THINGS THAT SEE

James L. Crowley, Joëlle Coutaz, and François Bérard

Exploring machine vision for human-computer interaction.

The exponential decrease in the costs of computation and of communication is rapidly leading to convergence and ubiquity. At the same time, inexpensive computing power is enabling a quiet revolution in the machine perception of human action. In the near future, we expect machine perception to converge with ubiquitous computing and communication.

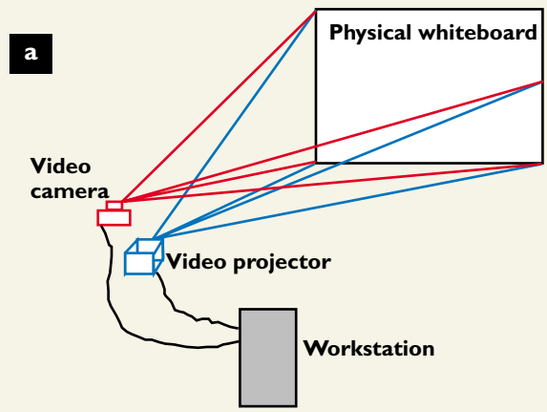
This convergence may lead to the widespread introduction of things and environments that see. However, reaping the benefits of ubiquitous perception will require consideration of human abilities and social needs as an integral part of system design.

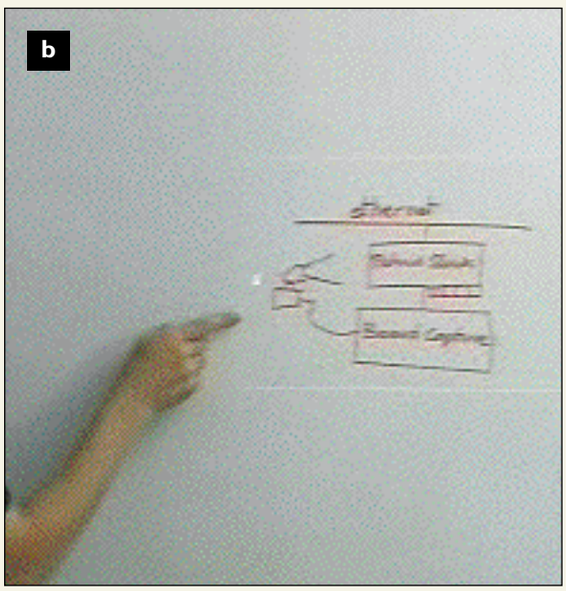
Research in human-computer interaction (HCI) has developed cognitive theories, design methods, and software tools for building useful and usable systems. Scientific results and empirical studies have led to sound principles such as that of direct manipulation. For nearly two decades, however, direct manipulation has been instantiated in the form of the "electronic desktop metaphor," jeopardizing the directness and the affordance of the physical world.

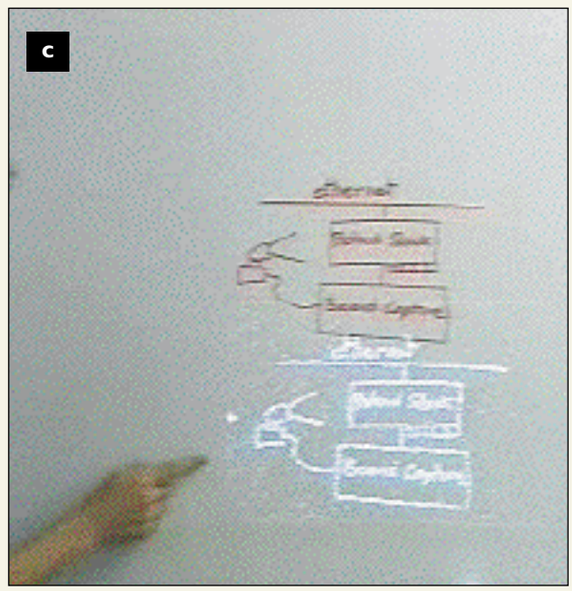
Recent efforts in HCI seek a seamless bridge between physical and electronic bits. Wellner's Digital Desk [12] and Fitzmaurice's Bricks [6] illustrate this trend. In the Digital Desk, physical office tools such as paper sheets and erasers are augmented with computation using video projection and machine vision. The Bricks allow direct manipulation of electronic objects using Lego™-like physical artifacts as handles to control the virtual world. The Digital Desk and other graspable iconic systems inspired from the Bricks have demonstrated the benefits of mixing physical and virtual entities. However, because they use naive machine vision techniques, their contribution cannot be tested in real-life conditions. Advances in machine vision provide an opportunity to make these new paradigms effective.

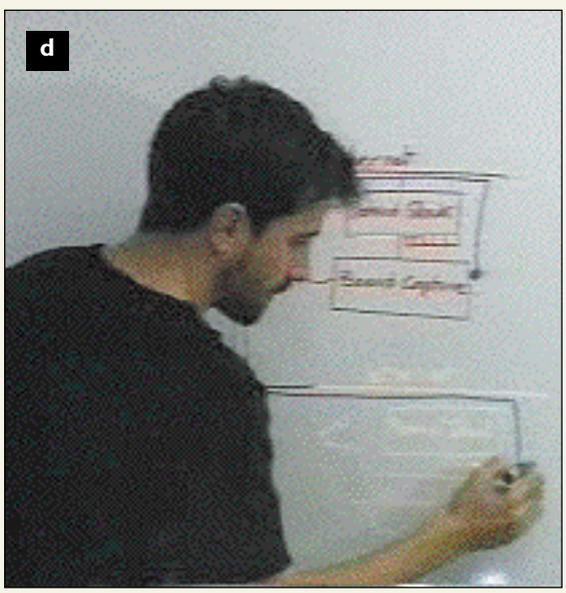
Figure 1. Interacting with the Magic Board (iihm.imag.fr/demos/magicboard/).

