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

Vladimir I. Pavlović

Rajeev Sharma

Thomas S. Huang

Department of Electrical and Computer Engineering, and
The Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign
405 N. Mathews Avenue, Urbana, IL 61801, USA

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. We survey the literature on vision-based hand gesture recognition within the context of its role in HCI. The number of approaches to video-based hand gesture recognition has grown in recent years. Thus, the need for systematization and analysis of different aspects of gestural interaction has developed. We discuss a complete model of hand gestures that possesses both spatial and dynamic properties of human hand gestures and can accommodate for all their natural types. Two classes of models that have been employed for interpretation of hand gestures for HCI are considered. The first utilizes 3D models of the human hand, while the second relies on the appearance of the human hand in the image. Investigation of model parameters and analysis features and their impact on the interpretation of hand gestures is presented in the light of the naturalness desired for HCI. 3D hand models offer a way for complete modeling of all hand gestures. However, they lack the simplicity and computational efficiency which is highly preferred, and currently feasible with the appearance-based models. We suggest some methods that can increase the effectiveness of gestural interface for HCI. Integration of hand gestures with other natural modes of communication can provide a potential answer to this problem. Nonetheless, further work that puts together advances in computer vision with understanding of human-computer interaction will be necessary to produce an effective and natural hand gesture interface.

1 *Introduction*

Human society lives through interaction among its entities and their environments. In our daily lives we interact with other people and objects to perform a variety of actions that are important to us. Computers and computerized machines have become a new element of our society. They increasingly influence many aspects of our lives: for example, the way we communicate, the way we perform our actions, and the way we interact with our environment. A new concept of interaction has, thus, emerged: *human-computer interaction* (HCI). Although the computers themselves have advanced tremendously, the common HCI still relies on simple mechanical devices - keyboards, mice and joysticks - that tremendously reduce the effectiveness and naturalness of such interaction. This limitation has become even more evident with the emergence of a new concept surrounding this interaction - *virtual reality*. Recent studies have shown that it is very natural to point at virtual space using the index finger and to explore virtual objects using one's hands [1]. It is also easier to understand other people in the same virtual environment if we can see them manipulate objects in it. The ideas of *virtual collaborative environments* (VCEs) in which researchers jointly design and test prototypes of new products are becoming a reality [2, 3]. However, new means of HCI have to be available for us to perform interactions in such environments in a more natural way.

Ever since the early days of computers we have been attempting to make them understand our speech. But only in the last several years has there been an increased interest in trying to introduce the other means of human-to-human interaction to the field of HCI. These new means include a class of devices based on the spatial motion of the human arm: *hand gestures*. Human hand gestures are a means of non-verbal interaction among people. They range from simple actions of pointing at objects and moving them around to the more complex ones that express our feelings or 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, sometimes, 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 so-called *glove-based devices* [4, 5, 6, 7, 8]. However, glove-based devices do not completely fulfill one important requirement on which HCI should be based: naturalness. Glove-based gestural interfaces force the user to 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.

To overcome the limitations imposed by the glove-based devices a vision-based approach to hand-centered HCI has been proposed in recent years. The approach suggests using a set of video cameras and computer vision techniques to interpret gestures. The non-obstructiveness of the resulting vision-based interface has resulted in a burst of recent activity in this area. Most of the work on vision-based gestural HCI has mainly been focused on the recognition of static hand gestures or *postures*. 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. Dynamics of the hand gestures were easily disregarded and interpreted in the light of object tracking. However, hand gestures are dynamic actions and should be interpreted as such: the motion of the hands conveys as much meaning as their posture does. The fusion of the dynamic characteristics of gestures into HCI has only recently been given the role it deserves. Numerous approaches, ranging from global hand motion to independent fingertip motion, have yet again been exploited, sometimes without the necessary justification. The rapid growth of gesture-based HCI has brought to surface the need for systematization and analysis of many aspects of such interaction, especially in the light of its naturalness. Unfortunately, this task has not yet been accomplished in the technical literature. This paper attempts to achieve that goal.

The paper is organized as follows: In Section 2 we formulate the problem of using vision-based hand gestures as a means of HCI. Next, we discuss the modeling methods for hand

gestures in HCI. We propose a definition and taxonomy of gestures that is suitable for the HCI framework. Section 4 and Section 5 establish a unified approach to analysis and recognition of hand gestures. We consider the choice of various features used in the analysis of the human hand/arm images and the influence of different models of gestures on the performance of the overall gesture interpretation scheme. To further stress the importance of hand gestures for HCI, we briefly discuss in Section 6 a number of their possible applications in light of the modeling, analysis and recognition techniques that were previously presented. We conclude this review with the discussion of limitations of the current approaches to visual interpretation of hand gestures and propose a number of possible solutions as well as some prospective research directions.

2 Hand Gestures in HCI

Hand gestures are a new mode for HCI. Visual interpretation of hand/arm movements carries a tremendous advantage over other techniques that require the use of mechanical transducers: it is non-obstructive. There are few restrictions imposed on the user's movements - restrictions that may otherwise be caused by the weight or discomfort of mechanical devices. Nevertheless, visual interpretation also carries a burden of complexity in implementation.

Numerous approaches have been applied to the problem of visual interpretation of gestures for HCI, as it 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: hand tracking, pose classification, or hand posture interpretation, for example. To effectively study the process of hand gesture interpretation, a global structure of the interpretation system needs to be established. For that purpose we propose the following global vision-based gesture interpretation system (Figure 1): The system requires that a mathematical model of gestures be established first. Such model is pivotal for the successful functioning of the system - we devote Section 3 to in-depth discussion of gesture modeling issues. Once the model is decided upon, the system

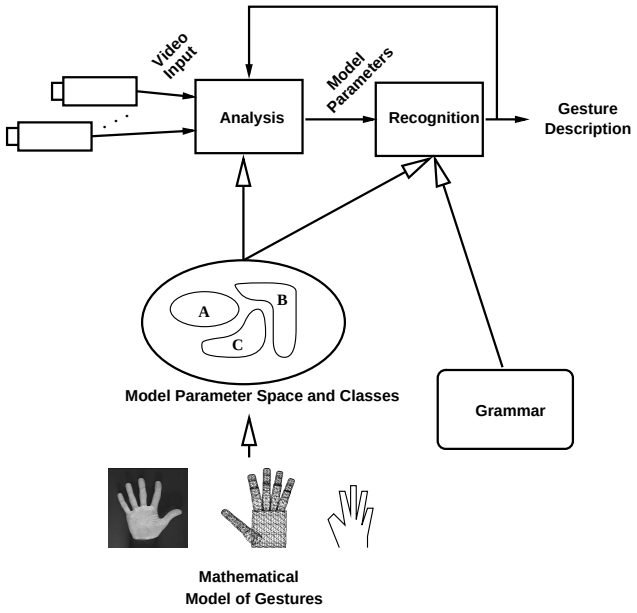

Figure 1: Block diagram of vision-based gesture interpretation system.

