Chapter 10. Gesture Recognition

Matthew Turk

1. Introduction 1
2. The Nature of Gesture 3
3. Representations of Gesture 6
  3.1 Pen-based gesture recognition 8
  3.2 Tracker-based gesture recognition 9
    3.2.1 Instrumented gloves 9
    3.2.2 Body suits 11
  3.3 Passive vision-based gesture recognition 12
    3.3.1 Head and face gestures 15
    3.3.2 Hand and arm gestures 16
    3.3.3 Body gestures 17
4. Suggestions for Systems Design 18
5. Conclusions 19
6. References 20

1. Introduction

A primary goal of virtual environments (VE) is to provide natural, efficient, powerful, and flexible interaction. Gesture as an input modality can help meet these requirements. Human gestures are certainly natural and flexible, and may often be efficient and powerful, especially as compared with alternative interaction modes. This chapter will cover automatic gesture recognition, particularly computer vision based techniques that do not require the user to wear extra sensors, clothing or equipment.

The traditional two-dimensional (2D), 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. Devices which sense body position and orientation, direction of gaze, speech and sound, facial expression, galvanic skin response, and other aspects of human behavior or state can be used to mediate communication between the human 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 appears promising.

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 task at hand, or imposes a high cognitive load on the user, it becomes a burden and an obstacle to a successful VE experience. Therefore, there is focused interest in technologies that are unobtrusive and passive.

To support gesture recognition, human position and movement must be tracked and interpreted in order to recognize semantically meaningful gestures. While tracking of 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 Figure 1. The output of position (and other) sensing must be interpreted to allow users to communicate more naturally and effortlessly through gesture.

_____

[Figure 1 goes here.]

_____

Gesture is 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. 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., visual or verbal behaviors 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 indicate agreement, participation, attention, conversation turn taking, etc.

Given that the human body can express a huge variety of gestures, what is appropriate to sense? 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. More subtle cues such as hand tension, overall muscle tension, locations of self-contact, and even pupil dilation may be of use.

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. This chapter covers interpretation of tracking data from such devices in order to recognize gestures. Additional attention is focused on passive sensing from cameras, using computer vision techniques. The chapter concludes with suggestions for gesture recognition system design.

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.

Gesture recognition is the process by which 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; here the interest is 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.

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

- navigating through a space;
- specifying items of interest;
- manipulating objects in the environment;
- changing object values;
- controlling virtual objects; and
- 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:

- analysis of usability;
- analysis of user tasks;
- monitoring of changes in a user's state;
- better understanding a user's intent or emphasis; and
- 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). And, like speech and handwriting, gestures vary among individuals, they vary from instance to instance for a given individual, and they are subject to the effects of co-articulation.

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.

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," depicted in Figure 2, defining five different kinds of gestures:

- *Gesticulation.* Spontaneous movements of the hands and arms that accompany speech.
- *Language-like 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, which are well defined.

As the list progresses (moving from left to right in Figure 2), the association with speech declines, language properties increase, spontaneity decreases, and social regulation increases.

_____

[Figure 2 goes 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.

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) describe a system that models the relationship between speech and gesture and generates interactive dialogs between three-dimensional (3D) animated characters that gesture as they speak.

These spontaneous gestures (*gesticulation* in Kendon's Continuum) make up some 90% of human gestures. 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 right side of Figure 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* system by Bolt (1980). The remainder of this chapter will focus on *symbolic gestures* (which includes emblematic gestures and predefined gesture languages) and *deictic gestures*.

3. Representations of Gesture

The concept of gesture is loosely defined, and depends on the context of the interaction. Recognition of natural, continuous gestures requires temporally segmenting gestures. Automatically segmenting gestures is difficult, and is often finessed or ignored in current systems by requiring a starting position in time and/or space. Similar to this is the problem of distinguishing intentional gestures from other "random" movements. There is no standard way to do gesture recognition – a variety of representations and classification schemes are used. However, most gesture recognition systems share some common structure.