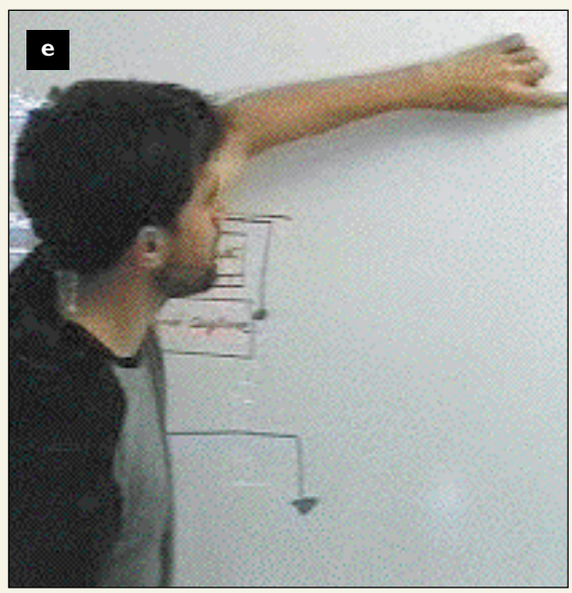
(a) The apparatus of the Magic Board; (b) Selecting a physical drawing with the finger; (c) Copying the selected drawing; (d) Completing the drawing with physical markers; (e) The menu at the top of the physical board to facilitate reinitialization.









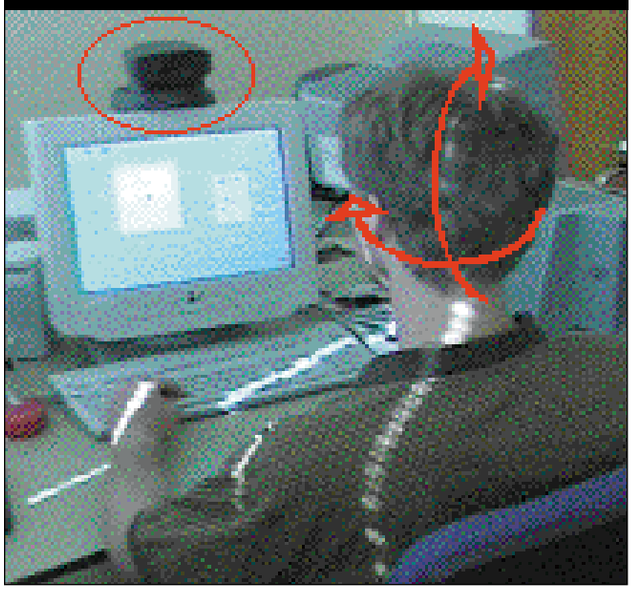


PUI

What Can Machine Vision Do For You?

Machine vision is the observation of an environment using cameras. It differs from image processing in that it extracts information from images that are relevant for a particular set of services. The basic services that machine vision can provide to HCI include detection, identification, and tracking. Detection determines the presence or the absence of an entity of a given type. For example, is there a cat in the scene? Identification is recognizing what entity of the class is present in the scene, for example, that my cat Garfield is in the scene. Tracking is determining the location of an entity over time.

Figure 2. The Perceptual Window uses small head motions as a second input stream to navigate within a document.



In the context of HCI, entities of interest include:

- Real-world objects used as instruments such as the physical icons inspired from the Bricks or the eraser of the Digital Desk;
- The human body including fingers, hands and feet of a dancer, the face of the Perceptual Window presented in a following section [1], or the whole body in Krueger's VideoPlace [9];
- Activities such as typing, standing up, walking, entering a doorway or shaking hands [3].

Techniques may be designed to observe a single entity, couples, or larger groups. For example, interaction envisioned for the Digital Desk requires the management of multiple types of entities such as an eraser, a finger, and a pen. In collaborative settings, the system may have to distinguish between the hands belonging to different persons or between the two hands of the same person.

Based on the preceding fundamental services, machine vision can be exploited in multiple ways. First, human action or objects can be tracked without being constrained by cumbersome wires as with the data glove, the body suit, or magnetic localizers such as an Ascension Flock of Birds™. Second, machine vision can extend human visual abilities by delegating those tasks to the system that are hard or impossible to perform, such as monitoring remote sites. Third, machine vision can be exploited to improve directness in the interaction process by suppressing mouse-like intermediate instruments: when tracked in real time, your finger becomes an input device. As opposed to the mouse, there is no need to grasp it. It is already in your hand! We illustrate the property of directness with two different systems: the Magic Board and the Perceptual Window.

The Magic Board

The Magic Board, shown in Figure 1, is a physical whiteboard combined with a video projector and a steerable camera to provide a simple augmented workspace. Like recent electronic commercial smart boards, it allows the combined usage of electronic ink with physical dry markers and conventional erasers. Unlike them, it does not provide sophisticated services. We wanted it to maintain the natural affordance of the existing tools. Therefore, the physical board is augmented with the minimal electronic editing functions used for brainstorming (such as select, copy, move, and save).

