10
Gesture Recognition

Matthew Turk
University of California
Computer Science Department
Santa Barbara, CA 93106-5110
mturk@cs.ucsb.edu

1. INTRODUCTION

A primary goal of virtual environments (VEs) is to provide natural, efficient, powerful, and flexible interaction for users while navigating, exploring, and communicating. Providing gestural capabilities as an input modality, particularly in the context of multimodal interaction (Jaimes and Sebe 2005), can help meet these requirements. Human gestures are natural and flexible and may often be efficient and powerful, especially as compared with alternative interaction modes. This chapter will cover automatic gesture recognition, focusing particularly on computer vision based techniques that do not require the user to wear extra sensors, clothing, or equipment.

The traditional two-dimensional (2-D), keyboard- and mouse-oriented graphical user interface (GUI) is not well suited for virtual environments. Synthetic environments provide the opportunity to utilize several different sensing modalities and technologies and to integrate them into the user experience; these may be alternative methods for utilizing mouse and/or keyboard-like interactions in VEs, or, more generally, natural interaction techniques that move beyond desktop interaction paradigms. Devices that sense body position and orientation, direction of gaze, speech and sound, facial expression, galvanic skin response, and other aspects of human state or behavior can be used to mediate communication between the user and the environment. Combinations of communication modalities and sensing devices can produce a wide range of unimodal and multimodal interface techniques. The potential for these techniques to support natural and powerful interfaces for communication in VEs is compelling, and advances in the past decade are bringing this closer to being a commonplace reality.

If interaction technologies are overly obtrusive, awkward, or constraining, the user's experience with the synthetic environment is severely degraded. If the interaction itself draws attention to the technology rather than the experience or the task at hand, or if it imposes a high cognitive load on the user, it becomes a burden and an obstacle to a successful VE experience. It is therefore critical that gesture recognition technologies are unobtrusive and passive, supporting the immersive experience rather than detracting from it.

To support gesture recognition, human position and movement must be tracked and interpreted in order to recognize semantically meaningful gestures. While tracking a user's head position or hand configuration may be quite useful for directly controlling objects or inputting parameters, people naturally express communicative acts through higher level constructs, as shown schematically in Fig. 10.1. The output of position (and other) sensing must be interpreted to allow users to communicate more naturally and effortlessly through gesture. Gesture recognition, then, includes not only low-level, bottom-up processing of image data, but also top-down processing that brings context and semantics into the equation.

Gesture has been used for control and navigation in CAVEs (Cave Automatic Virtual Environments; Pavlovic et al., 1996; see Chapter 11, this volume) and in other VEs, such as smart rooms, virtual work environments, and performance spaces. More recently, gesture recognition has become a popular consumer technology for use in gaming, driven largely by the introduction of the Microsoft Kinect in 2010 (Xbox, 2012), and in "smart televisions," with recent commercial offerings by Samsung and others. Sensors providing depth images (or RGB plus depth) at frame rate have made these systems possible.

In addition, gesture may be perceived by the environment in order to be transmitted elsewhere (e.g., as a compression technique to be reconstructed at the receiver). Gesture recognition may also influence—intentionally or unintentionally—a system's model of the user's state. For example, a look of frustration may cause a system to slow down its presentation of information, or the urgency of a gesture may cause the system to speed up. Gesture may also be used as a communication *backchannel* (i.e., a visual or verbal behavior such as nodding or saying, "uh-huh" to indicate "I'm with you, continue," or raising a finger to indicate the desire to interrupt) to communicate agreement, participation, attention, conversation turn taking, and so forth.

Given that the human body can express a huge variety of gestures, which are of interest for human-computer interaction, and what is appropriate to sense and recognize? Clearly the position and orientation of each body part—the parameters of an articulated body model—would be useful, as well as features that are derived from those measurements, such as velocity and acceleration. Facial expressions are very expressive and meaningful to people. More subtle cues such as hand tension, overall muscle tension, locations of self-contact, and even pupil dilation may be useful in sensing a person's intention or state.

Chapter 8 (this volume) covers technologies to track the head, hands, and body. These include instrumented gloves, body suits, and marker-based optical tracking. Most of the gesture recognition work applied to VEs has used these tracking technologies as input. Chapter 9 (this volume) covers eye-tracking devices and discusses their limitations in tracking gaze direction. The current chapter covers the representation and interpretation of tracked data from such devices in order to recognize gestures. Additional attention is focused on passive sensing from image-based sensors using computer vision techniques. The chapter concludes with suggestions for gesture-recognition system design.

*** Figure 10.1 goes around here ***

2. THE NATURE OF GESTURE

Gestures are expressive, meaningful body motions—i.e., physical movements of the fingers, hands, arms, head, face, or body with the intent to convey information or interact with the environment. Cadoz (1994) described three functional roles of human gesture:

- Semiotic—to communicate meaningful information.
- Ergotic—to manipulate the environment.
- Epistemic—to discover the environment through tactile experience.

In an HCI environment, gesture recognition is the process by which semiotic gestures made by the user are made known to the system. One could argue that in GUI-based systems, standard mouse and keyboard actions used for selecting items and issuing commands are gestures; however, we are interested in less trivial cases. While static position (also referred to as posture, configuration, or pose) is not technically considered gesture, it is included for the purposes of this chapter, as certain poses may be characteristic of the gestures that created them, and a gesture may be considered as a temporal sequence of poses.

In virtual environments, users need to communicate in a variety of ways, to the system itself and also to other local or remote users. Communication tasks include specifying commands and/or parameters for tasks such as:

- Navigating through a space
- Specifying items of interest
- Manipulating objects in the environment
- Changing object values
- Controlling virtual objects
- Issuing task-specific commands

In addition to user-initiated communication, a VE system may benefit from observing a user's behavior for purposes such as:

- Analyzing system usability

- Analyzing user behavior
- Monitoring changes in a user's state
- Better understanding a user's intent or emphasis
- Communicating user behavior to other users or environments

Messages can be expressed through gesture in many ways. For example, an emotion such as sadness can be communicated through facial expression, a lowered head position, relaxed muscles, and lethargic movement. Similarly, a gesture to indicate "Stop!" can be simply a raised hand with the palm facing forward or an exaggerated waving of both hands above the head. In general, there exists a many-to-one mapping from concept to gesture (i.e., gestures are ambiguous); there is also a many-to-one mapping from gesture to concept (i.e., gestures are not completely specified). Like speech and handwriting, gestures vary among individuals, from instance to instance for a given individual, and are subject to the effects of coarticulation. So as with many recognition tasks, it is important to consider both *within-class* and *between-class* variations in the data.

An interesting real-world example of the use of gestures in visual communications is a U.S. Army field manual (Anonymous, 1987) that serves as a reference and guide to commonly used visual signals, including hand and arm gestures for a variety of situations. The manual describes visual signals used to transmit standardized messages rapidly over short distances. In contrast is the training material for the Microsoft Kinect, where a small number of gestures are imprecisely defined and users are mostly learned by doing.

Despite the richness and complexity of gestural communication, researchers have made progress in beginning to understand and describe the nature of gesture. Kendon (1972) described a "gesture continuum," defining five different kinds of gestures:

- *Gesticulation*—spontaneous movements of the hands and arms that accompany speech.
- *Languagelike gestures*—gesticulation that is integrated into a spoken utterance, replacing a particular spoken word or phrase.
- *Pantomimes*—gestures that depict objects or actions, with or without accompanying speech.
- *Emblems*—familiar gestures such as "V for victory," "thumbs up," and assorted rude gestures (these are often culturally specific)
- *Sign languages*—linguistic systems, such as American Sign Language (ASL), which are well defined.

As the list progresses (moving from left to right in Fig. 10.2), the association with speech declines, language properties increase, spontaneity decreases, and social regulation increases.

*** Figure 10.2 goes around here ***

Within the first category—spontaneous, speech-associated gesture—McNeill (1992) defined four gesture types:

- *Iconic*—representational gestures depicting some feature of the object, action, or event being described.
- *Metaphoric*—gestures that represent a common metaphor, rather than the object or event directly.
- *Beat*—small, formless gestures, often associated with word emphasis.
- *Deictic*—pointing gestures that refer to people, objects, or events in space or time.

Fig. 10.2 depicts the relationships among Cadoz's functional roles of human gesture, Kendon's gesture continuum, and McNeill's gesture types.

These types of gesture modify the content of accompanying speech and may often help to disambiguate speech, similar to the role of spoken intonation. Cassell et al. (1994) described a system that models the relationship between speech and gesture and generates interactive dialogs between three-dimensional (3-D) animated characters that gesture as they speak.

These spontaneous gestures (*gesticulation* in Kendon's continuum) are estimated to make up some 90% of human gestures (McNeill 2000). People even gesture when they are on the telephone, and blind people regularly gesture when speaking to one another. Across cultures, speech-associated gesture is natural and common. For human-computer interaction (HCI) to be truly natural, technology to understand both speech and gesture together must be developed.

Despite the importance of this type of gesture in normal human-to-human interaction, most research to date in HCI, and most VE technology, focuses on the lower right side of Fig. 10.2, where gestures tend to be less ambiguous, less spontaneous and natural, more learned, and more culture-specific. Emblematic gestures and gestural languages, although perhaps less spontaneous and natural, carry more clear semantic meaning and may be more appropriate for the kinds of command-and-control interaction that VEs tend to support. The main exception to this is work in recognizing and integrating deictic (mainly pointing) gestures, beginning with the well-known *"Put That There"* demonstration by Bolt at MIT in the context of the Media Room project (Bolt 1980). The remainder of this chapter will focus on *symbolic gestures* (which include emblematic gestures and predefined gesture languages) and *deictic gestures,* keeping our attention largely on the bottom-up approaches.