Gestures can be static, where the user assumes a certain pose or configuration, or dynamic, defined by movement. McNeill (1992) defines three phases of a dynamic gesture: pre-stroke, stroke, and post-stroke. 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 *co-articulations* 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) describe 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, and datasuits, which are attached to the user, or indirectly using cameras and computer vision techniques. Each sensing technology differs along several dimensions, including accuracy, resolution, 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 of a virtual object, here the 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 describing positions, velocities, and angles of relevant body parts and 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 Figure 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.

[Figure 3 goes here.]

Because gestures are highly variable, from one person to another and from one example to another within a single person, 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 to be recognized only works for trivial systems; in general, a system needs to be trained through some kind of learning. As with speech recognition systems, there is often a tradeoff 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.

Static gesture, or pose, recognition can be accomplished by a straightforward implementation of Figure 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, dynamic time warping, hidden Markov models (HMMs), and Bayesian networks. Some examples will be presented in the following sections.

3.1 Pen-based gesture recognition

Recognizing gestures from 2D input devices 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. More recently, systems such as the OGI QuickSet system (Cohen et al. 1997) have demonstrated the utility of pen-based gesture recognition in concert with speech recognition to control a virtual environment. QuickSet recognizes 68 pen gestures, including map symbols, editing gestures, route indicators, area indicators, and taps. Oviatt (1996) has demonstrated significant benefits of using speech and pen gestures together in certain tasks. Zeleznik et al. (1996) and Landay and Myers (1995) developed interfaces that recognize gestures from pen-based sketching.

A significant benefit of pen-based gestural systems is that sensing and interpretation is relatively straightforward as compared with vision-based techniques. There have been commercially available Personal Digital Assistants (PDAs) for several years, starting with the Apple Newton, and more recently the 3Com PalmPilot and various Windows CE devices. These PDAs perform handwriting recognition and allow 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.

Although pen-based gesture recognition is promising for many HCI environments, it presumes the availability of, and proximity to, a flat 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 next two sections cover two primary technologies for gesture recognition in virtual environments: instrumented gloves and vision-based interfaces.

3.2 Tracker-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 useful in a gestural interface, the focus here is on gestures based on input from tracking the hands and body.

3.2.1 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.

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 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, 3D spatial information, wrist rotation);
- provides data at a high sampling frequency;
- easy to use;
- no line-of-sight occlusion problems;
- relatively low cost versions available; and
- data 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; and
- the user is forced to wear a somewhat cumbersome device.

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 multi-agent 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; Takahashi and Kishino (1991) for the Japanese Kana manual alphabet. The system of Lee and Xu (1996) could learn and recognize new gestures online.

Despite the fact that many, if not most, gestures involve two hands, most 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.

The HIT Lab at the University of Washington developed GloveGRASP, a C/C++ class library that allows software developers to add gesture recognition capabilities to SGI systems, including user-dependent training and one- or two-handed gesture recognition. A commercial version of this system is available from General Reality.

3.2.2 Body suits

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. One way to approach the recognition of human movements and postures is to optically measure the 3D position of several such markers attached to the body 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 systems only require dots or small balls to be placed on top of a user's clothing, all of these body motion capture systems are referred to herein generically as "body suits."

Body suits 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 typically non-trivial. The optical systems typically use several cameras and process their data offline – their major advantage is the lack of wires and a tether.

Body suits 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) have developed an instrumented jacket for an orchestral conductor that includes physiological monitoring to study the correlation between affect, gesture, and musical expression.

Although current optical and electromechanical tracking technologies are cumbersome and therefore contrary to the desire for more natural interfaces, it is likely that advances in sensor technology will 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, instrumented body suits, which are currently exceedingly cumbersome, may be displaced by sensing technologies embedded in belts, shoes, eyeglasses, and even shirts and pants. While sensing technology has a long way to go to reach these ideals, passive sensing using computer vision techniques is beginning to make headway as a user-friendly interface technology.

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.

3.3 Passive vision-based gesture recognition

The most significant disadvantage of the tracker-based systems in Section 3.2 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 suffer from these shortcomings, albeit not as severely. What many have wished for is a technology that provides real-time data useful for analyzing and recognizing human motion

that is passive and non-obtrusive. Computer vision techniques have the potential to meet these requirements.