follows a classical path. Model parameters are computed in the analysis stage (Section 4) from image features extracted from single or multiple video input streams. Selection of features is specific to the task of gesture interpretation and crucial for the effective model parameter computation - thus, it should not be overlooked, as we note in Section 4.2 and Section 4.3. The analysis stage is followed by the recognition block (Section 5). Here, the parameters are classified and interpreted in the light of the accepted model and the rules imposed by some adequate grammar. The grammar reflects 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. Naturalness of interpretation of gestures is measured at this point. It encompasses both accuracy, robustness, and speed, as well as the variability in the number of different classes of hand/arm movements it covers. This requirement is the milestone that remains to be fully reached in the future.

3 Gesture Modeling

The quality of 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. In some instances, for example, 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 focuses on the question of modeling of hand gestures for HCI.

3.1 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 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 the man's expression and social interaction [9]. 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 the 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" [9]).

For the purpose of establishing a hand-based means of interaction in the computer controlled virtual environment we propose the following definition of hand gestures:

Definition 1 *Let* $\mathbf{h}(t) \in \mathcal{S}$ *be a vector that describes the pose of the hands and/or arms and their spatial position within an environment at time* t *in the parameter space* $\mathcal{S}$*. A hand gesture is represented by a trajectory in the parameter space* $\mathcal{S}$ *over a suitably defined interval* $\mathcal{I}$*.*

Note that the definition suggested above allows the possibility of two-handed gestures. In spite of that, we should note that most of the gestures performed in a natural environment are of a single-hand type (the exceptions including some manipulations that use two hands or some modalizing gestures). Two questions remain, however. The first one is the construction of the gestural model over the parameter set $\mathcal{S}$. The other one is how to define the gesture interval $\mathcal{I}$. We now focus our attention on these two questions.

3.2 Gestural Taxonomy

The lack of a clear definition of gestures in general raises another issue: the taxonomy of gestures. Several taxonomies have been suggested in the literature that deals with psychological aspects of gestures. They vary from author to author. Kendon [9] distinguishes "autonomous gestures" (that occur independently of speech) from "gesticulation" (gestures that occur in association with speech). McNeill and Levy [10] recognize three groups of gestures: iconic and metaphoric gestures, and "beats". The taxonomy that seems most appropriate for HCI purposes was recently developed by Quek [11, 12]. We adopt and further generalize this taxonomy in the following proposition:

Proposition 1 *A taxonomy of gestures applicable to HCI is given in Figure 2.*

We first classify all hand/arm movements into two major classes: gestures and unintentional movements. Unintentional movements are those hand/arm movements that do not convey any gestural 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

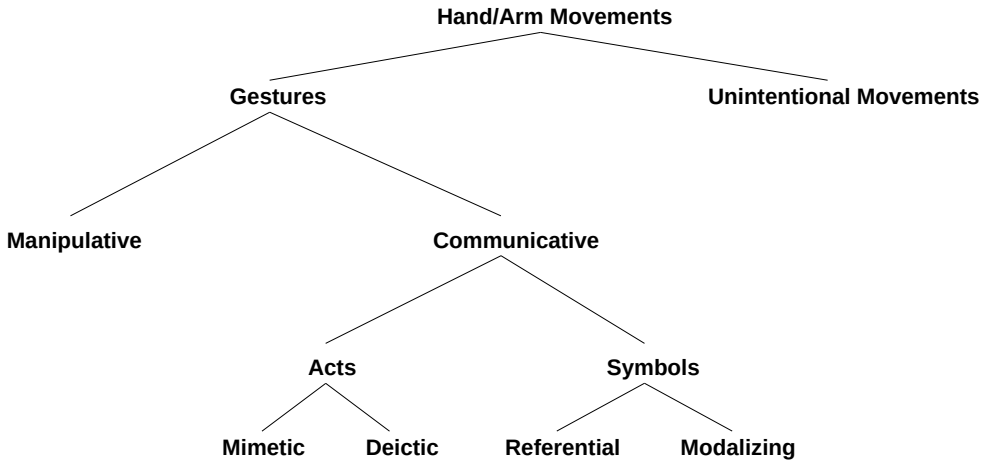

Figure 2: Gestural taxonomy for HCI.

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, the most commonly used gestures since they can often be represented by different static hand postures, as we will discuss further in Section 6. 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 $\mathcal{S}$ and gesture interval $\mathcal{I}$ are determined. Related to the gestural taxonomy is the classification of gestural dynamics. We consider this issue next.

3.3 Temporal Modeling of Gestures

Human gestures are a dynamic process. Therefore, the issue of temporal (dynamic) characteristics of gestures is of a very practical nature. It helps us resolve the problem of temporal

segmentation of gestures from other unintentional hand/arm movements. This is tantamount to the question of how to determine gesture interval $\mathcal{I}$.

Surprisingly, psychological studies of gestures provide us with a fairly consistent answer to the previous question. Kendon [9] calls this interval a "gesture phrase". It has been established that three phases make a gesture: preparation, nucleus (peak or stroke [10]), and retraction. 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" [9]. Finally, the hand either returns to the rest position or repositions for the new gesture phase. An exception to this rule are 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. However, a more useful set of rules can be developed that leads to the same temporal classification. This set of rules was suggested by Quek [11, 12]. We formulate a modified version of these rules in the form of the following proposition:

Proposition 2 *In a HCI environment the following set of rules determines the temporal segmentation of gestures:*

1. *Gesture interval consists of three phases: preparation, stroke, and retraction.*
2. *Hand pose during the stroke follows a classifiable path in the parameter space.*
3. *Gestures are confined to a specified spatial volume (workspace).*
4. *Repetitive hand movements are gestures.*
5. *Manipulative gestures have longer gesture interval lengths than communicative 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.