3. REPRESENTATIONS OF POSTURE AND GESTURE

Although most gesture recognition systems share some common approaches, there is no standard way to do gesture recognition—a variety of representations and classification schemes are used. The concept of gesture is not precisely defined, and the interpretation of gestures depends on the context of the interaction. Recognition of natural, continuous gestures requires temporally segmenting gestures, since gestures are fundamentally time-varying events. Automatically segmenting gestures is difficult and has often been finessed or ignored in systems by requiring a starting position in time and/or space, similar to "push to talk" speech recognition systems. Morency et al. (2007) proposed a solution to this problem using latent-dynamic discriminative models, which was also used by Song et al. (2012) based on extracted body and hand features for continuous gesture recognition. Alon et al. (2009) used spatiotemporal matching and offline learning to incorporate both top-down and bottom-up information flow in gesture segmentation. Similar to this is the problem of distinguishing intentional gestures from other "random" movements.

Gestures can be static, where the user assumes a certain pose, posture, or configuration; or dynamic, defined by movement. McNeill (1992) defined three phases of a dynamic gesture: prestroke, stroke, and poststroke. Some gestures have both static and dynamic elements, where the pose is important in one or more of the gesture phases; this is particularly relevant in sign languages. When gestures are produced continuously, each gesture is affected by the gesture that preceded it, and possibly by the gesture that follows it. These *coarticulations* may be taken into account as a system is trained.

There are several aspects of a gesture that may be relevant and therefore may need to be represented explicitly. Hummels and Stappers (1998) described four aspects of a gesture which may be important to its meaning:

- Spatial information—where it occurs, locations a gesture refers to.
- Pathic information—the path that a gesture takes.
- Symbolic information—the sign that a gesture makes.
- Affective information—the emotional quality of a gesture.

In order to infer these aspects of gesture, human position, configuration, and movement must be sensed. This can be done directly with sensing devices such as magnetic field trackers, instrumented gloves, inertial sensors, and datasuits, which are attached to or held by the user, or indirectly using techniques such as electric field sensing (as in a theremin) or cameras and computer vision techniques. Each sensing technology differs along several dimensions, including accuracy, resolution, mobility, latency, range of motion, user comfort, and cost. The integration of multiple sensors in gesture recognition is a complex task, since each sensing technology varies along these dimensions. Although the output from these sensors can be used to directly control parameters such as navigation speed and direction or movement through a virtual space, our interest is primarily in the interpretation of sensor data to recognize gestural information.

The output of initial sensor processing is a time-varying sequence of parameters typically describing positions, velocities, and angles of relevant body parts and features – ideally 3-D and view-independent, but in some cases 2-D, view-dependent features. These should (but often do not) include a representation of uncertainty that indicates limitations of the sensor and processing algorithms. Recognizing gestures from these parameters is a pattern recognition task that typically involves transforming input into the appropriate representation (feature space) and then classifying it from a database of predefined gesture representations, as shown in Fig. 10.3. The parameters produced by the sensors may be transformed into a global coordinate space, processed to produce sensor-independent features, or used directly in the classification step.

Because gestures are highly variable, from one person to another (inter-person variations) and from one example to another within a single person (intra-person variation), it is essential to capture the essence of a gesture—its invariant properties—and use this to represent the gesture. Besides the choice of representation itself, a significant issue in building gesture recognition systems is how to create and update the database of known gestures. Hand-coding gestures only works for trivial systems; in general, a gesture recognition system needs to be trained through some kind of learning procedure. As with speech recognition systems, there is often a trade-off between accuracy and generality—the more accuracy desired, the more user-specific training is required. In addition, systems may be fully trained when in use, or they may adapt over time to the current user. As should be apparent throughout this chapter, there is no one-size-fits-all solution to problems in gesture recognition.

Static gesture, or *pose*, recognition can be accomplished by a straightforward implementation of Fig. 10.3, using template matching, geometric feature classification, neural networks, or other standard pattern recognition techniques to classify pose. Dynamic gesture recognition, however, requires consideration of temporal events. This is typically accomplished through the use of techniques such as time-compressing templates, finite state machines (Hong et al. 2000), dynamic time warping (Darrell and Pentland 1993, Reyes et al. 2011), hidden Markov models (Wilson and Bobick 1999, Lee and Kim 1999), conditional random fields (Wang et al. 2006), and Bayesian networks (Suk et al. 2008). Elgammal et al. (2003) described learning dynamics for gestures, applied to sequences of body poses.

*** Figure 10.3 goes around here ***

4. APPROACHES FOR RECOGNIZING GESTURE

Gesture recognition is useful in a wide range of application areas (Wachs et al. 2011), and approaches in most of these are applicable to the use of gesture in virtual environments. For example, gesture recognition has been used extensively in surface computing (Wobbrock et al. 2009), immersive environments (Kehl and Van Gool 2004), smart home control (Kühnel et al. 2011), interactive gaming (Kang et al. 2004, Bleiweiss et al. 2010), interactive music (Overholt et al. 2009), and artistic interfaces (Faste et al. 2008). These different scenarios may assume the use of various devices or sensing capabilities.

4.1 Pen-based and Touch-based Gesture Recognition

Recognizing gestures from a 2-D input device such as a pen or mouse has been considered for some time. The early Sketchpad system in 1963 (Johnson, 1963) used light-pen gestures, for example. Some commercial systems have used pen gestures since the 1970s. There are examples of gesture recognition for document editing, for air traffic control, and for design tasks such as editing splines. In the 1990s, systems such as the OGI QuickSet system (Cohen et al. 1997) demonstrated the utility of pen-based gesture recognition and speech recognition to control a virtual environment. QuickSet recognized 68 pen gestures, including map symbols, editing gestures, route indicators, area indicators, and taps. Oviatt (1996) demonstrated significant benefits of using speech and pen gestures together in certain tasks. Zeleznik, Herndon, and Hughes (1996) and Landay and Myers (1995) developed interfaces that recognize gestures from pen-based sketching.

The introduction of inexpensive touch screens and multi-touch capabilities on mobile phones (with iOS, Android, Windows Phone, and other such devices) in recent years has made touch-based gesture a common experience for smart phone users, expanding the existing vocabulary of mouse- and pen-based gestures. A significant benefit of pen-based and touch-based gestural systems is that sensing and interpretation is relatively straightforward as compared with vision-based techniques. Early pen-based PDAs performed handwriting recognition and allowed users to invoke operations by various, albeit quite limited, pen gestures. Long, Landay, and Rowe (1998) survey problems and benefits of these gestural interfaces and provide insight for interface designers. A much wider range of more sophisticated gestures are now available with multi-touch devices and prototypes of various sizes and configurations (Wilson 2004, Wobbrock et al. 2009).

Although pen- and touch-based gesture recognition is promising for many HCI environments, it presumes the availability of, and proximity to, an interactive surface or screen. In VEs this is often too constraining; techniques that allow the user to move around and interact in more natural ways are more compelling. The following sections cover two primary technology approaches for gesture recognition in virtual environments: instrumented or device-based (active) and vision-based (passive) interfaces.

4.2 Device-based Gesture Recognition

There are a number of commercially available tracking systems (covered in Chapters 8 and 9), which can be used as input to gesture recognition, primarily for tracking eye gaze, hand configuration, and overall body position. Each sensor type has its strengths and weaknesses in the context of VE interaction. While eye gaze can be quite useful in a gestural interface, the focus here is on gestures based on input from tracking the hands and body.

4.2.1 Instrumented Gloves

People naturally use their hands for a wide variety of manipulation and communication tasks. Besides being quite convenient, hands are extremely dexterous and expressive, with approximately 29 degrees of freedom (including the wrist). In his comprehensive thesis on whole-hand input, Sturman (1992) showed that the hand can be used as a sophisticated input and control device in a wide variety of application domains, providing real-time control of complex tasks with many degrees of freedom. He analyzed task characteristics and requirements, hand action capabilities, and device capabilities, and discussed important issues in developing whole-hand input techniques. Sturman suggested a taxonomy of whole-hand input that categorizes input techniques along two dimensions:

- Classes of hand actions: continuous or discrete.
- Interpretation of hand actions: direct, mapped, or symbolic.

The resulting six categories describe the styles of whole-hand input. A given interaction task can be evaluated as to which style best suits the task. Mulder (1996) presented an overview of hand gestures in human-computer interaction, discussing the classification of hand movement, standard hand gestures, and hand-gesture interface design. LaViola (1999) provided a thorough survey on hand posture and gesture recognition techniques at the end of the decade; Dipietro et al. (2008) provide a more recent survey of glove-based systems.

For several years, commercial devices have been available which measure, to various degrees of precision, accuracy, and completeness, the position and configuration of the hand. These include “data gloves” and exoskeleton devices (Fig. 10.4) mounted on the hand and fingers (the term *instrumented glove* is used to include both types). Some advantages of instrumented gloves include:

- Direct measurement of hand and finger parameters (joint angles, 3-D spatial information, wrist rotation)
- Provide data at a high sampling frequency
- Ease of use
- No line-of-sight occlusion problems
- Relatively low cost versions available

- Data that is translation-independent (within the range of motion)

Disadvantages of instrumented gloves include:

- Calibration can be difficult.
- Tethered gloves reduce range of motion and comfort.
- Data from inexpensive systems can be very noisy.
- Accurate systems are expensive.
- The user is forced to wear a cumbersome device.

*** Figure 10.4 goes about here ***

*** Figure 10.5 goes about here ***

Many projects have used hand input from instrumented gloves for “point, reach, and grab” operations or more sophisticated gestural interfaces. Latoschik and Wachsmuth (1997) present a multiagent architecture for detecting pointing gestures in a multimedia application. Väänänen and Böhm (1992) developed a neural network system that recognized static gestures and allows the user to interactively teach new gestures to the system. Böhm et al. (1994) extend that work to dynamic gestures using a Kohohen Feature Map (KFM) for data reduction.

Baudel and Beaudouin-Lafon (1993) developed a system to provide gestural input to a computer while giving a presentation. This work included a gesture notation and set of guidelines for designing gestural command sets. Fels and Hinton (1995) used an adaptive neural network interface to translate hand gestures to speech. Kadous (1996) used glove input to recognize Australian sign language as did Takahashi and Kishino (1991) for the Japanese Kana manual alphabet. The system of Lee and Xu (1996) could learn and recognize new gestures online.