Vision-based interfaces use one or more cameras to capture images, at a frame rate of 30 Hz or more, and interpret those images to produce visual features that can be used to interpret human activity and recognize gestures. 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 decade, 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.

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 – e.g., 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 this adds correspondence and integration problems.

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 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 vision and 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 enable different analysis and interpretation. For example, tracking devices can in principle detect fast and subtle movements of the fingers while a user is waving his hands, while human vision in that case will 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, while tracking devices do not. From a research 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. There is a growing interest in CMOS-based cameras, which promise miniaturized, low cost, low power cameras integrated with processing circuitry on a single chip. With its integrated processing, such a sensor could conceivably output motion or gesture parameters to the virtual environment.

Currently, most computer vision systems for recognition look something like Figure 3. Analog cameras feed their signal into a digitizer board, or framegrabber, which may do a DMA transfer directly to host memory. Digital cameras bypass the analog-to-digital conversion and go straight to 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 2D or 3D 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, are they combined early (stereo) or late (multi-view)?
- 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, the lighting, the speed of movement, etc.?
- 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, etc.)?
- Two- or three-dimensional representation. Does the system construct a 3D 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)?

3.3.1 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; and
- looks of surprise, happiness, disgust, anger, sadness, etc.

People display a wide range of facial expressions. Ekman and Friesen (1978) developed a system called 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/face gestures. Moses et al. (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 spatio-temporal motion-energy templates of the whole face for each different expression – these templates were then used for expression recognition. Oliver et al. (1997) describe a real-time system for tracking the face and mouth that recognized facial expressions and head movements. Otsuka and Ohya (1998) model coarticulation in facial expressions and use an HMM for recognition.

Black and Yacoob (1995) used local parametric motion models to track and recognize both rigid and non-rigid 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 et al. (1998) extended this approach using texture-mapped surface models and non-planar parameterized motion models to better capture the facial motion.

3.3.2 Hand and arm gestures

Hand and arm gestures receive the most attention among those who study gesture – in fact, many (if not 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 produce precise hand and arm configuration while 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, while a sequence of these makes a gesture. Similarly, Maggioni and Kämmerer (1998) described the GestureComputer, which recognized both hand gestures and head movements. Other systems that recognize hand postures amidst complex visual backgrounds are reported by Weng and Cui (1998) and Triesch and von der Malsburg (1996).

There has been a lot of interest in creating devices to automatically interpret various sign languages to aid the deaf community. One of the first to use computer vision without requiring the user to wear anything special was built by Starner (1995), who used HMMs to recognize a limited vocabulary of ASL sentences. A more recent effort, which uses HMMs to recognize Sign Language of the Netherlands is described by Assan and Grobel (1997).

The recognition of hand and arm gestures has been applied to entertainment applications. Freeman et al. (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. A nice overview of work up to 1995 in hand gesture modeling, analysis, and synthesis is presented by Huang and Pavlovic (1995).

3.3.3 Body gestures

This section includes tracking full body motion, recognizing body gestures, and recognizing human activity. 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. Bobick (1997) proposed a taxonomy of motion understanding in terms of:

- 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 et al., 1996) developed at the MIT Media Lab has been used by a number of groups to do body tracking and gesture recognition. It forms a 2D representation of the body, using statistical models of color and shape. The body model provides an effective interface for applications such as video games, interpretive dance, navigation, and interaction with virtual characters. Lucente et al. (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 46, 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 et al. (1998) describe 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, an interactive, immersive, narrative environment for children. A nice online description of this project can be found at http://vismod.www.media.mit.edu/vismod/demos/kidsroom/.

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 et al., 1998) tracks people and detects patterns of activity.

4. Suggestions for Systems Design

There has been little work in evaluating the utility and usability of gesture recognition systems. 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 what they affect must be clear to the user.

Do give the user feedback. Feedback is essential to let the user know when a gesture has been 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 2D 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, segmentation of gestures can be quite difficult.

