by Massimo Piccardi and Tony Jan

Recent Advances in Computer Vision

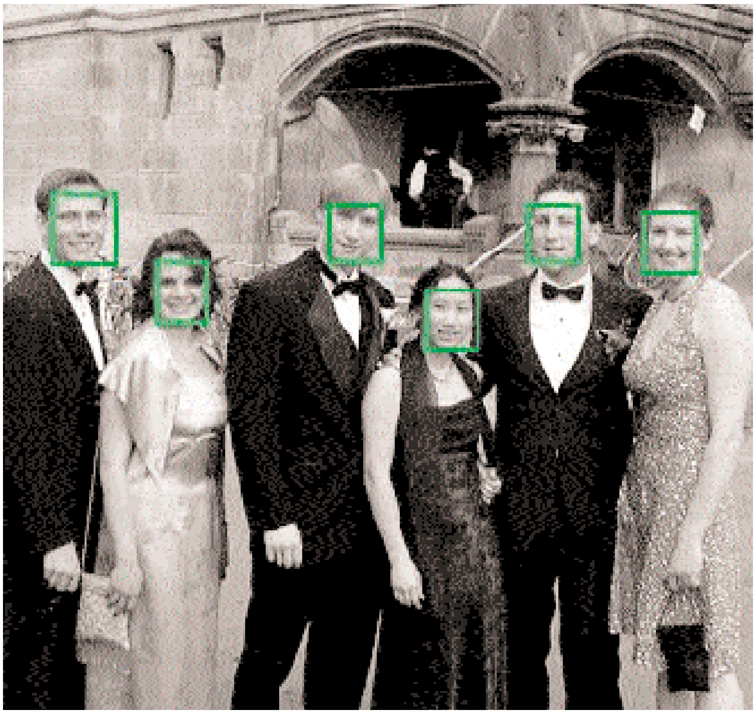


Figure 1. The Face Detection Project can automatically distinguish images containing faces from other images and put a box around each detected face—frontal, profile, or three-quarter images.

Henry Schneiderman and Takeo Kanade, Carnegie Mellon University

Computer vision is the branch of artificial intelligence that focuses on providing computers with the functions typical of human vision. To date, computer vision has produced important applications in fields such as industrial automation, robotics, biomedicine, and satellite observation of Earth. In the field of industrial automation alone, its applications include guidance for robots to correctly pick up and place manufactured parts, nondestructive quality and integrity inspection, and on-line measurements.

Until a few years ago, chronic problems affected computer-vision systems and prevented their widespread adoption. Since its start, computer vision has appeared as a computationally intensive and almost intractable field because its algorithms require a minimum of hundreds of MIPS (millions of instructions per second) to be executed in acceptable real time. Even the input–output of high-resolution images at video rate was traditionally a bottleneck for common computing platforms such as personal computers and workstations. To solve these problems, the research community has produced an impressive number of dedicated computer-vision systems. One such famous system was the Massively Parallel Processor (MPP), designed at the Goddard Space Flight Center in 1983 and operated there until 1991. The MPP used an array of 16,384 single-bit processors and was capable at peak performance of 250 million floating-point operations/s—an impressive feat at the time.

Dedicated computers such as the MMP have always received a cold reception from industry because they were expensive, cumbersome, and difficult to program. In recent years, however, increased performance at the system level—faster microprocessors, faster and larger memories, and faster and wider buses—has made computer vision affordable on a wide scale. Fast microprocessors and digital-signal processors are now available as off-the-shelf solutions, and some of them can execute calculations at rates of thousands of MIPS. The Texas Instruments C6414 processor, for example, runs at 600 MHz and can achieve a peak performance of 4,800 MIPS. High-speed serial buses such as the IEEE 1394 and USB 2.0 are capable of transferring hundreds of megabits per second, a rate that greatly exceeds the requirements of any common high-resolution video camera. These buses are already integrated into the most recent personal computer chipsets or are available as inexpensive daughterboards. Moreover, video cameras have gone almost completely to digital, and they come in several price ranges and types. Consumer camcorders are based on standards such as the Digital Video (DV), which provides videos with 720 × 480 pixels/frame at a rate of 30 frames/s. Even Webcams can now provide images of satisfactory quality at prices starting as low as $25.