Magic Board also does not capture physical ink on the fly but at specific points in the interaction process (such as copying and saving) supporting fast deletion of markings with the usual physical tools (your hand, a handkerchief, you name it). It is not limited by the resolution of the sampling system: drawing with markers can be done at any speed and with any pressure without any loss of information. Finally, because the working surface is observed by a steerable camera, it is not limited in size. In addition, any white (or black) surface can be used as a production space. It is also possible to capture and digitize material written on pieces of paper such as Post-Its.™

Figure 1 demonstrates the principles of the interaction supported by the Magic Board. Whereas the Magic Board uses machine vision and image processing for tracking a finger and capturing the content of the board at a high resolution, the Perceptual Window [1] uses machine vision for head tracking.

The Perceptual Window

The Perceptual Window (Figure 2) offers a novel interaction technique using head motions to control the 2D location of a window viewpoint within a document (see iihm.imag.fr/demos/pwindow/). The Perceptual Window is not an eye tracker. Indeed, eye movements are poorly adapted to motion control. Natural eye movements alternate between short periods of fixation and rapid saccades, which tend to respond to involuntary reflexes. While fixation can be used for selection, saccadic motions are too rapid and involuntary for motion control. We have found that head motions provide a much more natural form of command for scrolling.

On standard workstations, scrolling is typically allotted to the mouse and scroll bars. The Perceptual Window offers multiple forms of scrolling techniques. One possibility is to control the rate of scroll while the mouse can be used for another task (say, selection). As the head is tilted upward outside of a neutral area, the window content is scrolled down. The rate of scrolling is determined by the angle of the head. Returning the head inside the neutral area stops upward scrolling. Tilting the head downward, left or right, or even diagonally, induces a similar scrolling actions. Scrolling speed is governed by an exponential function of position, permitting both accurate adjustments and fast scrolling depending on the amount of head movement.

The novelty of the perceptual window results from the use of head motion to establish the context for interaction. In the mid-1980s, Guiard demonstrated that using two hands improves performance provided the hands are used asymmetrically [7]. In such motion, the nondominant hand defines the frame of reference for the dominant hand. The nondominant hand moves first, followed by the dominant hand. The nondominant hand executes coarse-grained motions whereas the dominant hand is allocated to fine-grained actions.

In the Perceptual Window, the hand and mouse form the dominant stream and the head is used as a nondominant stream: the head sets the window viewpoint for the mouse workspace, it moves first, and window viewpoint does not have to be set accurately. As predicted by Guiard's theory, head motion interaction significantly outperforms scrollbars (by an average improvement of 32% on task completion time) [1].

The Perceptual Window and the Magic Board illustrate how HCI can benefit from machine vision. However, this is made possible only if human-centered requirements are satisfied.

Human-centered Requirements

To be usable, machine vision must be robust and autonomous. Designing robust and autonomous interactive systems for real-world environments is much more difficult than constructing systems for controlled laboratory settings. In the real world, illumination and background conditions may change in an abrupt manner, and users may behave in unexpected ways. Furthermore, when tightly coupled to human actions, response time must conform to human action-perception skills.

Robustness and autonomy. A machine vision system is robust if it does not break in the presence of disturbances. It is autonomous if it is capable of detecting failures and correcting problems without the explicit intervention of the user. Robustness requires reconfiguration and reinitialization to accommodate new operating conditions. If reconfiguration or reinitialization requires human intervention, then the user will be interrupted from the central task, and the usability of the system will be seriously degraded. Thus usability requires both robustness and autonomy.

In current systems, acceptable compromises are devised on a case-per-case basis. For example, in the Magic Board, each time a user selects a menu with his finger, the finger tracker is initialized (see Figure 1e). In the media space CoMedi [4], vision processes for face tracking are reinitialized whenever the user blinks and reconfigured dynamically as explained here. In both cases, reinitialization is integrated into system operation so that adaptation is transparent (or nearly transparent) to the user. On the other hand, the Perceptual Window was designed as a laboratory experiment and must be initialized manually. The lack of autonomy for the current implementation of the Perceptual Window makes it unsuitable for use in the real world.

Tightly coupled interaction: It's latency that counts. When a human and an artificial system are bound in a continuous manner in the accomplishment of actions that are mutually dependent and mutually observable, they are said to be "tightly coupled." For example, in the Magic Board, the user and the finger tracker are tightly coupled while the user selects a mark with the finger. In the Perceptual Window, the user and the head tracker are tightly coupled when the user performs scrolling tasks.

In tightly coupled interaction, the artificial system and the human processor form a closed loop with behavior that can be formally analyzed using analytical tools developed for Control Theory. Latency (or lag) is a key parameter for closed-loop systems. Using the Model Human Processor [2], we have been able to estimate that the latency of machine perception must be less than 50ms for direct manipulation using finger tracking. This prediction is backed up by empirical results from Ware et al. performed on a Polhemus-

based head tracker developed for their Fish Tank system [11].

Improper system latency leads to redundant actions and oscillation. For example, in the absence of immediate system feedback, the user may attempt to make improper corrections. These corrections then lead to undesired system responses, which the user may further attempt to correct. This condition can rapidly drive a system to divergence or oscillation.

Although latency is paramount to the usability of tracking systems, stability, resolution, and precision are also important characteristics to consider when developing a perceptual system for HCI. Using a camera for improving directness in computer-human interaction is one thing. Being watched by remote peers brings to bear privacy issues.

Privacy protection. Privacy protection has been addressed in video communication used in media spaces. Media spaces use low bandwidth video communications to provide informal communication and group awareness among geographically dispersed members of a team [4]. To be socially acceptable, a media space must support privacy. Privacy filters include reduced resolution, shadowing, temporal difference images, and the Eigen-space filter.

