroughly and doing it quickly enough to be interactive. There is just too much video data coming in to do sophisticated processing in real time. We need better algorithms, faster hardware, and smarter ways to decide what to compute and what

to ignore. (Digital cameras that provide image streams already processed will help a lot!)

- *Initialization.* Many techniques track well after initial model acquisition, but the initialization step is often very slow and demands user participation. Systems must initialize quickly and transparently.
- *Usability.* Demonstrations of vision technology often work well for the person who developed the system (who spent many hours figuring out its intricacies), but fail for the novice who wasn't "trained by the system." Instead, these systems need to adapt to users and deal with unanticipated user behavior. In addition, they need to provide simple mechanisms for correcting or overriding misinterpretations and to provide feedback to the user, to avoid unintended, catastrophic results.
- *Contextual integration.* A vision-based interaction technology is not an end in itself, but it is part of a larger system. Gestures and activity need to be understood in the appropriate application context, not as isolated behavior. In the long run, this requires a deep understanding of human behavior in the context of various applications.

The first three of these issues are being addressed daily in research labs and product development groups around the world; usability and contextual integration are occasionally considered, but as more applications are developed they will need to come to the forefront of the research agenda.

Conclusion

Computer vision is a very difficult problem still far from being solved in the general case after several decades of determined research, largely driven by a few main applications. However, in the past dozen or so years, there has been growing interest in turning the camera around and using computer vision to "look at people," that is, to detect and recognize human faces, track heads, faces, hands, and bodies, analyze facial expressions and body movements, and recognize gestures. There has been significant progress toward building real-time, robust vision techniques, helped partly by advances in hardware performance driven by Moore's Law. Certain subproblems (for example, face detection and face recognition) have had notable commercial success, while others (for example, gesture recognition) have not yet found a large commercial niche. In all of these areas, there remain significant speed and robustness issues, as fast approaches tend to be brittle, while more principled and thorough approaches tend to be excruciatingly slow. Compared to speech recognition technology, which has seen years of commercial viability and has been improving steadily for decades, computer vision technology for HCI is still in the Stone Age.

However, there are many reasons to be optimistic about the future of computer vision in the interface. Individual component technologies have progressed significantly in the past decade; some of the areas are finally becoming commercially viable, and others should soon follow. Basic research in computer vision continues to progress, and new ideas will be speedily applied to vision-based interaction technologies. There are currently many conferences and workshops devoted to this area of research and also to its integration with other modalities. The area of face recognition has been a good model in terms of directed funding, shared data sets, head-to-head competition, and commercial application—these have greatly pushed the state of the art. Other technologies are likely to follow this path, until there is a critical mass of research, working technology, and commercial application to help bring computer vision technology to the forefront of multimodal human-computer interaction.

References

1. Bobick, A., Intille, S., Davis, J., Baird, F., Pinhanez, C., Campbell, L., Ivanov, Y., Sch¸tte, A., Wilson, A. The KidsRoom: A perceptually-based interactive and immersive story environment. *PRESENCE: Teleoperators and Virtual Environments 8*, 4 (Aug. 1999), *367–391*.
2. Forsyth, D., and Ponce, J. *Computer Vision: A Modern Approach.* Prentice Hall, Upper Saddle River, NJ, 2003.
3. Marr, D. *Vision*. W.H. Freeman, NY, June 1982.
4. Moeslund, T.B., and Granum, E. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding 18* (2001), 231–268.
5. Phillips, P.J., Grother, P., Micheals, R.J., Blackburn, D.M., Tabassi, E., and Bone, M. Face recognition vendor test 2002: Overview and summary; www.frvt.org.
6. Quek, F., McNeill, D., Bryll, R., Kirbas, C., Arslan, H., McCullough, K.E., Furuyama, N., and Ansari, R. Gesture, speech, and gaze cues for discourse segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (Hilton Head Island, South Carolina, June 13–15, 2000), 247–254.
7. Starner, T., and Pentland, A. Visual recognition of American Sign Language using Hidden Markov Models. *International Workshop on Automatic Face and Gesture Recognition,* (Zurich, Switzerland, 1995), 189–194.
8. Tian, Y.L., Kanade, T., and Cohn, J.F. Recognizing action units for facial expression analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence 23*, 2 (2001), 97–115.
9. Turk, M. and Robertson, G. Perceptual Interfaces. *Commun. 43,* 3 (2000), 32-24.
10. Wren, C., Azarbayejani, A., Darrell, T., and Pentland, A. Pfinder: Real-time tracking of the human body. *IEEE Trans. on Pattern Analysis and Machine Intelligence 19,* 7 (July 1997), 780–785.
11. Wu, Y. and Huang, T.S. Human hand modeling, analysis and animation in the context of human computer interaction. *IEEE Signal Processing, special issue on Immersive Interactive Technology 18,* 3 (2001), 51–60.
12. Zhao, W., Chellappa, R., Rosenfeld, A., and Phillips, J. Face recognition: A literature survey. Technical Report, CS-TR4167R. University of Maryland, College Park, MD, 2002.

Matthew Turk (mturk@cs.ucsb.edu) is an associate professor in the Computer Science Department and the Media Arts and Technology Program at the University of California, Santa Barbara.