DOI:10.1145/1897816.1897838

Body posture and finger pointing are a natural modality for human-machine interaction, but first the system must know what it's seeing.

BY JUAN PABLO WACHS, MATHIAS KÖLSCH, HELMAN STERN, AND YAEL EDAN

Vision-Based Hand-Gesture Applications

THERE IS STRONG evidence that future human-computer interfaces will enable more natural, intuitive communication between people and all kinds of sensor-based devices, thus more closely resembling human-human communication. Progress in the field of human-computer interaction has introduced innovative technologies that empower users to interact with computer systems in increasingly natural and intuitive ways; systems adopting them show increased efficiency, speed, power, and realism. However, users comfortable with traditional interaction methods like mice and keyboards are often unwilling to embrace new, alternative interfaces. Ideally, new interface technologies should be more accessible without requiring long periods of learning and adaptation. They should also provide more natural human-machine communication. As described in Myron Krueger's pioneering 1991 book *Artificial Reality*,[27] "natural interaction" means voice and gesture. Pursuing this vision requires tools and features that mimic the principles of human communication. Employing hand-gesture communication, such interfaces have been studied and developed by many researchers over the past 30 years in multiple application areas. It is thus worthwhile to review these efforts and identify the requirements needed to win general social acceptance.

Here, we describe the requirements of hand-gesture interfaces and the challenges in meeting the needs of various application types. System requirements vary depending on the scope of the application; for example, an entertainment system does not need the gesture-recognition accuracy required of a surgical system.

We divide these applications into four main classes—medical systems and assistive technologies; crisis management and disaster relief; entertainment; and human-robot interaction—illustrating them through a set of examples. For each, we present the human factors and usability considerations needed to motivate use. Some techniques are simple, often lacking robustness in cluttered or dynamic scenarios, indicating the potential for further improvement. In each, the raw data is real-time video streams of hand gestures (vision-based), requiring effective methods for capturing and processing images. (Not covered is the literature related to voice recognition and gaze-tracking control.)

» key insights

- **Gestures are useful for computer interaction since they are the most primary and expressive form of human communication.**
- **Gesture interfaces for gaming based on hand/body gesture technology must be designed to achieve social and commercial success.**
- **No single method for automatic hand-gesture recognition is suitable for every application; each gesture-recognition algorithm depends on user cultural background, application domain, and environment.**

Basic Communication Form

We humans use gestures to interact with our environment during the earliest stages of our development. We also communicate using such gestures as body movement, facial expression, and finger pointing. Though much has been written about gesture interfaces, interface technology rarely adopts this media; consequently, expressiveness and naturalness elements are missing from most user interfaces. Hand-gesture applications provide three main advantages over conventional human-machine interaction systems:

Accessing information while maintaining total sterility. Touchless interfaces are especially useful in healthcare environments;

Overcoming physical handicaps. Control of home devices and appliances for people with physical handicaps and/or elderly users with impaired mobility; and

Exploring big data. Exploration of large complex data volumes and manipulation of high-quality images through intuitive actions benefit from 3D interaction, rather than constrained traditional 2D methods.

Human-robot interaction is another application where the main motivation for gesture-based systems is to have this communication resemble natural human dialogue as much as possible. For example, imagine how intuitive it could be to use hand gestures to tell a robot what to do or where to go. Pointing to a dust spot to indicate "Clean that spot," users would be able to tell a Roomba robotic vacuum cleaner what to do next. Finally, gestures provide a source of expressiveness when immersed in realistic video games. Some notable technologies (such as Microsoft Kinect, Sony PSP, and Nintendo DS and Wii) include gesture recognition in their consoles. Unfortunately, only dynamic gestures (such as waving and fist hitting) are recognized so far. Dynamic hand-shape recognition, as in American Sign Language, remains a challenge.

Costs/Benefits

The appeal of gesture interfaces derives partly from their flexibility and customizability. Still, many requirements as to their functionality and performance are the same throughout most classes of use. As devices and hand-gesture interfaces proliferate as a result of inexpensive cameras and computational power, questions concerning market acceptance also become more frequent. Here are the basic requirements, though they are likely to vary depending on application:

Price. Better camera quality, framerate, distortion, and auto-shutter speed yield better performance but higher cost. Some inexpensive methods for achieving 3D reconstruction (such as flashing IR LED illuminators from multiple angles) can replace stereo cameras. But the sum of the prices for discrete hardware components can add up for the typical consumer, as well as for a manufacturer. The cost of more advanced sensors and sensor setups must be weighed against any potential performance benefit.

Challenges. Given a fixed budget, the challenge for the developer is to decide how the development budget should be spent and, especially, which hardware the system cannot do without.

Responsiveness. The system should be able to perform real-time gesture recognition. If slow, the system will be unacceptable for practical purposes. In 1963, Sheridan and Ferrell[43] found maximum latency between "event occurrence" and "system response" of 45ms was experienced by most of their human test subjects as "no delay." Starting at 300ms, an interface feels sluggish, possibly provoking oscillations and causing a symptom known as "move and wait."

Challenges. Simple, computationally efficient features are of great interest to machine-vision researchers, though more effective techniques must still be developed.

User adaptability and feedback. Some systems are able to recognize only a fixed number of gestures selected by the system designer; others adapt to a nuanced spectrum of user-selected gestures. The type of gesture selected depends on the application; for example, in video games, learning gestures is part of a gratifying experience playing the game. In either case, feedback indicating the correctness of the gesture performed is necessary for successful interaction.

Challenges. Most hand-gesture systems have a core algorithm trained offline (not in real time). Training a classifier online requires a fast, flexible online learning algorithm capable of generalizing from a few training samples. Presenting feedback to the user without increasing cognitive load is an additional problem.

Learnability. Gesture patterns (the lexicon) used to control applications must be easy to perform and remember. These factors are strongly associated with learning rate and "memorability" indices, as reported by Wachs.[51]

Challenges. The learning rate depends on task, user experience, and user cognitive skills. Hardly any literature exists on user performance as a function of gesture vocabulary size or user experience. Two exceptions are by Nielsen[34] and by Kela et al.[23] focusing on acceleration-based gestures. A possible solution is to adopt gestures that are natural and intuitive to the user; users are also more likely to remember them.

Accuracy (detection, tracking, and recognition). Among these three main criteria affecting the performance of hand-gesture systems, detection describes whether a hand is in the camera's view. Tracking describes the ability to follow the hand from frame to frame. And recognition is based on how close the hand's trajectories are to learned templates, based on distance metrics, and indicates the level of confusion of the given gesture with other gestures. For this article, we limit ourselves to performance measures for per-frame-analysis as opposed to activity-recognition systems where more complex performance measures are considered.[32]

Challenges. The main challenges for the three performance measures are at the forefront of research in machine vision. Detection is an extremely complex problem due to hand shape, variable lighting conditions, skin color, and hand size. Tracking complications arise from occlusions, cluttered environments, and rapid motions causing motion blur. Addressing these challenges allows good recognition accuracy to follow.

