Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review

Vladimir I. Pavlovic, *Student Member, IEEE*,
Rajeev Sharma, *Member, IEEE*,
and Thomas S. Huang, *Fellow, IEEE*

Abstract—The use of hand gestures provides an attractive alternative to cumbersome interface devices for human-computer interaction (HCI). In particular, visual interpretation of hand gestures can help in achieving the ease and naturalness desired for HCI. This has motivated a very active research area concerned with computer vision-based analysis and interpretation of hand gestures. We survey the literature on visual interpretation of hand gestures in the context of its role in HCI. This discussion is organized on the basis of the method used for modeling, analyzing, and recognizing gestures. Important differences in the gesture interpretation approaches arise depending on whether a *3D model* of the human hand or an image *appearance model* of the human hand is used. 3D hand models offer a way of more elaborate modeling of hand gestures but lead to computational hurdles that have not been overcome given the real-time requirements of HCI. Appearance-based models lead to computationally efficient "purposive" approaches that work well under constrained situations but seem to lack the generality desirable for HCI. We also discuss implemented gestural systems as well as other potential applications of vision-based gesture recognition. Although the current progress is encouraging, further theoretical as well as computational advances are needed before gestures can be widely used for HCI. We discuss directions of future research in gesture recognition, including its integration with other natural modes of human-computer interaction.

Index Terms—Vision-based gesture recognition, gesture analysis, hand tracking, nonrigid motion analysis, human-computer interaction.

1 Introduction

With the massive influx of computers in society, *human-computer interaction*, or HCI, has become an increasingly important part of our daily lives. It is widely believed that as the computing, communication, and display technologies progress even further, the existing HCI techniques may become a bottleneck in the effective utilization of the available information flow. For example, the most popular mode of HCI is based on simple mechanical devices—keyboards and mice. These devices have grown to be familiar but inherently limit the speed and naturalness with which we can interact with the computer. This limitation has become even more apparent with the emergence of novel display technology such as virtual reality [2], [78], [41]. Thus in recent years there has been a tremendous push in research toward novel devices and techniques that will address this HCI bottleneck.

One long-term attempt in HCI has been to migrate the "natural" means that humans employ to communicate with each other into HCI. With this motivation automatic speech recognition has been a topic of research for decades. Tremendous progress has been made in speech recognition, and several commercially successful speech interfaces have been deployed [75]. However, it has only been in recent years that there has been an increased interest in trying to introduce other human-to-human communication modalities into HCI. This includes a class of techniques based on the movement of the human arm and hand, or *hand gestures*. Human hand gestures are a means of non-verbal interaction among people. They range from simple actions of using our hand to point at and move objects around to the more complex ones that express our feelings and allow us to communicate with others.

To exploit the use of gestures in HCI it is necessary to provide the means by which they can be interpreted by computers. The HCI interpretation of gestures requires that dynamic and/or static configurations of the human hand, arm, and even other parts of the human body, be measurable by the machine. First attempts to solve this problem resulted in mechanical devices that directly measure hand and/or arm joint angles and spatial position. This group is best represented by the so-called *glove-based devices* [9], [32], [88], [70], [101]. Glove-based gestural interfaces require the user to wear a cumbersome device, and generally carry a load of cables that connect the device to a computer. This hinders the ease and naturalness with which the user can interact with the computer controlled environment. Even though the use of such specific devices may be justified by a highly specialized application domain, for example simulation of surgery in a virtual reality environment, the "everyday" user will certainly be deterred by such cumbersome interface tools. This has spawned active research toward more "natural" HCI techniques.

- *V. Pavlovic and T.S. Huang are with The Beckman Institute and Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801. E-mail: vladimir@ifp.uiuc.edu.*
- *R. Sharma is with the Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802. E-mail: rsharma@cse.psu.edu.*

Manuscript received 17 Nov. 1995; revised 5 May 1997. Recommended for acceptance by R. Kasturi.
For information on obtaining reprints of this article, please send e-mail to: transpami@computer.org, and reference IEEECS Log Number 105027.

0162-8828/97/$10.00 © 1997 IEEE

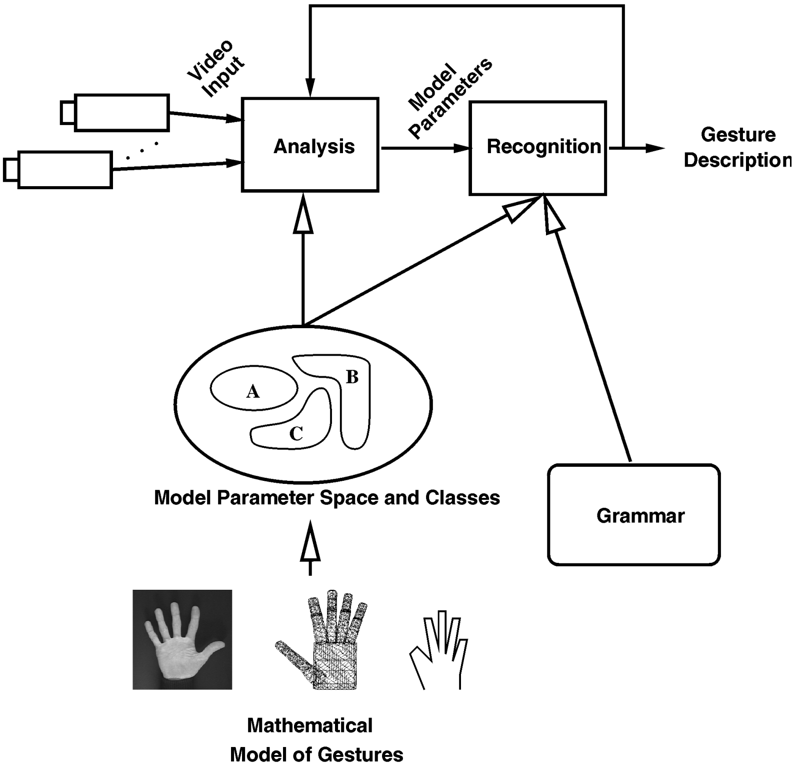


Fig. 1. Vision-based gesture interpretation system. Visual images of gesturers are acquired by one or more video cameras. They are processed in the analysis stage where the gesture model parameters are estimated. Using the estimated parameters and some higher level knowledge, the observed gestures are inferred in the recognition stage.

Potentially, any awkwardness in using gloves and other devices can be overcome by using video-based *noncontact* interaction techniques. This approach suggests using a set of video cameras and computer vision techniques to interpret gestures. The nonobstructiveness of the resulting vision-based interface has resulted in a burst of recent activity in this area. Other factors that may have contributed to this increased interest include the availability of fast computing that makes real-time vision processing feasible, and recent advances in computer vision techniques. Numerous approaches have been applied to the problem of visual interpretation of gestures for HCI, as will be seen in the following sections. Many of those approaches have been chosen and implemented so that they focus on one particular aspect of gestures, such as, hand tracking, hand posture estimation, or hand pose classification. Many studies have been undertaken within the context of a particular application, such as using a finger as a pointer to control a TV, or interpretation of American Sign Language.

Until recently, most of the work on vision-based gestural HCI has been focused on the recognition of static hand gestures or *postures*. A variety of models, most of them taken directly from general object recognition approaches, have been utilized for that purpose. Images of hands, geometric moments, contours, silhouettes, and 3D hand skeleton models are a few examples. In recent year, however, there has been an interest in incorporating the dynamic characteristics of gestures. The rationale is that hand gestures are dynamic actions and the motion of the hands conveys as much meaning as their posture does. Numerous approaches, ranging from global hand motion analysis to independent fingertip motion analysis, have been proposed for gesture analysis. There has thus been rapid growth of various studies related to vision-based gesture analysis fueled by a need to develop more natural and efficient human-computer interfaces. These studies are reported in disparate literature and are sometimes confusing in their claims and their scope. Thus there is a growing need to survey the state-of-the-art in vision-based gesture recognition and to systematically analyze the progress toward vision-based gestural human-computer interface. This paper attempts to bring together the recent progress in visual gesture interpretation within the context of its role in HCI.

We organize the survey by breaking the discussion into the following main components based on the general view of a gesture recognition system as shown in Fig. 1:

- *Gesture Modeling* (Section 2)
- *Gesture Analysis* (Section 3)
- *Gesture Recognition* (Section 4)
- *Gesture-Based Systems and Applications* (Section 5)

The first phase of a recognition task (whether considered explicitly or implicitly in a particular study) is choosing a model of the gesture. The mathematical model may consider both the spatial and temporal characteristic of the hand and hand gestures. We devote Section 2 to an in-depth discussion of gesture modeling issues. The approach used for modeling plays a pivotal role in the nature and performance of gesture interpretation.

Once the model is decided upon, an analysis stage is used to compute the model parameters from the image features that are extracted from single or multiple video input streams. These parameters constitute some description of the hand pose or trajectory and depend on the modeling approach used. Among the important problems involved in the analysis are that of hand localization, hand tracking, and selection of suitable image features. We discuss these and other issues of gesture analysis in Section 3.

The computation of model parameters is followed by gesture recognition. Here, the parameters are classified and interpreted in the light of the accepted model and perhaps the rules imposed by some grammar. The grammar could reflect not only the internal syntax of gestural commands but also the possibility of interaction of gestures with other communication modes like speech, gaze, or facial expressions. Evaluation of a particular gesture recognition approach encompasses both accuracy, robustness, and speed, as well as the variability in the number of different classes of hand / arm movements it covers. We survey the various gesture recognition approaches in Section 4.

A major motivation for the reported studies on gesture recognition is the potential to use hand gestures in various applications aiming at a natural interaction between the human and various computer-controlled displays. Some of these applications have been used as a basis for defining gesture recognition, using a "purposive" formulation of the underlying computer vision problem. In Section 5 we survey the reported as well as other potential applications of visual interpretation of hand gestures.

Although the current progress in gesture recognition is encouraging, further theoretical as well as computational advances are needed before gestures can be widely used for HCI. We discuss some of the directions of research for gesture recognition, including its integration with other natural modes of human-computer interaction in Section 6. This is followed by concluding remarks in Section 7.

2 Gesture Modeling

In order to systematically discuss the literature on gesture interpretation, it is important to first consider what model the authors have used for the hand gesture. In fact, the scope of a gestural interface for HCI is directly related to the proper modeling of hand gestures. How to model hand gestures depends primarily on the intended application within the HCI context. For a given application, a very coarse and simple model may be sufficient. However, if the purpose is a natural-like interaction, a model has to be established that allows many if not all natural gestures to be interpreted by the computer. The following discussion addresses the question of modeling of hand gestures for HCI.

2.1 Definition of Gestures

Outside the HCI framework, hand gestures cannot be easily defined. The definitions, if they exist, are particularly related to the communicational aspect of the human hand and body movements. Webster's Dictionary, for example, defines gestures as "...the use of motions of the limbs or body as a means of expression; a movement usually of the body or limbs that expresses or emphasizes an idea, sentiment, or attitude." Psychological and social studies tend to narrow this broad definition and relate it even more to man's expression and social interaction [48]. However, in the domain of HCI the notion of gestures is somewhat different. In a computer controlled environment one wants to use the human hand to perform tasks that mimic both the natural use of the hand as a manipulator, and its use in human-machine communication (control of computer / machine functions through gestures). Classical definitions of gestures, on the other hand, are rarely, if ever, concerned with the former mentioned use of the human hand (so called practical gestures [48]).

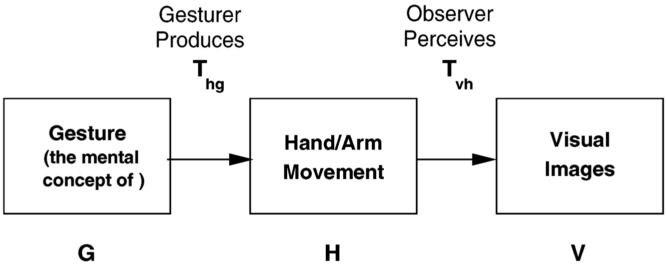


Fig. 2. Production and perception of gestures. Hand gestures originate as a mental concept G, are expressed (T_{hg}) through arm and hand motion H, and are perceived (T_{vh}) as visual images V.