More recent examples of glove-based gesture research can be found in (Parvini et al. 2009) who compare the performance of different approaches; data glove gestures were recently studied in (Huang et al. 2011).

Despite the fact that many or most gestures involve two hands, a large portion of the research efforts in glove-based gesture recognition use only one glove for input. The features that are used for recognition and the degree to which dynamic gestures are considered vary quite a bit in the research literature.

4.2.2 Body Suits and Motion Tracking Systems

It is well-known that by viewing only a small number of strategically placed dots on the human body, people can easily perceive complex movement patterns such as the activities, gestures, identities, and other aspects of bodies in motion (Johansson 1973). One way to approach the recognition of human movements and postures is to optically measure the 3-D position of several such markers attached to the body (Fig. 10.5) and then recover the time-varying articulated structure of the body. The articulated structure may also be measured more directly by sensing joint angles and positions using electromechanical body sensors. Although some of the optical tracking systems require a full body suit to be donned, others only require dots or small balls to be placed on top of a subject’s clothing. Both systems are referred to generically here as “motion capture systems.”

Motion capture systems have advantages and disadvantages that are similar to those of instrumented gloves: they can provide reliable data at a high sampling rate (at least for electromagnetic devices), but they are expensive and very cumbersome. Calibration is often nontrivial. While the optical systems typically use several cameras and may have to process the data offline, their major advantage is the lack of wires and a tether.

Motion capture systems have been used, often along with instrumented gloves, in several gesture recognition systems. Wexelblat (1994) implemented a continuous gesture analysis system using a data suit, data gloves, and an eye tracker. In this system, data from the sensors is segmented in time (between movement and inaction), key features are extracted, motion is analyzed, and a set of special-purpose gesture recognizers look for significant changes.

Marrin and Picard (1998) developed an instrumented jacket for an orchestral conductor that includes physiological monitoring to study the correlation between affect, gesture, and musical expression.

Although many optical and electromechanical tracking technologies are cumbersome and therefore contrary to the desire for more natural interfaces, advances in sensor technology may well enable a new generation of devices (including stationary field-sensing devices, gloves, watches, and rings) that are just as useful as current trackers but much less obtrusive. Similarly, the currently cumbersome instrumented body suits may be displaced by sensing technologies embedded in belts, shoes, eyeglasses, cell phones, and even shirts and pants. For example, inertial sensors in cell phones have been used to classify human activity, including gestures (Pylvänäinen 2005, Liu et al. 2009, Wu et al. 2009, Kratz 2010, Wang 2012).

Note that although some of the body tracking methods in this section use cameras and computer vision techniques to track joint or limb positions, they require the user to wear special markers. In the next section only passive techniques that do not require the user to wear any special markers or equipment are considered.

4.3 Passive Vision-based Gesture Recognition

The most significant disadvantage of the tracker-based systems in Section 4.2 is that they are cumbersome. This detracts from the immersive nature of a VE by requiring the user to don an unnatural device that cannot easily be ignored and which often requires significant effort to put on and calibrate. Even optical systems with markers applied to the body or clothing suffer from these shortcomings, albeit not as severely. Computer vision techniques have the potential to provide real-time data useful for analyzing and recognizing human motion in a manner that is passive and unobtrusive.

Vision-based interfaces use one or more cameras to capture images, typically at a frame rate of 30 Hz or more, and interpret those images to produce visual features that can be used to represent human activity and recognize gestures (Wu and Huang 1999, Mitra and Acharya 2007)). Typically the camera locations are fixed in the environment, although they may also be mounted on moving platforms or on other people. For the past two decades, there has been a significant amount of research in the computer vision community on detecting and recognizing faces, analyzing facial expression, extracting lip and facial motion to aid speech recognition, interpreting human activity, and recognizing particular gestures, most of which is relevant to this topic.

Unlike sensors worn on the body, vision approaches to body tracking have to contend with occlusions. From the point of view of a given camera, there are always parts of the user's body that are occluded and therefore not visible; for example, the backside of the user is not visible when the camera is in front. More significantly, self-occlusion often prevents a full view of the fingers, hands, arms, and body from a single view. Multiple cameras can be used, but at the cost of higher complexity in both setup, calibration, and processing.

The occlusion problem makes full-body tracking difficult, if not impossible, without a strong model of body kinematics and perhaps dynamics. However, recovering all the parameters of body motion may not be a prerequisite for gesture recognition. The fact that people can reliably and robustly recognize gestures leads to three possible conclusions: (1) the parameters that cannot be directly observed are inferred; (2) these parameters are not needed to accomplish the task; or (3) some are inferred and others are ignored.

It is a mistake to consider passive vision approaches and direct tracking devices (such as instrumented gloves and body suits) as alternative paths to the same end. Although there is overlap in what they can provide, these technologies in general produce qualitatively and quantitatively different output which enables different analysis and interpretation. For example, tracking devices can in principle detect and measure fast and subtle movements of the fingers while a user is waving his or her hands, whereas human vision in that case may at best get a general sense of the type of finger motion. Similarly, vision can use properties like texture and color in its analysis of gesture, whereas tracking devices do not. From a practical perspective, these observations imply that it may not be an optimal strategy to merely substitute vision at a later date into a system that was developed to use an instrumented glove or a body suit—or vice versa.

Unlike special devices that measure human position and motion, vision uses a multipurpose sensor; the same device used to recognize gestures can be used to recognize other objects in the environment and also to transmit video for teleconferencing, surveillance, and other purposes. On the other hand, advances in miniaturized, low-cost, low-power cameras integrated with processing circuitry on a single chip increasingly make possible a special-purpose “gesture sensor” that outputs motion or gesture parameters to the virtual environment.

Currently, most computer vision systems for recognition look something like Fig. 10.3. A digital camera feeds video frames to a computer’s memory. There may be a preprocessing step, where images are normalized, enhanced, or transformed in some manner, and then a feature extraction step. The features—which may be any of a variety of 2-D or 3-D features, statistical properties, or estimated body parameters—are analyzed and classified as a particular gesture if appropriate.

Vision-based systems for gesture recognition vary along a number of dimensions, most notably:

- Number of cameras. How many cameras are used? If more than one, is their information combined early (at the pixel or feature level) or late (after recognition of body parts or basic movements)?
- Speed and latency. Is the system real-time (i.e., fast enough, with low enough latency, to support interaction)?
- Structured environment. Are there restrictions on the background, lighting, speed of movement, and so forth?
- User requirements. Must the user wear anything special (e.g., markers, gloves, long sleeves)? Is anything disallowed (e.g., glasses, beard, rings)?
- Primary features. What low-level features are computed (edges, regions, silhouettes, moments, histograms, depth maps, etc.)?
- Two- or three-dimensional representation. Does the system construct a 3-D model of the body part(s), or is classification done on some other (view-based) representation?
- Representation of time. How is the temporal aspect of gesture represented and used in recognition (e.g., via a state machine, dynamic time warping, HMMs, time-compressed template)?

*** Figure 10.6 goes here ***

4.3.1 Head and Face Gestures

When people interact with one another, they use an assortment of cues from the head and face to convey information. These gestures may be intentional or unintentional, they may be the primary communication mode or backchannels, and they can span the range from extremely subtle to highly exaggerated. Some examples of head and face gestures include:

- Nodding or shaking the head
- Direction of eye gaze
- Raising the eyebrows
- Opening the mouth to speak
- Winking
- Flaring the nostrils
- Looks of surprise, happiness, disgust, anger, sadness, etc.

People display a wide range of facial expressions. Ekman and Friesen (1978) developed the Facial Action Coding System (FACS) for measuring facial movement and coding expression; this description forms the core representation for many facial expression analysis systems.

A real-time system to recognize actions of the head and facial features was developed by Zelinsky and Heinzmann (1996), who used feature template tracking in a Kalman filter framework to recognize thirteen head and face gestures. Moses, Reynard, and Blake (1995) used fast-contour tracking to determine facial expression from a mouth contour. Essa and

Pentland (1997) used optical flow information with a physical muscle model of the face to produce accurate estimates of facial motion. This system was also used to generate spatiotemporal motion-energy templates of the whole face for each different expression. These templates were then used for expression recognition. Oliver, Pentland, and Bérard (1997) described a real-time system for tracking the face and mouth that recognized facial expressions and head movements. Otsuka and Ohya (1998) modeled coarticulation in facial expressions and used an HMM for recognition.

Black and Yacoob (1995) used local parametric motion models to track and recognize both rigid and nonrigid facial motions. Demonstrations of this system show facial expressions being detected from television talk show guests and news anchors (in non–real time). La Cascia, Isidoro, and Sclaroff (1998) extended this approach using texture mapped surface models and nonplanar parameterized motion models to better capture the facial motion.

4.3.2 Hand and Arm Gestures

Hand and arm gestures receive the most attention among those who study gesture; in fact, many (perhaps most) references to gesture recognition only consider hand and arm gestures. The vast majority of automatic recognition systems are for deictic gestures (pointing), emblematic gestures (isolated signs), and sign languages (with a limited vocabulary and syntax). Some are components of bimodal systems, integrated with speech recognition. Some estimate precise hand and arm configuration, whereas others only coarse motion.

Stark and Kohler (1995) developed the ZYKLOP system for recognizing hand poses and gestures in real time. After segmenting the hand from the background and extracting features such as shape moments and fingertip positions, the hand posture is classified. Temporal gesture recognition is then performed on the sequence of hand poses and their motion trajectory. A small number of hand poses comprises the gesture catalog, whereas a sequence of these makes a gesture. Similarly, Maggioni and Kämmerer (1998) described the GestureComputer, which recognized both hand gestures and head movements. Other early systems that recognize hand postures amidst complex visual backgrounds are reported by Weng and Cui (1998) and Triesch and von der Malsburg (1996). Oka et al. (2002) tracked fingertips, Bretzner et al. (2002) used multi-scale skin color features and particle filtering, and Yang et al. (2002) used motion trajectories to represent and recognize hand gestures.