Proposition 2 holds in the case of general gestures for HCI. However, as it will be seen in Section 5, the complexity of gestural interpretation usually imposes more stringent constraints on the allowed temporal variability of hand gestures. Hence, most of the work in vision-based gesture HCI that has been done so far often reduces gestures to their static equivalents - hand poses.

3.4 Spatial Modeling of Gestures

Hand/arm movements are actions in a 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 influenced by the kind of application the gestural interface is intended for. For example, some applications require simple models (like static image templates of the human hand in TV set control in [13]), while some others require more sophisticated ones (3D hand model used by [14], for instance). This gives rise to the following question: is there a model of hand/arm movements that can provide a complete description of gestures for HCI? We propose the following answer:

Proposition 3 *A complete gesture model for HCI is the one whose parameters belong to the parameter space $\mathcal{S}$ constructed in the following manner:*

$$\mathcal{S} = \{\mathbf{x} \ : \ \mathbf{x} = \textit{ position of all hand and arm segment joints and fingertips in a 3D space}\}.$$

Proposition 3 relies on the assumption that the human hand and arm can be thought of as an articulated object. This is valid in HCI since the deformations of the human hand skin do not convey any additional information needed to interpret gestures for HCI.

The model proposed above could provide all the information required for correct analysis of hand gestures for HCI. However, there are two hindrances in this approach. First, the dimensionality of the parameter space is high (more than 23×3 parameters per arm). Second

and more important, to obtain the parameters of this model via computer vision techniques proves to be extremely complex.

To overcome this obstacle two major approaches in gesture modeling have been utilized so far (see Figure 3). One is to model gestures using a 3D hand and/or arm model. The other approach is *appearance-based*. We examine the two approaches more closely in the following subsections.

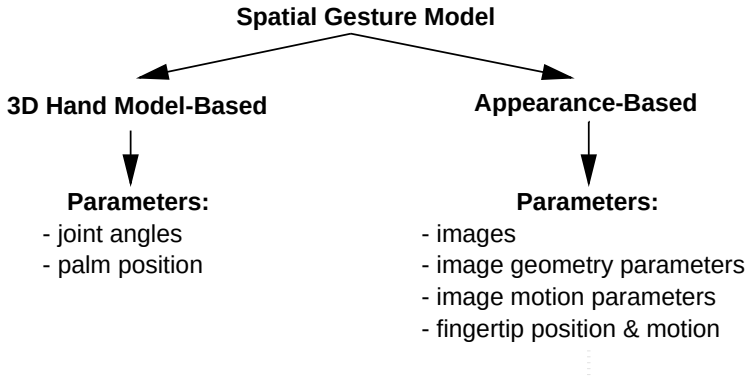

Figure 3: Spatial gesture models.

3.4.1 3D Hand/Arm Model

The employment of a 3D hand/arm model for the purpose of gesture modeling is a direct consequence of Proposition 3. Instead of dealing with all the parameters mentioned in Proposition 3, a reduced set of equivalent joint angle parameters together with segment lengths is usually used. The reduction is accomplished using sets of assumptions that generally hold. Such assumptions, for example, introduce dependencies between different joints and also impose bounds on the moving ranges of joint angles.

Most of the 3D hand/arm models are based on the simplified skeletons of the human hand/arm. Researchers concerned more with global body/arm motion use cylindrical models of the human arms or body segments [15, 16, 17, 18]. For the modeling of the human hands, skeleton models are more common (see Figure 4). Such models mimic the human hand skeleton

kinematics. Examples of the studies of the human hand morphology and biomechanics are found in [19, 20]. We briefly describe the basic notions relevant to our discussion.

The human hand skeleton consists of 27 bones, divided in three groups: carpals (wrist bones - 8), metacarpals (palm bones - 5), 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 Figure 4). Finger joints show the most flexibility: for instance, the MCP and the TM joint have two DoF (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 [21] in his 26 DoF hand model:

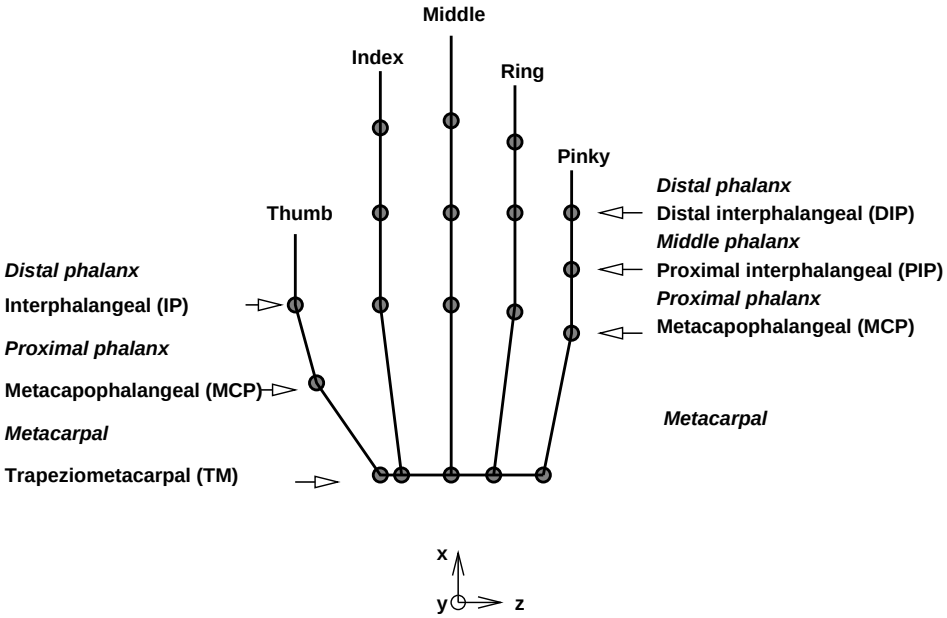

Figure 4: Skeleton-based model of the human hand.