Hand gestures are a means of communication, similar to spoken language. The production and perception of gestures can thus be described using a model commonly found in the field of spoken language recognition [85], [100]. An interpretation of this model, applied to gestures, is depicted in Fig. 2. According to the model, gestures originate as a gesturer's mental concept, possibly in conjunction with speech. They are expressed through the motion of arms and hands, the same way speech is produced by air stream modulation through the human vocal tract. Also, observers perceive gestures as streams of visual images which they interpret using the knowledge they possess about those gestures. The production and perception model of gestures can also be summarized in the following form:

$$H = T_{hg}G \tag{1}$$

$$V = T_{vh}H \tag{2}$$

$$V = T_{vh}(T_{hg}G) = T_{vg}G \tag{3}$$

Transformations $T_{\cdot}$ can be viewed as different *models*: T_{hg} is a model of hand or arm motion given gesture G, T_{vh} is a model of visual images given hand or arm motion H, and T_{vg} describes how visual images V are formed given some gesture G. The models are parametric, with the parameters belonging to their respective parameter spaces $\mathcal{M}_{T_{\cdot}}$. In light of this notation, one can say that the aim of visual interpretation of hand gestures is to infer gestures G from their visual images V using a suitable gesture model T_{vg}, or

$$\hat{G} = T_{vg}^{-1}V \tag{4}$$

In the context of visual interpretation of gestures, it may then be useful to consider the following definition of gestures:

> *A hand gesture is a stochastic process in the gesture model parameter space $\mathcal{M}_T$ over a suitably defined time interval $\mathcal{I}$.*

Each realization of one gesture can then been seen as a *trajectory* in the model parameter space. For example, in performing a gesture the human hand's position in 3D space describes a trajectory in such space, Fig. 3. The stochastic property in the definition of gestures affirms their natural character: no two realizations of the same gesture will result in the same hand and arm motion or the same set of visual images. The presence of the time interval I suggests the gesture's dynamic nature.

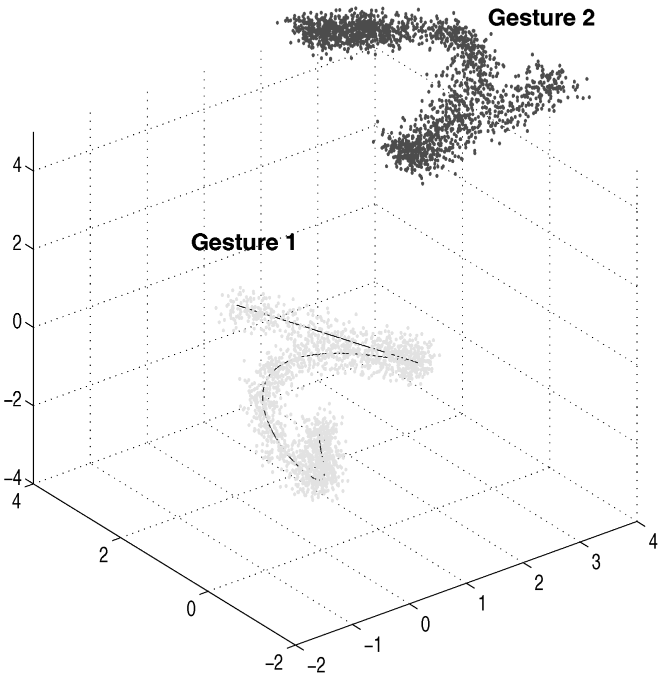


Fig. 3. Gesture as a stochastic process. Gestures can be viewed as random trajectories in parameter spaces which describe hand or arm spatial states. In this example, two different gestures are shown in a three dimensional parameter space. One realization of Gesture 1 is a trajectory in that space (solid line).

The gesture analysis and gesture recognition problems can then be posed in terms of the parameters involved in the above definition. For example, the problem of constructing the gestural model T over the parameter set $\mathcal{M}_T$, or the problem of defining the gesture interval I.

2.2 Gestural Taxonomy

Several alternative taxonomies have been suggested in the literature that deal with psychological aspects of gestures. Kendon [48] distinguishes "autonomous gestures" (that occur independently of speech) from "gesticulation" (gestures that occur in association with speech). McNeill and Levy [65] recognize three groups of gestures: iconic and metaphoric gestures, and "beats." The taxonomy that seems most appropriate within the context of HCI was recently developed by Quek [71], [72]. A slightly modified version of the taxonomy is given in Fig. 4.

All hand/arm movements are first classified into two major classes:

- gestures and
- unintentional movements.

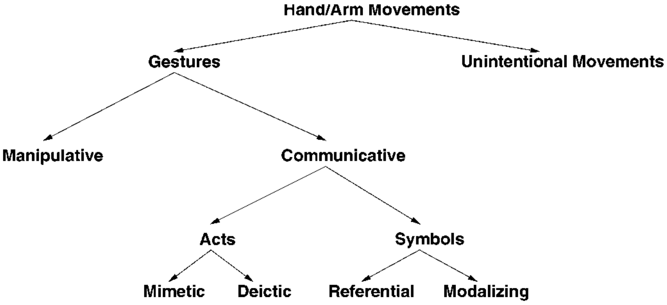


Fig. 4. A taxonomy of hand gestures for HCI. Meaningful gestures are differentiated from unintentional movements. Gestures used for manipulation (examination) of objects are separated from the gestures which possess inherent communicational character.

Unintentional movements are those hand/arm movements that do not convey any meaningful information. Gestures themselves can have two modalities:

- communicative and
- manipulative.

Manipulative gestures are the ones used to act on objects in an environment (object movement, rotation, etc.) Communicative gestures, on the other hand, have an inherent communicational purpose. In a natural environment they are usually accompanied by speech. Communicative gestures can be either acts or symbols. Symbols are those gestures that have a linguistic role. They symbolize some referential action (for instance, circular motion of index finger may be a referent for a wheel) or are used as modalizers, often of speech ("Look at that wing!" and a modalizing gesture specifying that the wing is vibrating, for example). In HCI context these gesture are, so far, one of the most commonly used gestures since they can often be represented by different static hand postures, as we will discuss further in Section 5. Finally, acts are gestures that are directly related to the interpretation of the movement itself. Such movements are classified as either mimetic (which imitate some actions) or deictic (pointing acts).

Taxonomy of gestures largely influences the way parameter space $\mathcal{M}_T$ and gesture interval I are determined. A related issue is the classification of gestural dynamics, which we consider next.

2.3 Temporal Modeling of Gestures

Since human gestures are a dynamic process, it is important to consider the temporal characteristics of gestures. This may help in the temporal segmentation of gestures from other unintentional hand/arm movements. In terms of our general definition of hand gestures, this is equivalent to determining the gesture interval I. Surprisingly, psychological studies are fairly consistent about the temporal nature of hand gestures. Kendon [48] calls this interval a "gesture phrase." It has been established that three phases make a gesture:

- preparation,
- nucleus (peak or stroke [65]), and
- retraction.

The preparation phase consists of a preparatory movement that sets the hand in motion from some resting position.

The nucleus of a gesture has some "definite form and enhanced dynamic qualities" [48]. Finally, the hand either returns to the resting position or repositions for the new gesture phase. An exception to this rule is the so called "beats" (gestures related to the rhythmic structure of the speech).

The above discussion can guide us in the process of temporal discrimination of gestures. The three temporal phases are distinguishable through the general hand/arm motion: "Preparation" and "retraction" are characterized by the rapid change in position of the hand, while the "stroke," in general, exhibits relatively slower hand motion. However, as it will be seen in Section 4, the complexity of gestural interpretation usually imposes more stringent constraints on the allowed temporal variability of hand gestures. Hence, a work in vision-based gesture HCI sometimes reduces gestures to their static equivalents, ignoring their dynamic nature.

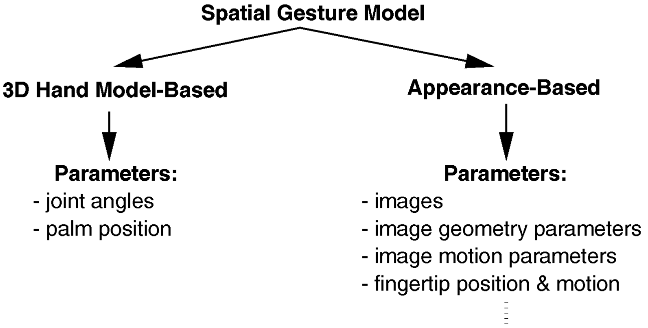


Fig. 5. Spatial models of gestures. 3D hand model-based models of gestures use articulated models of the human hand and arm to estimate the hand and arm movement parameters. Such movements are later recognized as gestures. Appearance-based models directly link the appearance of the hand and arm movements in visual images to specific gestures.

2.4 Spatial Modeling of Gestures

Gestures are observed as hand and arm movements, actions in 3D space. The description of gestures, hence, also involves the characterization of their spatial properties. In a HCI domain this characterization has so far been mainly influenced by the kind of application for which the gestural interface is intended. For example, some applications require simple models (like static image templates of the human hand in TV set control in [35]), while some others require more sophisticated ones (3D hand model used by [56], for instance).

If one considers the gesture production and perception model suggested in Section 2.1, two possible approaches to gesture modeling may become obvious. One approach may be to try to infer gestures directly from the visual images observed, as stated by (4). This approach has been often used to model gestures, and is usually denoted as *appearance-based* modeling. Another approach may result if the intermediate tool for gesture production is considered: the human hand and arm. In this case, a two step modeling process may be followed:

$$\hat{H} = T_{vh}^{-1} V \tag{5}$$

$$\hat{G} = T_{hg}^{-1} \hat{H} \tag{6}$$

In other words, one can first model the motion and posture of the hand and arm $\hat{H}$ and then infer gestures $\hat{G}$ from the motion and posture model parameters. A group of models which follows this approach is known as *3D-model-based*.

Fig. 5 shows the two major approaches used in the spatial modeling of gestures. We examine the two approaches more closely in the following subsections.

2.4.1 3D Hand/Arm Model

The 3D hand and arm models have often been a choice for hand gesture modeling. They can be classified in two large groups:

- *volumetric models* and
- *skeletal models*.

Volumetric models are meant to describe the 3D visual appearance of the human hand and arms. They are commonly found in the field of computer animation [64], but have recently also been used in computer vision applications. In the field of computer vision volumetric models of the human body are used for *analysis-by-synthesis* tracking and recognition of the body's posture [52], [105]. Briefly, the idea behind the analysis-by-synthesis approach is to analyze the body's posture by synthesizing the 3D model of the human body in question and then varying its parameters until the model and the real human body appear as the same visual images. Most of the volumetric models used in computer animation are complex 3D surfaces (NURBS or nonuniform rational B-splines) which enclose the parts of the human body they model [64]. Even though such models have become quite realistic, they are too complex to be rendered in real-time. A more appealing approach, suitable to real-time computer vision, lies in the use of simple 3D geometric structures to model the human body [68]. Structures like *generalized cylinders* and *super-quadrics* which encompass cylinders, spheres, ellipsoids and hyper-rectangles are often used to approximate the shape of simple body parts, like finger links, forearm, or upperarm [6], [20], [29], [31], [37]. The parameters of such geometric structures are quite simple. For example, a cylindrical model is completely described with only three parameters: height, radius, and color. The 3D models of more complex body parts, like hands, arms, or legs, are then obtained by connecting together the models of the simpler parts [46]. In addition to the parameters of the simple models, these structures contain the information on connections between the basic parts. The information may also include constraints which describe the interaction between the basic parts in the structure. There are two possible problems in using such elaborate hand and arm models. First, the dimensionality of the parameter space is high (more than 23×3 parameters per hand). Second, and more importantly, obtaining the parameters of those models via computer vision techniques may prove to be quite complex.