The Eigen-space filter uses principal component analysis (PCA). An orthogonal set of basis images are determined by PCA of a set of "socially correct" images. Live images are coded by computing the inner product with the basis images. This technique is made possible by keeping the user centered in the image using automatic face tracking to drive a steerable camera.

An interesting property of Eigen-space coding is that only information within the original image set will be captured by the coding and reconstructed in the resulting images. For example, in Figure 3a, the source image (left) shows François with his finger in his nose. This socially incorrect gesture does not belong to the basis space and is not displayed in the reconstructed image (right). Similarly, persons appearing in the background are not communicated unless present in the basis.

Eigen-space coding also allows a user to animate a database composed of images of a face of a different person or character, raising ethical issues. For example, the image of François (left in Figure 3b) is reconstructed as Fred (to the right). This property can be useful when a user wishes to always appear properly groomed, but can also raise ethical problems if users masquerade as other people.

Human-centered requirements may be incompatible with available computational resources and vision techniques. Robustness and response time may be mutually satisfied at the cost of accepting constraints on the operating conditions. Such constraints are acceptable provided they preserve the ontology (that is, the raison d'être) of the system.

As a consequence of the requirement for real-time response, machine vision for HCI must use simple, minimal vision techniques. One method for simplifying machine vision is to exploit explicit constraints on the user or on the operating environment. Constraints are acceptable provided they satisfy the following criteria:

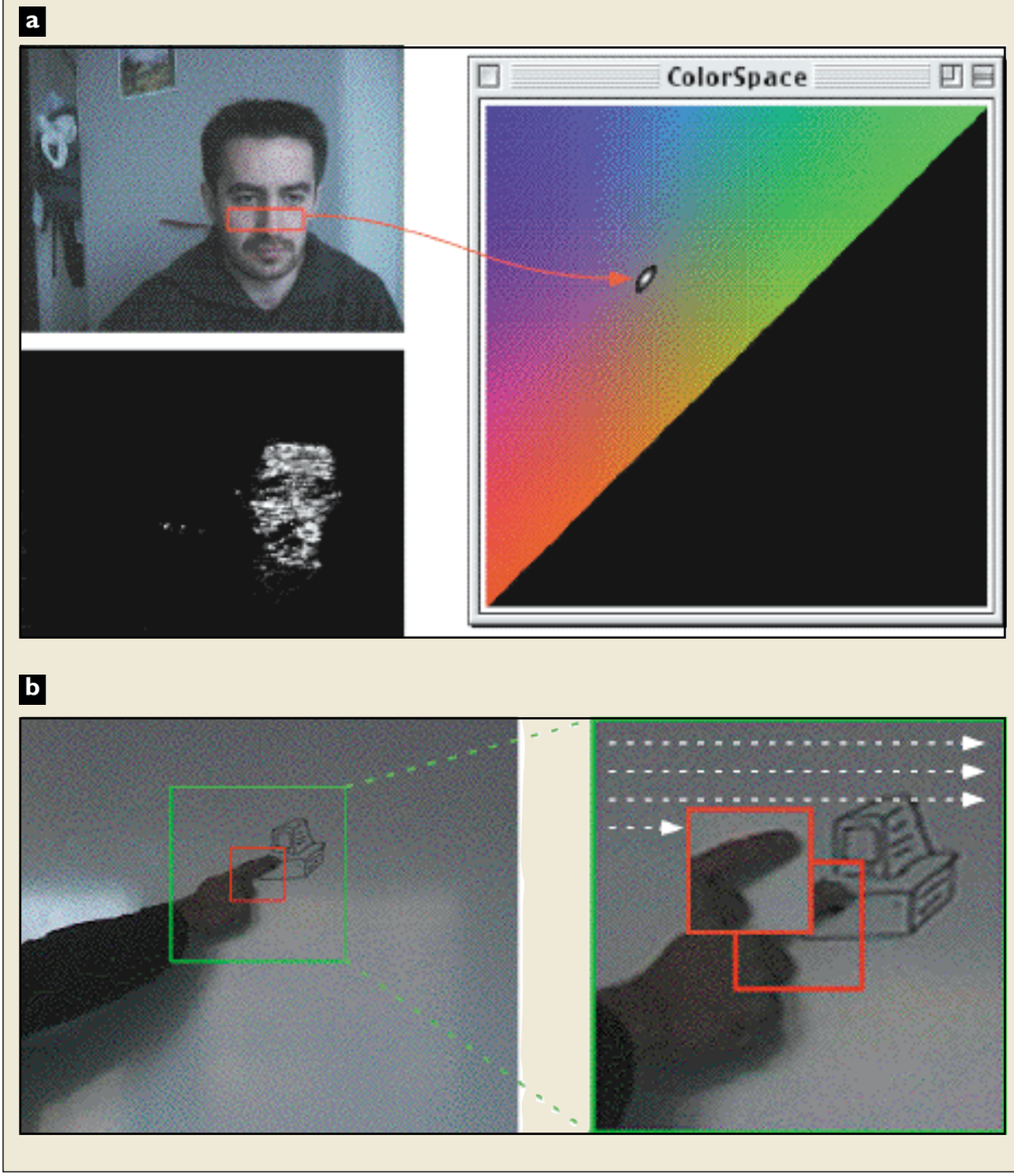
- The constraint should simplify the processing required, thereby speeding up the system response; and
- The constraint must not render the system unusable. Constraints must conform to the ontology of the system.

Figure 3. Eigen-space filtering for supporting privacy.