Low mental load. Having to recall gesture trajectories, finger configurations, and associated actions is likely to add to a user's mental load. Another source of mental (and physical) load

is when users' hands cover the display, preventing them from seeing the graphics being guided.

Challenges. The gestures should be simple, temporally short, and natural. For a given set of tasks, users should have to remember at most only a few postures. Iconic representations of gesture-command associations may also help relieve users' mental load.

Intuitiveness. The gesture types selected by interface developers should have a clear cognitive association with the functions they perform. For example, an open palm can represent a "stop" command, a closed fist with thumb up can represent "OK," and a pointing finger can represent the direction to move an object. Few users are able to remember complex shapes and unnatural finger configurations. Intuitiveness is associated with other usability terms (such as learnability and "easy to remember"). Other factors affecting user-gesture choices are general knowledge, cultural environment, and linguistic capability.[51]

Challenges. Intuitiveness is strongly associated with cultural background and experience. A gesture natural to one user may be unnatural to others. Moreover, Stern et al.[46] showed there is no consensus among users regarding gesture-function associations. This problem can be overcome by letting users decide which gesture best represents their intentions. The "Wizard of Oz" paradigm[34] and analytical structured approaches[51] help achieve this representation.

Comfort. Lexicon design should avoid gestures that require intense muscle tension over long periods, a syndrome commonly called "Gorilla arm." Gestures must be concise and comfortable while minimizing stress on the hand. Awkward, repetitive postures can strain tissues and result in pressure within the carpal tunnel. Two types of muscular stress are found: static, the effort required to maintain a posture for a fixed amount of time, and dynamic, the effort required to move a hand through a trajectory.

Challenges. Measuring stress produced by hand gestures is very difficult. For stress-index measures, experiments vary from subjective questionnaires to electronic devices (such as electromyograms) that measure muscle activity. The main obstacle with physiological methods is that muscle potentials are highly variable within subjects and depend on external factors like positioning, temperature, and physiologic state. Instead, analytical approaches help assess stress based on the dynamics of musculoskeletal models.

Lexicon design should avoid gestures that require intense muscle tension over long periods, a syndrome commonly called "Gorilla arm."

Lexicon size and multi-hand systems. For sign languages (such as American Sign Language), hand-gesture-recognition systems must be able to recognize a large lexicon of both single-handed and two-handed gestures. For multi-touch systems, lexicon size plays a minor role. In either case, the challenge is to detect (and recognize) as many hands as possible.

Challenges. The different types of gestures to be recognized must be weighed against the system's robustness. A classifier that recognizes a small number of gestures generally outperforms the same system trained on more gestures. The challenge for the vision algorithm is to select robust features and classifiers such that the system's performance is barely affected by lexicon size. Multi-hand systems pose additional challenges (such as disambiguation of mutual hand occlusions and correctly associating hands and people).

Come as you are.[48] This phrase refers to an HCI design that poses no requirement on the user to wear markers, gloves, or long sleeves, fix the background, or choose a particular illumination. Many methods encumber the user in order to track and recognize gestures by standardizing the appearance of the hands (markers, gloves, long sleeves) but make interaction cumbersome. The challenge for the vision algorithm is to recognize hand gestures without requiring the user wear additional aids or being wired to a device.

Challenges. This flexibility constraint suggests a machine-vision-based solution that is not invasive. The drawback reveals itself with varied environments and user appearance. Assumptions about user characteristics and illumination affect system robustness. Near-IR illuminators can help. Far-IR cameras, ultrasonic, IR laser scanners, and capacitive imagers are also possible approaches for maintaining a system that lets users come as you are.

Reconfigurability. Hand-gesture systems are used by many different types of users, and related hand-gesture interfaces are not "one size fits all." Location, anthropometric characteristics, and type and number of gestures are some of the most common features that vary among users.

Challenges. This requirement is not technically challenging; the main problem is the choice of functionalities within the interface that can change and those that cannot. The designer should avoid overwhelming the user by offering infinite tunable parameters and menus. On the other hand, users should have enough flexibility that they can freely set up the system when a major component is replaced or extended.

Interaction space. Most systems assume users are standing in a fixed place with hands extended (limited by a virtual interaction envelope) and within the envelope recognize gestures. But these assumptions do not hold for mobile ubiquitous hand-gesture-recognition systems where the interaction envelope surrounds only the mobile device.

Figure 1. Head and hand detection using depth from stereo, illumination-specific color segmentation, and knowledge of typical body characteristics.[17]

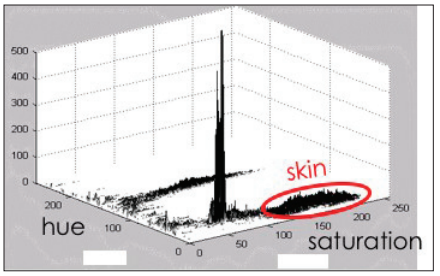


Figure 2. Hue-Saturation histogram of skin color. The circled region contains the hand pixels in the photo; the high spike is caused by grayish and white pixels.

Challenges. Recognition of 3D body-arm configurations is usually achieved through at least two cameras with stereo vision, a setup requiring previous calibration and usually slower response than single-camera-based systems. Monocular vision can be used to disambiguate 3D location using accurate anthropomorphic models of the body, but fitting such a model to the image is computationally expensive.

Gesture spotting and the immersion syndrome. Gesture spotting consists of distinguishing useful gestures from unintentional movement related to the immersion-syndrome phenomenon,[2] where unintended movement is interpreted against the user's will. Unintended gestures are usually evoked when the user interacts simultaneously with other people and devices or just resting the hands.

Challenges. The main challenge here is cue selection to determine the temporal landmarks where gesture interaction starts and ends; for example, hand tension can be used to find the "peak" of the gesture temporal trajectory, or "stroke," while voice can be used to mark the beginning and culmination of the interaction. However, recognition alone is not a reliable measure when the start and end of a gesture are unknown, since irrelevant activities often occur during the gesture period. One solution is to assume that relevant gestures are associated with activities that produce some kind of sound; audio-signal analysis can therefore aid the recognition task.[52]

While responsiveness, accuracy, intuitiveness, come as you are, and gesture spotting apply to all classes of gesture interface, other requirements are more specific to the context of the application. For mobile environments in particular, ubiquity and wearability represent special requirements:

Ubiquity and wearability. For mobile hand-gesture interfaces, these requirements should be incorporated into every aspect of daily activity in every location and every context; for example, small cameras attached to the body or distributed, networked sensors can be used to access information when the user is mobile.