Instead of dealing with all the parameters of a volumetric hand and arm model, models with a reduced set of equivalent joint angle parameters together with segment lengths are often used. Such models are known as *skeletal models*. Skeletal models are extensively studied in the human hand morphology and biomechanics [92], [95]. We briefly describe the basic notions relevant to our discussion. The human hand skeleton consists of 27 bones, divided in three groups:

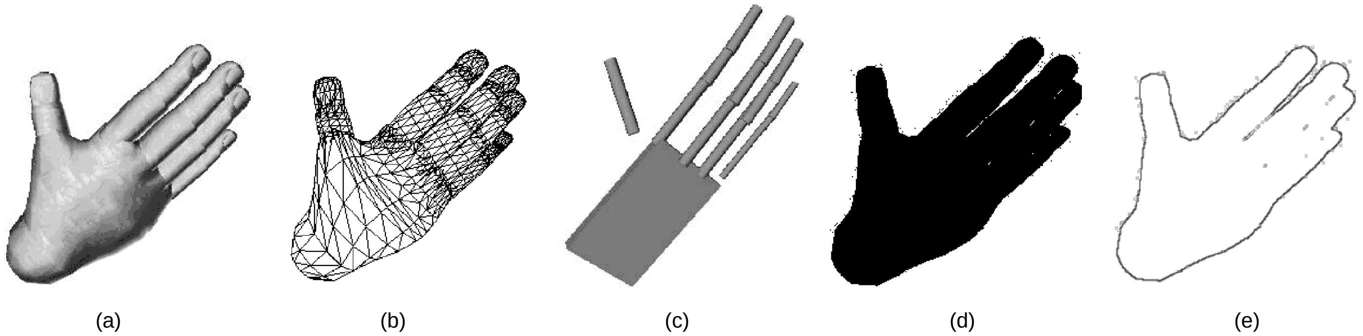


Fig. 6. Hand models. Different hand models can be used to represent the same hand posture. (a) 3D Textured volumetric model. (b) 3D wireframe volumetric model. (c) 3D skeletal model. (d) Binary silhouette. (e) Contour.

- carpals (wrist bones—eight),
- metacarpals (palm bones—five), and
- phalanges (finger bones—14).

The joints connecting the bones naturally exhibit different *degrees of freedom* (DoF). Most of the joints connecting carpals have very limited freedom of movement. The same holds for the carpal-metacarpal joints (except for the TM, see Fig. 7). Finger joints show the most flexibility: For instance, the MCP and the TM joint have two DoFs (one for extension/flexion and one for adduction/abduction), while the PIP and the DIP joints have one DoF (extension/flexion). Equally important to the notion of DoF is the notion of dependability between the movements in neighboring joints. For instance, it is natural to most people to bend (flex/extend) their fingers such that both PIP and DIP joints flex/extend. Also, there is only a certain range of angles that the hand joints can naturally assume. Hence, two sets of constraints can be placed on the joint angle movements: static (range) and dynamic (dependencies). One set of such constraints was used by Kuch [55] in his 26 DoF hand model:

<table>
<tr><th colspan="2">Static Constraints</th></tr>
<tr><th>Fingers</th><th>Thumb</th></tr>
<tr><td>$0 \le \theta^y_{MCP,s} \le 90°$
$-15° \le \theta^x_{MCP,s} \le 15°$</td><td></td></tr>
<tr><th colspan="2">Dynamic Constraints</th></tr>
<tr><td>$\theta^y_{PIP} = \frac{3}{2}\theta^y_{DIP}$
$\theta^y_{MCP} = \frac{1}{2}\theta^y_{PIP}$
$\theta^x_{MCP} = \frac{\theta^y_{MCP}}{90}\left(\theta^x_{MCP,converge} - \theta^x_{MCP,s}\right) + \theta^x_{MCP,s}$</td><td>$\theta^y_{IP} = \theta^y_{MCP}$
$\theta^y_{TM} = \frac{1}{3}\theta^y_{MCP}$
$\theta^x_{TM} = \frac{1}{2}\theta^x_{MCP}$</td></tr>
</table>

where superscripts denote flexions/extensions ("y") or adduction/abduction ("x") movements in local, joint centered coordinate systems. In another example, Lee and Kunii [59], [60] developed a 27 degree of freedom hand skeleton model with an analogous set of constraints. Similar skeleton-based models of equal or lesser complexity have been used by other authors [4], [66], [76], [77], [97].

2.4.2 Appearance-Based Model

The second group of models is based on appearance of hands/arms in the visual images. This means that the model parameters are not directly derived from the 3D spatial description of the hand. The gestures are modeled by relating the appearance of any gesture to the appearance of the set of predefined, template gestures.

A large variety of models belong to this group. Some are based on deformable 2D templates of the human hands, arms, or even body [18], [21], [45], [49], [58]. Deformable 2D templates are the sets of points on the outline of an object, used as interpolation nodes for the object outline approximation. The simplest interpolation function used is a piecewise linear function. The templates consist of the average point sets, point variability parameters, and so-called external deformations. Average point sets describe the "average" shape within a certain group of shapes. Point variability parameters describe the allowed shape deformation (variation) within that same group of shapes. These two types of parameters are usually denoted as internal. For instance, the human hand in open position has one shape on the average, and all other instances of any open posture of the human hand can be formed by slightly varying the average shape. Internal parameters are obtained through *principal component analysis* (PCA) of many of the training sets of data. External parameters or deformations are meant to describe

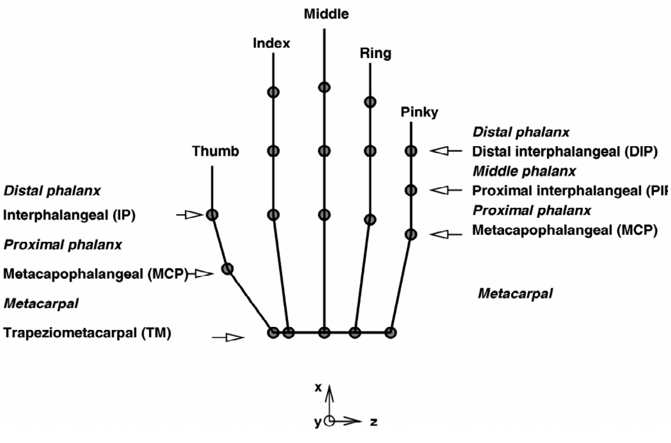


Fig. 7. Skeleton-based model of the human hand. The human hand skeleton consists of 27 bones. This model, on the other hand, approximates the anatomical structure using five serial link chains with 19 links.

the global motion of one deformable template. Rotations and translations are used to describe such motion. Template-based models are used mostly for hand-tracking purposes [18], [49]. They can also be used for simple gesture classification based on the multitude of classes of templates [58]. Trajectories of external parameters of deformable templates have also been used for simple gesture recognition [45]. Extensions of the 2D template approach to 3D deformable models have also been recently explored. For example, *3D point distribution model* has been employed for gesture tracking [42].

A different group of appearance-based models uses 2D hand image sequences as gesture templates. Each gesture from the set of allowed gestures is modeled by a sequence of representative image n-tuples. Furthermore, each element of the n-tuple corresponds to one view of the same hand or arm. In the most common case, only one (monoscopic) or two (stereoscopic) views are used. Parameters of such models can be either images themselves or some features derived from the images. For instance, complete image sequences of the human hands in motion can be used as templates per se for various gestures [25], [26]. Images of fingers only can also be employed as templates [22] in a finger tracking application. Another recently pursued approach has been to model different gestural actions by *motion history images* or MHIs [12]. MHIs are 2D images formed by accumulating the motion of every single pixel in the visual image over some temporal window. This way the intensity of the pixel in the MHI relates to how much prolonged motion is observed at that pixel.

The majority of appearance-based models, however, use parameters derived from images in the templates. We denote this class of parameters as *hand image property parameters*. They include: contours and edges, image moments, and image eigenvectors, to mention a few. Many of these parameters are also used as features in the analysis of gestures (see Section 3). Contours as a direct model parameter are often used: simple edge-based contours [17], [81] or “signatures” (contours in polar coordinates) [14] are some possible examples. Contours can also be employed as the basis for further eigenspace analysis [23], [67]. Other parameters that are sometimes used are image moments [80], [86]. They are easily calculated from hand / arm silhouettes or contours. Finally, many other parameters have been used: Zernike moments [79] and orientation histograms [34], for example.

Another group of models uses fingertip positions as parameters. This approach is based on the assumption that the position of fingertips in the human hand, relative to the palm, is almost always sufficient to differentiate a finite number of different gestures. The assumption holds in 3D space under several restrictions; some of them were noted by Lee and Kunii [59], [60]: The palm must be assumed to be rigid, and the fingers can only have a limited number of DoFs. However, most of the models use only 2D locations of fingertips and the palm [3], [28], [57]. Applications that are concerned with deictic gestures usually use only a single (index) fingertip and some other reference point on the hand or body [36], [57], [73].

3 Gesture Analysis

In the previous section, we discussed different approaches for modeling gestures for HCI. In this section we consider the analysis phase where the goal is to estimate the parameters of the gesture model using measurements from the video images of a human operator engaged in HCI. Two generally sequential tasks are involved in the analysis (see Fig. 8). The first task involves “detecting” or extracting relevant image features from the raw image or image sequence. The second task uses these image features for computing the model parameters. We discuss the different approaches used in this analysis.

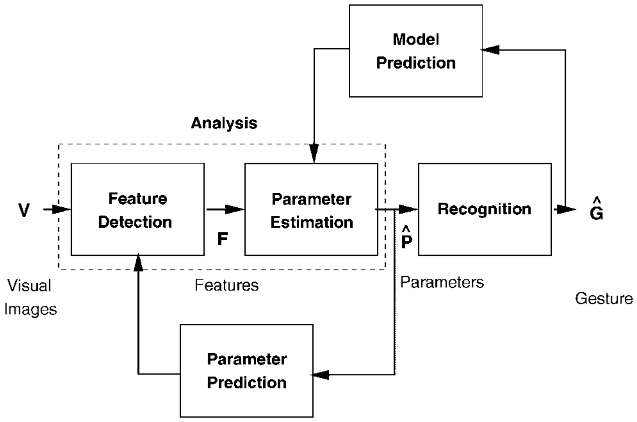


Fig. 8. Analysis and recognition of gestures. In the analysis stage, features F are extracted from visual images V. Model parameters $\hat{P}$ are estimated and possibly predicted. Gestures $\hat{G}$ are recognized in the recognition stage. Recognition may also influence the analysis stage by predicting the gesture model at the next time instance.

3.1 Feature Detection

Feature detection stage is concerned with the detection of features which are used for the estimation of parameters of the chosen gestural model. In the detection process it is first necessary to localize the gesturer. Once the gesturer is localized, the desired set of features can be detected.

3.1.1 Localization

Gesturer localization is a process in which the person who is performing the gestures is extracted from the rest of the visual image. Two types of cues are often used in the localization process:

- *color cues* and
- *motion* cues.

Color cues are applicable because of the characteristic color footprint of the human skin. The color footprint is usually more distinctive and less sensitive to illumination changes in the hue-saturation space than in the standard (camera capture) RGB color space. Most of the color segmentation techniques rely on histogram matching [4] or employ a simple look-up table approach [51], [73] based on the training data for the skin and possibly its surrounding areas. The major drawback of color-based localization techniques is the variability of the skin color footprint in different lighting conditions. This frequently results in undetected skin regions or falsely detected nonskin textures. The problem can be somewhat alleviated by considering only the regions of a certain size (scale filtering) or at certain spatial position (positional filtering). Another common solution to the problem is the use of restrictive backgrounds and clothing (uniform black background and long dark sleeves, for example.) Finally, many of the gesture recognition applications resort to the use of uniquely colored gloves or markers on hands/fingers [19], [28], [57], [60], [62]. The use of background restriction or colored gloves makes it possible to localize the hand efficiently and even in real-time, but imposes the obvious restriction on the user and the interface setup. On the other hand, without these restrictions some of the color-based localization techniques such as the ones that use histogram matching are computationally intensive and currently hard to implement in real-time.

Motion cue is also commonly applied for gesturer localization and is used in conjunction with certain assumptions about the gesturer. For example, in the HCI context, it is usually the case that only one person gestures at any given time. Moreover, the gesturer is usually stationary with respect to the (also stationary) background. Hence, the main component of motion in the visual image is usually the motion of the arm/hand of the gesturer and can thus be used to localize her/him. This localization approach is used in [35], [72]. The disadvantage of the motion cue approach is in its assumptions. While the assumptions hold over a wide spectrum of cases, there are occasions when more than one gesturer is active at a time (active role transition periods) or the background is not stationary.

To overcome the limitations of the individual cues for localization, several approaches have been suggested. One approach is the *fusion* of color, motion and other visual cues [7] or the fusion of visual cues with nonvisual cues like speech or gaze [83]. The potential advantage of the so-called *multimodal* approach has not yet been fully exploited for hand localization though it has been explored for face localization in video [38]. We discuss the multimodal approach further in Section 6. Another way in which the localization problem can be substantially eased is by the use of *prediction* techniques. These techniques provide estimates of the future feature locations based on the model dynamics and the previously known locations. We will discuss this further in Section 3.2.