(a) Eigen-space filtering for private video space. An original image (left) with its reconstruction using a socially correct image set;

(b) An original image of François (left) and the resulting reconstruction (right) using Fred's image set.

Figure 4. Popular machine-vision techniques for HCI used in the appearance-based approach to computer vision: (a) Skin color detection. A sample of skin (small red rectangle in the top left image) is used to construct a color space (right). The oval area in the color space denotes pixels that are skin color. Pixels in the bottom left image denote the probability of skin. White pixels indicate skin color region; (b) Cross-correlation. A template is acquired of the item to be tracked (for example, the finger for the Magic Board, red rectangle). The template is compared to the image at positions within a search region (green rectangle, englarged on the left side). The image position at which the template most matches the image is selected and defines the position of the search region in Figure 5.



The KidsRoom

Aaron F. Bobick, Stephen S. Intille, James W. Davis, Freedom Baird, Claudio S. Pinhanez, Lee W. Campbell, Yuri A. Ivanov, Arjan Schütte, and Andrew Wilson

Computer vision sensing technologies turn a child's bedroom into a dreamy wonderland.

The KidsRoom is a fully automated and interactive narrative playspace for children developed at the MIT Media Laboratory. Built to explore the design of perceptually based interactive interfaces, the KidsRoom uses computer vision action recognition simultaneously with computerized control of images, video, light, music, sound, and narration to guide children through a storybook adventure. Unlike most previous work in interactive environments, the KidsRoom does not require people in the space to wear any special clothing or hardware, and the KidsRoom can accommodate up to four people simultaneously. The system was designed to use computational perception to keep most interaction in the real, physical space even as participants interacted with virtual characters and scenes.

The KidsRoom, designed in the spirit of several popular children's books, is an interactive child's bedroom that stimulates imagination by responding to actions with images and sound to transform itself into a storybook world. Two of the bedroom walls resemble the real walls in a child's room, complete with real furniture, posters, and windows. The other two walls are large, back-projected video screens used to transform the appearance of the room environment. Four speakers and one amplifier project steerable sound effects, music, and narration into the space. Three video cameras overlooking the space provide input to computer vision people-tracking and action recognition algorithms. Computer-controlled theatrical lighting illuminates the space, and a microphone detects the volume of enthusiastic screams. The room is fully automated.

During the story, children interact with objects in the room, with one another, and with virtual creatures projected onto the walls. Perceptual recognition makes it possible for the room to respond to the physical actions of the children by appropriately moving the story forward thereby creating a compelling interactive narrative experience. Conversely, the narrative context of the story makes it easier to develop context-dependent (and therefore more robust) action recognition algorithms.

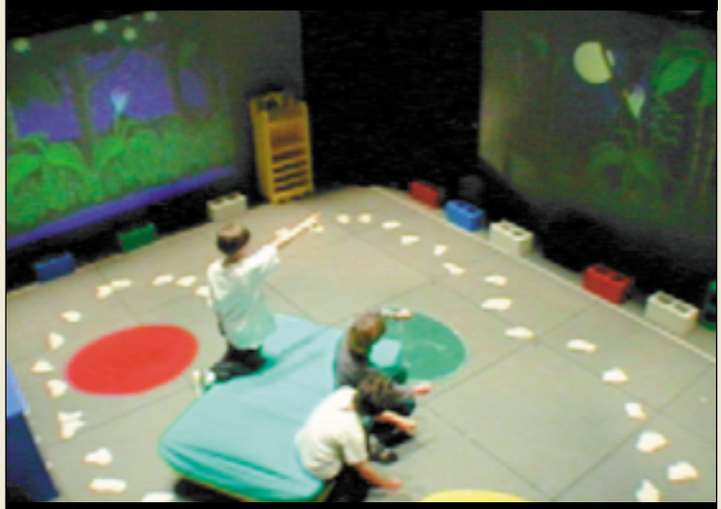
The story developed for the KidsRoom begins with a normal-looking bedroom. Children enter after being told to find out the magic word by asking the talking furniture that speaks when approached. When the children scream the magic word loudly, sounds and images transform the room into a mystical forest. The story narration prods the children to stay in a group and follow a path to a river (see the stone path (a) in the figure). Along the way, they encounter roaring monsters and must hide behind the bed to make the roars subside. After a short walk, the children reach the river world, and the narrator informs them the bed has become a magic boat that will take them on an adventure. The children climb on the "boat" and paddle to make it move, which is represented by images of the river flowing by on the screens. To avoid obstacles in the river, the children must row collaboratively on the appropriate side of the bed. Finally, the children reach the monster world. The monsters appear and teach the children some dance steps, and then the monsters mimic the children as the children perform these steps. The story ends when an insistent, motherly voice off in the distance urges the children to return to bed, at which point the room transforms back to a normal bedroom. A typical interaction runs nearly 12 minutes.

Throughout the adventure, the computer system tracks the positions of the movable bed and up to four children. The system detects and responds to events like "Is everyone on the bed?" "Is everyone near the chest?" "Are the children in a group?" and "Are the children following the path?" The music, sound, and narrative of the story change depending upon what the children are doing. For example, if the children fail to get on the bed, characters in the story encourage them to do so. The vision systems use the context established by the story (for example, that everyone is on the bed) for robust initialization and performance. Although the storyline is linear, the room continually reacts to the children's actions, giving the environment an interactive feel. During the river scene, the vision system determines the side of the bed with the highest motion energy and uses this information to "steer" the bed as the children use their arms to row down the virtual river. In the monster world, the still-frame animated cartoon monsters teach the children four different dance moves (for example, "spin around like a top"), after which the children can

perform any step. The vision system is trained to recognize these dance moves, which then triggers the corresponding animations of the monsters with encouraging character narrations. When the vision processing requires constraints (for example, people in certain positions), they were built naturally into the storyline. For instance, the monsters tell the kids to stand on particular rugs "so's we can see ya;" this storyline device actually ensures that each camera has a nonoccluded view of each child.