Challenges. Hand-gesture systems that are spatially versatile and adaptable to changing environments and users require self-calibration. Small programmable sensors are expensive, and cross-platform environments have yet to be developed.

In a literature review we undertook as we wrote this article, we found that the requirements outlined here are acknowledged by only a few scientists, including Baudel and Beaudouin-Lafon[2] and Triesch and Malsburg.[48]

Hand-Gesture Recognition

Hand gestures can be captured through a variety of sensors, including "data gloves" that precisely record every digit's flex and abduction angles, and electromagnetic or optical position and orientation sensors for the wrist. Yet wearing gloves or trackers, as well as associated tethers, is uncomfortable and increases the "time-to-interface," or setup time. Conversely, computer-vision-based interfaces offer unencumbered interaction, providing several notable advantages:

- Computer vision is nonintrusive;
- Sensing is passive, silent, possibly stealthy;
- Installed camera systems can perform other tasks aside from hand-gesture interfaces; and
- Sensing and processing hardware is commercially available at low cost.

However, vision-based systems usually require application-specific algorithm development, programming, and machine learning. Deploying them in everyday environments is a challenge, particularly for achieving the robustness necessary for user-interface acceptability: robustness for camera sensor and lens characteristics, scene and background details, lighting conditions, and user differences. Here, we look at methods employed in systems that have overcome these difficulties, first discussing feature-extraction methods (aimed at gaining information about gesture position, orientation, posture, and temporal progression), then briefly covering popular approaches to feature classification (see Figure 1).

Motion. Frame-to-frame comparison against a learned background model is an effective and computationally efficient method for finding foreground objects and for observing their position and movement. This comparison requires several assumptions (such as a stationary camera or image pre-processing to stabilize the video) and a static background; for example, Kang et al.[21] employed the Lucas-Kanade tracking method.[29]

Depth. Range data from a calibrated camera pair[40] or direct range sensors (such as LiDAR) is a particularly useful cue if the user is expected to face the camera(s) and the hands are considered the closest object. Depth from stereo is usually coarse-grain and rather noisy, so it is often combined with other image cues (such as color[17,22,33]). Well-calibrated stereo cameras are costly, and depth can be calculated accurately only if the scene contains sufficient texture. If texture is lacking, artificial texture can be projected into the scene through a digital light projector injecting structured light patterns.[39]

Color. Heads and hands are found with reasonable accuracy based purely on their color.[24,40] Skin color occupies a rather well-defined area in color spaces (such as Hue, Saturation, and Intensity, L*a*b*, and YIQ) so can be used for segmentation (see Figure 2 and Hasanuzzaman et al.,[19] Rogalla et al.,[41] and Yin and Zhu[53]). Combined histogram-matching and blob-tracking with Camshift[7] or the Viterbi algorithm[54] is a popular approach due to its speed, ease of implementation, and performance. Shortcomings stem from confusion with similar-colored objects in the background and limitations with respect to posture recognition. Better optics and sensors often improve color saturation, therefore color-based algorithms; another accuracy boost can be achieved through user-worn markers (such as colored gloves and bright LEDs). While simplifying the interface implementation, these aids do not permit users to "come as you are," so IR illumination can be used instead of markers. The IR light source illuminates users' hands, allowing an IR camera to capture the images of the illuminated parts.[44] In addition, reflective material affixed to a body part can increase the part's reflection properties.

Shape. Many objects can be distinguished by their shape, or silhouette. Different object orientations are often also revealed based on shape alone. Shape is available if the object is clearly segmented from the background scenery, achievable in controlled environments (such as with chroma keying), often for stationary-camera systems (using a background model) and a bit less reliably with a good hand-color model.[53] Popular methods include statistical moments,[13] rule-based methods (see Figure 3 and Kawarazaki et al.[22] and Yin and Zhu[53]), active shape models,[12] and shape context.[4]

Appearance. Methods that consider the intensity and/or color values across a region of interest are more powerful and robust than methods that consider shape alone. Since they do not rely on segmentation, they are generally able to handle situations with no intensity/color distinction between foreground and background. The theoretical upper bound on lexicon size is much greater for appearance-based methods than for purely depth- and shape-based methods. The drawback is increased computational cost during training and recognition; for example, detecting heads in all possible orientations or hands in all possible configurations is not currently possible at interactive frame rates. Examples of appearance-based methods (such as by Viola and Jones[49]) have been employed for various vision-based interfaces, including those reported by Hasanuzzaman.[19]

Multi-cue. Rather than rely on a single image cue, a number of schemes combine information from multiple cues. Motion-based region-of-interest designation, combined with appearance-based hand, face, or body detection, improves speed and accuracy. Appearance and color for detection and motion cues, together with color, were used for hand-gesture interfaces (see Figure 4) by Kölsch et al.[24] and Rauschert et al.[40] Removal of any of these cues was shown to hurt performance. Methods that segment a gesture in an image based on color, then classify the shape, do not fall into this multi-cue category, since the cues are used sequentially, not cooperatively; that is, if the first cue fails, the second cue is useless. True multi-cue systems face similar difficulties with data combination as classic sensor fusion: intra-cue confidence is often unavailable for weighting; the data domains and/or ranges are often distinct; and the combination function may be highly nonlinear.

The extracted features are then subjected to various classifiers, from generic support vector machines[10] to highly customized shape classifiers, as in Yin and Zhu.[53] Some features perform classification implicitly; for example, the Lucas-Kanade-based tracker discards "unreliable" patches, and Camshift[7] determines a decision boundary in space and color histograms.

Classification is sometimes externally combined with feature extraction, as in the boosting approach involving a combination of weak detectors.[49] Other methods involve a distinct translation step into feature space and subsequent classification; for example, consider the motion track of a hand gesture, with its spatial location over time serving as feature vector and a hidden Markov model classifying hand trajectory into various temporal/dynamic gestures[26,33,40,42] (see Figure 5).

As with speech recognition, dy-

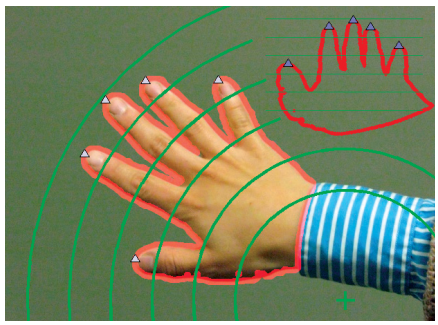


Figure 3. Hand-gesture recognition using color segmentation, conversion into polar coordinates, and maxima detection to identify and count fingers.