3.1.2 Features and Detection

Even though different gesture models are based on different types of parameters, the image features employed to compute those parameters are often very similar. For example, some 3D hand/arm models and models that use finger trajectories all require fingertips to be extracted first. Color or gray scale images which encompass hands and arms or gesturers themselves are often used as the features. This choice of features is very common in the appearance-based models of gestures where sequences of images are used to form *temporal templates* of gestures [25]. The computational burden of the detection of these features is relatively low and is associated mostly with the gesturer localization phase. Another approach to using whole images as features is related to building of the so-called *motion energy (history) images* or MEI (MHI). MEIs are 2D images which unify the motion information of a sequence of 2D images by accumulating the motion of some characteristic image points over the sequence [30]. One simple yet effective choice of characteristic points is the whole image itself [12]. As discussed in Section 3.2, such features can represent a valid choice for the recognition of communicative gestures. However, their applicability to hand and arm tracking and recognition of manipulative gestures seems to be limited.

Hand and arm silhouettes are among the simplest, yet most frequently used features. Silhouettes are easily extracted from local hand and arm images in the restrictive background setups. In the case of complex backgrounds, techniques that employ color histogram analyses, as described in the gesturer localization phase, can be used. Examples of the use of silhouettes as features are found in both 3D hand model-based analyses [56] as well as in the appearance-based techniques (as in [54], [69]). Naturally, the use of such binary features results in a loss of information which can effect the performance especially for 3D hand posture estimators. For example, in the 3D hand posture estimation problem of [56], the binary silhouette prevents the accurate estimation of the positions of some fingers.

Contours represent another group of commonly used features. Several different edge detection schemes can be used to produce contours. Some are extracted from simple hand-arm silhouettes, and thus, are equivalent to them, while the others come from color or gray-level images. Contours are often employed in 3D model-based analyses. In such cases, contours can be used to select finger and arm link candidates through the clustering of the sets of parallel edges [29], [31], or through image-contour-to-model-contour matching [37], for example. In appearance-based models, on the other hand, many different parameters can be associated with contours: for instance “signatures” (description in polar coordinates of the points on the contour [14]) and “size functions” [96].

A frequently used feature in gesture analysis is the fingertip. Fingertip locations can be used to obtain parameters of both the 3D hand models and the 2D appearance-based gestural models (see Section 3.2). However, the detection of fingertip locations in either 3D or 2D space is not trivial. A simple and effective solution to the fingertip detection problem is to use marked gloves or color markers to designate the characteristic fingertips (see [19], [28], [57], [60], [93], for instance). Extraction of fingertip location is then fairly simplified and can be performed using color histogram-based techniques. A different way to detect fingertips is to use pattern matching techniques: templates can be images of fingertips [22] or fingers [77] or generic 3D cylindrical models [27]. Such pattern matching techniques can be enhanced by using additional image features, like contours [76]. Some fingertip extraction algorithms are based on the characteristic properties of fingertips in the image. For instance, curvature of a fingertip outline follows a characteristic pattern (low-high-low) which can be used for the feature detection [63], [97]. Other heuristics can be used as well. For example, for deictic gestures it can be assumed that the finger represents the foremost point of the hand [63], [73]. Finally, many other indirect approaches in detection of fingertips have be employed in some instances, like image analysis using specially tuned Gabor kernels [66]. The main hindrance in the use of fingertips as features is their susceptibility to occlusions. Very often one or more fingers are occluded by the palm from a given camera viewpoint and direction. The most obvious solution to this occlusion problem involves the use of multiple cameras [60], [76]. Other solutions are based on the estimation of the occluded fingertip positions based on the knowledge of the 3D model of the gesture in question [77]. More often, however, restrictions are placed on the user to posture her / his hand so that the occlusions are minimized.

3.2 Parameter Estimation

Computation of the model parameters is the last stage of the gesture analysis phase. In the gesture recognition systems, this is followed by the recognition stage, as shown in Fig. 8. For hand or arm tracking systems, however, the parameter computation stage usually produces the final output. The type of computation used depends on both the model parameters and the features that were selected.

3.2.1 Estimation of 3D Model Parameters

As mentioned in Section 2.4.1, two sets of parameters are used in 3D hand models—angular (joint angles) and linear (phalangae lengths and palm dimensions). The estimation of these *kinematic* parameters from the detected features is a complex and cumbersome task. The process involves two steps:

- the initial parameter estimation and
- the parameter update as the hand gesture evolves in time.

All of the 3D hand models employed so far assume that all the linear parameters are known a priori. This assumption reduces the problem of finding the hand joint angles to an *inverse kinematics* problem. Given a 3D position of the end-effectors and the base of a kinematic chain, the inverse kinematic's task is to find the joint angles between the links in the chain. The 3D model of the hand can then be viewed as a set of five serial kinematic chains (finger links) attached to a common base (palm). The finger tips now play the role of the end-effectors in the chains. Inverse kinematic problems are in general ill-posed, allow for multiple solutions, and are computationally expensive. The use of *constraints on parameter values* (see Section 2.4.1) somewhat alleviates those problems. Nevertheless, alternative approaches to 3D hand parameter estimation have been often sought. One automated solution to the initial parameter estimation problem was proposed by [59] through a two phase procedure using the accumulated displacement torque approach. The first phase involves the initial wrist positioning while the second phase deals with palm / finger adjustment. The procedure is applied recursively until the accumulated torque excerpted on all links reaches a local minimum, constrained on a set of static and dynamic joint angle constraints. Even though this approach produces accurate parameter estimates, it is computationally very expensive and thus not applicable to real-time problems. Some simpler solutions involve a user interactive model parameter initialization [56]. Another approach is to use interpolation of the discretized *forward kinematics* mappings to approximate the inverse kinematics [4]. Given a table of the discrete values of the joint angles and the resulting fingertip positions it is possible to estimate the values of the joint angles for a nontable value of the fingertip position.

Once the hand model parameters are initially estimated, the parameter estimates can be updated using some kind of prediction / smoothing scheme. A commonly used scheme is *Kalman filtering and prediction*. This scheme works under the assumption of small motion displacements and a known parameter update (motion) model. Such a model can be derived from a known hand kinematics model, using the inverse Jacobian mapping from the space of measurable linear displacements into the space of desired angular displacements. A variation of this approach was used by [76] in a real-time 27 degree of freedom hand tracker. On the other hand, when the dynamics are not explicitly available a simple scheme like the one reported in [56] may be employed. In this scheme, a simple silhouette matching between the 3D hand model and the real hand image was used to obtain satisfactory parameter estimation and update.

It is necessary to stress three major drawbacks associated with the mentioned 3D hand model parameter estimation approach. One has to do with the obvious computational complexity of any task involving the inverse kinematics. The other, potentially more serious problem, is due to occlusions of the fingertips used as the model features. An obvious, yet expensive, solution is to use multiple cameras. Another possible solution was developed by [77], and involves the use of finger links as features built upon a set of rules designed to resolve the finger occlusions. The last drawback stems from the employed assumption that the linear dimensions of the hand are known, which is necessary in the inverse kinematics problems. Thus, any change in scale of the hand images always results in inaccurate estimates of the hand joint angles. Finally, it should be pointed out that the knowledge of the exact hand posture parameters seem unnecessary for the recognition of communicative gestures [71] although the exact role of 3D hand parameter in gesture recognition is not clear.

The motion of the arm and hand also plays a role in gesture recognition although again the exact nature of this role is controversial [71]. The estimates of such motion can be made using either 3D space or 2D space. The 3D arm parameters are similar to the ones used in the 3D hand model description—joint angles and links. Hence, similar techniques could be used for the 3D arm parameter estimation. However, because of the simpler macro structure of the arm (the arm can be viewed as a serial kinematic chain with only three links) and fewer occlusions, it is possible to use less complex approaches to the arm parameter estimation. Most of the approaches match simplified geometrical 3D models of the arm (see Section 2.4.1) to the visual images of a real arm. The commonly used features are edges and contours which are used to estimate the link axes. For example, [29], [31] used sets of symmetry axes of line segments to estimate the axes of generalized cylinders which modeled the arm links and the upper body. In another example, [37] used chamfer matching to align the 3D tapered super-quadrics model of the upper body to two camera visual images. In a slightly different approach [105] used fusion of color "blob" features and contours to detect elements of the "blob" representation of the human body.

As is the case in the 3D hand model parameter estimation, a good initialization of arm parameters is crucial for many of these techniques to work. This is because these techniques often rely on dynamic updates of the parameters through a Kalman-based filtering/prediction scheme rather than a global initial search. Also, the introduction of constraints on the position and motion of the arm links, as in [37], can greatly improve the estimation process.

3.2.1 Estimation of Appearance Parameters

Many different appearance-based models have been reported. The estimation of the parameters of such models usually coincides with the estimation of some compact description of the image or image sequence.

Appearance models based on the visual images per se are often used to describe gestural actions. These models are often known as the *temporal* models. Various different parameters of such models are used. In the simplest case the parameters can be selected as the sets of key visual frames, as in [25]. Another possibility is to use the eigendecomposition representation of visual images in the sequence with respect to an average image [104]. A promising direction has recently been explored: accumulation of *spatio/temporal* information of a sequence of visual images into a single 2D image, a so-called *motion history image* (MHI) [12]. Such a 2D image can then be easily parameterized using one of 2D image description techniques, such as the geometric moment description or eigendecomposition. A major advantage of using these appearance models is the inherent simplicity of their parameter computation. However, this advantage may be outweighed by the loss of precise spatial information which makes them especially less suited for manipulative gestures.

Deformable 2D template-based models are often employed as the *spatial* models of hand and arm contours or even the whole human body [45]. They are usually specified through a pair of mean values of the template nodes **m** and their covariances **v** [21], [49]. The parameter estimates are obtained through *principal component analysis* (PCA) on sets of training data. Different parameters are then used to describe individual gestures. The variation of the node parameters allows for the same gesture to be recognized despite the fact that it takes on slightly different appearance when performed by different gesturers. An extension of this approach to 3D deformable templates or *point distribution models* (PDM) was recently suggested in [42]. Associated with the deformable template model parameters are also the so called external deformations or global motion parameters (rotation and translation of the hand or body in the workspace). The updates of the model parameters can then be estimated in a framework similar to the one used for rigid motion estimation. The main difference is that in the case of deformable templates an additional displacement due to the template variability $d\mathbf{v}$ also needs to be estimated [42], [49]. While the parameter computation for such deformable models is not extensive in the parameter update phase, it can be overwhelming during the initialization. On the other hand, deformable models can provide sufficient information for the recognition of both classes of gestures: manipulative and communicative.

Finally, a wide class of appearance models uses silhouettes or gray level images of the hands. In such cases, the model parameters attempt to capture a description of the shape of the hand while being relatively simple. A very commonly employed technique is built upon the geometric moment description of hand shapes [14], [69], [86]. Usually, moments of up to the second order are used. Some other techniques use Zernike moments [79] whose magnitudes are invariant to rotation, thus allowing for rotation invariant shape classification. Many other shape descriptors have also been tested—orientation histograms [34], for example, represent summary information of small patch orientations over the whole image. This parameter tends to be invariant under changes in the lighting conditions which often occur during the hand motion. Even though the parameters of the above mentioned models are easy to estimate, they are also very sensitive to the presence of other, nonhand objects in the same visual image. This means that tight "bounding boxes" around the hand need to be known at all times during the hand motion. This in turn implies either the use of good motion prediction or restriction to the hand postures. Like the other parameter estimation tasks, the reported estimation of motion parameters are usually based on simple Newtonian dynamics models and Kalman-based predictors.

4 Gesture Recognition

Gesture recognition is the phase in which the data analyzed from the visual images of gestures is recognized as a specific gesture. Analogously, using the notation we established in Section 2, the trajectory in the model parameter space (obtained in the analysis stage) is classified as a member of some meaningful subset of that parameter space. Two tasks are commonly associated with the recognition process:

- Optimal partitioning of the parameter space and
- Implementation of the recognition procedure.

The task of optimal partitioning is usually addressed through different *learning-from-examples* training procedures. The key concern in the implementation of the recognition procedure is computational efficiency. We discuss each of the above issues in more detail.