There has been a great deal of interest in creating devices to automatically interpret various sign languages to aid the deaf community (Ong and Ranganath, 2005). One of the first to use computer vision without requiring the user to wear specialized devices on the hands was built by Starner and Pentland (1995), who used HMMs to recognize a limited vocabulary of ASL sentences. A similar effort, which uses HMMs to recognize Sign Language of the Netherlands, was described by Assan and Grobel (1997). Vogler and Metaxas (2001) dealt with recognizing simultaneous aspects of complex sign language; Wang et al. (2007) focused on viewpoint-invariant recognition, while Zaki and Shaheen (2011) proposed a combination of new image features for sign language recognition.

The recognition of hand and arm gestures has been applied to entertainment applications. Freeman, Tanaka, Ohta, and Kyuma (1996) developed a real-time system to recognize hand poses using image moments and orientation histograms, and applied it to interactive video games. Cutler and Turk (1998) described a system for children to play virtual instruments and interact with lifelike characters by classifying measurements based on optical flow; more recent optical flow-based gesture recognition is presented in (Holte et al. 2010). A nice overview of work up to 1995 in hand gesture modeling, analysis, and synthesis is presented by Huang and Pavlovic (1995).

The Leap (Leap Motion 2012), a gesture control system introduced in 2012 initially focused on desktop use, is a small device that, according to early reports, accurately tracks the 3D positions of fingertips, hands, and lower arms, capturing subtle movements in an interaction space of about eight cubic feet. The device senses individual hand and finger movements independently, providing touch-free motion sensing and control. While there are few details available for this device as of Q1 2013, the demonstrations are impressive, and there is great interest among researchers and developers to try this device in a variety of interactive environments.

4.3.3 Body Gestures

Full-body motion is needed to recognize large-scale gestures and to recognize or analyze human activity (Gavrila 1999, Moeslund et al. 2011). Activity may be defined over a much longer period of time than what is normally considered a gesture; for example, two people meeting in an open area, stopping to talk, and then continuing on their way may be considered a recognizable activity. A different view of these terms was proposed by Bobick (1997) in his hierarchy of motion perception:

- Movement—the atomic elements of motion.
- Activity—a sequence of movements or static configurations.
- Action—high-level description of what is happening in context.

Most research to date has focused on the first two levels.

The Pfinder system (Wren, Azarbayejani, Darrell, & Pentland, 1996) developed at the MIT Media Lab has been used by a number of groups to do body tracking and gesture recognition. It formed a 2-D representation of the body, using statistical models of color and shape. The body model provided an effective interface for applications such as video games, interpretive dance, navigation, and interaction with virtual characters. Lucente, Zwart, and George (1998) combined Pfinder with speech recognition in an interactive environment called Visualization Space, allowing a user to manipulate virtual objects and navigate through virtual worlds. Paradiso and Sparacino (1997) used Pfinder to create an interactive performance space where a dancer can generate music and graphics through their body movements, for example, hand and body gestures can trigger rhythmic and melodic changes in the music.

Systems that analyze human motion in VEs may be quite useful in medical rehabilitation (see Chapter 49, this volume) and athletic and military training (see Chapter 43, this volume). For example, a system like the one developed by Boyd and Little (1998) to recognize human gaits could potentially be used to evaluate rehabilitation progress. Yamamoto, Kondo, and Yamagiwa (1998) described a system that used computer vision to analyze body motion in order to evaluate the performance of skiers.

Davis and Bobick (1997) used a view-based approach by representing and recognizing human action based on “temporal templates,” where a single image template captures the recent history of motion. This technique was used in the KidsRoom system (Bobick et al. 1999), an interactive, immersive, narrative environment for children.

Video surveillance and monitoring of human activity has received significant attention in recent years. For example, the W^4 system developed at the University of Maryland (Haritaoglu, Harwood, & Davis, 1998) tracked people and detected patterns of their activity. Although partly relevant to virtual environments – especially in understanding context and in multi-person environments – the topic is beyond the scope of this chapter.

4.4 Depth Cameras

There has been a recent proliferation in the use of depth sensors to detect, track, and recognize the gestures and activity of people, largely due to the introduction and rapid popularity of the Microsoft Kinect device (Fig. 10.6), which is based on the PrimeSense depth sensor. Others have used stereo camera rigs, time-of-flight depth sensors, and other technologies to obtain 3-D information about the scene for use in gesture recognition. Many of these current approaches use both RGB and depth data, which are available from the Kinect and other similar RGBD cameras.

Breuer et al. (2007) and Hackenberg et al. (2011) explored hand gesture recognition using a time-of-flight range camera. Girshick et al. (2011) have used the Kinect for modeling body pose and determining human activity; Ren et al. (2011) and Van den Bergh and Van Gool (2011) have used RGB and depth data for robustly recognizing hand gestures. Zafrulla et al. (2011) used a Kinect to recognize American Sign Language. Doliotis et al. (2011) have studied gesture recognition accuracy using an RGBD sensor.

Although current depth sensors have significant limitations for general use (such as a limited range of distances from the sensor and mostly indoor lighting conditions), improvements are underway to improve on these, and there is great promise for continued progress in real-time, robust gesture recognition in virtual environments using such devices. Table 10.1 summarizes some of the main benefits and drawbacks of various categories of gesture recognition devices.

Table 10.1. Pros and cons of gesture recognition devices

Device	**Pros**	**Cons**
Pen- or stylus-based	High resolution and precision; inexpensive; high familiarity	Inconvenience of additional device (pen or stylus) and surface; proximity to surface; limited 2D recognition space
Touch-based	Ease of use; very high familiarity; multi-touch; inexpensive	Inconvenience of additional surface; proximity to surface; limited 2D recognition space
Haptic device (e.g., Phantom)	Relatively high resolution and precision; 3D gestures	Expensive; proximity to devices
Instrumented glove	High input dimensionality; whole-hand input; natural use; no line-of-sight occlusion; two-hand gesture possible	Expensive; need to don extra equipment; limited range if tethered; complex processing/interpretation
Motion tracking	No direct contact required; relatively wide range of motion possible; full body gestures	Very expensive; calibration needed; obtrusive to wear; fixed area of use
Vision-based tracking and recognition	No direct contact required; wide range of motion (esp. with mobile sensors); high resolution possible; 3D gestures possible; full body gestures; ease of use; may be inexpensive; synergy with gaming and other commercial markets	Must be aware of camera occlusion, field-of-view; calibration may be needed; IR illumination may interfere with other sensors; complex processing

5. GUIDELINES FOR GESTURE RECOGNITION SYSTEMS

There has been surprising little work in evaluating the utility and usability of gesture recognition systems in realistic settings. However, those developing gestural systems have learned a number of lessons along the way. Here a few guidelines are presented in the form of “dos and don’ts” for gestural interface designers:

- **Do inform the user.** As discussed in Section 2, people use different kinds of gestures for many purposes, from spontaneous gesticulation associated with speech to structured sign languages. Similarly, gesture may play a number of different roles in a virtual environment. To make compelling use of gesture, the types of gestures allowed and their effects must be clear to the user or easily discoverable.
- **Do give the user feedback.** Feedback is essential to let the user know when a gesture has been recognized (or has been detected but not recognized). This could be inferred from the action taken by the system, when that action is obvious, or by more subtle visual or audible confirmation methods.

- **Do take advantage of the uniqueness of gesture.** Gesture is not just a substitute for a mouse or keyboard. It may not be as useful for 2-D pointing or text entry but great for more expressive input.
- **Do understand the benefits and limits of the particular technology.** For example, precise finger positions are better suited to instrumented gloves than vision-based techniques. Tethers from gloves or body suits may constrain the user's movement.
- **Do usability testing on the system.** Don't just rely on the designer's intuition (see Chapter 34, this volume).
- **Do avoid temporal segmentation if feasible.** At least with the current state of the art, an initial, bottom-up segmentation of gestures can be quite difficult and error prone
- **Don't tire the user.** Gesture is seldom the primary mode of communication. When a user is forced to make frequent, awkward, or precise gestures, the user can become fatigued quickly. For example, holding one's arm in the air to make repeated hand gestures becomes tiring very rapidly.
- **Don't make the gestures to be recognized too similar.** For ease of classification and to help the user more easily make distinguishable gestures.
- **Don't use gesture as a gimmick.** If something is better done with a mouse, keyboard, speech, or some other device or mode, use it—extraneous use of gesture should be avoided.
- **Don't unduly increase the user's cognitive load.** Having to remember the whats, wheres, and hows of a gestural interface can be a burden to the user. The system's gestures should be as intuitive and simple as possible. The learning curve for a gestural interface is more difficult than for a mouse and menu interface because it requires recall rather than just recognition among a visible list of options.
- **Don't require precise motion.** Especially when motioning in space with no tactile feedback, it is difficult to make highly accurate or repeatable gestures.
- **Don't create new, unnatural gestural languages.** If it is necessary to devise a new gesture language, make it as intuitive as possible.

6. CONCLUSIONS AND FUTURE DIRECTIONS IN GESTURE RECOGNITION

Although several research efforts have been referenced in this chapter, these are just a sampling; many more have been omitted for the sake of brevity. Good sources for much of the work in gesture recognition can be found in the proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG) and in a number of workshops devoted to various aspects of gesture recognition.

There is still much to be done before rich and fully natural gestural interfaces, which track and recognize local human activities, become pervasive and cost-effective for the masses. However, much progress has been made in the past decade and with the continuing march toward computers and sensors that are faster, smaller, and more ubiquitous, there is cause for optimism. As touch-based and pen-based computing continues to proliferate, 2-D gestures have become more common, and some of the technology will transfer to 3-D hand, head, and body gestural interfaces. Gaming and entertainment is, of course, driving much of the progress in gesture recognition. Similarly, technology developed in surveillance and security areas will also find uses in gesture recognition for virtual environments.