The availability of affordable hardware and software has opened the way for new, pervasive applications of computer vision. These applications have one factor in common. They tend to be *human-centered*; that is, either humans are the targets of the vision system or they wander about wearing small cameras, or sometimes both. Vision systems have become the central sensor in applications such as

- human-computer interfaces (HCIs), the links between computers and their users;
- augmented perception, tools that increase normal perception capabilities of humans;
- automatic media interpretation, which provides an understanding of the content of modern digital media, such as videos and movies, without the need for human intervention or annotation; and
- video surveillance and biometrics.

Human-computer interfaces

The basic idea behind the use of computer vision in HCIs is that in several applications, computers can be instructed more naturally by human gestures than by the use of a keyboard or mouse. In one interesting application, computer scientist James L. Crowley of the National Polytechnical Institute of Grenoble in France and his colleagues used human eye movements to scroll a computer screen up and down. A camera located on top of the screen tracked the eye movements. The French researchers reported that a trained operator could com-

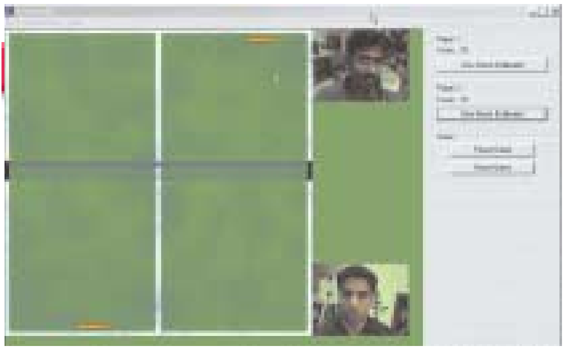


Figure 2. A camera tracks the point of each player's nose closest to the camera and links it to the red "bat" at the top (or bottom) of the table to return the computer ball across the "net."

Institute for Information Technology
National Research Council Canada

plete a given task 32% faster by using his eyes rather than a keyboard or mouse to direct screen scrolling. In general, using cameras to sense human gestures is much easier than making users wear cumbersome peripherals such as digital gloves.

Another interesting example of an HCI application can be downloaded from http://www.cv.iit.nrc.ca/research/Nouse/ for personal testing, provided a Webcam is plugged into your personal computer. This application—called Nouse, for nose as a mouse—tracks the movements of your nose, and was developed by Dmitry Gorodnichy. You can play NosePong, a nose-driven version of the Pong video game (Figure 2), or test your ability to paint with your nose or to write with your nose. Although this application is slanted toward fun, it is a convincing demonstration of the potential uses of cameras as natural interfaces. In industry, for example, an operator might quickly stop a conveyor belt with a specific gesture detected by a camera without needing to physically push a button, pull a lever, or carry a remote control.

Cameras could also become powerful peripherals for the so-called intelligent home. A camera located in your living room would perform several tasks, starting with sensing a human presence and then turning the lights on and the heat up. Indeed, cameras could replace the many hard-to-find remote controls around today's homes, provide environmental surveillance, and turn the TV off when you fall asleep in your favourite armchair.

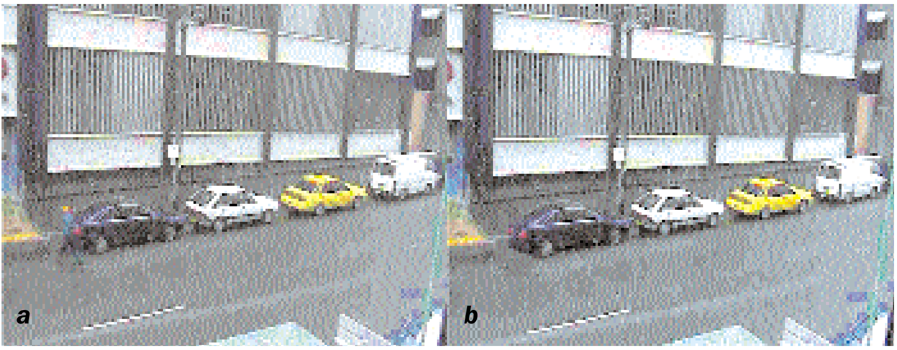


Figure 3. This parking-lot surveillance system subtracts the static background image, distinguishes a person from moving vehicles, locates the head, and calculates the speed of the head in each frame.