Static constraints	
Fingers	Thumb
$0 \leq \theta^y_{MCP,s} \leq 90^o$ $-15^o \leq \theta^x_{MCP,s} \leq 15^o$	
Dynamic constraints	
$\theta^y_{PIP} = \frac{3}{2}\theta^y_{DIP}$ $\theta^y_{MCP} = \frac{1}{2}\theta^y_{PIP}$ $\theta^x_{MCP} =$ $\frac{\theta^y_{MCP}}{90}\left(\theta^x_{MCP,converge} - \theta^x_{MCP,s}\right) +$ $\theta^x_{MCP,s}$	$\theta^y_{IP} = \theta^y_{MCP}$ $\theta^y_{TM} = \frac{1}{3}\theta^y_{MCP}$ $\theta^x_{TM} = \frac{1}{2}\theta^x_{MCP}$

where superscripts denote flexions/extensions ("y") or adduction/abduction ("x") movements in local, joint centered coordinate systems. In another example, Lee and Kunii [22, 23] developed a 27 degree of freedom (DoF) hand skeleton model with an analogous set of constraints. Similar skeleton-based models of equal or lesser complexity have been used by other authors [24, 25, 26, 27, 28].

3.4.2 Appearance-Based Model

The second group of models is based on appearance of hands/arms in the image. This means that the model parameters themselves do not encompass any of the parameters mentioned in Proposition 3 or the ones directly derived from them. They model gestures 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 hand and/or arm [29, 30, 31]. Deformable 2D templates are the sets of points on the outline of an object that are used as interpolation nodes for the object outline approximation. The simplest interpolation function used is a piecewise linear function. The template sets and their corresponding variability parameters (that describe variability of elements within the set)

are obtained through principal component analysis (PCA) of many of the training sets of data. Template-based models are used mostly for hand-tracking purposes [30]. They can also be used for simple gesture classification based on the multitude of classes of templates [31].

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 [32, 33]. Images of fingers only can also be employed as templates [34] in a finger tracking application.

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 4). Contours as a direct model parameter are often used: simple edge-based contours [35, 36] or "signatures" (contours in polar coordinates) [37] are some possible examples. Contours can also be employed as the basis for further eigenspace analysis [38, 39]. Other parameters that are sometimes used are image moments [40, 41]. They are easily calculated from hand/arm silhouettes or contours. Finally, many other parameters have been used: Zernike moments [42] and orientation histograms [43], 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 [22, 23]: 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 [44, 45, 46]. Applications that are concerned with deictic gestures usually use only a single (index) fingertip and some other reference point on the hand or body [47, 46, 48].

4 *Gesture Analysis*

In the previous section we discussed different ways to model gestures in HCI. The purpose of the analysis stage is to estimate the parameters (trajectory in parameter space) of the gesture model based on the number of low level features extracted from images of human operators acting on a HCI environment. Parameters of gesture models are acquired through a multistage analysis of mono or multi camera video input sequences or still images. Three steps constitute such analysis: 1) hand/arm localization, 2) hand/arm feature extraction, and 3) hand/arm model parameter computation from features (see Figure 5). We further analyze each of these steps.

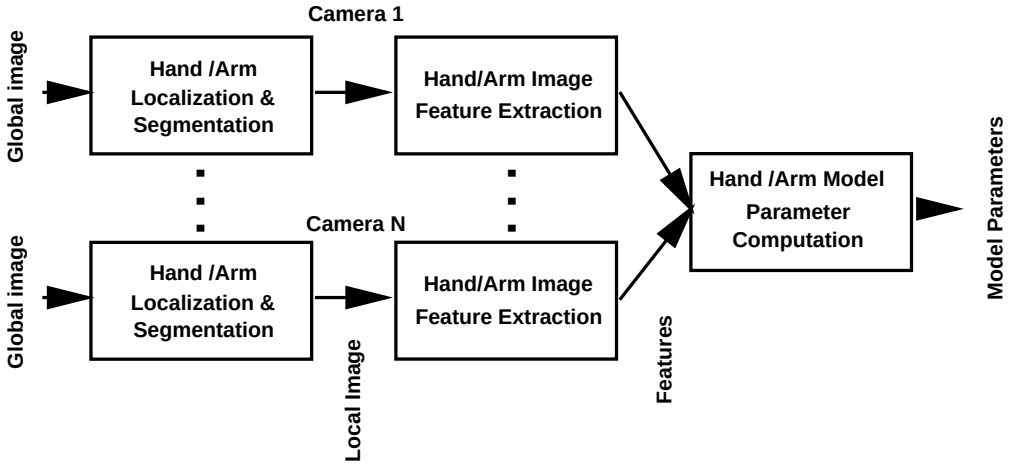

Figure 5: Gesture Analysis in HCI.

4.1 *Hand Localization and Segmentation*

Hand/arm localization and segmentation is a process in which the hands and/or arm are extracted from the rest of the image. In general, this is a complex task. Hence, to lower the burden

of the localization and segmentation analysis a variety of restrictions are usually used: 1) restrictions on background, 2) restrictions on user, and 3) restrictions on imaging. Restrictions on background are the most commonly used ones: a uniform, distinctive (dark) background greatly simplifies the segmentation task. Additional restriction on the user (the requirement to wear long dark sleeves for example) simplify the localization problem and so do the restrictions on imaging (on-hand focused cameras, for instance). Extraction of the hands from the background is then performed by thresholding the image directly. Less restrictive setups usually employ color histogram analysis. The color space-based analysis is applicable because of the characteristic histogram footprint (usually in HSV color space) of the human skin (see [26, 49, 48], for instance). Since both of the mentioned approaches may require additional processing steps (exclusion of false candidates, for instance) several applications resort to the use of uniquely colored gloves or markers on hands/fingers [50, 45, 46, 23, 51]. Although, from the computational point of view, these methods are easier to implement they tend to reduce the naturalness of the interaction. Finally, some other techniques take advantage of motion analysis of the scene: moving artifacts, under certain restrictions, are mostly produced by hand/arm movements and can thus be used to segment out the hand from other static objects ([13, 11], for example).

4.2 Features

The extraction of low level image features depends on the model of the gestures in use. Even though different models use different types of parameters, the features employed to calculate the parameters are often very similar. For example, some 3D hand/arm models and models that use finger trajectories all require fingertips to be extracted as their features.

Images of hands/arms are often used as features by themselves. A wide scope of parameters (previously described in Section 3.4.2) as well as all other features can be obtained from images.

Hand/arm silhouettes are one of the simplest features, yet they are widely used. Silhouettes are easily extracted from local hand/arm images in restricted background setups. In the case of

complex backgrounds, techniques that employ color histogram analysis (similar to the one used for hand localization) can be used. Examples of the silhouettes as features are found in both 3D hand model-based analysis (for example [14]) as well as in the appearance-based techniques (as in [52]). Furthermore, many of the other, higher level features and parameters can be extracted from silhouettes (like image moments, contours, or fingertips).