The KidsRoom demonstrated that nonencumbering, computer-vision sensing technologies can be used to automatically create new types of physical interactive experiences in real environments by integrating sensing and narrative control. We believe the KidsRoom is the first multiperson, fully automated, interactive, narrative playspace ever constructed, and the experience we acquired designing and building the space has allowed us to identify some major questions and to propose a few solutions to simplify construction of more complex spaces in the future. C

(a) A view of the KidsRoom showing the two projection screens and the movable bed.



(b) A child and mother rowing the boat together. Rowing was detected using story context and motion energy.

For sound, image, and video clips of the KidsRoom, see vismod.www.media.mit.edu/vismod/demos/kidsroom. For more information on the KidsRoom and the sensing technologies that were employed see [1]. A simplified reimplementation of the KidsRoom is on display at the Millennium Dome in London.

References

1. A.F. Bobick, S.S. Intille, J.W. Davis, F. Baird, L.W. Campbell, Y. Ivanov, C.S. Pinhanez, A. Schütte, and A. Wilson. The KidsRoom: A perceptually-based interactive and immersive story environment. *PRESENCE: Teleoperators and Virtual Environments 8*, 4 (Aug. 1999), 367–391.

Aaron Bobick (afb@cc.gatech.edu) is an associate professor of Computer Science in the College of Computing and Associate Director of the GVU Center at the Georgia Institute of Technology in Atlanta, Ga.

Stephen Intille (intille@mit.edu) is a research scientist at the Massachusetts Institute of Technology in Cambridge, Mass.

Jim Davis (jdavis@media.mit.edu) is a Ph.D. student in the MIT Media Lab's Perceptual Computing section in Cambridge, Mass.

Lee Campbell (elwin@media.mit.edu), is a Ph.D. student in the MIT Media Lab's Perceptual Computing section in Cambridge, Mass.

Claudio Pinhanez (pinhanez@us.ibm.com) is a research scientist at the IBM TJ Watson Research Center in Yorktown Heights, New York.

Freedom Baird (baird@media.mit.edu) designs and builds instruments and songs for a band in Cambridge, Mass.

Yuri Ivanov (yivanov@media.mit.edu) is a Ph.D. student in the MIT Media Lab's Perceptual Computing section in Cambridge, Mass.

Arjan Schütte (arjan@mulchmedia.com) is partner with Mulch Media and a consultant to the Internet education industry.

Andy Wilson (drew@media.mit.edu) is a Ph.D. student in the MIT Media Lab's Perceptual Computing section in Cambridge, Mass.

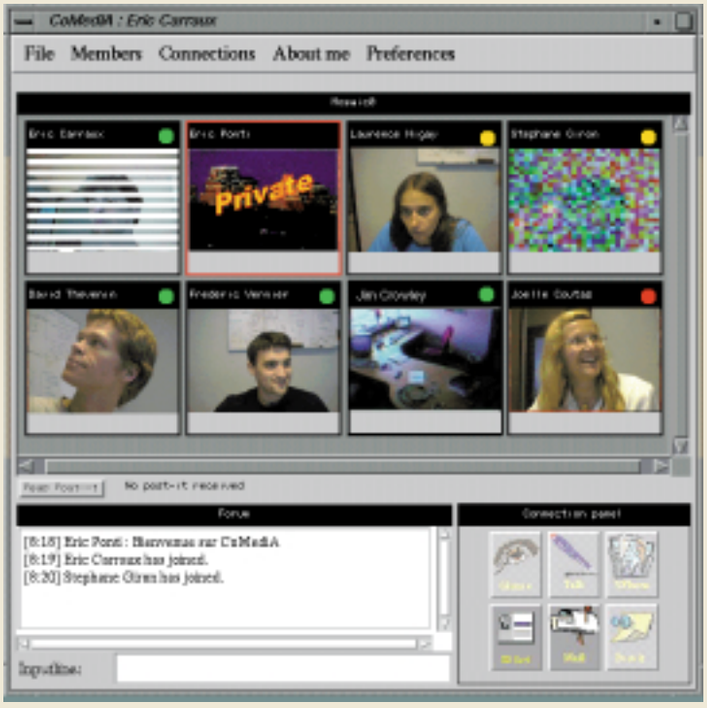
These principles are illustrated in VideoPlace and the Magic Board. VideoPlace requires a luminous background and is limited to a single user at a time. In the Magic Board, a single finger is tracked at a time. These restrictions simplify image processing without threatening the ontology of the system: users can come as they are. Other possible constraints include limitations on the speed of movements or the wearing of special clothing. These constraints, which limit participant's skills and freedom, would not be acceptable options in the context of VideoPlace and the Magic Board, but can be useful in other applications.