The task of optimal partitioning of the model parameter space is related to the choice of the gestural models and their parameters, as mentioned in Section 2. However, most of the gestural models are not implicitly designed with the recognition process in mind. This is especially true for the models of static gestures or hand postures. For example, most of the static models are meant to accurately describe the visual appearance of the gesturer's hand as they appear to a human observer. To perform recognition of those gestures, some type of parameter clustering technique stemming from *vector quantization* (VQ) is usually used. Briefly, in vector quantization, an *n*-dimensional space is partitioned into convex sets using *n*-dimensional hyperplanes, based on training examples and some metric for determining the nearest neighbor. If the parameters of the model are chosen especially to help with the recognition, as for example in [23], [90], the separation of classes belonging to different gestures can be done easily. However, if the model parameters are not chosen to properly describe the desired classes, the separation of the classes, and thus, accurate recognition in that parameter space may not be possible. For example, with contour descriptors, several hand postures would be confused during classification and recognition. Therefore, contours are often used for hand or arm tracking rather than for the recognition of hand postures. Parameters of other appearance-based static hand models often suffer from the same problem. For example, it is known that geometric moment parameters are not rotationally invariant. Thus, a small change in rotation of the same hand posture can cause it to be classified as a different posture. This problem can be somewhat alleviated if the chosen training hand postures classes are either very distinct or somehow normalized with respect to rotation. Another approach is to introduce different model parameters, such as Zernike moments [79] or orientation histograms [34], which posses 2D rotational invariance property. Some other models, based on eigenspace decompositions, are more discriminant and hence produce higher recognition accuracies under classical clustering techniques [103]. The problem of accurate recognition of postures which use model parameters that cluster in nonconvex sets can also be solved by selecting nonlinear clustering schemes. Neural networks are one such option, although their use for gesture recognition has not been fully explored [50]. Such nonlinear schemes are often sensitive to training and may be computationally expensive. Further, there is an inherent limitation in the discrimination capability by considering a 2D projection (or appearance) of a 3D hand when trying to capture a wide class of natural gestures. On the other hand, the use of 3D hand and gesture models offers the possibility of improving recognition, but because of the complexity of model parameter computation they are not often used for hand posture recognition.

Gestural actions, as opposed to static gestures, involve both the temporal and the spatial context [16]. As in the case of static posture recognition, the recognition of gestural actions depends on the choice of gestural models. Most of the gestural models, as seen in Section 2, produce trajectories in the model's parameter space. Since gestural action possess temporal context, the main requirement for any clustering technique used in their classification is that it be *time instance invariant* and *time scale invariant*. For example, a clapping gesture should be recognized as such whether it is performed slowly or quickly, now, or in 10 minutes. Numerous signal recognition techniques deal with such problems, the most prominent of these being automatic speech recognition (ASR). Since both speech as well as gestures are a means of natural human communication, an analogy is drawn between them and computational tools developed for ASR are frequently used in gesture recognition.

In speech recognition problems, a long standing task has been to recognize spoken words independent of their duration and variation in pronunciation. A tool called the *Hidden Markov Models* or HMM [74] has shown tremendous success is such tasks. HMM is a doubly stochastic process, a probabilistic network with *hidden* and *observable states*. The hidden states "drive" the model dynamics—at each time instance the model is in one of its hidden states. Transitions between the hidden states are governed by probabilistic rules. The observable states produce outcomes during hidden state transitions or while the model is in one of its hidden states. Such outcomes are measurable by an outside observer. The outcomes are governed by a set of probabilistic rules. Thus, an HMM can be represented as a triplet (A, b, π), where A is called the (hidden) state transition matrix, b describes the probabilities of the observation states, and π is the initial hidden state distribution. It is common to assume that the hidden state space is discrete, and that the observables are allowed to assume a continuum of values. In such cases, b is usually represented as a mixture of Gaussian (MOG) probability density functions. In automatic speech recognition, one HMM is associated with each different unit of speech (phoneme or sometimes word). Analogously, in the recognition of gestural actions, one HMM can be associated with each different gesture. In speech, the observables take on values of the linear prediction cepstrum coefficients (LPC cepstrum). In gestures, the observable is a vector of the spatial model parameters, like geometric moments [69], Zernike moments [79], or eigen image coefficients [103]. The process of association of different HMMs with different gestures (speech) units is denoted as training. In this process the parameters of the HMM (A, b, π) are modified so that the chosen model "best" describes the spatio/temporal dynamics of the desired gestural action. The training is usually achieved by optimizing the *maximum likelihood* measure $\log(Pr(observation \mid model))$ over a set of training examples for the particular gesture associated with the model. Such optimization involves the use of computationally expensive *expectation-maximization* or EM procedures, like the Baum-Welch algorithm [74]. However, any such training procedure involves a step based on *dynamic programming* or DP which in turn has a *dynamic time warping* or DTW property. This means that the variability in duration of training samples is accounted for in the model. The same is true for the recognition or model evaluation process. In that process, a gesture trajectory is tested over the set of trained HMMs in order to decide which one it belongs to. A probability of the gesture being produced by each HMM is evaluated using the *Viterbi* algorithm [74]. Obviously, the larger the number of trained HMMs (gestures) is, the more computationally demanding the recognition procedure. Problems like this one have successfully been solved by imposing an external set of rules or *grammar* which describes the language sentence structure or how the trained units (gestures or spoken) can be "connected" in time [69], [86]. Several problems are related to the use of the HMM as a recognition tool. For example, in its original formulation, an HMM is a first order stochastic process. This implies that the (hidden) state of the model at time instance i depends only on the state at time $i - 1$. While this model may be sufficient for some processes, it often results in lower recognition rates for the processes which do not follow the first order Markov property. As in speech, such problems can be somewhat reduced by extending the parameter vectors with the time derivatives of the original parameters [15]. It is also possible to derive a higher order HMMs, however such models do not share the computational efficiency of the first order models [75]. Another possible drawback of classical HMMs is the assumption that probability distribution functions or pdfs of the observables can be modeled as mixtures of Gaussians. The main reason for modeling the observables as MOGs is in training. In such cases, the HMM parameters can be efficiently computed using the Baum-Welch algorithm. Extensions in this direction have been achieved in ASR by using neural networks to model the observation pdfs [13]. Unfortunately, the training procedure in that case is computationally overwhelming. Also, in the original formulation a HMM is assumed to be *stationary*. This means that the observation probabilities do not vary in time. Such assumption may hold over short time intervals. However, since a complete gestural action is often modeled as a single HMM, the stationarity of observation pdfs may not hold true in this case. Nonstationary HMMs have been formulated for ASR [89], but have not yet been used for gesture recognition. Finally, it is interesting to note that hidden states of the HMM may possibly be viewed as the temporal phases known from the psychological studies of gesture (Section 2.3).

Another approach to recognition of gestural actions proposed recently is based on *temporal templates, so called motion energy* [30] or *motion history images* (MHIs) [12] (see Section 3.2). Such motion templates accumulate the motion history of a sequence of visual images into a single 2D image. Each MHI is parameterized by the length of the time history window that was used for its computation. To achieve time duration invariance, the templates are calculated for a set of history windows of different durations, ranging between two predefined values. The recognition is then simply achieved using any of the 2D image clustering techniques, based on the sets of trained templates. An advantage of a such temporal template approach is in its extreme computational simplicity. However, the fact that the motion is accumulated over the *entire* visual image can result in artifacts being introduced by motions of unrelated objects or body parts present in the images.

A successful recognition scheme should also consider the time-space context of any specific gesture. This can be established by introducing a grammatical element into recognition procedure. The grammar should reflect the linguistic character of communicative gestures as well as spatial character of manipulative gestures. In other words, only certain subclasses of gestural actions with respect to the current and previous states of the HCI environment are (naturally) plausible. For example, if a user reaches (performs a valid manipulative gesture) for the coffee cup handle and the handle is not visible from the user's point of view, the HCI system should discard such a gesture. Still, only a small number of the systems so far exploits this. The grammars are simple and usually introduce artificial linguistic structures: they build their own "languages" that have to be learned by the user [36], [50], [73], [80].

The computational complexity of a recognition approach is important in the context of HCI. The trade-offs involved across various approaches is a classical one—model complexity, versus richness of the gesture classes, versus recognition time. The more complex the model is, the wider class of gestures to which it can be applied. The computational complexity increases, and, hence, the recognition time. Most of the 3D model-based gesture models are characterized by more than 10 parameters. Their parameter calculation (gesture analysis) requires computationally expensive successive approximation procedures (the price of which is somewhat lowered using prediction-type analysis). The systems based on such models rarely show close to real-time performance. For example, the time performance ranges from 45 minutes per single frame in [59] (although it does not use any prediction element) to 10 frames per second in [76]. Yet potentially, the 3D models can be used to capture the richest sets of hand gestures for HCI. The appearance-based models are usually restricted in their applicability to a narrower subclass of HCI applications, enhancements of the computer mouse concept [22], [35], [36], [50], [73], or hand posture classification [43], [63], [66], [67], [81], [86]. On the other hand, because of the lower complexity of the appearance-based models they are easier to implement in real-time and more widely used.

5 Applications and Systems

Recent interest in gestural interface for HCI has been driven by a vast number of potential applications (Fig. 9). Hand gestures as a mode of HCI can simply enhance the interaction in "classical" desktop computer applications by replacing the computer mouse or similar hand-held devices. They can also replace joysticks and buttons in the control of computerized machinery or be used to help the physically impaired to communicate more easily with others. Nevertheless, the major impulse to the development of gestural interfaces has come from the growth of applications situated in *virtual environments* (VEs) [2], [53].

Hand gestures in natural environments are used for both manipulative actions and communication (see Section 2). However, the communicative role of gestures is subtle, since hand gestures tend to be a supportive element of speech (with the exception of deictic gestures, which play a major role in human communication). Manipulative aspect of gestures also prevails in their current use for HCI. However, some applications have emerged recently which take advantage of the communicative role of gestures. We present a brief overview of several application driven systems with interfaces based on hand gestures.

Most applications of hand gestures portray them as the manipulators of *virtual objects* (VOs). This is depicted in Fig. 9. VOs can be computer generated graphics, like simulated 2D and 3D objects [14], [22], [44], [36] or windows [50], [73], or abstractions of computer-controlled physical objects, such as device control panels [4], [35], [36] or robotic arms [18], [43], [47], [94]. To perform manipulations of such objects through HCI a combination of coarse tracking and communicative gestures is currently being used. For example, to direct the computer to rotate an object a user of such an interface may issue a two-step command: *<select object> <rotate object>*. The first action uses coarse hand tracking to move a pointer in the VE to the vicinity of the object.

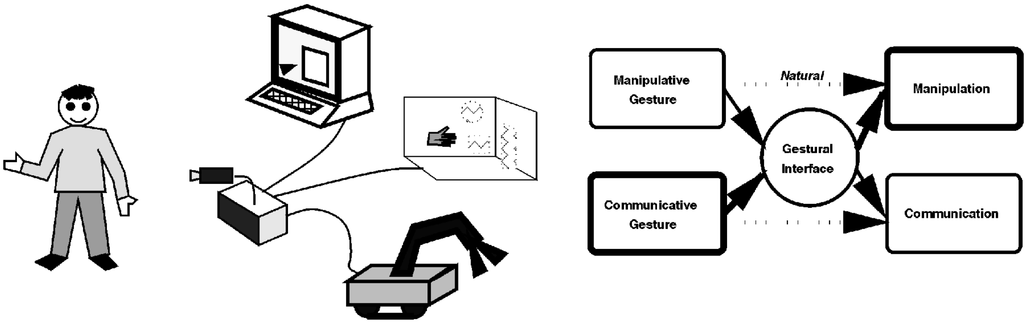


Fig. 9. Applications of gestural interface for HCI. Unlike the gestures in a natural environment, both manipulative and communicative gestures in HCI can be employed to direct manipulations of objects or to convey messages.