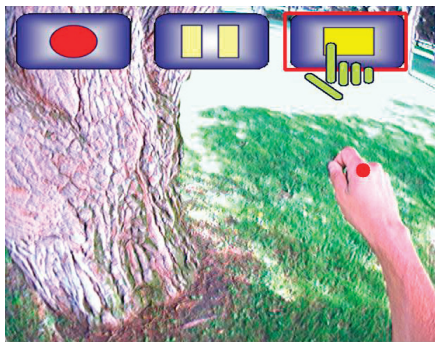
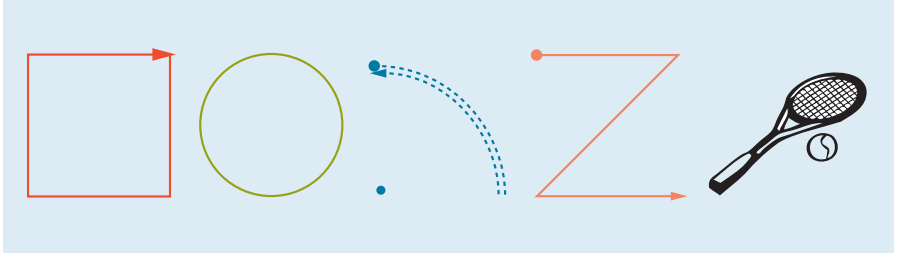


Figure 4. Multi-cue hand tracking and posture recognition.[24]

Figure 5. Motions reliably distinguished by hidden Markov models,[42] from moving the WiiMote in a square to swinging an arm, as in serving with a tennis racquet. The third gesture from the left describes a 90-degree roll angle around the *z*-axis (back and forth).



namic gesture segmentation (when the gesture starts and ends) is a challenge; in gesture research, such temporal segmentation is often called "gesture spotting" (see the section on requirements and challenges). Spotting is not only difficult but necessitates a lag between gesture start and finish (or later), limiting the responsiveness of the user interface. Other successful classification methods are dynamic time warping, Hough transforms, mean-shift and Camshift, and Bayesian approaches.

Applications

The first application of hand-gesture control we review—medical systems and assistive technologies—provides the user sterility needed to help avoid the spread of infection. The second—entertainment—involves naturalness of the interface as part of the user experience. The next—crisis management and disaster relief—involves a performed task requiring quick user feedback. Finally, human-robot interaction must be natural and intuitive for the personal robot of the future. Here, we cover hand-gesture control interfaces for each category and discuss the related design considerations.

Medical systems and assistive technologies. Gestures can be used to control the distribution of resources in hospitals, interact with medical instrumentation, control visualization displays, and help handicapped users as part of their rehabilitation therapy.[35,50] Some of these concepts have been exploited to improve medical procedures and systems; for example, Face MOUSe[35] satisfied the "come as you are" requirement, where surgeons control the motion of a laparoscope by making appropriate facial gestures without hand or foot switches or voice input. Graetzel et al.[16] covered ways to incorporate hand gestures into doctor-computer interfaces, describing a computer-vision system that enables surgeons to perform standard mouse functions, including pointer movement and button presses, with hand gestures that satisfy the "intuitiveness" requirement. Wachs et al.[50] developed a hand-gesture-tracking device called Gestix that allows surgeons to browse MRI images in an operating room (see Figure 6), using a natural interface to satisfy both "come as you are" and "intuitiveness."

Figure 6. Surgeon using Gestix to browse medical images.

A European Community Project called WearIT@work[30] satisfies the "comfort" requirement by encouraging physicians to use a wrist-mounted RFID reader to identify the patient and interact through gestures with the hospital information system to document exams and write prescriptions, helping ensure sterility. However, since this is an encumbered interface, the "come as you are" requirement is violated. We expect to see some of these new technologies (based on "smart instruments") introduced directly into the operating room, where the direction/activation of a robotic end effecter could be performed through gesture recognition.[35]

For the impaired, the critical requirements of a hand-gesture interface system are "user adaptability and feedback" and "come as you are." In this context, wheelchairs, as mobility aids, have been enhanced through robotic/intelligent vehicles able to recognize hand-gesture commands (such as in Kuno et al.[28]). The Gesture Pendant[44] is a wearable gesture-recognition system used to control home devices and provide additional functionality as a medical diagnostic tool. The Staying Alive[3] virtual-reality-imagery-and-relaxation tool satisfies the "user adaptability and feedback" requirement, allowing cancer patients to navigate through a virtual scene using 18 traditional T'ai Chi gestures. In the same vein, a telerehabilitation system[18] for kinesthetic therapy—treatment of patients with arm-motion coordination disorders—uses force-feedback of patient gestures. Force-feedback was also used by Patel and Roy[36] to guide an attachable interface for individuals with severely dysarthric speech. Also, a hand-worn haptic glove was used to help rehabilitate post-stroke patients in the chronic phase by Boian et al.[5] These systems

ROBERT IRVING, INSTITUTE FOR MEDICAL INFORMATICS, WASHINGTON HOSPITAL CENTER, WASHINGTON, D.C.

illustrate how medical systems and rehabilitative procedures promise to provide a rich environment for the potential exploitation of hand-gesture systems. Still, additional research and evaluation procedures are needed to encourage system adoption.

Entertainment. Computer games are a particularly technologically promising and commercially rewarding arena for innovative interfaces due to the entertaining nature of the interaction. Users are eager to try new interface paradigms since they are likely immersed in a challenging game-like environment.[45] In a multi-touch device, control is delivered through the user's fingertips. Which finger touches the screen is irrelevant; most important is where the touch is made and the number of fingers used.

In computer-vision-based, hand-gesture-controlled games,[13] the system must respond quickly to user gestures, the "fast-response" requirement. In games, computer-vision algorithms must be robust and efficient, as opposed to applications (such as inspection systems) with no real-time requirement, and where recognition performance is the highest priority. Research efforts should thus focus on tracking and gesture/posture recognition with high-frame-rate image processing (>10 fps).

Another challenge is "gesture spotting and immersion syndrome," aiming to distinguish useful gestures from unintentional movement. One approach is to select a particular gesture to mark the "start" of a sequence of gestures, as in the "push to talk" approach in radio-based communication where users press a button to start talking. In touchscreen mobile phones, the user evokes a "swipe" gesture to start operating the device. To "end" the interaction, the user may evoke the "ending" gesture or just "rest" the hands on the side of the body. This multi-gesture routine may be preferable to purely gaze-based interaction where signaling the end of the interaction is a difficult problem, since users cannot turn off their eyes. The problem of discriminating between intentional gestures and unintentional movement is also known as the Midas Touch problem (http://www.diku.dk/hjemmesider/ansatte/panic/eyegaze/node27.html).

For sign languages (such as American Sign Language), hand-gesture-recognition systems must be able to recognize a large lexicon of single-handed and two-handed gestures.