University of Technology, Sydney, Australia

The Voice

Another application is The vOICe, developed at Philips Research Laboratories (Eindhoven, The Netherlands) by Peter B. L. Meijer and available online for testing at http://www.seeingwithsound.com. The vOICe provides a simple yet effective means of *augmented perception* for people with partially impaired vision. In the virtual demonstration, the camera accompanies you in your wanderings. The camera periodically scans the scene in front of you and turns images into sounds, using different pitches and lengths to encode objects' position and size.

Media interpretation

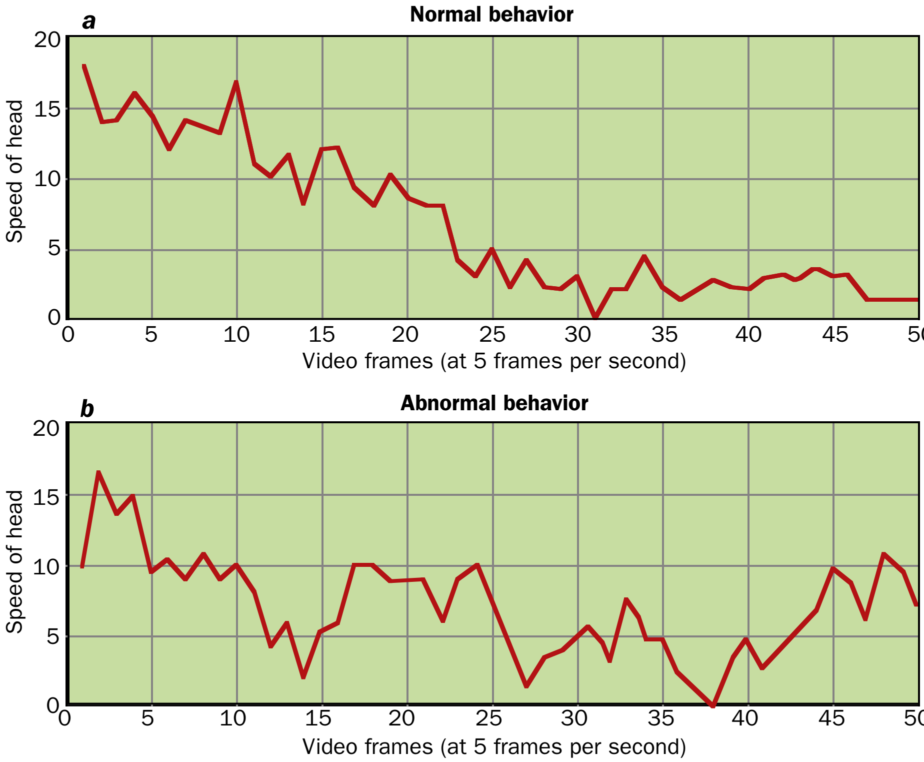
The use of computer vision for *automatic media interpretation* assists users in searching for specific scenes and shots otherwise not annotated in the video-scene indexes. For example, images containing faces can be automatically distinguished from other images, as the results of the Face Detection Project led by Henry Schneiderman and Takeo Kanade at Carnegie Mellon University (CMU) prove. The CMU face detector is considered the most accurate for frontal face detection and is also reliable for facial profiles and three-quarter images. Many examples are available at http://vasc.ri.cmu.edu/demos/faceindex—one is shown in Figure 1—and anyone can submit an image to http://www.vasc.ri.cmu.edu/cgi-bin/demos/findface.cgi, which will process the image overnight and depict all detected faces with a box around them.

However, computer vision can do much more for multimedia. For example, it is an invaluable support to recent multimedia standards aimed at compressing digital videos—reducing their size in bytes—while still retaining acceptable visual quality. One such standard is MPEG-4 from the Moving Picture Expert Group, which allows the compression of different objects in a scene with specific compression levels in such a way as to adjust the trade-off between space reduction and visual quality on a per-object basis. The basic idea is that important objects such as actors should retain the highest visual quality, while objects in the background can be encoded with lower quality to save bytes. Nonetheless, MPEG-4 is silent on how to separate a video into the objects of which it is composed. Here again, computer vision can help with a variety of techniques that perform the task automatically.