So far, we have presented the benefits HCI designers can expect from machine vision and have made explicit the requirements machine vision developers must address in order to provide usable techniques for real-life settings

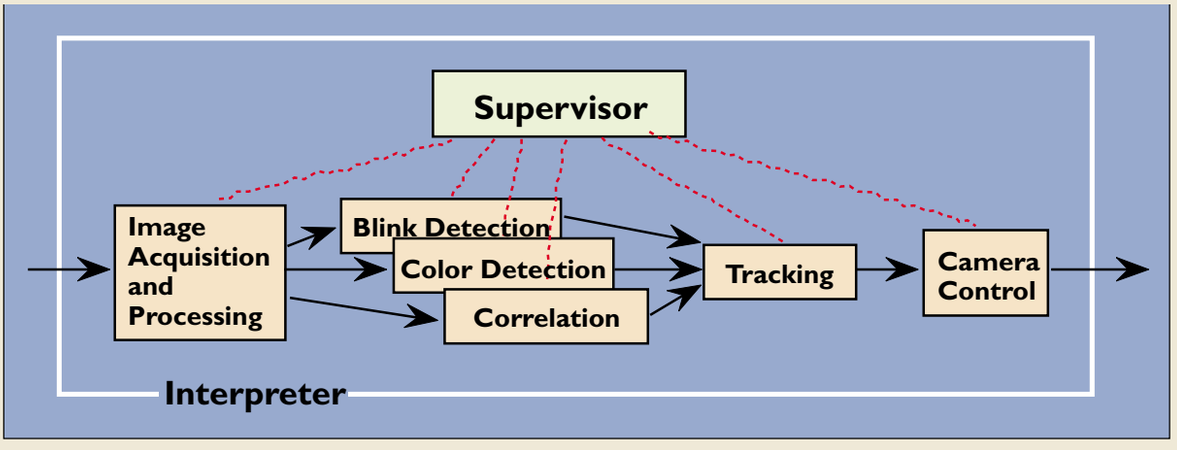
Machine Vision Techniques for HCI

Since the early 1990s, machine perception of human action has been driven along a learning curve by several factors. The dominant driving force has been

Figure 5. The media space CoMedi—Its porthole and the reactive system used for tracking faces in a robust and (nearly) autonomous manner:



(a) The user's porthole for CoMedi. The user can observe that Eric Carraux is hiding behind the Venetian blind whereas Eric Ponti has selected complete privacy. Stephane, top right, is using the blurring filter to protect his privacy, while Cyril has departed the Media Space. Eric, David, and Frederic are available; the dot in their slot is green. Laurence and Stephane display yellow dots indicating they are busy. A red dot indicates Joëlle must not be disturbed. At the bottom of the display, the forum window displays events relevant to awareness and can be used to chat by typing messages. The connection panel gives access to additional communication means such as email or telephone.



(b) A supervisory controller selects and controls the sequencing of perceptual processes. Multiple processes can be active at the same time.

inexpensive computing power. An additional influence has been the introduction of image acquisition hardware in personal computers, which has dramatically lowered the investment required to experiment with real-time computer vision systems.

Exponential growth in experimentation has led to the development of techniques that provide reliable and reproducible results at close to video rates. An important advance in image analysis and description has been provided by appearance-based methods. Appearance-based methods measure scene information directly from images without attempting 3D reconstruction.

Appearance-based vision. Most appearance-based methods use previously observed images as models or templates. Such techniques tend to be fast and simple to compute, making them popular tools for building systems for observing human action. Popular techniques include active contours [8], skin color detection [10], and cross-correlation.

An active contour iteratively computes a balance between external forces that attract it to high contrast and internal forces that maintain connectivity. Applying principal components analysis to the points of active contours under motion leads to simple models that can track faces, lips, and hands in real time.

Skin color detection using a ratio of histograms of normalized color can be programmed using table look-up, making it possible to detect and track skin colored regions in real time (see Figure 4a). Cross-correlation uses small regions of an image as templates for searching in later images (see Figure 4b).

The recognition of face expressions, gestures, or pedestrian movements can be formulated as a process of recognizing trajectories. Hidden Markov models (HMMs) provide a formalism for recognizing trajectories that represent gestures, facial expressions, or human activity. Recent progress has extended this model to coupled trajectories of ensembles of objects or people.

Autonomy and robustness require methods for integration and control of continuously operating perceptual processes. Integration and control may be provided by reactive systems.

Reactive systems. A reactive perception system can be composed from a set of perceptual processes integrated in an event-driven architecture. Perceptual processes are formalized as cyclic transformations from sensor data to symbolic events and property vectors. They are initiated, controlled, and terminated by a supervisor that reacts to events and serves as a scheduler and resource allocator.

Symbolic events are messages that assert information about the environment or about the state of a sensor or a perceptual process. Examples of symbolic events include the arrival of a person at a door, the grasping of an object on a table, and assertions about the failure of a sensor. Property vectors are used to control devices, to communicate commands, to enter information in digital form, or to adapt perceptual processes to environmental conditions. An example of a property vector is the estimate of the position, orientation, and size of a face in an image, which can then be used to steer the pan, tilt, and zoom of a camera. Another example is a property vector that gives the position and identity for a set of persons in a room. A third example is the vector of position, orientation, and velocity of an object such as a Brick [6].

The face tracking system [5], developed for the media space CoMedi is an example of a reactive system in which three complementary visual processes are initialized and controlled by a supervisor based on events. These processes are eye blink detection, skin color detection, and cross-correlation, as shown in Figure 5b. Eye blink detection provides an estimated position of the face in the image that can be used to initialize procedures for skin detection and cross-correlation. Cross-correlation with a reference template provides fast and accurate tracking, but fails when the head turns or moves too fast. Skin color detection provides robust face tracking, but is slower and less precise than correlation. Both processes are easily reinitialized by blink detection, providing a system that continually adapts to users and their environments and is sufficiently robust and rapid to allow "natural" displacements.

Trends and Challenges

So, where does convergence of perception, communication, and computation lead? Although details are impossible to predict, we can predict general tendencies based on the forces driving innovation and technological evolution.