Contours represent another group of features. Several different edge detection schemes can be used to produce contours. Some are based on simple hand/arm silhouettes and the others use color or grey-level images. Contours are used in both 3D model and appearance-based model analysis. In 3D model-based gesture recognition they can be used to form sets of finger link candidates. In appearance-based models many different parameters can be associated with contours: for instance "signatures" (polar functions of points on the contour [37]) and "size functions" [53].

A very commonly used feature in gesture analysis is the fingertip. Fingertip locations can be used to obtain parameters of 3D hand models or 2D appearance-based models. However, the detection of fingertip locations in either 3D or 2D space is not trivial. A simple, yet effective solution to the detection problem is to use marked gloves or color markers to designate the characteristic fingertips (see [50, 54, 46, 23, 55], 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 [34] or fingers [25] or generic 3D cylindrical models [56]. These techniques can be enhanced by using additional image features, like contours [24]. Some fingertip extraction algorithms are based on the characteristic properties of fingertips in the image. For instance, curvature of a fingertip follows a characteristic pattern (low-high-low) [57, 27]. Heuristics (the fact that the finger represents the foremost point of the hand in deictic gestures, for instance) can also be used [57, 48]. Finally, many other indirect approaches in detection of fingertips can be employed in some instances, like image analysis using specially tuned Gabor kernels [28].

The discussed features reflect only some of the possible features that can be used. In fact, one could, for example, combine different features to form a more robust set used for effective parameter computation.

4.3 Parameter computation

Computation of the model parameters is the last stage in hand/arm gesture analysis. In the gesture recognition systems, this stage is followed by the recognition block. For hand/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.

Most of the 3D hand/arm model-based gesture models employ successive approximation methods for their parameter computation. The basic idea is to vary model parameters until the features extracted from the model match the ones obtained from the data images (see Figure 6). The matching procedure usually begins with the palm and ends with the matching of fingers. Initial model parameters are usually selected as either the ones that match a generic hand position (open hand, for example) or the ones obtained from the prediction analysis of parameters in the previous images in the sequence. A multitude of features has been used for the hand parameter computation: Kuch and Huang [14] used hand silhouettes. They varied their 3D volumetric model parameters until the model silhouette matched the one of the hand image. Many applications rely on fingertip locations to calculate model parameters. Lee and Kunii [22] proved that 3D locations of five fingertips together with two additional characteristic points on the palm uniquely define a hand pose (under several assumptions similar to the ones discussed in Section 3.4.1). Models that use the similar approach were derived by many other authors: Ahmad [26], Kang et al. [58], and Vaillant and Darmon [27], for example. Other approaches use contours and edges to guide successive adjustments of 3D model parameters (in selecting possible candidates for finger or arm links or the palm): for example, Clergue et al. [15], Downton and Drouet [16], and Etoh et al. [17], Gavrila and Davis [18].

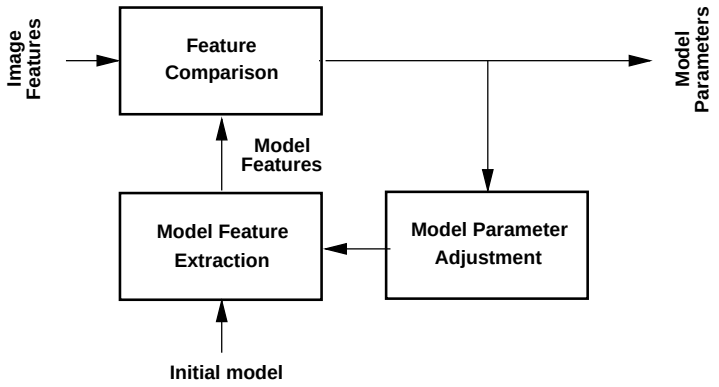

Figure 6: 3D Hand/arm model parameter computation through successive approximation technique.

Deformable 2D template-based models usually use a similar successive approximation approach: model position, orientation and principal components are adjusted in successive steps, until a satisfactory match between the model and the image is achieved [30, 29, 31].

Many of the simpler models use direct mappings between the feature and the parameter spaces. Most of the mappings are explicitly defined (image moments from image silhouettes, for instance), while the others employ interpolation on the feature-parameter correspondence tables (usually obtained through some training procedure).

5 *Gesture Recognition*

Gesture recognition is the phase in which the trajectory in the parameter space (obtained in the analysis stage) is classified as a member of some meaningful subset of the parameter space. Two problems are associated with the recognition process: 1) optimal partitioning of the time-model parameter space and 2) implementation of the recognition procedure. The problem of partitioning further raises two new questions: 1) how to design a meaningful partition of the parameter space such that it reflects the "natural" (human) perception of different gesture, and 2) how to determine class membership mappings in the parameter space. Related to the implementation issue of the recognition process is the question of a more practical nature: how

computationally intensive is the recognition procedure. We discuss each of the above questions in more detail.

Partitioning of the gesture parameter space is influenced primarily by the kind of application the gesture-based HCI system is intended for. An optimal partitioning should be such that it produces a single class in the parameter space corresponding to each allowed gesture that minimally intersect with any other gesture class. To provide a complete description of any *general* gesture, the partitioning has to reflect taxonomy of hand/arm movements and temporal properties of gestures. Temporal properties of gestures, presented in Proposition 2, suggest that temporal partitioning has to obey the three phase model. Furthermore, the model parameter space partitioning is meaningful only if it is based on the second temporal phase ("stroke") model parameters, since the initial and the final phase do not contain any gestural information. And then, it should follow the general taxonomy of gestures from Proposition 1 and also provide subclassifications within the main taxonomical groups, such as differentiation between different mimetic gestures.

The complete gesture model from Proposition 3 induces the parameter space that should allow these partitioning rules to produce a set of distinctive classes associated with each natural gesture. 3D hand model-based gesture models are the closest to the complete gesture model and therefore provide for the possibility of general gesture recognition. The more complete the 3D models are, the wider the class of gestures they can cover. On the other hand, the systems that use the appearance-based models fall short of the general gesture recognition. The parameter space of such models is insufficient to describe general gestures (i.e. several different general gestures can have very similar, difficult to distinguish trajectories). Most of those models, therefore, assume two restrictions. First, they restrict themselves to a particular taxonomical group (deictic gestures, for instance) where the particular parameter space can be optimally partitioned (with respect to the gestures in the group). Second, they often disregard dynamic properties of gestures and analyze only static postures. In such restrictive setups, many of the

simple models exhibit satisfactory recognition performance.