In the Mind-Warping augmented-reality fighting game,[45] where users interact with virtual opponents through hand gestures, gesture spotting is solved through voice recognition. The start and end of a temporal gesture is "marked" by voice—the start and end of a Kung Fu yell; Kang et al.[21] addressed the problem of gesture spotting in the first-person-shooter *Quake II*. Such games use contextual information like gesture velocity and curvature to extract meaningful gestures from a video sequence. Bannach et al.[41] addressed gesture spotting through a sliding window and bottom-up approach in a mixed-reality parking game. Schlömer et al.[42] addressed accelerometer-based gesture recognition for drawing and browsing operations in a computer game. Gesture spotting in many Nintendo Wii games is overcome by pressing a button on the WiiMote control through the "push to talk" analogy.

Intuitiveness is another important requirement in entertainment systems. In the commercial arena, most Nintendo Wii games are designed to mimic actual human motions in sports games (such as golf, tennis, and bowling). Wii games easily meet the requirement of "intuitiveness," even as they violate the "come as you are" requirement, since users must hold the WiiMote, instead of using a bare hand. Sony's EyeToy for the Playstation and the Kinect sensor for Microsoft's Xbox360 overcome this limitation while achieving the same level of immersion through natural gestures for interaction. These interfaces use hand-body gesture recognition (also voice recognition in Kinect) to augment the gaming experience.

In the research arena, the intuitive aspect of hand-gesture vocabulary is addressed in a children's action game called QuiQui's Giant Bounce[20] where control gestures are selected through a "Wizard of Oz" paradigm in which a player interacts with a computer application controlled by an unseen subject, with five full-body gestures detected through a low-cost USB Web camera.

"User adaptability and feedback" is the most remarkable requirement addressed in these applications. In entertainment systems, users profit from having to learn the gesture vocabularies employed by the games. A

training session is usually required to teach them how the gestures should be performed, including speed, trajectory, finger configuration, and body posture. While beginners need time to learn the gesture-related functions, experienced users navigate through the games at least as quickly as if they were using a mechanical-control device or attached sensors.[8,37]

Intelligent user interfaces that rely on hand/body gesture technology face special challenges that must be addressed before future commercial systems are able to gain popularity. Aside from technical obstacles like reliability, speed, and low-cost implementation, hand-gesture interaction must also address intuitiveness and gesture spotting.

Crisis management and disaster relief. Command-and-control systems help manage public response to natural disasters (such as tornados, floods, wildfires, and epidemic diseases) and to human-caused disasters (such as terrorist attacks and toxic spills). In them, the emergency response must be planned and coordinated by teams of experts with access to large volumes of complex data, in most cases through traditional human-computer interfaces. One such system, the "Command Post of the Future,"[47] uses pen-based gestures.[11] Such hand-gesture interface systems must reflect the requirements of "fast learning," "intuitiveness," "lexicon size and number of hands," and "interaction space" to achieve satisfactory performance.[26] The first two involve natural interaction with geospatial information (easy to remember and common gestures); the last two involve the system's support of collaborative decision making among individuals. Multimodality (speech and gesture), an additional requirement for crisis-management systems, is not part of our original list of requirements since it includes modalities other than gestures. The pioneering work was Richard A. Bolt's "Put-That-There" system,[6] providing multimodal voice input plus gesture to manipulate objects on a large display.

DAVE_G,[40] a multimodal, multi-user geographical information system (GIS), has an interface that supports decision making based on geospatial data to be shown on a large-screen display. Potential users are detected as soon as they enter the room (the "come as you are" requirement) through a face-detection algorithm; the detected facial region helps create a skin-color model applied to images to help track the hands and face. Motion cues are combined with color information to increase the robustness of the tracking module. Spatial information is conveyed using "here" and "there" manipulative gestures that are, in turn, recognized through a hidden Markov model. The system was extended to operate with multiple users in the "XISM" system at Pennsylvania State University[26] where users simultaneously interface with the GIS, allowing a realistic decision-making process; however, Krahnstoever et al.[26] provided no detail as to how the system disambiguates tracking information of the different users.

Aside from technical obstacles like reliability, speed, and low-cost implementation, hand-gesture interaction must also address intuitiveness and gesture spotting.

Other approaches to multi-user hand-gesture interfaces have adopted multi-touch control through off-the-shelf technology,[15,31] allowing designers to focus on collaborative user performance rather than on hand-gesture-recognition algorithms, These systems give multiple users a rich hand-gesture vocabulary for image manipulation, including zoom, pan, line drawing, and defining regions of interest, satisfying the "lexicon size and number of hands" requirement. Spatial information about objects on the GIS can be obtained by clicking (touching) the appropriate object.

These applications combine collaborative hand-gesture interaction with large visual displays. Their main advantage is user-to-user communication, rather than human-computer interaction, so the subjects use their usual gestures without having to learn new vocabularies; for example, sweeping the desk can be used to clean the surface.

Human-robot interaction. Hand-gesture recognition is a critical aspect of fixed and mobile robots, as suggested by Kortenkamp et al.[25] Most important, gestures can be combined with voice commands to improve robustness or provide redundancy and deal with "gesture spotting." Second, hand gestures involve valuable geometric properties for navigational robot tasks; for example, the pointing gesture can symbolize the "go there" command for

mobile robots. For a robotic arm, human users may use the "put it there" command while pointing to the object and then the place. Hand actions can be used to manipulate operations (such as grasp and release), since a human hand is able to simulate the form of the robot gripper. All these aspects of robot interaction help satisfy the "intuitiveness" requirement. Third, people with physical handicaps are able to control robots through gestures when other channels of interaction are limited or impossible without special keyboards and teach-pendants, or robot controls, satisfying the "come as you are" requirement. Fourth, such an interface brings operability to beginners who find it difficult to use sophisticated controls to command robots. Hand-gesture control of robots faces several constraints specific to this category of interfaces, including "fast," "intuitive," "accuracy," "interaction space," and "reconfigurability." While most systems succeed to some extent in overcoming the technical requirements ("accuracy"), the interaction aspects of these systems involve many unsolved challenges.

Using stereo vision to develop a cooperative work system, Kawarazaki[22] combined robotic manipulators and human users with hand-gesture instructions to recognize four static gestures; when users point at an object on a table with their forefinger the robot must be able to detect it. Chen and Tseng[10] described human-robot interaction for game playing in which three static gestures at multiple angles and scales are recognized by a computer-vision algorithm with 95% accuracy, satisfying the "accuracy" requirement.

Using Sony's AIBO entertainment robot, Hasanuzzaman[19] achieved interaction by combining eight hand gestures and face detection to identify two nodding gestures and the hand (left or right) being used, allowing for a larger lexicon than hand gestures alone.