There are many open questions in this area. There has been little activity in evaluating usability (see Chapter 34, this volume) and understanding performance requirements and limitations of gestural interaction. Error rates are reported from 1% to 50%, depending on the difficulty and generality of the scenario. There are currently few common databases or metrics with which to compare research results. Can gesture recognition systems adapt to variations among individuals, or will extensive individual training be required? What about individual variation due to fatigue and other factors? How good do gesture recognition systems need to be to become truly useful in mass applications beyond simple games?

Each technology discussed in this chapter has its benefits and limitations. Devices that are worn or held—pens, gloves, body suits—are currently more advanced, as evidenced by the fact that there are many commercial products available. However, passive sensing (using cameras or other sensors) promises to be more powerful, more general, and less obtrusive than other technologies. It is likely that both camps will continue to improve and coexist and that new sensing technologies will arise to give even more choice to VE developers.

7. REFERENCES

Alon, J., Athitsos, V., Quan Yuan, Sclaroff, S. (2009). A unified framework for gesture recognition and spatiotemporal gesture segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.31, No.9, pp.1685-1699, September.

Anonymous (1987). Visual Signals. U.S. Army Field Manual FM-2160 [online]. Available: http://155.217.58.58/atdls.html

Assan, M., Grobel, K. (1997). Video-based sign language recognition using hidden Markov models. In I. Wachsmuth M. Fröhlich (Eds.), Gesture and sign language in human–computer interaction (Proceedings of the International Gesture Workshop). Bielefeld, Germany: Springer-Verlag, Berlin.

Baudel, T., Beaudouin-Lafon, M. (1993). CHARADE: remote control of objects using free-hand gestures. Communications of the ACM, 36(7), 28–35.

Black, M., Yacoob, Y. (1995). Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. International Conference on Computer Vision, pp. 374–381, Cambridge, MA.

Bobick, A. (1997). Movement, activity, and action: The role of knowledge in the perception of motion. Royal Society Workshop on Knowledge-based Vision in Man and Machine. London, England.

Bobick, A. F., Intille, S. S.,Davis, J. W., Baird, F., Pinhanez, C. S., Campbell, L. W., Ivanov, Y. A., Schütte, A., Wilson, A. (1999). The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment. Presence: Teleoperators and Virtual Environments, Vol. 8, No. 4, pp. 367-391, August.

Bleiweiss, A., Eshar, D., Kutliroff, G., Lerner, A., Oshrat, Y., Yanai, Y. (2010). Enhanced interactive gaming by blending full-body tracking and gesture animation. ACM SIGGRAPH ASIA 2010 Sketches, Article 34, 2 pages.

Böhm, K., Broll, W., Solokewicz, M. (1994). Dynamic gesture recognition using neural networks: A fundament for advanced interaction construction. In S. Fisher, J. Merrit, M. Bolan (Eds.), Stereoscopic displays and virtual reality systems. (SPIE Conference on Electronic Imaging Science and Technology, Vol. 2177). San Jose, CA.

Bolt, R. A. (1980). Put-That-There: Voice and gesture at the graphics interface. Computer Graphics, 14(3), 262–270.

Boyd, J., Little, J. (1998). Shape of motion and the perception of human gaits. IEEE Workshop on Empirical Evaluation Methods in Computer Vision. Santa Barbara, CA.

Bretzner, L., Laptev, I., Lindeberg, T. (2002). Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. IEEE International Conference on Automatic Face and Gesture Recognition, pp.423-428, 21-21 May.

Breuer, P., Eckes, C., Müller, S. (2007). Hand Gesture Recognition with a Novel IR Time-of-Flight Range Camera–A Pilot Study. In A. Gagalowicz and W. Philips (Eds.): MIRAGE 2007, LNCS 4418, pp. 247–260.

Cadoz, C. (1994). Les réalités virtuelles. Dominos, Flammarion,

Cassell, J., Steedman, M., Badler, N., Pelachaud, C., Stone, M., Douville, B., Prevost, S., Achorn, B. (1994). Modeling the interaction between speech and gesture. Proceedings of the Sixteenth Conference of the Cognitive Science Society.

Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., Clow, J. (1997). QuickSet: Multimodal interaction for distributed applications. In Proceedings of the Fifth Annual International Multimodal Conference, pp. 31-40, Seattle, WA.

Cutler, R., Turk, M. (1998). View-based interpretation of real-time optical flow for gesture recognition. International Conference on Automatic Face and Gesture Recognition. Nara, Japan.

Darrell, T., Pentland, A. (1993). Space-time gestures IEEE Conference on Computer Vision and Pattern Recognition, pp. 335-340, June.

Davis, J., Bobick, A. (1997). The representation and recognition of human movement using temporal trajectories. IEEE Conference on Computer Vision and Pattern Recognition. Puerto Rico.

Dipietro, L., Sabatini, A., Dario, P. (2008). A survey of glove-based systems and their applications, IEEE Transactions on Systems, Man and Cybernetics 38 (4), pp. 461-482.

Doliotis, P., Stefan, A., McMurrough, C., Eckhard, D., Athitsos, V. (2011). Comparing gesture recognition accuracy using color and depth information. ACM International Conference on Pervasive Technologies Related to Assistive Environments (PETRA '11). Article 20 , 7 pages.

Ekman, P., Friesen, W. V. (1978). Facial action coding system: A technique for the measurement of facial movement. Palo Alto, CA: Consulting Psychologists Press.

Elgammal, A., Shet, V., Yacoob, Y., Davis, L.S. (2003). Learning dynamics for exemplar-based gesture recognition. IEEE Conference on Computer Vision and Pattern Recognition, pp. I-571- I-578 June.

Essa, I., Pentland, A. (1997). Coding, analysis, interpretation and recognition of facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 19, No. 7.

Faste, H., Ghedini, F., Avizzano, C. A., Bergamasco, M. (2008). Passages: an artistic 3D interface. CHI 2008, Florence, Italy, April.

Fels, S., Hinton, G. (1995). Glove-Talk II: An adaptive gesture-to-formant interface. CHI ’95. Denver, CO.

Freeman, W., Tanaka, K., Ohta, J., Kyuma, K. (1996). Computer vision for computer games. International Conference on Automatic Face and Gesture Recognition. Killington, VT.

Gavrila, D. M., (1999). The visual analysis of human movement: a survey. Computer Vision and Image Understanding, Volume 73, Issue 1, pp. 82-98, January.

Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A. (2011). Efficient regression of general-activity human poses from depth images. IEEE International Conference on Computer Vision, October.

Hackenberg, G., McCall, R., Broll, W. (2011). Lightweight palm and finger tracking for real-time 3D gesture control. IEEE Virtual Reality Conference (VR 2011), vol., no., pp.19-26, 19-23 March.

Haritaoglu, I., Harwood, D., Davis, L. (1998). W4: Who? When? Where? What? A real-time system for detecting and tracking people. International Conference on Automatic Face and Gesture Recognition. Nara, Japan.

M.B. Holte, T.B. Moeslund, P. Fihl, (2010). View-invariant gesture recognition using 3D optical flow and harmonic motion context, Computer Vision and Image Understanding, Volume 114, Issue 12, December, Pages 1353-1361.

Hong, P., Turk, M., Huang, T.S. (2000). Gesture modeling and recognition using finite state machines. IEEE International Conference on Automatic Face and Gesture Recognition. pp.410-415.

Huang, T., Pavlovic, V. (1995). Hand-gesture modeling, analysis, and synthesis. International Workshop on Automatic Face- and Gesture-Recognition. Zurich, Switzerland.

Huang, Y., Monekosso, D., Wang, H., Augusto, J. C. (2011). A concept grounding approach for glove-based gesture recognition. International Conference on Intelligent Environments, pp. 358-361.

Hummels, C., Stappers, P. (1998). Meaningful gestures for human-computer interaction: Beyond hand gestures. International Conference on Automatic Face and Gesture Recognition. Nara, Japan.

Jaimes, A., Sebe, N., (2005). Multimodal human computer interaction: a survey. in N. Sebe, M.S. Lew, and T.S. Huang (Eds.): HCI/ICCV 2005, LNCS 3766, pp. 1-15.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. Perception and Psychophysics, 14:201–211.

Johnson, T. (1963). Sketchpad III: Three-dimensional graphical communication with a digital computer. AFIPS Spring Joint Computer Conference, 23, 347–353

Kadous, W. (1996). Computer recognition of Auslan signs with PowerGloves. Workshop on the Integration of Gesture in Language and Speech, Wilmington, DE.

Kirk, A. G., O'Brien, J. F., Forsyth, D. A. (2005). Skeletal parameter estimation from optical motion capture data. IEEE Conference on Computer Vision and Pattern Recognition.

Kang, H., Lee C. W., Jung, K. (2004). Recognition-based gesture spotting in video games. Pattern Recognition Letters, Volume 25, Issue 15, November, pp. 1701-1714.

Kehl, R. Van Gool, L. (2004). Real-time pointing gesture recognition for an immersive environment. IEEE Conference on Automatic Face and Gesture Recognition, Seoul, Korea, May.

Kendon, A. (1972). Some relationships between body motion and speech. In A. W. Siegman and B. Pope (Eds.), Studies in dyadic communication. New York: Pergamon Press.

Kratz, S. Rohs, M. (2010). A $3 gesture recognizer: simple gesture recognition for devices equipped with 3D acceleration sensors. International Conference on Intelligent User Interfaces (IUI '10). ACM, New York, NY, USA, 341-344.

Kühnel, C., Westermann, T., Hemmert, F., Kratz, S., Müller, A., Möller, S. (2011). I'm home: Defining and evaluating a gesture set for smart-home control. International Journal of Human-Computer Studies, Volume 69, Issue 11, October, pp. 693-704.