To rotate the object, the user rotates his/her hand back and forth producing a *metaphor* for rotational manipulation [14]. One may then pose the question: "Why use the communicative gestures for manipulative actions?" Communicative gestures imply a finite (and usually small) vocabulary of gestures that has to be learned, whereas the manipulative ones are natural hand/arm movements. To answer this question, one has to consider the complexity of analysis and recognition of each type of gestural models (Section 3 and Section 4). The 3D hand model-based gestural models are well suited for modeling of both manipulative and communicative gestures, while the appearance-based models of gestures are mostly applicable to the communicative ones. However, the use of 3D hand model-based gesture models are computationally more expensive than that of the appearance-based models (see Section 4). Therefore, to achieve a usable (real-time) performance one has to usually resort to the less desirable appearance-based models of gestures. Recently, however, with the increase in computing power, simplified hand/head blob models [6], [105] have been considered for applications which use communicative gesture recognition [10]. Such models are simple enough to be analyzed in real-time and are used for recognition of a small set of communicative gestures. For example, [10] used such a model followed by a HMM classifier to recognize eighteen *T'ai Chi* gestures. The system was intended to provide a virtual environment for the relaxation of cancer patients.

A brief summary of characteristics of some of the systems aimed at the application of hand gestures for HCI is given in Table 1. It summarizes the basic modeling technique used for the gestures, the class of gesture commands that are interpreted, and the reported performance in terms of the speed of processing.

Not all of the applications of hand gestures for HCI are meant to yield manipulative actions. Gestures for HCI can also be used to convey messages for the purpose of their analysis, storage or transmission. Video-teleconferencing (VTC) and processing of American sign language (ASL) provide such opportunities. In VTC applications, reduction of bandwidth is one of the major issues. A typical solution is to use different coding techniques. One such technique is model-based coding where image sequences are described by the states (e.g., position, scale, and orientation) of all physical objects in the scene (human participants in the case of VTC) [1], [40]. Only the updates of descriptors are sent while at the receiving end a computer generated model of physical objects is driven using the received data. Model-based coding for VTC, therefore, requires that the human bodies be modeled appropriately. Depending on the amount of detail desired, this can be achieved by only coarse models of the upper body and limbs [20], or finely tuned models of human faces or hands. Modeling of hand/arm gestures can then be of substantial value for such applications.

Recognition of ASL is often considered as another application that naturally employs human gestures as means of communication. Such applications could play a vital role in communication with people with a communication impairment like deafness. A device which could automatically translate ASL hand gestures into speech signals would undoubtedly have a positive impact on such individuals. However, the more practical reason for using the ASL as a test bed for the present hand gesture recognition systems is its well-defined structure compared to other natural gestures humans use. This fact implies that the appearance-based modeling techniques are particularly suited for such ASL interpretation, as was proven in several recent applications [86], [96].

There are numerous prospective applications of vision-based hand gesture analysis. The applications mentioned so far are only the first steps toward using hand gestures in HCI. The need for further development is thus quite clear. We discuss several important research issues that need to be addressed toward incorporating natural hand gestures into the HCI.

6 Future Directions

To fully exploit the potential of gestures in HCI environments, the class of recognizable gestures should be as broad as possible. Ideally, any and every gesture performed by the user should be unambiguously interpretable, thus allowing for *naturalness* of the interface. However, the state of the art in vision-based gesture recognition does not provide a satisfactory solution for achieving this goal. Most of the gesture-based HCI systems at the present time address a very narrow group of applications: mostly symbolic commands based on hand postures or 3D-mouse type of pointing (see Section 5). The reason for this is the complexity associated with the analysis (Section 3) and recognition (Section 4) of gestures. Simple gesture models make it possible to build real-time gestural interfaces—for example, pointing direction can be quickly found from the silhou-

TABLE 1
SYSTEMS THAT EMPLOY HAND GESTURES FOR HCI

Application	Gestural Modeling Technique	Gestural Commands	Complexity (Speed)
CD Player Control Panel [4]	Hand silhouette moments	Tracking only	30 fps[1]
Staying Alive [10]	3D hand / head blob model [6]	Tracking & HMM-based recognition	real time
Virtual Squash [14]	Hand silhouette moments & contour "signature"	Tracking & three metaphors	10.6 fps
FingerPaint [22]	Fingertip template	Tracking only	n.a.[2]
ALIVE [26]	Template correlation	Tracking combined with recognition of facial expressions	real-time
Computer Game Control [33]	Image moments using dedicated hardware	Hand & body posture recognition	real-time
TV Display Control [35]	Template correlation	Tracking only	5 fps
FingerPointer [36]	Heuristic detection of pointing action	Tracking and one metaphor combined with speech	real-time
Window Manager [50]	Hand pose recognition using neural networks	Tracking & four metaphors	real-time
GestureComputer [63]	Image moments & fingertip position	Tracking and six metaphors	10-25 fps
FingerMouse [73]	Heuristic detection of pointing action	Tracking only	real-time
DigitEyes [76]	27 DoF 3D hand model	Tracking only	10 fps
Robot manipulator guidance [18]	active contour	pointing	real-time
ROBOGEST [43]	Silhouette Zernike moments	Six metaphors	1/2 fps
Automatic robot instruction	Fingertip position in 2D	Grasp tracking	n.a.
Robot manipulator control [94]	Fingertip positions in 3D	Six metaphors	real-time
Hand sign recognition [24]	Most expressive featur cameres (MEF) of images	40 signs	n.a.
ASL recognition [86]	Silhouette moments & grammar	40 words	5 fps

1. Frames per second.
2. Not available.

We choose speed as the measure of complexity of interpretation given the lack of any other accurate measure. Note, however, that different applications may be implemented on different computer systems with different levels of optimization.

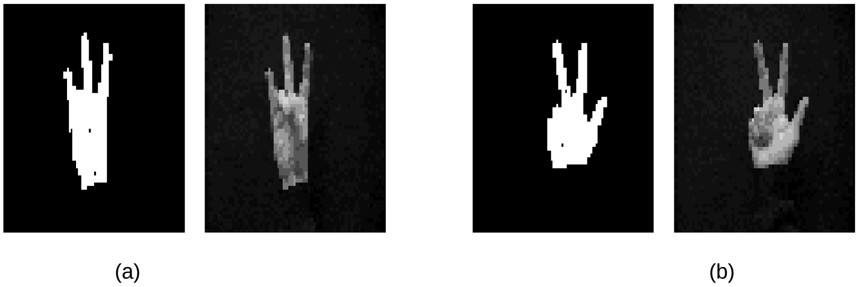


Fig. 10. Silhouettes and gray-scale images of two different hand postures. The silhouette in (a) can also be interpreted as the reflection about the vertical axes of the silhouette in (b). Hence, the two silhouettes do not unambiguously define the hand posture.

ettes of the human hand in relatively nonrestrictive environments ([36],[73]). However, as it can be seen from Fig. 10 to find the hand posture and thus distinguish between the two gestures from the simple image appearance (silhouettes) is sometimes quite difficult.

Real-time interaction based on 3D hand model-based gesture analysis is yet to be demonstrated. The use of 3D models are mostly confined to hand tracking and hand posture analysis. Yet, the analysis of the parameters of the 3D hand model-based models can result in a wider class of hand gestures that can be identified than the analysis linked with the appearance-based models. This leads us to the conclusion that, from the point of the naturalness of HCI, the 3D hand model-based approaches offer more promise than the appearance-based models. However, this prospect is presently hindered by a lack of speed and the restrictiveness of the background in the 3D hand model-based approaches. The first problem is associated with the complexity of the model and the feature extraction. Fingertip positions seem to be a very useful feature (see Section 3.1.2), yet sometimes difficult to extract. A possible solution to this problem may employ the use of skin and nail texture to distinguish the tips of the fingers. Additionally, the computational complexity of estimating the model parameters (Section 3.2) can be reduced by choosing an optimal number of parameters that satisfies a particular level of naturalness and employing parallelization of the computations involved.

Several other aspects that pertain to the construction of a natural HCI need to be adequately addressed in the future. One of the aspects involves the two-handed gestures. Human gestures, especially communicative, naturally employ actions of both hands. Yet, many of the vision-based gesture systems focus their attention on single-hand gestures. Until recently, the single-hand gesture approach has been almost inevitable. First, many analysis techniques require that the hands be extracted from global images. If the two-handed gestures are allowed, several ambiguous situations that do not occur in single-hand case may occur that have to be dealt with (occlusion of hands, distinction between or indexing of left/right hand). Second, the most versatile gesture analysis techniques (namely, 3D model-based techniques) currently exhibit one major drawback: speed. Some 3D model-based techniques which use coarse upper body and arm models [6] have reached the near real-time speeds and have been utilized for basic two-hand gesture analysis. Appearance-based techniques can, in principle, handle two-handed gestures. Nevertheless, their applicability has been usually restricted to simple (symbolic) gestures that do not require two hands. A more recent work on appearance-based motion templates [12] has indirectly addressed the issue of two-handed gestures. Another notable exception is an early system developed by Krueger [54]. Thus, to adequately address the issue of two-handed gestures in the future, more effective analysis techniques should be considered. These techniques should not only rely on the improvements of the classical techniques used in single-hand gestures, but also exploit the interdependence between the two hands performing a gesture since in many case the two hands performing a single gesture assume symmetrical postures.

An issue related to two-handed gestures is the one of multiple gesturers. Successful interaction in HCI-based environments has to consider multiple users. For example, a virtual modeling task can benefit enormously if several designers simultaneously participate in the process. However, the implementation of the multi-user interface has several difficult issues to face, the foremost one being the analysis of gestures. The analysis at the present assumes that there is a well-defined workspace associated with the gesturer. However, in the case of multiple users the intersection of workspaces is a very probable event. The differentiation between the users can then pose a serious problem. The use of active computer vision [11], [82], [5], [8], in which the cameras adaptively focus on some area of interest, may offer a solution to this problem. Another approach would be to optimize the parameters of the stationary camera(s) for a given interface; related issues are studied under *sensor planning* [39], [91].

Hand gestures are, like speech, body movement, and gaze, a means of communication (see Section 2.1). Moreover, almost any natural communication among humans concurrently involves several modes of communication that accompany each other. For instance, the "come here" gesture is usually accompanied by the words "Come here." Another example is the sentence "Notice *this* control panel." and a deictic gesture involving an index finger pointing at the particular control panel and a gaze directed at the panel. As seen from the above examples, the communicative gestures can be used both to *affirm* and to *complement* the meaning of a speech message. In fact, in the literature that reports psychological studies of human communication, the interaction between the speech and gestures as well as the other means of communication is often explored [48], [61], [87]. This leads to the conclusion that any such *multimodal* interaction can also be rendered useful for HCI (see Fig. 11 and [84]). The affirmative hand gesture (speech) can be used to reduce the uncertainty in speech (hand gesture) recognition and, thus, provide a more robust interface. Gestures that complement speech, on the other hand, carry a complete communicational message only if they are interpreted together with speech and, possibly, gaze. The use of such multimodal messages can help reduce the complexity and increase the naturalness of the interface for HCI (see Fig. 12). For example, instead of designing a complicated gestural command for the object selection which may consist of a deictic gesture followed by a symbolic gesture (to symbolize that the object that was pointed at by the hand is supposed to be selected) a simple concurrent deictic gesture and verbal command "this" can be used [84]. The number of studies that explore the use of multimodality in HCI has been steadily increasing over the past couple of years [36], [83], [98], [99], [102]. At the present time, the integration of communication modes in such systems is performed after the commands portions of different modes have been independently recognized. Although the interface structure is simplified in this way, the information pertaining to the interaction of the modes at lower levels is probably lost. To utilize the multimodal interaction at all levels, new approaches that fuse the multimodal input analysis as well as recognition should be considered in the future.

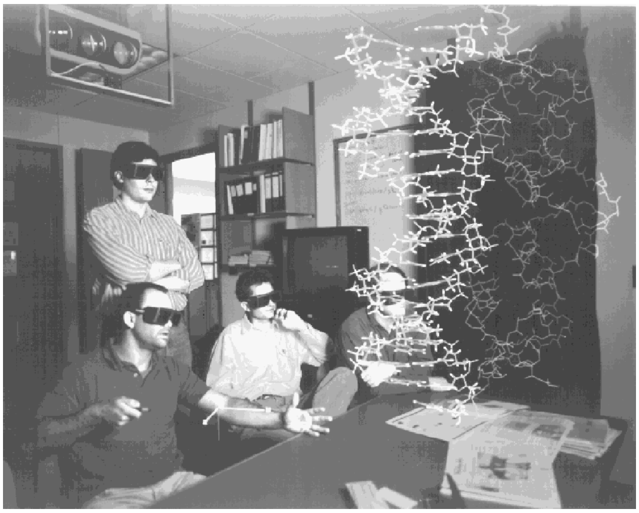


Fig. 11. A possible situation where speech/gesture integration may be particularly effective: A 3D visualization facility for structural biologist where researchers could be examining and discussing the results of a simulation.