Rogalla et al.[41] developed a robotic-assistant interaction system using both gesture recognition and voice that first tracks gestures, then combines voice and gesture recognition to evoke a command. Once the hand is segmented, six gestures are trained using a hand contour as the main feature of each gesture. Since the user and robot interact with objects on a table, the interaction space is large enough to include both user and objects. Rogella et al. [41] reported 95.9% recognition accuracy.

Nickel and Stiefelhagen[33] developed a system that recognizes dynamic pointing gestures that rely on head and arm orientation for human-robot interaction. The system uses a hidden Markov model to recognize trajectories of the segmented hands and up to 210 gestures, satisfying the requirement of "lexicon size and number of hands."

Yin and Zhu[53] implemented a programming-by-demonstration approach in which the robot learns gestures from a human user (the instructor), satisfying the requirement of "user adaptability and feedback." The system uses eight static gestures to control a hybrid service robot system called HARO-1. Calinon and Billard[9] also used a programming-by-demonstration paradigm, allowing users to help the robot reproduce a gesture through kinesthetic teaching; in it, the user teaches the robot 10 dynamic gestures acquired through sensors attached to the torso and upper and lower arm, hence violating the "come as you are" and "comfort" requirements.

Most approaches we've reviewed here employ a stereo camera to acquire hand gestures. Some systems also add voice detection, thereby solving the "gesture spotting" problem and improving recognition accuracy. Most of them detect static hand gestures but are not robust enough to recognize more than 10 gestures, so do not satisfy the requirement of "lexicon size and number of hands." Two-handed dynamic-gesture multimodal interaction is thus a promising area for future research.

Conclusion

Hand-gesture implementation involves significant usability challenges, including fast response time, high recognition accuracy, quick to learn, and user satisfaction, helping explain why few vision-based gesture systems have matured beyond prototypes or made it to the commercial market for human-computer devices. Nevertheless, multi-touchscreens and non-joystick and -keyboard interaction methods have found a home in the game-console market, commercial appeal suggesting that hand-gesture-based interactive applications could yet become important players in next-generation interface systems due to their ease of access and naturalness of control.

Four recommended guidelines help evaluate future hand-gesture interfaces to increase the likelihood of their widespread commercial/social acceptance:

Validation. Rigorous statistical validation procedures for gesture-based systems on public, standard test sets. A system's performance can be demonstrated through several statistical measures[32]: sensitivity/recall, precision/positive predictive value, specificity, negative predictive value, f-measure, likelihood ratio, and accuracy;

User independence. User independence while permitting customizability enhances acceptability;

Usability criteria. Provide usability criteria to evaluate learnability, efficiency, ease of remembering, likelihood of errors, and user satisfaction; performance can be evaluated through task completion time and subjective workload assessment through, say, the NASA Task Load Index (http://human-systems.arc.nasa.gov/groups/TLX/) and the Subjective Workload Assessment Technique[38]; and

Qualitative/quantitative assessment. Provide qualitative and quantitative assessments of this modality compared to other modalities (such as voice recognition); for example, user performance when using alternative modalities can be compared with the metrics outlined in the guideline concerning usability criteria.

Questions. Reviewing the HCI literature as we wrote this article revealed increasing adoption of certain principles and heuristics that contribute to the design of hand-gesture-recognition systems:

Context support in hand-gesture recognition. Gestures are context-dependent. Gestures and their types and uses are determined by the context in which they are applied. Task domain analysis helps identify users' intended actions, goals, and means. Previously, HCI researchers adopted task analysis to help determine suitable features for natural HCI.[26,40]

The trade-off between increasing the number of gestures to be recog-

nized and the performance of the recognition is a well-known obstacle in the design of gesture-based interfaces. The more freely a system allows users to express themselves, the less accurate it gets; conversely, the greater the rigor in specifying gestures, the greater the likelihood the system will perform accurately.

A common approach toward achieving this trade-off is to create a set of specific grammars or vocabularies for different contexts. The system dynamically activates different subsets of vocabularies and grammars according to the context, instead of maintaining a single large lexicon. This built-in feature reduces complexity in gesture-recognition systems, as separate gesture-recognition algorithms are used for smaller gesture subsets. Context is captured in many ways, including hand position, interaction log, task, type of gesture, and how the user interacts with devices in the environment.

Methods for hand-gesture recognition. No single algorithm for hand-gesture recognition favors every application. The suitability of each approach depends on application, domain, and physical environment. Nevertheless, integration of multiple methods lends robustness to hand-tracking algorithms; for example, when a tracker loses track of a hand due to occlusion, a different tracker using a different tracking paradigm can still be active. Occlusion is usually disambiguated through the stereo cameras to create depth maps of the environment. Common approaches for hand-gesture tracking use color and motion cues. Human skin color is distinctive and serves to distinguish the human face and hand from other objects. Trackers sensitive to skin color and motion can achieve a high degree of robustness.[40]

Regarding classification, gestures are the outcome of stochastic processes. Thus, defining discrete representations for patterns of spatio-temporal gesture motion is a complicated process. Gesture templates can be determined by clustering gesture training sets to produce classification methods with accurate recognition performance; Kang et al.[21] described examples of such methods, including hidden Markov models, dynamic time warping, and finite state machines.

Two-handed dynamic-gesture multimodal interaction is thus a promising area for future research.

Finally, Kang et al.[21] also addressed the problem of gesture spotting through sliding windows, distinguishing intentional gestures from captured gestures through recognition accuracy of the observed gestures.

Intuitive gestures (selection and teaching) in interface design. Ideally, gestures in HCI should be intuitive and spontaneous. Psycholinguistics and cognitive sciences have produced a significant body of work involving human-to-human communication that can help find intuitive means of interaction for HCI systems. A widely accepted solution for identifying intuitive gestures was suggested by Baudel et al.,[2] and in Höysniemi et al.'s "Wizard-of-Oz" experiment, an external observer interprets user hand movement and simulates the system's response.[20] Called "teaching by demonstration," it is widely used for gesture learning. Rather than pick the gestures during the design stage of the interface, they are selected during real-time operation while interacting with the user, thus mimicking the process of parents teaching gestures to a toddler.[9] First, the parents show the toddler a gesture, then assist the toddler to imitate the gesture by moving the toddler's own hands. The toddler learns the skill of producing the gesture by focusing on his or her own active body parts. Hand gestures play an important role in human-human communication. Analysis of these gestures based on experimental sociology and learning methodologies will lead to more robust, natural, intuitive interfaces.

Acknowledgments

This research was performed while the first author held a National Research Council Research Associateship Award at the Naval Postgraduate School, Monterey, CA. It was partially supported by the Paul Ivanier Center for Robotics Research & Production Management at Ben-Gurion University of the Negev.