La Cascia, M., Isidoro, J., Sclaroff, S. (1998). Head tracking via robust registration in texture map images. IEEE Conference on Computer Vision and Pattern Recognition. Santa Barbara, CA.

Landay, J. A., Myers, B. A. (1995). Interactive sketching for the early stages of user interface design. ACM CHI '95, pp. 43-50.

Latoschik, M., Wachsmuth, I. (1997). Exploiting distant pointing gestures for object selection in a virtual environment. In I. Wachsmuth M. Fröhlich (Eds.), International Gesture Workshop: Gesture and sign language in human–computer interaction. Bielefeld, Germany.

LaViola, J. J. Jr. (1999). A survey of hand posture and gesture recognition techniques and technology. Technical Report CS-99-11, Department of Computer Science, Brown University.

Leap Motion (2012). http://live.leapmotion.com/about/.

Lee, C., Xu, Y. (1996). Online, interactive learning of gestures for human/robot interfaces. IEEE International Conference on Robotics and Automation, Vol. 4, pp. 2982-2987, Minneapolis.

Lee, H.-K., Kim, J.-H. (1999). An HMM-based threshold model approach for gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, No. 10, Oct..

Liu, J., Zhong, L.,Wickramasuriya, J., Vasudevan, V. (2009). uWave: Accelerometer-based personalized gesture recognition and its applications. Pervasive and Mobile Computing, Volume 5, Issue 6, pp. 657-675, December.

Long, A., Landay, J., Rowe, L. (1998).PDA and gesture uses in practice: Insights for designers of pen-based user interfaces (Report CSD-97-976). Berkeley, CA: University of California, Berkeley, CS Division, EECS Department.

Lucente, M., Zwart, G., George, A. (1998). Visualization space: A testbed for deviceless multimodal user interface. Intelligent Environments Symposium. Stanford, CA: AAAI Spring Symposium Series.

Maggioni, C., Kämmerer, B. (1998). GestureComputer—history, design and applications. In R. Cipolla A. Pentland (Eds.), Computer vision for human–machine interaction. Cambridge University Press, Cambridge, U.K.

Marrin, T., Picard, R. (1998). The conductor's jacket: A testbed for research on gestural and affective expression. Twelth Colloquium for Musical Informatics, Gorizia, Italy.

McNeill. D. (1992). Hand and mind: What gestures reveal about thought. Chicago: University of Chicago Press.

McNeill, D. (Ed.) (2000). Language and Gesture, Cambridge, New York, Melbourne and Madrid: Cambridge University Press.

Mitra, S., Acharya, T. (2007). Gesture recognition: a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Vol.37, No. 3, pp. 311-324, May.

Moeslund, Th.B., Hilton, A., Krüger, V., Sigal, L. (Eds.) (2011). Visual Analysis of Humans: Looking at People, Springer.

Morency, L.-P., Quattoni, A., Darrell, T. (2007). Latent-dynamic discriminative models for continuous gesture recognition. IEEE Conference on Computer Vision and Pattern Recognition, June.

Moses, Y., Reynard, D., Blake, A. (1995). Determining facial expressions in real time. International Conference on Computer Vision. Cambridge, MA.

Mulder, A. (1996). Hand gestures for HCI (Tech. Rep. No. 96-1). Simon Fraser University, School of Kinesiology.

Oka, K., Sato, Y., Koike, H. (2002). Real-time fingertip tracking and gesture recognition. IEEE Computer Graphics and Applications, Vol.22, No.6, pp. 64-71, Nov/Dec.

Oliver, N., Pentland, A., Bérard, F. (1997). LAFTER: Lips and face real-time tracker. IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico.

Ong, S.C.W., Ranganath, S. (2005). Automatic sign language analysis: a survey and the future beyond lexical meaning. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.27, no.6, pp. 873- 891, June.

Otsuka, T., Ohya, J. (1998). Recognizing abruptly changing facial expressions from time-sequential face images. IEEE Conference on Computer Vision and Pattern Recognition. Santa Barbara, CA.

Overholt, D., Thompson, J., Putnam, L., Bell, B., Kleban, J., Sturm, B., Kuchera-Morin, J. (2009). A multimodal system for gesture recognition in interactive music performance. Comput. Music J. 33, 4, December, pp. 69-82.

Oviatt, S. L. (1996). Multimodal interfaces for dynamic interactive maps. ACM CHI: Human Factors in Computing Systems, pp. 95-102.

Paradiso, J.,Sparacino, F. (1997). Optical tracking for music and dance Performance. Fourth Conference on Optical 3-D Measurement Techniques. Zurich, Switzerland.

Parvini, F., Mcleod, D., Shahabi, C., Navai, B., Zali, B., Ghandeharizadeh, S. (2009). An approach to glove-based gesture recognition. International Conference on Human-Computer Interaction. Part II: Novel Interaction Methods and Techniques, Julie A. Jacko (Ed.). Springer-Verlag, Berlin, Heidelberg, 236-245.

Pavlovic. V, Sharma, R., Huang, T. (1996). Gestural interface to a visual computing environment for molecular biologists. International Conference on Automatic Face and Gesture Recognition. Killington, VT.

Pylvänäinen, T. (2005). Accelerometer based gesture recognition using continuous HMMs. LNCS: Pattern Recognition and Image Analysis, vol.3522, pp. 413-430, Springer Berlin.

Ren, Z., Meng, J., Yuan, J., Zhang, Z. (2011). Robust hand gesture recognition with kinect sensor. ACM International Conference on Multimedia, pp.759-760.

Reyes, M., Dominguez, G., Escalera, S. (2011). Feature weighting in dynamic timewarping for gesture recognition in depth data. IEEE Workshop on Consumer Depth Cameras for Computer Vision, Barcelona, November.

Song, Y., Demirdjian, D., Davis, R. (2012). Continuous body and hand gesture recognition for natural human-computer interaction. ACM Transactions on Interactive and Intelligent Systems 2, 1, Article 5 (March), 28 pages.

Stark, M., Kohler, M. (1995). Video-based gesture recognition for human-computer interaction. In W. D. Fellner(Ed.), Modeling—Virtual Worlds—Distributed Graphics.

Starner, T., Pentland, A. (1995). Visual recognition of American Sign Language using hidden Markov models. International Workshop on Automatic Face- and Gesture-Recognition. Zurich, Switzerland.

Sturman, J. (1992). Whole-hand Input. Unpublished doctoral dissertation, MIT Media

Labortory.

Suk, H., Sin, B., Lee, S. (2008).Recognizing hand gestures using dynamic bayesian network, IEEE Conference on Automatic Face and Gesture Recognition.

Takahashi, T., Kishino, F. (1991). Gesture coding based in experiments with a hand-gesture interface device. SIGCHI Bulletin, 23(2), 67–73.

Triesch, J., von der Malsburg, C. (1996). Robust classification of hand postures against complex backgrounds. International Conference on Automatic Face and Gesture Recognition. Killington, VT.

Väänänen, K., Böhm, K. (1992). Gesture-driven interaction as a human factor in virtual environments—An approach with neural networks. Virtual Reality Systems. British Computer Society.

Van den Bergh, M., Van Gool, L., (2011). Combining RGB and ToF cameras for real-time 3D hand gesture interaction. IEEE Workshop on Applications of Computer Vision (WACV 2011), pp.66-72, 5-7 January.

Vogler, C., Metaxas, D. (2001). A framework for recognizing the simultaneous aspects of American sign language. Computer Vision and Image Understanding, Volume 81, Issue 3, March, pp. 358-384.

Wachs, J. P., Kölsch, M., Stern, H., Edan, Y. (2011). Vision-based hand-gesture applications. Commun. ACM 54, 2 (February), 60-71.

Wang, Q., Chen, X., Zhang, L.-G., Wang, C., Gao, W. (2007). Viewpoint invariant sign language recognition. Computer Vision and Image Understanding, Volume 108, Issues 1–2, October–November, pp. 87-97.

Wang, S. B., Quattoni, A., Morency, L.-P., Demirdjian, D., Darrell, T. (2006). Hidden conditional random fields for gesture recognition. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1521- 1527.

Wang, X., Tarrío, P.,Metola, E., Bernardos, A. M., Casar, J. R. (2012). Gesture recognition using mobile phone’s inertial sensors. In S. Omatu et al. (Eds.): Distributed Computing and Artificial Intelligence, AISC 151, pp. 173–184, Springer-Verlag Berlin Heidelberg.

Weng, J., Cui, Y. (1998). Recognition of hand signs from complex backgrounds. In R. Cipolla and A. Pentland (Eds.), Computer vision for human-machine interaction. Cambridge University Press.

Wexelblat, A. (1994). A feature-based approach to continuous-gesture analysis. Unpublished master’s thesis, MIT Media Labortory.

Wilson, A., Bobick, A. (1999). Parametric hidden markov models for gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, No. 9, pp. 884-900, September.

Wilson, A. D. (2004). TouchLight: an imaging touch screen and display for gesture-based interaction. International Conference on Multimodal Interfaces, State College, PA, October.

Wobbrock, J. O., Morris, M. R., Wilson, A. D. (2009). User-defined gestures for surface computing. International Conference on Human Factors in Computing Systems (Boston, MA, USA, April 04 - 09, 2009), pp. 1083-1092.

Wren, C., Azarbayejani, A., Darrell, T., Pentland, A. (1996). Pfinder: Real-time tracking of the human body. International Conference on Automatic Face and Gesture Recognition. Killington, VT.

Wu, J.,Pan, G., Zhang, D., Qi, G., Li, S. (2009). Gesture recognition with a 3-D accelerometer. In D. Zhang et al. (Eds.): UIC 2009, LNCS 5585, pp. 25-38.

Wu, Y., Huang, T. S. (1999). Vision-based gesture recognition: a review. In A. Braffort et al. (Eds.): Gesture Workshop (GW'99), LNAI 1739, pp. 103-115, Springer-Verlag.

Yamamoto, J., Kondo, T., Yamagiwa, T., Yamanaka, K. (1998). Skill recognition. International Conference on Automatic Face and Gesture Recognition. Nara, Japan.