To perform the actual partitioning, several methods can be used. Time partitioning requires that the global hand/arm motion be known, since that is what distinguishes the three temporal phases (see Proposition 2). The model parameter space partitioning itself can be performed using a number of different classification methods. The methods require at least one representative gesture per class to be known. The class representative can either be given ad hoc or determined through some learning-from-examples procedure, like averaging, K-means, hidden Markov models, and neural networks. The membership mappings for the classes are then based on minimum distance measure from a class representative. Hidden Markov models (HMM) is one technique that is particularly appropriate in this case. The states of the HMM can easily be associated with the temporal gesture phases. Therefore, the gesture HMM should contain at least, and usually more than, three (hidden) states. The HMM training procedure is built on learning-from-examples based classification of time-parameter space, while the recognition procedure uses dynamic time warping (DTW) for temporally invariant classification. So far the gesture models that use HMM have been employed in appearance-based recognition with notable success [40, 41] .

A successful recognition scheme should in general be based not only on classification through distance-based membership functions but also on 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 gesture. Still, only a small number of the systems so far exploits this fact. The grammars are simple and usually introduce artificial linguistic structures: they build their own "languages"

that have to be learned by the user [47, 49, 48, 40].

Finally, the question of computational effectiveness in recognition arises. The trade-off is classical: model complexity versus recognition applicability versus recognition time. The more complex the model is, the wider class of gestures it can, in general, be applied to. However, the computational complexity increases, and, hence, the recognition time. Most of the 3D model-based gesture models are characterized by more than ten 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 [22] (although it does not use any prediction element) to 10 frames per second in [24]. Yet the applicability of the systems in the general HCI gesture recognition arena is superior to the one of the simple appearance-based models. The appearance-based models are usually restricted in their applicability to a narrow subclass of HCI applications: enhancements of the computer mouse concept [34, 13, 47, 49, 48], or hand posture classification [59, 57, 28, 39, 36, 41], for instance. On the other hand, they are of the lower complexity and, thus, computationally more affordable and easier to implement in real time applications.

6 Applications and Systems

Recent interest in gestural interface for HCI has been driven by a vast number of potential applications (Figure 7). 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 control of computerized machineries or be used to help physically impaired communicate 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) [60, 61].

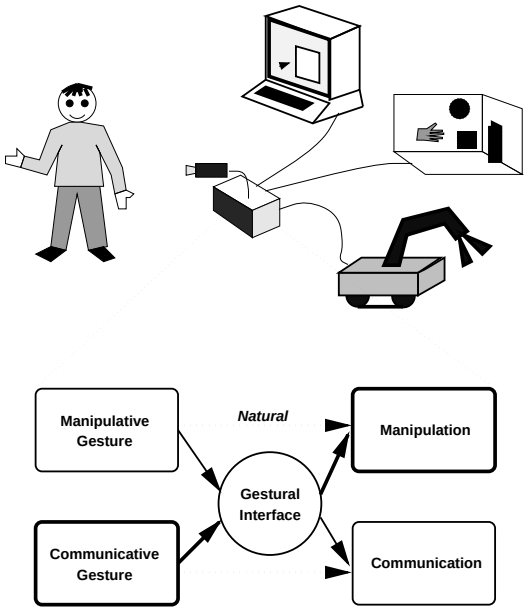

Figure 7: Applications of gestural interface for HCI. Unlike the gestures in a natural environment, both manipulative and communicative gestures in HCI can be employed to direct manipulations of objects or convey messages.

Hand gestures in natural environments are used for both manipulative actions and communication (see Section 3). The communicative role of gestures is, however, very subtle as hand gestures tend to be a supportive element of speech (with the exception of deictic gestures, which play a major role in human communication). That is why the manipulative aspect of gestures prevails in their current use for HCI: most applications of hand gestures portray them as the manipulators of *virtual objects* (VOs). This is depicted in Figure 7. VOs can be computer generated graphics, like simulated 2D and 3D objects [37, 34, 62, 47] or windows [49, 48], or abstractions of computer-controlled physical objects, such as device control panels [26, 13, 47] or robotic arms [59, 58, 63]. 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: $< \textit{select object} >< \textit{rotate object} >$. The first action uses coarse hand tracking to move a pointer in the VE to the vicinity of the object. To rotate the object the user rotates his/her

hand back and forth producing a *metaphor* for rotational manipulation [37]. 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 4 and Section 5). 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, 3D hand model-based gesture models are computationally more expensive then the appearance-based models (see Section 5). Therefore, to achieve a usable (real-time) performance one has to resort to, for this purpose, the less desirable appearance-based models of gestures. A brief summary of characteristics of some of the systems aimed at such applications is given in Table 1.

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 to the problem uses different coding techniques. One of such techniques is model-based coding: image sequences are described by the states (position, scale, and orientation, for instance) of all physical objects in the scene (human participants in the case of VTC) [64, 65]. 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 [15], or finely tuned models of human faces or hands. Modeling of hand/arm gestures can then be of substantial value for such applications.

Application	Gestural modeling technique	Gestural commands	Complexity (speed)
CD Player Control Panel [26]	Hand silhouette moments	Tracking only	30 fps[a]
Virtual Squash [37]	Hand silhouette moments & contour "signature"	Tracking & three metaphors	10.6 fps
FingerPaint [34]	Fingertip template	Tracking only	n.a.[b]
ALIVE [33]	Template correlation	Tracking combined with recognition of facial expressions	real-time
TV Display Control [13]	Template correlation	Tracking only	5 fps
FingerPointer [47]	Heuristic detection of pointing action	Tracking and one metaphor combined with speech	real-time
Window Manager [49]	Hand pose recognition using neural networks	Tracking & four metaphors	real-time
GestureComputer [57]	Image moments & fingertip position	Tracking and six metaphors	10-25 fps
FingerMouse [48]	Heuristic detection of pointing action	Tracking only	real-time
DigitEyes [24]	27 DoF 3D hand model	Tracking only	10 fps
ROBOGEST [59]	Silhouette Zernike moments	Six metaphors	1/2 fps
Automatic robot instruction	Fingertip position in 2D	Grasp tracking	n.a.
Robot manipulator control [63]	Fingertip positions in 3D	Six metaphors	real-time
Hand sign recognition [38]	Most discriminating features (MDF) of images	28 signs	n.a.
ASL recognition [41]	Silhouette moments & grammar	40 words	5 fps

[a] Frames per second.
[b] Not available

Table 1: Systems that employ hand gestures for HCI. We choose speed as the measure of complexity of interpretation given the lack of any other accurate measure even though different applications may be implemented on different computer systems with different levels of optimization.