Perception of human action is expected to be a key component in the next generation of tools for man-machine interaction. Such methods may permit humans to interact with machines in a natural manner similar to the way that humans interact between themselves. The key to perceptual user interfaces is usability. Usability determines requirements for technological innovation. Interaction based on machine perception will most likely evolve most rapidly in areas where traditional GUI interfaces are inappropriate. Kiosks for commercial and informational services are an obvious example. Another area is intelligent spaces that may evolve from video surveillance.

Socioeconomic conditions will drive initial development in perceptual environments to areas where

performance gains are most easily obtained, and where resources are most easily rewarded with return on investment. Commercial and office environments provide a particularly fertile area for this technology. Currently, the highest growth is in video surveillance for security. Video surveillance can potentially provide operational feedback to commercial and business managers permitting more effective design of product presentations and pedestrian passages. It can also provide important information to product designers. However, such applications will require that users be assured of privacy protection. Perception of human action can provide the means to extract commercially interesting information about the way people interact with products and displays without revealing identity and without storing or communicating images.

Another area with near-term growth potential occurs with the convergence of communications and perception. In the media space, low-bandwidth video communications allow geographically distant workers to dynamically form work teams by providing continual informal communications. However privacy protection and restrictions on movement make such applications impractical. Perception of human actions allows users to freely move about their environment while also providing tools for protecting privacy. Eventually, as such technology matures, media spaces will allow geographically separated families, including the aged, to share presence while protecting individual privacy. The long-term market size for domestic applications of media spaces is a large percentage of the human population.

In the long run, media spaces and commercial surveillance services will be dwarfed by applications related to intelligent spaces. When your office, car and home know your habits and observe your activities, many common chores can be provided automatically. For example, your car can tell your home to turn on the heat, and begin preparations for dinner when you head for home at the end of the day. Your home can coordinate delivery of automatically ordered products to occur while you are at home. Automated cleaning devices can be triggered during your absence. At work or play, any physical device can be used to communicate to the digital world simply by the way it is manipulated.

Conclusion

We have described existing machine vision techniques for detecting events, measuring properties, recognizing and tracking of humans and their actions, and shown how such processes can be integrated into an event-driven architecture. Nevertheless, an important gap remains between these techniques and "machines that see." This gap can be summarized by the word *awareness.*

A machine can be said to be aware when it maintains a description of the location, identity, and roles of objects and actors in its environment. Awareness goes beyond perception to include autonomy, adaptation, and man-machine interaction. A machine perception system will only be accepted as aware when it can communicate in a functionally useful manner with users. Machine awareness represents one of the grand challenges for information technology, on a scale with electric light, the telephone, or powered flight. The long-term impact on human quality of life may be enormous. C

References

1. Bérard, F. The perceptual window: Head motion as a new input stream. In *Proceedings of the IFIP Conference on Human-Computer Interaction (INTERACT99).* A.M. Sasse and C. Johnson, Eds. IOS Press (1999), 238–244.
2. Card, S., Moran, T., and Newell, A. *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum, 1983.
3. Chomat, O., and Crowley, J.L. Probabilistic recognition of activity using local appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR '99.* (June 1999, Fort Collins, CO) IEEE Press, NY.
4. Coutaz, J., Bérard, F., Carraux, E., Astier, W., and Crowley, J.L. CoMedi: Using computer vision to support awareness and privacy in mediaspaces. In *Proceedings of ACM conference on Computer-Human Interaction (CHI).* Extended abstracts (Video demo), (1999), 13–14.
5. Crowley, J.L. and Bérard, F. MultiModal tracking of faces for video communications. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR '97.* (June 1997, San Juan) IEEE Press, NY.
6. Fitzmaurice, G., Ishii, H., and Buxton, W. Bricks: laying the foundations for graspable user interfaces. In *Proceedings of CHI95* (1995) ACM Press, NY, 442–449.
7. Guiard, Y. Asymmetric division of labor in human skilled bimanual action: The kinematic chain as a model. *J. Motor Behavior 19*, 4 (1987), 486–517.
8. Kass, M., Witkin, A., and Terzopoulos, D. Snakes: Active contour models. In *Proceedings of the First International Conference on Computer Vision.* (1987), 259–268.
9. Krueger, M. *Artificial Reality II.* Addison Wesley, Reading, PA, 1991.
10. Schiele, B., and Waibel, A. Gaze tracking based on face color. In *Proceedings of the International Workshop on Automatic Face and Gesture Recognition.* (1995, Zurich).
11. Ware, C. and Balakrishnan, R. Reaching for objects in VR displays: Lag and frame rate. *ACM Trans. on Computer-Human Interaction (TOCHI) 1,* 4 (1994), 331–356.
12. Wellner, P., Mackay, W., and Gold, R. Computer-augmented environments: Back to the real world. *Commun. ACM 36,* 7 (July 1993).

James L. Crowley (James.Crowley@imag.fr) is a professor of computer science at the National Polytechnique Institute of Grenoble France and directs the research group PRIMA, working on computer vision at the GRAVIR laboratory of IMAG.

Joëlle Coutaz (Joelle.Coutaz@imag.fr) is a professor of computer science at the University Joseph Fourier in Grenoble France and directs the "Engineering HCI" research group at the CLIPS-IMAG laboratory.

François Bérard (Francois.Berard@media.mit.edu) is a post-doctoral researcher at the MIT Media Lab. He completed his doctoral dissertation at CLIPS-IMAG under the direction of Joëlle Coutaz and James L. Crowley.