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 quickly.

Don't make the gestures to be recognized too similar. For ease of classification and to help the user.

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 increase the user's cognitive load. Having to remember the whats, wheres, and hows of a gestural interface can make it oppressive 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, since it requires *recall* rather than just *recognition* among a list.

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.

5. Conclusions

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 *Gesture Workshops* and the *International Conference on Automatic Face and Gesture Recognition*.

There is still much to be done before gestural interfaces, which track and recognize human activities, can become pervasive and cost-effective for the masses. However, much progress has been made in the past decade and with the continuing march towards computers and sensors that are faster, smaller, and more ubiquitous, there is cause for optimism. As PDAs and pen-based computing continue to

proliferate, pen-based 2D gestures should become more common, and some of the technology will transfer to 3D hand, head, and body gestural interfaces. 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 no 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?

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 co-exist, often be used together in systems, and that new sensing technologies will arise to give even more choice to VE developers.

6. References

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

Assan, M. and Grobel, K. (1997). Video-based sign language recognition using hidden Markov models. In I. Wachsmuth and M. Fröhlich (Eds.), Gesture and Sign Language in Human-Computer Interaction. Proc. International Gesture Workshop, Bielefeld, Germany.

Baudel, T. and Beaudouin-Lafon, M. (1993). CHARADE: remote control of objects using free-hand gestures. Communications of the ACM (pp. 28-35), Vol. 36, No. 7.

Black, M. and Yacoob, Y. (1995). Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. Proc. ICCV (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.

Böhm, K., Broll, W., and Solokewicz, M. (1994). Dynamic gesture recognition using neural networks; a fundament for advanced interaction construction. In S. Fisher, J. Merrit, and 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 (pp. 262-270).

Boyd, J. and Little, J. (1998). Shape of Motion and the Perception of Human Gaits. IEEE Workshop on Empirical Evaluation Methods in Computer Vision. CVPR 98, Santa Barbara, CA.

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

Cassell, J., Steedman, M., Badler, N., Pelachaud, C., Stone, M., Douville, B., Prevost, S., and 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., and Clow, J. (1997). QuickSet: Multimodal interaction for distributed applications, Proceedings of the Fifth Annual International Multimodal Conference (pp 31-40), Seattle, WA, ACM Press.

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

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

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

Essa, I. and Pentland, A. (1997). Coding, Analysis, Interpretation and Recognition of Facial Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence,Vol. 19 (7), IEEE Computer Society Press.

Fels, S. and Hinton, G. (1995). Glove-TalkII: An Adaptive Gesture-to-Formant Interface. CHI'95. Denver, CO.

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

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

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

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

Johnson, T. (1963). Sketchpad III: Three Dimensional Graphical Communication with a Digital Computer. AFIPS Spring Joint Computer Conference, 23. (pp. 347-353).

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

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.

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

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

Latoschik, M. and Wachsmuth, I. (1997). Exploiting distant pointing gestures for object selection in a virtual environment. In I. Wachsmuth, and M. Fröhlich (Eds.), Gesture and Sign Language in Human-Computer Interaction. International Gesture Workshop, Bielefeld, Germany.

Lee, C. and Xu, Y. (1996). Online, Interactive Learning of Gestures for Human/Robot Interfaces. 1996 IEEE International Conference on Robotics and Automation, Vol. 4 (pp. 2982-2987). Minneapolis, MN.

Long, A., Landay, J., and Rowe, L. (1998). PDA and Gesture Uses in Practice: Insights for Designers of Pen-Based User Interfaces. Report #CSD-97-976, CS Division, EECS Department. UC Berkeley. Berkeley, CA.

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

Maggioni, C. and Kämmerer, B. (1998). GestureComputer – history, design and applications. In R. Cipolla and A. Pentland (Eds.), Computer Vision for Human-Machine Interaction. Cambridge University Press.

Marrin, T. and Picard, R. (1998). The Conductor's Jacket: a Testbed for Research on Gestural and Affective Expression. XII Colloquium for Musical Informatics, Gorizia, Italy.