Recognition of ASL is often considered as another application that naturally employs human gestures as a 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 testbed 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 [41, 53].

Prospects for the use of hand gesture for HCI are vast. The applications mentioned in this section are only the first steps in introducing the hand gestures to HCI. The need for their further development is, thus, quite natural. We point to several important developmental issues as well as to some exciting new applications of hand gestures for HCI in more detail in the following section.

7 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 the *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 6). The reason for this is the complexity associated with the analysis (Section 4) and recognition (Section 5) of gestures. Simple gesture models usually result in the real-time gestural interfaces: for example, pointing direction can be quickly found from the silhouettes of

the human hand in relatively non-restrictive environments ([47, 48]). However, as it can be seen from Figure 8, to find the hand posture, and thus distinguish gesture, from the simple image appearance (silhouettes) is sometimes quite difficult. Interface systems based on 3D hand model-

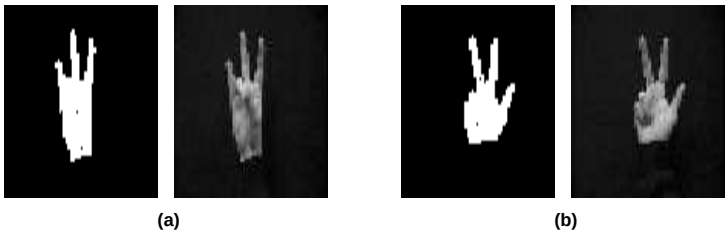

Figure 8: Silhouettes and gray-scale images of two different hand postures. The silhouette in (a) can also be interpreted as the reflection about the vertical axes of the silhouette in (b). Hence, the two silhouettes do not unambiguously define the hand posture.

based gesture models are currently non-existent. These models are most commonly employed for 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 gesture models offer more prospect 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 latter issue can be adequately solved given the solution to the first problem. 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 4.2), 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 the model parameters (Section 4.3) 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 addressed in the future. One of the aspects involves the two-handed gestures. Human gestures

naturally employ actions of both hands. Yet, almost all of the vision-based gesture systems focus their attention on single-hand gestures (with the exception of the systems developed by M. Krueger [52]). This approach, however, seems inevitable at the present time. First, every analysis technique requires 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. Other techniques (appearance based) can, in principle, handle two-handed gestures. However, their applicability is usually restricted to simple (symbolic) gestures that do not require two hands. Hence, 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 (see Proposition 3). 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 [66, 67], in which the cameras adaptively focus on some area of interest, may offer a solution to this problem.

Finally, we address the issue of interaction of multiple communication modes in HCI related

to hand gestures. Hand gesture is, like speech, body movement, and gaze, a means of communication (see Section 3.1). 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 [9, 68, 69]. This leads to the conclusion that any such *multimodal* interaction can also be rendered useful for HCI (see Figure 9). The affirmative hand gesture (speech) can

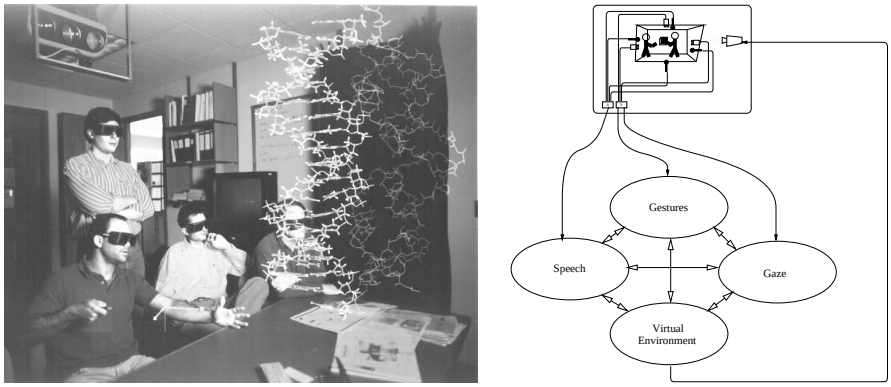

Figure 9: Multimodal gesture-speech HCI system. (Photograph courtesy of Rich Saal of the Illinois State Journal-Register, Springfield, Illinois.)

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 reduce the complexity and increase the naturalness of the interface for HCI. 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. The number of studies that explore the use of multimodality in HCI has been steadily increasing over the past couple of years [47, 70, 71, 72, 73]. 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.

8 Conclusions

Study of interaction between humans and machines has attracted the principal interest of researchers in the past several years. This coincides with the burst of activity surrounding the applications situated in the abstractions of our natural surroundings - virtual environments. However, most widely used interaction devices at this time, keyboards, mice, or joysticks, lower the ease and naturalness of interaction and, thus, hinder the effectiveness of use of such environments. The more direct use of natural means of interaction like speech, hand gestures, or gaze, has proven to play an essential role in the solution to this problem. In this paper we focus our discussion on one of the means: hand gestures.

Visual interpretation of hand gestures yields an interesting, non-obstructive potential solution to the problem of their interpretation for HCI in computer controlled environments. The number of different approaches to video-based hand gesture recognition has grown tremendously in recent years. The need for systematization and analysis of many aspects of gestural interaction has, thus, emerged. This paper addresses the task of presenting a unified approach to modeling, analysis and recognition of hand gestures for visual interpretation. An effective model of hand gestures for HCI should take into account characteristics of natural hand gestures. Therefore,

we propose a complete model of hand gestures that reflects both spatial and dynamic properties of human hand gestures and can accommodate for all their natural types. We then consider two classes of models that have been employed so far for the purpose of interpretation of hand gestures: the first one relies on 3D models of the human hand, while the second utilizes the appearance of the human hand in the image. Investigation of model parameters and analysis features and their influence on the recognition of hand gestures is then presented in the light of the naturalness preferred for HCI. The 3D hand models offer a way for complete modeling of all hand gestures - yet, at this stage, they lack the simplicity and computational efficiency which is highly desirable and possible to accomplish with the appearance-based models.