Yang, M.-H., Ahuja, N., Tabb, M. (2002). Extraction of 2D motion trajectories and its application to hand gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.24, no.8, pp. 1061- 1074, August.

Xbox (2012). http://support.xbox.com/en-US/kinect/body-tracking/body-controller

(September).

Zaki, M. M., Shaheen, S. I. (2011). Sign language recognition using a combination of new vision based features. Pattern Recognition Letters, Volume 32, Issue 4, 1 March, pp. 572-577.

Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H., Presti, P. (2011). American sign language recognition with the kinect. ACM International Conference on Multimodal Interfaces, pp. 279-286.

Zeleznik, R. C., Herndon, K. P., Hughes J. F. (1996). Sketch: An interface for sketching 3D scenes. ACM SIGGRAPH, pp. 163-170.

Zelinsky, A., Heinzmann, J. (1996). Real-time visual recognition of facial gestures for human–computer interaction. International Conference on Automatic Face and Gesture Recognition. Killington, VT.

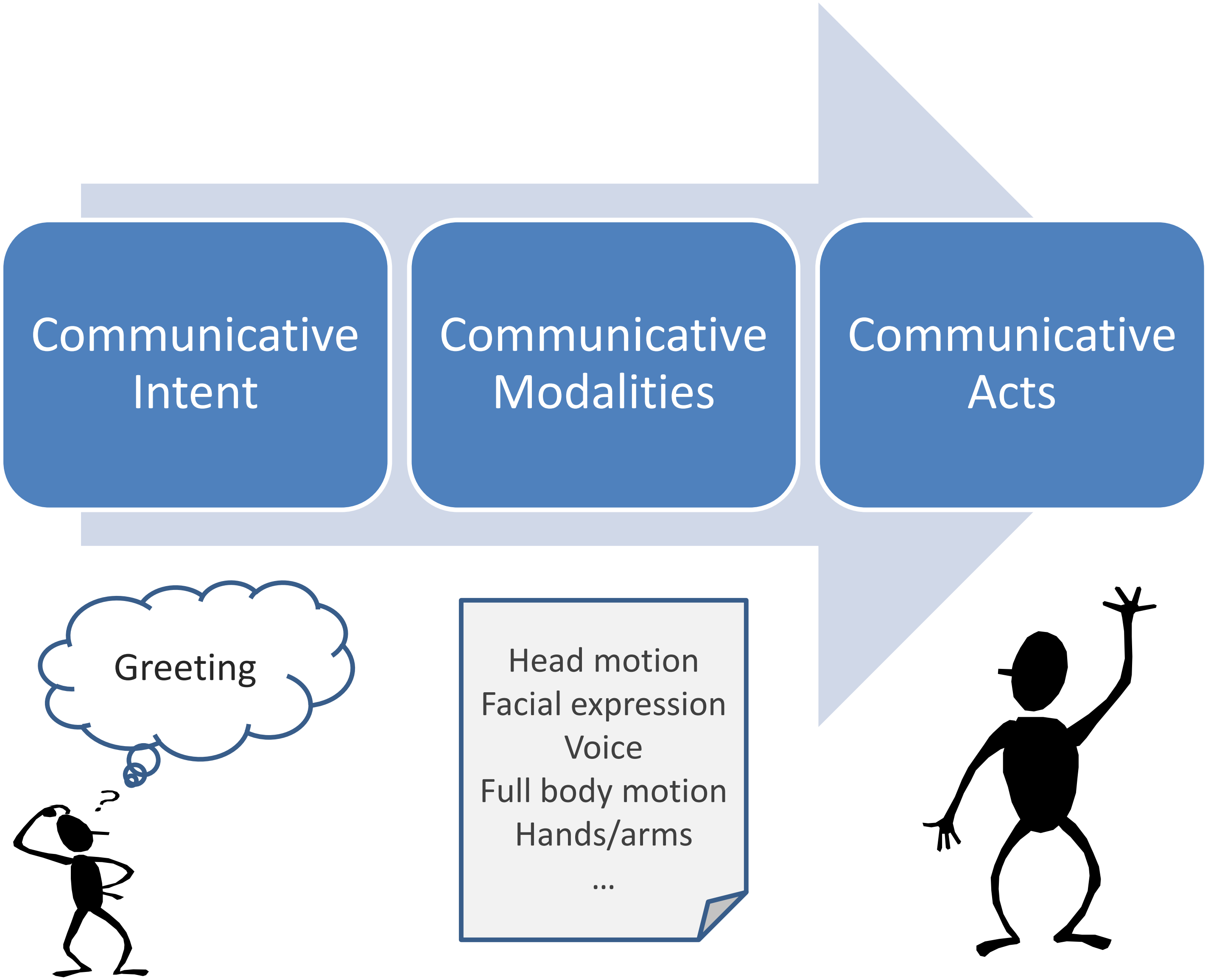

FIG. 10.1. Observable communication acts (such as gestures) are the result of expressing intent via communication modalities.

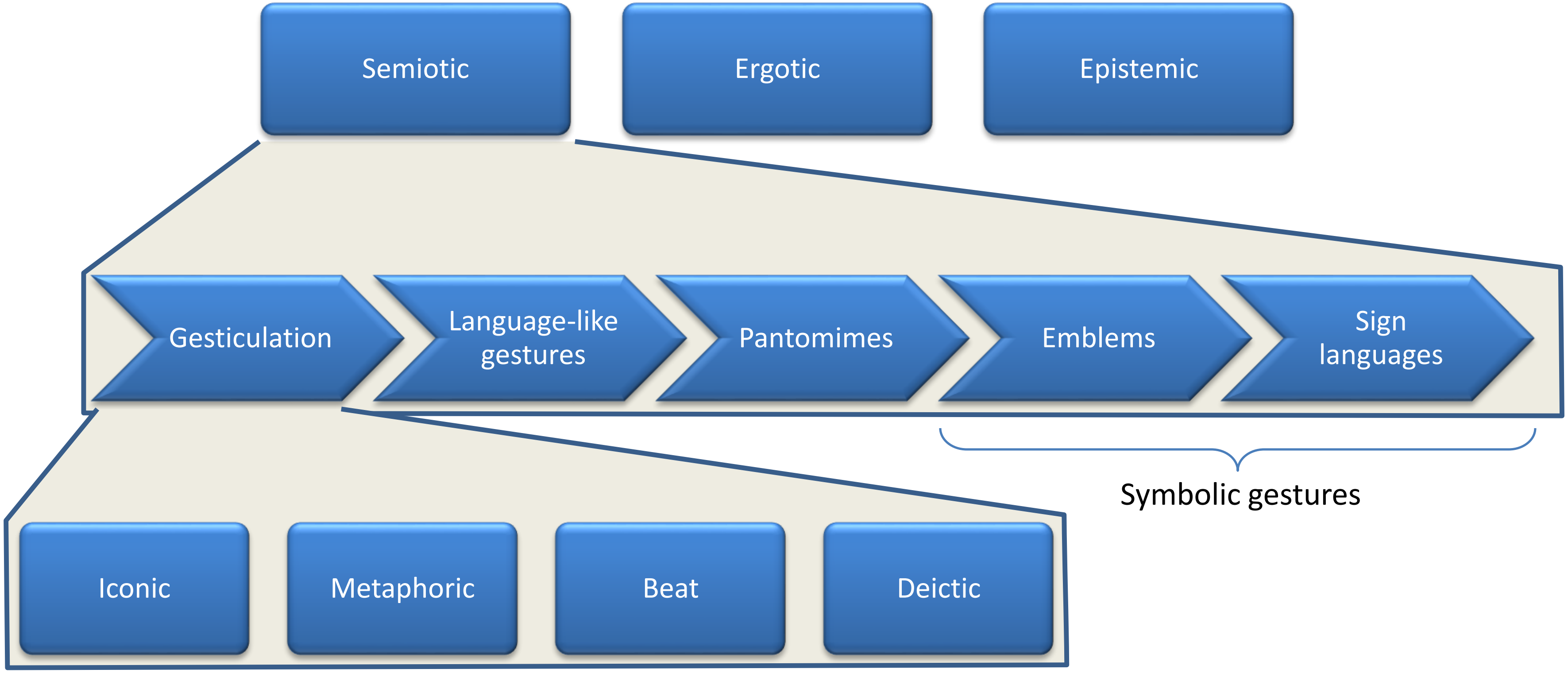

FIG. 10.2. Top: Cadoz's functional roles of human gesture. Middle: Kendon's gesture continuum. Bottom: McNeill's four gesture types.