Photograph courtesy of Rich Saal, Illinois State Journal-Register, Springfield, Ill.

7 Conclusions

Human-computer interaction is still in its infancy. Visual interpretation of hand gestures would allow the development of potentially natural interfaces to computer controlled environments. In response to this potential, the number of different approaches to video-based hand gesture recognition has grown tremendously in recent years. Thus there is a growing need for systematization and analysis of many aspects of gestural interaction. This paper surveys the different approaches to modeling, analysis, and recognition of hand gestures for visual interpretation. The discussion recognizes two classes of models employed in the visual interpretation of hand gestures. The first relies on 3D models of the human hand, while the second utilizes the appearance of the human hand in the image. The 3D hand models offer a rich description and discrimination capability that would allow a wide class of gestures to be recognized leading to natural HCI. However, the computation of 3D model parameters from visual images under real-time constraints remains an elusive goal. Appearance-based models are simpler to implement and use for real-time gesture recognition, but suffer from inherent limitations which could be a drawback for natural HCI.

Several simple HCI systems have been proposed that demonstrate the potential of vision-based gestural interfaces. However, from a practical standpoint, the development of such systems is in its infancy. Though most current systems employ hand gestures for the manipulation of objects, the complexity of the interpretation of gestures dictates the achievable solution. For example, the gestures used to convey manipulative actions today are usually of the communicative type. Further, hand gestures for HCI are mostly restricted to single-handed and produced only by a single user in the system. This consequently downgrades the effectiveness of the interaction. We suggest several directions of research for raising these limitations toward gestural HCI. For example, integration of hand gestures with speech, gaze and other naturally related modes of communication in a multimodal interface. However, substantial research effort that connects advances in computer vision with the basic study of human-computer interaction will be needed in the future to develop an effective and natural hand gesture interface.

Acknowledgments

This work was supported in part by the U. S. Army Research Laboratory under Cooperative Agreement No. DAAL01-96-2-0003 and in part by the National Science Foundation Grant IRI-96-34618. The authors would like to acknowledge Yusuf Azoz and Lalitha Devi for their help with the references, and Karin Pavlovic for her help in reviewing the manuscript. They would also like to thank the anonymous reviewers for their comments which greatly helped in improving this survey.

References

[1] J.F. Abramatic, P. Letellier, and M. Nadler, “A Narrow-Band Video Communication System for the Transmission of Sign Language Over Ordinary Telephone Lines,” *Image Sequences Processing and Dynamic Scene Analysis*, T.S. Huang, ed., pp. 314-336. Berlin and Heidelberg: Springer-Verlag, 1983.

[2] J.A. Adam, “Virtual Reality,” *IEEE Spectrum*, vol. 30, no. 10, pp. 22-29, 1993.

[3] S. Ahmad and V. Tresp, “Classification With Missing and Uncertain Inputs,” *Proc. Int’l Conf. Neural Networks*, vol. 3, pp. 1,949-1,954, 1993.

[4] S. Ahmad, “A Usable Real-Time 3D Hand Tracker,” *IEEE Asilomar Conf.*, 1994.

[5] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, “Active Vision,” *Int’l J. Computer Vision*, vol. 1, pp. 333-356, 1988.

[6] A. Azarbayejani, C. Wren, and A. Pentland, “Real-Time 3D Tracking of the Human Body,” *Proc. IMAGE’COM 96*, Bordeaux, France, 1996.

[7] Y. Azoz, L. Devi, and R. Sharma, “Vision-Based Human Arm Tracking for Gesture Analysis Using Multimodal Constraint Fusion,” *Proc. 1997 Advanced Display Federated Laboratory Symp.*, Adelphi, Md., Jan. 1997.

[8] R. Bajcsy, “Active Perception,” *Proc. IEEE*, vol. 78, pp. 996-1,005, 1988.

[9] T. Baudel and M. Baudouin-Lafon, “Charade: Remote Control of Objects Using Free-Hand Gestures,” *Comm. ACM*, vol. 36, no. 7, pp. 28-35, 1993.

[10] D.A. Becker and A. Pentland, “Using a Virtual Environment to Teach Cancer Patients T’ai Chi, Relaxation, and Self-Imagery,” *Proc. Int’l Conf. Automatic Face and Gesture Recognition*, Killington, Vt. , Oct. 1996.

[11] A. Blake and A. Yuille, *Active Vision*. Cambridge, Mass.: MIT Press, 1992.

[12] A.F. Bobick and J.W. Davis, “Real-Time Recognition of Activity Using Temporal Templates,” *Proc. Int’l Conf. Automatic Face and Gesture Recognition*, Killington, Vt., Oct. 1996.

[13] H.A. Boulard and N. Morgan, *Connectionnist Speech Recognition. A Hybrid Approach*. Norwell, Mass.: Kluwer Academic Publishers, 1994.

[14] U. Bröckl-Fox, “Real-Time 3D Interaction With Up to 16 Degrees of Freedom From Monocular Image Flows,” *Proc. Int’l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 172-178, June 1995.

[15] L.W. Campbell, D.A. Becker, A. Azarbayejani, A.F. Bobick, and A. Pentland, “Invariant Features for 3D Gesture Recognition,” *Proc. Int’l Conf. Automatic Face and Gesture Recognition*, Killington, Vt., pp. 157-162, Oct. 1996.

[16] C. Cedras and M. Shah, “Motion-Based Recognition: A Survey,” *Image and Vision Computing*, vol. 11, pp. 129-155, 1995.

[17] K. Cho and S.M. Dunn, “Learning Shape Classes,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, pp. 882-888, Sept. 1994.

[18] R. Cipolla and N.J. Hollinghurst, “Human-Robot Interface by Pointing With Uncalibrated Stereo Vision,” *Image and Vision Computing*, vol. 14, pp. 171-178, Mar. 1996.

[19] R. Cipolla, Y. Okamoto, and Y. Kuno, “Robust Structure From Motion Using Motion Parallax,” *Proc. IEEE Int’l Conf. Computer Vision*, pp. 374-382, 1993.

[20] E. Clergue, M. Goldberg, N. Madrane, and B. Merialdo, “Automatic Face and Gestural Recognition for Video Indexing,” *Proc. Int’l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 110-115, June 1995.

[21] T. F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, “Active Shape Models—Their Training and Application,” *Computer Vision and Image Understanding*, vol. 61, pp. 38-59, Jan. 1995.

[22] J.L. Crowley, F. Berard, and J. Coutaz, “Finger Tacking As an Input Device for Augmented Reality,” *Proc. Int’l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 195-200, June 1995.

[23] Y. Cui and J. Weng, “Learning-Based Hand Sign Recognition,” *Proc. Int’l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 201-206, June 1995.

[24] Y. Cui and J. J. Weng, “Hand Segmentation Using Learning-Based Prediction and Verification for Hand Sign Recognition,” *Proc. Int’l Conf. Automatic Face and Gesture Recognition*, Killington, Vt., pp. 88-93, Oct. 1996.

[25] T. Darrell, I. Essa, and A. Pentland, "Task-Specific Gesture Analysis in Real-Time Using Interpolated Views," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 12, pp. 1,236-1,242, Dec. 1996.

[26] T. Darrell and A.P. Pentland, "Attention-Driven Expression and Gesture Analysis in an Interactive Environment," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 135-140, June 1995.

[27] J. Davis and M. Shah, "Determining 3D Hand Motion," *Proc. 28th Asilomar Conf. Signals, Systems, and Computer*, 1994.

[28] J. Davis and M. Shah, "Recognizing Hand Gestures," *Proc. European Conf. Computer Vision*, Stockholm, Sweden, pp. 331-340, 1994.

[29] A. C. Downton and H. Drouet, "Image Analysis for Model-Based Sign Language Coding," *Progress in Image Analysis and Processing II: Proc. Sixth Int'l Conf. Image Analysis and Processing*, pp. 637-644, 1991.

[30] I. Essa and S. Pentland, "Facial Expression Recognition Using a Dynamic Model and Motion Energy," *Proc. IEEE Int'l Conf. Computer Vision*, 1995.

[31] M. Etoh, A. Tomono, and F. Kishino, "Stereo-Based Description by Generalized Cylinder Complexes From Occluding Contours," *Systems and Computers in Japan*, vol. 22, no. 12, pp. 79-89, 1991.

[32] S.S. Fels and G.E. Hinton, "Glove-Talk: A Neural Network Interface Between a Data-Glove and a Speech Synthesizer," *IEEE Trans. Neural Networks*, vol. 4, pp. 2-8, Jan. 1993.

[33] W.T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma, "Computer Vision for Computer Games," *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, Killington, Vt., pp. 100-105, Oct. 1996.

[34] W.T. Freeman and M. Roth, "Orientation Histograms for Hand Gesture Recognition," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, June 1995.

[35] W.T. Freeman and C.D. Weissman, "Television Control by Hand Gestures," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 179-183, June 1995.

[36] M. Fukumoto, Y. Suenaga, and K. Mase, "Finger-Pointer": Pointing Interface by Image Processing," *Computers and Graphics*, vol. 18, no. 5, pp. 633-642, 1994.

[37] D.M. Gavrila and L.S. Davis, "Towards 3D Model-Based Tracking and Recognition of Human Movement: A Multi-View Approach," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 272-277, June 1995.

[38] H.P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan, "Multi-Modal System for Locating Heads and Faces," *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, Killington, Vt., pp. 88-93, Oct. 1996.

[39] G. D. Hager, *Task Directed Sensor Fusion and Planning*. Kluwer Academic Publishers, 1990.

[40] H. Harashima and F. Kishino, "Intelligent Image Coding and Communications With Realistic Sensations—Recent Trends," *IEICE Trans.*, vol. E 74, pp. 1,582-1,592, June 1991.

[41] A.G. Hauptmann and P. McAvinney, "Gesture With Speech for Graphics Manipulation," *Int'l J. Man-Machine Studies*, vol. 38, pp. 231-249, Feb. 1993.

[42] T. Heap and D. Hogg, "Towards 3D Hand Tracking Using a Deformable Model," *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, Killington, Vt., pp. 140-145, Oct. 1996.

[43] E. Hunter, J. Schlenzig, and R. Jain, "Posture Estimation in Reduced-Model Gesture Input Systems," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, June 1995.

[44] K. Ishibuchi, H. Takemura, and F. Kishino, "Real Time Hand Gesture Recognition Using 3D Prediction Model," *Proc. 1993 Int'l Conf. Systems, Man, and Cybernetics*, Le Touquet, France, pp. 324-328, Oct. 17-20, 1993.

[45] S.X. Ju, M.J. Black, and Y.Y. oob, "Cardboard People: A Parameterized Model of Articulated Image Motion," *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, Killington, Vt., pp. 38-43, Oct. 1996.

[46] I.A. Kakadiaris, D. Metaxas, and R. Bajcsy, "Active Part-Decomposition, Shape and Motion Estimation of Articulated Objects: A Physics-Based Approach," *Proc. IEEE C.S. Conf. Computer Vision and Pattern Recognition*, pp. 980-984, 1994.

[47] S.B. Kang and K. Ikeuchi, "Toward Automatic Robot Instruction for Perception—Recognizing a Grasp From Observation," *IEEE Trans. Robotics and Automation*, vol. 9, pp. 432-443, Aug. 1993.

[48] A. Kendon, "Current Issues in the Study of Gesture," *The Biological Foundations of Gestures: Motor and Semiotic Aspects*, J.-L. Nespoulous, P. Peron, and A. R. Lecours, eds., pp. 23-47. Lawrence Erlbaum Assoc., 1986.

[49] C. Kervrann and F. Heitz, "Learning Structure and Deformation Modes of Nonrigid Objects in Long Image Sequences," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, June 1995.

[50] R. Kjeldsen and J. Kender, "Visual Hand Gesture Recognition for Window System Control," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 184-188, June 1995.

[51] R. Kjeldsen and J. Kender, "Finding Skin in Color Images," *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, Killington, Vt., pp. 312-317, Oct. 1996.

[52] R. Koch, "Dynamic 3D Scene Analysis Through Synthetic Feedback Control," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 556-568, 1993.

[53] M.W. Krueger, *Artificial Reality II*. Addison-Wesley, 1991.