Video surveillance

Perhaps the most developed modern application of computer vision is *video surveillance*. Long gone are the days when video surveillance meant low-resolution, black-and-white, analog closed-circuit television. Nowadays, computer vision enables the integration of views from many cameras into a single, consistent "superimage." Such an image automatically detects scenes with people and/or vehicles or other targets of interest, classifies them in categories such as people, cars, bicycles, or buses, extracts their trajectories, recognizes limb and arm positions, and provides some form of behavior analysis.

The analysis relies on a list of previously specified behaviors or on statistical observations



University of Technology, Sydney, Australia

Figure 4. Examples of the speed of the head (in pixels per frame) of a person in the parking lot exhibiting normal behavior (a) and abnormal behavior (b). Such video surveillance might alert a security guard to a possible car thief.

such as frequent-versus-infrequent behaviors. The basic goal is not to completely replace security personnel but to assist them in supervising wider areas and focusing their attention on events of interest. Although the critical issue of privacy must be addressed before society widely adopts these video surveillance systems, the recent need for increased security has made them more likely to win general acceptance. In addition, several technical countermeasures can be taken to prevent privacy abuses, such as protecting access to video footage by way of passwords and encryption.

At the University of Technology in Sydney, Australia, we have developed and tested a system that can detect suspicious pedestrian behavior in parking lots. Our approach is based on the assumption that a suspicious behavior corresponds to an individual's erratic walking trajectory. The rationale behind this assumption is that a potential offender will wander about and stop between different cars to inspect their contents, whereas normal users will maintain a more direct path of travel.

The first step consists of detecting all the moving objects in the scene by subtracting an estimated "background image"—one that represents only the static objects in the scene—from the current frame (Figures 3a and 3b). The next step is to distinguish people from moving vehicles on the basis of a form factor, such as the height:width ratio, and to locate their heads as the top region in their silhouette. In this way, the head's speed at each frame is automatically determined. Then, a series of speed samples are repeatedly measured for each person in the scene. Each series covers an interval of about 10 s, which is enough to detect suspicious behavior patterns (Figure 4).

Finally, a neural network classifier, trained to recognize the suspicious behaviors, provides the behavior classification. In the experiments we performed, the system achieved good accuracy, with a reasonably limited number of false dismissals and false alarms—4% and 2%, respectively, among more than 100 test samples. Although manufacturers and operators of surveillance systems have often been reluctant to accept innovation, recent results from research laboratories of major companies prove that these systems are now reliable, economical, and ready for commercialization. One example is DETER from Honeywell Labs, a prototype urban-surveillance system.

For those who want to build their own surveillance systems, an enormous amount of equipment is available. Web sites of manufacturers such as Sony, Axis, Pelco, and many others offer a wide range of cameras. You can find network cameras starting at less than $500 that can be simply plugged into any network, such as a TCP-IP, which can carry a full Web server and allow camera frames to be downloaded and processed. Adjustable pan–tilt–zoom cameras can be used to point and focus on specific targets over wide survey areas. And if cabling poses a problem because of camera location, wireless versions are available off-the-shelf.

Computer vision, already a useful aid in several industrial processes, will find increasing uses as companies develop new applications in areas such as HCI, augmented perception, and automatic media interpretation. Its potential to improve plant and public safety is attracting increasing attention in today's security-conscious world.

Further reading

Crowley, J. L; Coutaz, J.; Bérard, F. Perceptual user interfaces: things that see. *Commun. ACM* **2000**, *43* (3), 54–64.

Jan, T.; Piccardi, M.; Hintz, T. Automated Human Behaviour Classification using Modified Probabilistic Neural Network. In *Proc. Int. Conf. Computational Intelligence for Modelling, Control and Automation*; CIMCA 2003, Vienna, Austria, Feb. 12–14, 2003.

National Instruments Corp. (Austin, TX), markets a range of computer-vision products. Its LabView-based Vision line focuses on industrial and scientific uses. http://sine.ni.com/apps/we/nioc.vp?cid=1286&lang=US.

Pavlidis, I.; Morellas, V.; Tsiamyrtzis, P.; Harp, S. Urban surveillance systems: from the laboratory to the commercial world. *Proc. IEEE* **2001**, *89* (10), 1478–1496.

BIOGRAPHY

Massimo Piccardi (massimo@it.uts.edu.au) is an associate professor of computer science and Tony Jan (jant@it.uts.edu.au) is a lecturer in the department of computer systems at the University of Technology in Sydney, Australia.