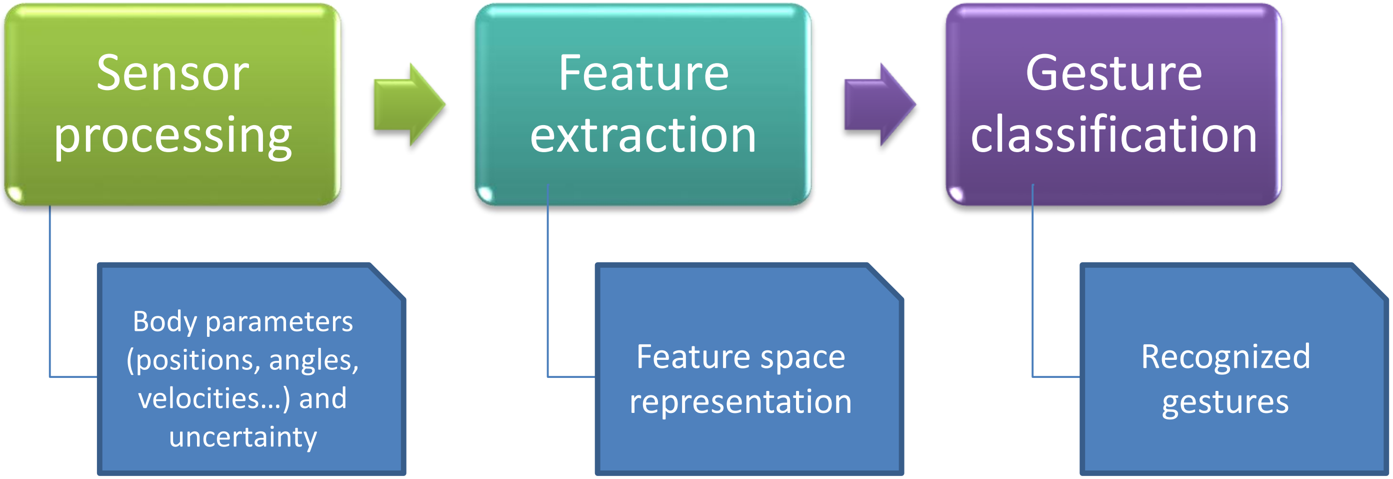
Sensor processing
Feature extraction
Gesture classification
Body parameters (positions, angles, velocities…) and uncertainty
Feature space representation
Recognized gestures

FIG. 10.3. Pattern recognition systems

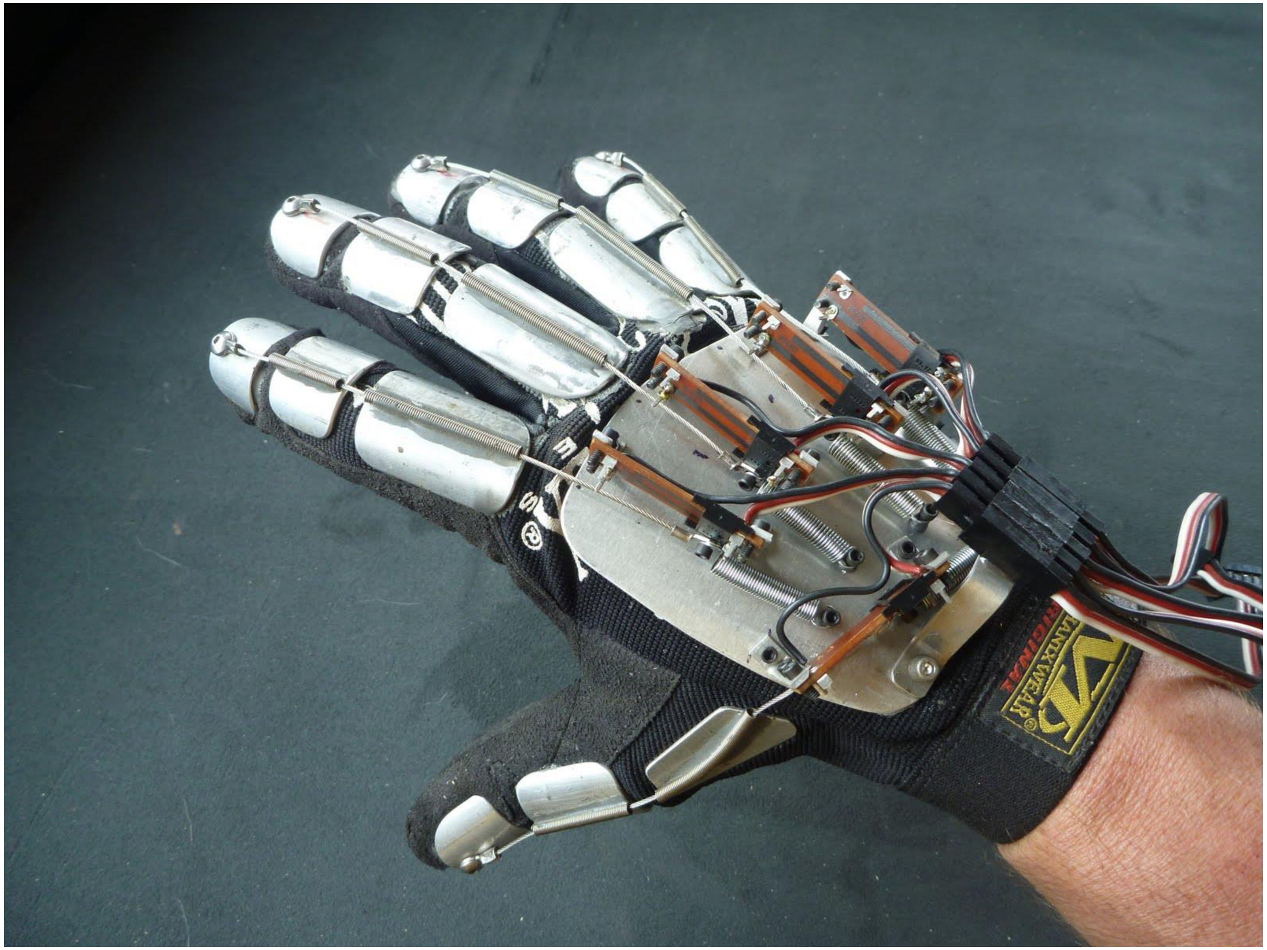

FIG. 10.4. An exoskeleton data glove (Courtesy of Christian Ristow)

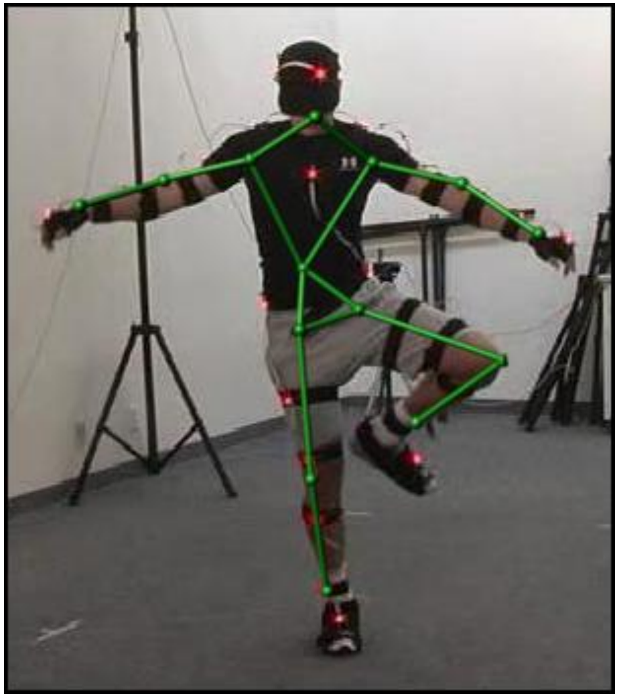

FIG. 10.5. An optical motion capture system in use (Kirk et al. 2005) (Copyright 2005 Adam Kirk, James O'Brien, and David Forsyth. Used with permission.**)**

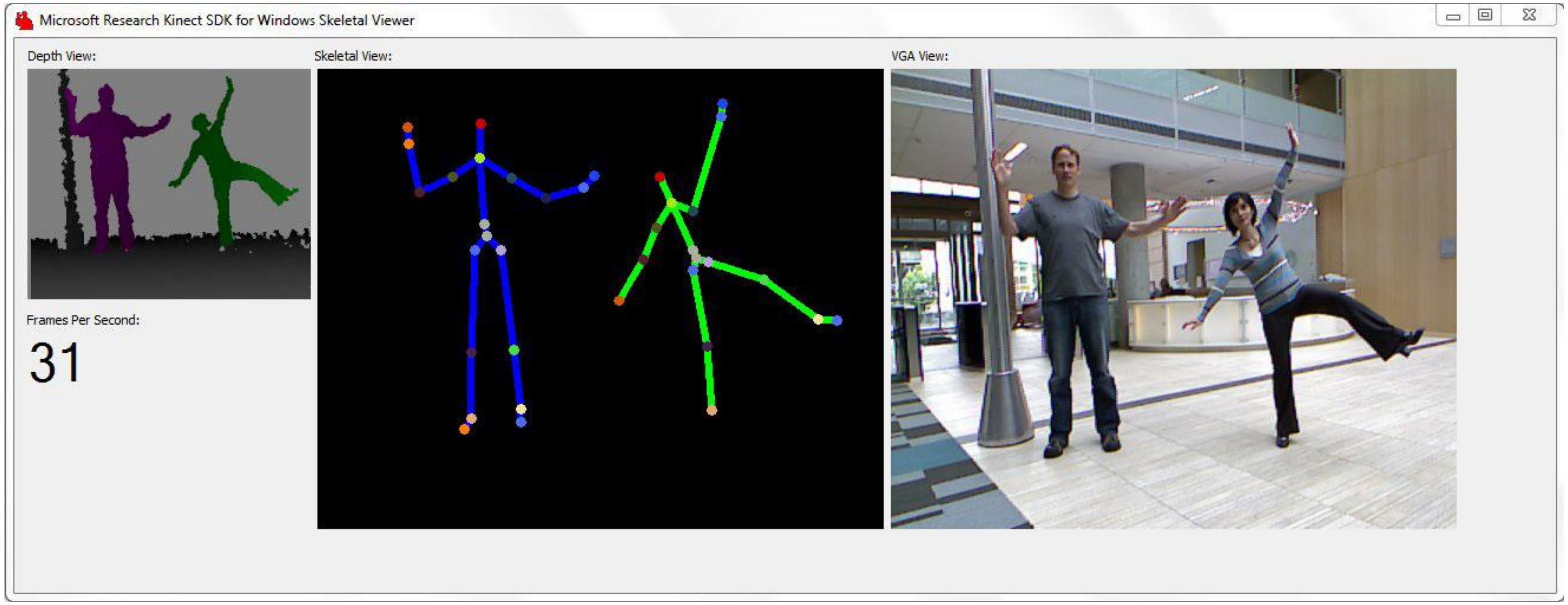

FIG. 10.6. The Microsoft Kinect in action (depth image, skeleton model, RGB image)

Publicly available at:
http://www.microsoft.com/en-us/news/ImageDetail.aspx?id=3EF3F47FF87ACC1CE3CFF8E03C1151AA6C4CC004