References

1. Bannach, D., Amft, O., Kunze, K.S., Heinz, E.A., Tröster, G., and Lukowicz, P. Waving real-hand gestures recorded by wearable motion sensors to a virtual car and driver in a mixed-reality parking game. In *Proceedings of the Second IEEE Symposium on Computational Intelligence and Games* (Honolulu, Apr. 1–5, 2007), 32–39.
2. Baudel, T. and Beaudouin-Lafon, M. Charade: Remote control of objects using FreeHand gestures. *Commun. ACM 36*, 7 (July 1993), 28–35.

3. Becker, D.A. and Pentland, T. Staying alive: A virtual reality visualization tool for cancer patients. In *Proceedings of the AAAI Workshop on Entertainment and Alife/AI*. AAAI Technical Report WS-96-03, 1996.
4. Belongie, S., Malik, J., and Puzicha, J. Shape matching and object recognition using shape contexts. *Transactions on Pattern Analysis and Machine Intelligence 24*, 24 (Apr. 2002), 509–522.
5. Boian, R., Sharma, R., Han, C., Merians, A., Burdea, G., Adamovich, S., Recce, M., Tremaine, M., and Poizner, H. Virtual reality-based post-stroke hand rehabilitation. *Studies in Health and Technology Information* (2002), 64–70.
6. Bolt, R.A. 'Put-That-There': Voice and gesture at the graphics interface. In *Proceedings of the Seventh International Conference on Computer Graphics and Interactive Techniques*. ACM Press, New York, 1980, 262–270.
7. Bradski, G.R. Computer-vision face tracking for use in a perceptual user interface. *Intel Technology Journal* (Q2 1998).
8. Brashear, H., Henderson, V., Park, K., Hamilton, H., Lee, S., and Starner, T. American Sign Language recognition in game development for deaf children. In *Proceedings of ACM SIGACCESS Conference on Assistive Technologies* (Portland, OR, Oct. 23–25). ACM Press, New York, 2006, 79–86.
9. Calinon, S. and Billard, A. Incremental learning of gestures by imitation in a humanoid robot. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (Arlington, VA, 2007), 255–262.
10. Chen, Y.T. and Tseng, K.T. Developing a multiple-angle hand-gesture-recognition system for human-machine interactions. In *Proceedings of the 33rd Annual Conference of the IEEE Industrial Electronics Society* (Taipei, Nov. 5–8, 2007), 489–492.
11. Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., and Clow, J. QuickSet: Multimodal interaction for distributed applications. In *Proceedings of the Fifth ACM international Conference on Multimedia* (Seattle, WA, Nov. 9–13). ACM Press, New York, 1997, 10–13.
12. Cootes, T.F. and Taylor, C.J. Active shape models: 'smart snakes.' In *Proceedings of the British Machine-Vision Conference* (Leeds, Sept. 22–24). Springer, Berlin, 1992, 266–275.
13. Freeman, W. and Roth, M. Orientation histograms for hand-gesture recognition. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition* (Zurich, June 1995).
14. Freeman, W.T., Tanaka, K., Ohta, J., and Kyuma, K. Computer vision for computer games. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition* (Zurich, June 1995), 296–301.
15. Fruijtier, S., Dulk, P.D., and Dias, E. Collaborative interaction and integrated spatial information and services in disaster management. In *Proceedings of the 2008 IEEE International Workshop on Horizontal Interactive Human Computer System* (Amsterdam, Oct. 1–3, 2008), 43–45.
16. Graetzel, C., Fong, T.W., Grange, C., and Baur, C. A non-contact mouse for surgeon-computer interaction. *Technology and Health Care 12*, 3 (Aug. 24, 2004), 245–257.
17. Grange, S., Fong, T., and Baur, C. M/ORIS: A medical/operating room interaction system. In *Proceedings of the ACM International Conference on Multimodal Interfaces* (State College, PA, 2004). ACM Press, New York, 2004, 159–166.
18. Gutierrez, M., Lemoine, P., Thalmann, D., and Vexo, F. Telerehabilitation: Controlling haptic virtual environments through handheld interfaces. In *Proceedings of ACM Symposium on Virtual Reality Software and Technology* (Hong Kong, Nov. 10–12), ACM Press, New York, 2004, 195–200.
19. Hasanuzzaman, M., Ampornaramveth, V., Zhang, T., Bhuiyan, M.A., Shirai, Y., and Ueno, H. Real-time vision-based gesture recognition for human-robot interaction. In *Proceedings of the IEEE International Conference on Robotics and Biomimetics* (Shenyang, China, Aug. 22–26, 2004), 413–418.
20. Höysniemi, J., Hämäläinen, P., Turkki, L., and Rouvi, T. Children's intuitive gestures in vision-based action games. *Commun. ACM 48*, 1 (Jan. 2005), 44–50.
21. Kang, H., Lee, C., and Jung, K. Recognition-based gesture spotting in video games. *Pattern Recognition Letters 25*, 15 (Nov. 2004), 1701–1714.
22. Kawarazaki, N. Hoya, I., Nishihara, K., and Yoshidome, T. Cooperative welfare robot system using hand-gesture instructions. *Lecture Notes in Control and Information Sciences 306*, Springer, Berlin, 2004, 143–153.
23. Kela, J., Korpipää, P., Mäntyjärvi, J., Kallio, S., Savino, G., Jozzo, L., and Marca, D. Accelerometer-based gesture control for a design environment. *Personal and Ubiquitous Computing 10*, 5 (July 2006), 285–299.
24. Kölsch, M., Turk, M., and Höllerer, T. Vision-based interfaces for mobility. In *Proceedings of the International Conference on Mobile and Ubiquitous Systems* (Boston, Aug. 22–26, 2004), 86–94.
25. Kortenkamp, D., Huber, E., and Bonasso, R. Recognizing and interpreting gestures on a mobile robot. In *Proceedings of the 13th Conference on Artificial Intelligence* (Portland, OR, Aug. 4–8, 1996), 915–921.
26. Krahnstoever, N., Kettebekov, S., Yeasin, M., and Sharma, R. A real-time framework for natural multimodal interaction with large-screen displays. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces* (Pittsburgh, PA Oct. 14–16). IEEE Computer Society, Washington, D.C., 2002, 349.
27. Krueger, M.W. *Artificial Reality, Second Ed.* Addison-Wesley, Redwood City, CA, 1991.
28. Kuno, Y., Murashima, T., Shimada, N., and Shirai, Y. Intelligent wheelchair remotely controlled by interactive gestures. In *Proceedings of 15th International Conference on Pattern Recognition* (Barcelona, Sept. 3–7, 2000), 672–675.
29. Lucas, B.D. and Kanade, T. An iterative image-registration technique with an application to stereo vision. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers, Inc., San Francisco, 1981, 674–679.