McNeill. D. (1992). Hand and Mind: What Gestures Reveal about Thought. Chicago: University of Chicago Press.

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

Mulder, A. (1996). Hand gestures for HCI. Technical Report 96-1, School of Kinesiology. Simon Fraser University.

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

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

Oviatt, S. L. (1996). Multimodal interfaces for dynamic interactive maps. Proceedings of CHI'96 Human Factors in Computing Systems (pp. 95-102). ACM Press, NY.

Paradiso, J. and Sparacino, F. (1997). Optical Tracking for Music and Dance Performance. Fourth Conference on Optical 3-D Measurement Techniques. Zurich, Switzerland.

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

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

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

Sturman, J. (1992). Whole-hand Input. Ph.D. Thesis, MIT Media Laboratory. Cambridge, MA.

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

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

Väänänen, K. and Böhm, K. (1992). Gesture driven interaction as a human factor in virtual environments – an approach with neural networks. Proc. Virtual Reality Systems. British Computer Society, Academic Press.

Weng, J. and 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. M.S. Thesis, MIT Media Laboratory, Cambridge, MA.

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

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

Zeleznik, R.C., Herndon, K.P., and Hughes J.F. (1996). Sketch: An interface for sketching 3D scenes. Computer Graphics (Proceedings of SIGGRAPH '96).

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

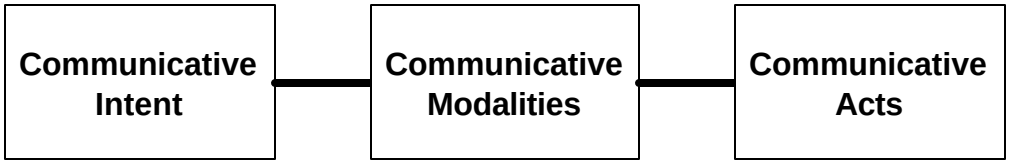

Figure 1. Observable communication acts (such as gestures) are the result of expressing intent via communication modalities.

Figure 2. Kendon's gesture continuum

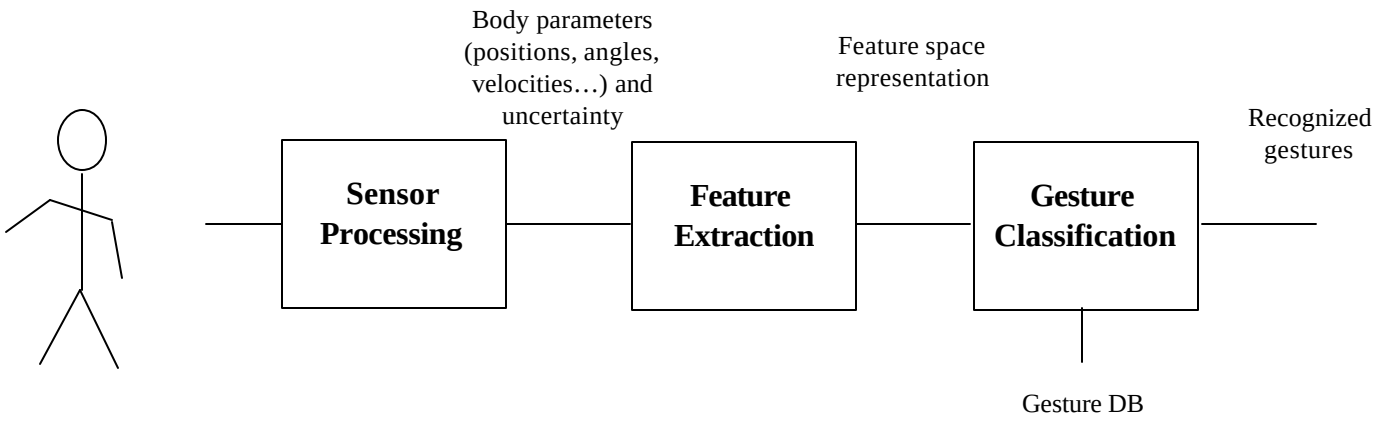

Figure 3. Pattern recognition systems