[54] M.W. Krueger, "Environmental Technology: Making the Real World Virtual," *Comm. ACM*, vol. 36, pp. 36-37, July 1993.

[55] J.J. Kuch, "Vision-Based Hand Modeling and Gesture Recognition for Human Computer Interaction," master's thesis, Univ. of Illinois at Urbana-Champaign, 1994.

[56] J.J. Kuch and T.S. Huang, "Vision-Based Hand Modeling and Tracking," *Proc. IEEE Int'l Conf. Computer Vision*, Cambridge, Mass., June 1995.

[57] Y. Kuno, M. Sakamoto, K. Sakata, and Y. Shirai, "Vision-Based Human Computer Interface With User Centered Frame," *Proc. IROS'94*, 1994.

[58] A. Lanitis, C.J. Taylor, T.F. Cootes, and T. Ahmed, "Automatic Interpretation of Human Faces and Hand Gestures Using Flexible Models," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 98-103, June 1995.

[59] J. Lee and T.L. Kunii, "Constraint-Based Hand Animation," *Models and Techniques in Computer Animation*, pp. 110-127. Tokyo: Springer-Verlag, 1993.

[60] J. Lee and T.L. Kunii, "Model-Based Analysis of Hand Posture," *IEEE Computer Graphics and Applications*, pp. 77-86, Sept. 1995.

[61] E.T. Levy and D. McNeill, "Speech, Gesture, and Discourse," *Discourse Processes*, no. 15, pp. 277-301, 1992.

[62] C. Maggioni, "A Novel Gestural Input Device for Virtual Reality," *1993 IEEE Annual Virtual Reality Int'l Symp.*, pp. 118-124, IEEE, 1993.

[63] C. Maggioni, "GestureComputer—New Ways of Operating a Computer," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 166-171, June 1995.

[64] N. Magnenat-Thalmann and D. Thalman, *Computer Animation: Theory and Practice*. New York: Springer-Verlag, 2nd rev. ed., 1990.

[65] D. McNeill and E. Levy, "Conceptual Representations in Language Activity and Gesture," *Speech, Place and Action: Studies in Deixis and Related Topics*, J. Jarvella and W. Klein, eds. Wiley, 1982.

[66] A. Meyering and H. Ritter, "Learning to Recognize 3D-Hand Postures From Perspective Pixel Images," *Artificial Neural Networks 2*, I. Alexander and J. Taylor, eds. North-Holland: Elsevier Science Publishers B.V., 1992.

[67] B. Moghaddam and A. Pentland, "Maximum Likelihood Detection of Faces and Hands," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 122-128, June 1995.

[68] O'Rourke and N.L. Badler, "Model-Based Image Analysis of Human Motion Using Constraint Propagation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 2, pp. 522-536, 1980.

[69] V.I. Pavlovic, R. Sharma, and T.S. Huang, "Gestural Interface to a Visual Computing Environment for Molecular Biologists," *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, Killington, Vt., pp. 30-35, Oct. 1996.

[70] D.L. Quam, "Gesture Recognition With a DataGlove," *Proc. 1990 IEEE National Aerospace and Electronics Conf.*, vol. 2, 1990.

[71] F.K.H. Quek, "Toward a Vision-Based Hand Gesture Interface," *Virtual Reality Software and Technology Conf.*, pp. 17-31, Aug. 1994.

[72] F.K.H. Quek, "Eyes in the Interface," *Image and Vision Computing*, vol. 13, Aug. 1995.

[73] F.K.H. Quek, T. Mysliwiec, and M. Zhao, "Finger Mouse: A Freehand Pointing Interface," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 372-377, June 1995.

[74] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, pp. 257-286, Feb. 1989.

[75] L.R. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, N.J.: Prentice Hall, 1993.

[76] J.M. Rehg and T. Kanade, "DigitEyes: Vision-Based Human Hand Tracking," Technical Report CMU-CS-93-220, School of Computer Science, Carnegie Mellon Univ., 1993.

[77] J.M. Rehg and T. Kanade, "Model-Based Tracking of Self-Occluding Articulated Objects," *Proc. IEEE Int'l Conf. Computer Vision*, Cambridge, Mass., pp. 612-617, June 20-23 1995.

[78] H. Rheingold, *Virtual Reality*. Summit Books, 1991.

[79] J. Schlenzig, E. Hunter, and R. Jain, "Vision-Based Hand Gesture Interpretation Using Recursive Estimation," *Proc. 28th Asilomar Conf. Signals, Systems, and Computer*, 1994.

[80] J. Schlenzig, E. Hunter, and R. Jain, "Recursive Identification of Gesture Inputs Using Hidden Markov Models," *Proc. Second IEEE Workshop on Applications of Computer Vision*, Sarasota, Fla., pp. 187-194, Dec. 5-7, 1994.

[81] J. Segen, "Controlling Computers With Gloveless Gestures," *Proc. Virtual Reality Systems*, Apr. 1993.

[82] R. Sharma, "Active Vision for Visual Servoing: A Review," *IEEE Workshop on Visual Servoing: Achievements, Applications and Open Problems*, May 1994.

[83] R. Sharma, T.S. Huang, and V.I. Pavlovic, "A Multimodal Framework for Interacting With Virtual Environments," *Human Interaction With Complex Systems*, C.A. Ntuen and E.H. Park, eds., pp. 53-71. Kluwer Academic Publishers, 1996.

[84] R. Sharma, T.S. Huang, V.I. Pavlovic, Y. Zhao, Z. Lo, S. Chu, K. Schulten, A. Dalke, J. Phillips, M. Zeller, and W. Humphrey, "Speech / Gesture Interface to a Visual Computing Environment for Molecular Biologists," *Proc. Int'l Conf. Pattern Recognition*, 1996.

[85] K. Shirai and S. Furui, "Special Issue on Spoken Dialogue," *Speech Communication*, vol. 15, pp. 3-4, 1994.

[86] T.E. Starner and A. Pentland, "Visual Recognition of American Sign Language Using Hidden Markov Models," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 189-194, June 1995.

[87] J. Streeck, "Gesture as Communication I: Its Coordination With Gaze and Speech," *Communication Monographs*, vol. 60, pp. 275-299, Dec. 1993.

[88] D.J. Sturman and D. Zeltzer, "A Survey of Glove-Based Input," *IEEE Computer Graphics and Applications*, vol. 14, pp. 30-39, Jan. 1994.

[89] D.X. Sun and L. Deng, "Nonstationary Hidden Markov Models for Speech Recognition," *Image Models (and Their Speech Model Cousins)*, S.E. Levinson and L. Shepp, eds., pp. 161-182. New York: Springer-Verlag, 1996.

[90] D.L. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831-836, 1996.

[91] K.A. Tarabanis, P.K. Allen, and R.Y. Tsai, "A Survey of Sensor Planning in Computer Vision," *IEEE Trans. Robotics and Automation*, vol. 11, pp. 86-104, 1995.

[92] D. Thompson, "Biomechanics of the Hand," *Perspectives in Computing*, vol. 1, pp. 12-19, Oct. 1981.

[93] Y.A. Tijerino, K. Mochizuki, and F. Kishino, "Interactive 3D Computer Graphics Driven Through Verbal Instructions: Previous and Current Activities at ATR," *Computers and Graphics*, vol. 18, no. 5, pp. 621-631, 1994.

[94] A. Torige and T. Kono, "Human-Interface by Recognition of Human Gestures With Image Processing. Recognition of Gesture to Specify Moving Directions," *IEEE Int'l Workshop on Robot and Human Communication*, pp. 105-110, 1992.

[95] R. Tubiana, ed., *The Hand*, vol. 1. Philadelphia, Penn.: Sanders, 1981.

[96] C. Uras and A. Verri, "Hand Gesture Recognition From Edge Maps," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 116-121, June 1995.

[97] R. Vaillant and D. Darmon, "Vision-Based Hand Pose Estimation," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 356-361, June 1995.

[98] M.T. Vo, R. Houghton, J. Yang, U. Bub, U. Meier, A. Waibel, and P. Duchnowski, "Multimodal Learning Interfaces," *ARPA Spoken Language Technology Workshop 1995*, Jan. 1995.

[99] M.T. Vo and A. Waibel, "A Multi-Modal Human-Computer Interface: Combination of Gesture and Speech Recognition," *Adjunct Proc. InterCHI'93*, Apr. 26-29 1993.

[100] A. Waibel and K.F. Lee, *Readings in Speech Recognition*. Morgan Kaufmann, 1990.

[101] C. Wang and D.J. Cannon, "A Virtual End-Effector Pointing System in Point-and-Direct Robotics for Inspection of Surface Flaws Using a Neural Network-Based Skeleton Transform," *Proc. IEEE Int'l Conf. Robotics and Automation*, vol. 3, pp. 784-789, May 1993.

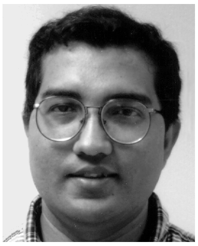
[102] K. Watanuki, K. Sakamoto, and F. Togawa, "Multimodal Interaction in Human Communication," *IEICE Trans. Information and Systems*, vol. E78-D, pp. 609-614, June 1995.

[103] A.D. Wilson and A.F. Bobick, "Configuration States for the Representation and Recognition of Gesture," *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 129-134, June 1995.

[104] A.D. Wilson and A.F. Bobick, "Recovering the Temporal Structure of Natural Gestures," *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, Killington, Vt., pp. 66-71, Oct. 1996.

[105] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, Killington, Vt., pp. 51-56, Oct. 1996.

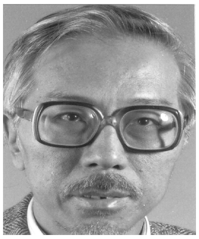
Vladimir I. Pavlovic received the Dipl Eng degree in electrical engineering from the University of Novi Sad, Yugoslavia, in 1991. In 1993, he received the MS degree in electrical engineering and computer science from the University of Illinois at Chicago. He is currently a doctoral student in electrical engineering at the Beckman Institute and the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign. His research interests include vision-based computer interaction, multimodal signal fusion, and image coding.



Rajeev Sharma is an assistant professor in the Department of Computer Science and Engineering at the Pennsylvania State University, University Park. After receiving a PhD in computer science from the University of Maryland, College Park, in 1993, he spent three years at the University of Illinois, Urbana-Champaign as a Beckman Fellow and adjunct assistant professor in the Department of Electrical and Computer Engineering.

He is a recipient of the Association of Computing Machinery Samuel Alexander Doctoral Dissertation Award and the IBM pre-doctoral fellowship.

His research interests lie in studying the role of computer vision in robotics and advanced human-computer interfaces.



T.S. Huang received his B.S. Degree in electrical engineering from National Taiwan University, Taipei, Taiwan, China; and his MS and ScD degrees in Electrical Engineering from the Massachusetts Institute of Technology, Cambridge, Massachusetts. He was on the faculty of the Department of Electrical Engineering at MIT from 1963 to 1973; and on the faculty of the School of Electrical Engineering and director of its Laboratory for Information and Signal Processing at Purdue University from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana-Champaign, where he is now William L. Everitt Distinguished Professor of Electrical and Computer Engineering, Research Professor at the Coordinated Science Laboratory, and head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology.

During his sabbatical leaves Dr. Huang has worked at the MIT Lincoln Laboratory, the IBM Thomas J. Watson Research Center, and the Rheinishes Landes Museum in Bonn, West Germany, and held visiting professor positions at the Swiss Institutes of Technology in Zurich and Lausanne, University of Hannover in West Germany, INRS-Telecommunications of the University of Quebec in Montreal, Canada, and University of Tokyo, Japan. He has served as a consultant to numerous industrial firms and government agencies both in the US and abroad.

Dr. Huang's professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. He has published 11 books, and over 300 papers in network theory, digital filtering, image processing, and computer vision. He is a Fellow of the International Association of Pattern Recognition, IEEE and the Optical Society of America and has received a Guggenheim Fellowship, an A.V. Humboldt Foundation Senior U.S. Scientist Award, and a Fellowship from the Japan Association for the Promotion of Science. He received the IEEE Acoustics, Speech, and Signal Processing Society's Technical Achievement Award in 1987 and the Society Award in 1991. He is a Founding Editor of the *International Journal Computer Vision, Graphics, and Image Processing*; and Editor of the *Springer Series in Information Sciences*, published by Springer Verlag.