30. Lukowicz, P., Timm-Giel, A., Lawo, M., and Herzog, O. WearIT@Work: Toward real-world industrial wearable computing. *IEEE Pervasive Computing 6*, 4 (Oct. 2007), 8–13.
31. Micire, M.J. and Yanco, H.A. Improving disaster response with multi-touch technologies. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (San Diego, Oct. 29–Nov. 2, 2007), 2567–2568.
32. Minnen, D., Westeyn, T., and Starner, T. Performance metrics and evaluation issues for continuous activity recognition. In *Proceedings of Performance Metrics in Intelligent Systems Workshop* (Aug. 21–23). NIST, Gaithersburg, MD, 2008.
33. Nickel, K. and Stiefelhagen, R. Visual recognition of pointing gestures for human-robot interaction. *Image and Vision Computing 25*, 12 (Dec. 2007), 1875–1884.
34. Nielsen, M., Storring, M., Moeslund, T.B., and Granum, E. *A Procedure for Developing Intuitive and Ergonomic Gesture Interfaces for Man-Machine Interaction. Technical Report CVMT 03-01*. Aalborg University, Aalborg, Denmark, Mar. 2003.
35. Nishikawa, A., Hosoi, T., Koara, K., Negoro, D., Hikita, A., Asano, S., Kakutani, H., Miyazaki, F., Sekimoto, M., Yasui, M., Miyake, Y., Takiguchi, S., and Monden, M. FAce MOUSe: A novel human-machine interface for controlling the position of a laparoscope. *IEEE Transactions on Robotics and Automation 19*, 5 (Oct. 2003), 825–841.
36. Patel, R. and Roy, D. Teachable interfaces for individuals with dysarthric speech and severe physical disabilities. In *Proceedings of the AAAI Workshop on Integrating Artificial Intelligence and Assistive Technology* (Madison, WI, July 26–30, 1998), 40–47.
37. Pentland, A. and Becker, D. *Sensei: A Real-Time Recognition, Feedback, and Training System for T'ai Chi Gestures*. Masters Thesis. Harvard University, Cambridge, MA, 1997.
38. Potter, S.S. and Bressler, J.R. *Subjective Workload Assessment Technique (SWAT): A User's Guide*. Interim Report, 1998.
39. Raskar, R., Welch, G., Cutts, M., Lake, A., Stesin, L., and Fuchs, H. The office of the future: A unified approach to image-based modeling and spatially immersive displays. In *Proceedings of SIGGRAPH* (Orlando, FL, July 19–24). ACM Press, New York, 1998, 179–188.
40. Rauschert, I., Agrawal, P., Sharma, R., Fuhrmann, S., Brewer, I., and MacEachren, A.M. Designing a human-centered, multimodal GIS interface to support emergency management. In *Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems* (McLean, VA, Nov. 8–9). ACM Press, New York, 2002, 119–124.
41. Rogalla, O., Ehrenmann, M., Zöllner, R., Becher, R., and Dillmann, R. Using gesture and speech control for commanding a robot assistant. In *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication* (Berlin, Sept. 25–27, 2002), 454–459.
42. Schlömer, T., Poppinga, B., Henze, N., and Boll, S. Gesture recognition with a Wii controller. In *Proceedings of the Second International Conference on Tangible and Embedded Interaction* (Bonn, Germany, Feb. 18–20). ACM Press, New York, 2008, 11–14.
43. Sheridan, T. and Ferrell, W. Remote manipulative control with transmission delay. *IEEE Transactions on Human Factors in Electronics 4*, 1 (1963), 25–29.
44. Starner, T., Auxier, J., Ashbrook, D., and Gandy, M. The gesture pendant: A self-illuminating, wearable, infrared computer-vision system for home-automation control and medical monitoring. In *Proceedings of the Fourth International Symposium on Wearable Computers* (Atlanta, Oct. 2000), 87–94.
45. Starner, T., Leibe, B., Singletary, B., and Pair, J. Mind-warping: Towards creating a compelling collaborative augmented reality game. In *Proceedings of the Fifth International Conference on Intelligent User Interfaces* (New Orleans, Jan. 9–12). ACM Press, New York, 2000, 256–259.
46. Stern, H.I., Wachs, J.P., and Edan, Y. Designing hand-gesture vocabularies for natural interaction by combining psycho-physiological and recognition factors (special issue on gesture in multimodal systems). *International Journal of Semantic Computing 2*, 1 (Mar. 2008), 137–160.
47. Thomas, J.J. and Cook, K.A., Eds. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE CS Press, 2005.
48. Triesch, J. and Malsburg, C.V.D. Robotic gesture recognition by cue combination. *Gesture and Sign Language in Human-Computer Interaction*. Lecture Notes in Computer Science. Springer, Berlin, 1998, 233–244.
49. Viola, P. and Jones, M. Robust real-time object detection. *International Journal of Computer Vision 57*, 2 (May 2004), 137–154.
50. Wachs, J., Stern, H., Edan, Y., Gillam, M., Feied, C., Smith, M., and Handler, J. A hand-gesture sterile tool for browsing MRI images in the OR. *Journal of the American Medical Informatics Association 15*, 3 (May–June 2008), 321–323.
51. Wachs, J.P. *Optimal Hand-Gesture Vocabulary Design Methodology for Virtual Robotic Control*. Ph.D. Thesis, Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Be'er She'eva, Israel, 2007.
52. Ward, J.A., Lukowicz, P., Troster, G., and Starner, T.E. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*, 10 (Oct. 2006), 1553–1567.
53. Yin, X. and Zhu, X. Hand-posture recognition in gesture-based human-robot interaction. In *Proceedings of the IEEE Conference on Industrial Electronics and Applications* (Singapore, May 24–26, 2006), 1–6.
54. Yuan, Q., Sclaroff, S., and Athitsos, V. Automatic 2D Hand tracking in video sequences. In *Proceedings of the IEEE Workshop on Applications of Computer Vision* (Breckenridge, CO, Jan. 5–7). IEEE Computer Society Press, 2005, 250–256.

Juan Pablo Wachs (jpwachs@purdue.edu) is an assistant professor in the School of Industrial Engineering of Purdue University, West Lafayette, IN.

Mathias Kölsch (kolsch@nps.edu) is an assistant professor in the Computer Science Department and Modeling, Virtual Environments, and Simulation Institute of the Naval Postgraduate School, Monterey, CA.

Helman Stern (helman@bgumail.bgu.ac.il) is Professor Emeritus in the Department of Industrial Engineering and Management of Ben-Gurion University of the Negev, Be'er She'eva, Israel.

Yael Edan (yael@bgu.ac.il) is a professor in the Department of Industrial Engineering and Management of Ben-Gurion University of the Negev, Be'er She'eva, Israel.