This study was motivated by the potential for application of hand gestures as a more natural human-computer interface. However, the applications of hand gesture interaction systems today are in their infancy. We find that even though most of the current systems employ hand gestures for manipulation of objects, the complexity of the interpretation of gestures dictates the achievable solution: gestures used to convey manipulative actions today are usually of the communicative type. Additionally, hand gestures for HCI are mostly restricted to be single-handed and produced only by a single user in the system. This consequently downgrades the effectiveness of the interaction. Hence, we suggest several methods which can elevate the effectiveness of gestural interface for HCI: we show that integration of hand gestures with speech, gaze and other naturally related modes of communication can provide us with an attractive potential solution to this problem. Nevertheless, 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.

9 Acknowledgments

This work was supported in part by National Science Foundation Grant IRI-89-08255 and in part by a grant from Sumitomo Electric Industries.

References

[1] A. G. Hauptmann and P. McAvinney, "Gesture with speech for graphics manipulation," *International Journal of Man-Machine Studies*, vol. 38, pp. 231–249, Feb. 1993.

[2] E. J. Haug, J. G. Kuhl, and F. F. Tsai, "Virtual prototyping for mechanical system concurrent engineering," *Concurrent Engineering: Tools and Technologies for Mechanical System Design*, vol. 108, pp. 851–879, 1993.

[3] S. C.-Y. Lu and K. S. et al, "Swift: System workbench for integrating and facilitating teams," *International Journal of Intelligent and Cooperative Information Systems*, 1994.

[4] T. Baudel and M. Baudouin-Lafon, "Charade: Remote control of objects using free-hand gestures," *Communications of the ACM*, vol. 36, no. 7, pp. 28–35, 1993.

[5] S. S. Fels and G. E. Hinton, "Glove-talk: A neural network interface between a data-glove and a speech synthesizer," *IEEE Transactions on Neural Networks*, vol. 4, pp. 2–8, Jan. 1993.

[6] D. J. Sturman and D. Zeltzer, "A survey of glove-based input," *IEEE Computer Graphics and Applications*, vol. 14, pp. 30–39, Jan. 1994.

[7] D. L. Quam, "Gesture recognition with a dataglove," in *Proceedings of the 1990 IEEE National Aerospace and Electronics Conference*, vol. 2, 1990.

[8] 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," in *Proceedings of IEEE International Conference on Robotics and Automation*, vol. 3, pp. 784–789, May 1993.

[9] A. Kendon, "Current issues in the study of gesture," in *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.

[10] D. McNeill and E. Levy, "Conceptual representations in language activity and gesture.," in *Speech, palce and action: Studies in deixis and related topics* (J. Jarvella and W. Klein, eds.), Wiley, 1982.

[11] F. K. H. Quek, "Eyes in the interface," *Image and Vision Computing*, vol. 13, August 1995.

[12] F. K. H. Quek, "Toward a vision-based hand gesture interface," in *Virtual Reality Software and Technology Conference*, pp. 17–31, Aug. 1994.

[13] W. T. Freeman and C. D. Weissman, "Television control by hand gestures," in *Proc. IWAFGR'95*, (Zurich), pp. 179–183, June 1995.

[14] J. J. Kuch and T. S. Huang, "Vision based hand modeling and tracking," in *Proceedings of International Conference on Computer Vision*, (Cambridge, MA), June 1995.

[15] E. Clergue, M. Goldberg, N. Madrane, and B. Merialdo, "Automatic face and gestural recognition for video indexing," in *Proc. of IWAFGR'95*, (Zurich), pp. 110–115, June 1995.

[16] A. C. Downton and H. Drouet, "Image analysis for model-based sign language coding," in *Progress in image analysis and processing II: proceedings of the 6th International Conference on Image Analysis and Processing*, pp. 637–644, 1991.

[17] 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.

[18] D. M. Gavrila and L. S. Davis, "Towards 3-d model-based tracking and recognition of human movement: a multi-view approach," in *Proc. of IWAFGR'95*, (Zurich), pp. 272–277, June 1995.

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

[20] D. Thompson, "Biomechanics of the hand," *Perspectives in Computing*, vol. 1, pp. 12–19, Oct. 1981.

[21] J. J. Kuch, "Vision-based hand modeling and gesture recognition for human computer interaction," Master's thesis, University of Illinois at Urbana-Champaign, 1994.

[22] J. Lee and T. L. Kunii, "Constraint-based hand animation," in *Models and techniques in computer animation*, pp. 110–127, Tokyo: Springer-Verlag, 1993.

[23] J. Lee and T. L. Kunii, "Model-based analysis of hand posture," *IEEE Computer Graphics and Applications*, pp. 77–86, September 1995.

[24] J. M. Rehg and T. Kanade, "Digiteyes: Vision-based human hand tracking," Tech. Rep. CMU-CS-93-220, School of Computer Science, Carnegie Mellon University, 1993.

[25] J. M. Rehg and T. Kanade, "Visual tracking of self-occluding articulated objects," Tech. Rep. CMU-CS-94-224, Carnegie Mellon University, School of Computer Science, CMU, Pittsburgh, PA 15213, December 1994.

[26] S. Ahmad, "A usable real-time 3d hand tracker," in *IEEE Asilomar Conference*, 1994.

[27] R. Vaillant and D. Darmon, "Vision based hand pose estimation," in *Proc. of IWAFGR'95*, (Zurich), pp. 356–361, June 1995.

[28] A. Meyering and H. Ritter, "Learning to recognize 3d-hand postures from perspective pixel images," in *Artificial neural networks 2* (I. Alexander and J. Taylor, eds.), (North-Holland): Elsevier Science Publishers B.V., 1992.

[29] 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, January 1995.

[30] C. Kervrann and F. Heitz, "Learning structure and deformation modes of nonrigid objects in long image sequences," in *International Workshop on Automatic Face- and Gesture-Recognition IWAFGR95*, June 1995.

[31] A. Lanitis, C. J. Taylor, T. F. Cootes, and T. Ahmed, "Automatic interpretation of human faces and hand gestures using flexible models," in *Proc. of IWAFGR'95*, (Zurich), pp. 98–103, June 1995.

[32] T. Darrell and A. Pentland, "Space-time gestures," in *Proceedings of Computer Vision and Pattern Recognition Conference*, 1993.

[33] T. Darrell and A. P. Pentland, "Attention-driven expression and gesture analysis in an interactive environment," in *Proc. of IWAFGR'95*, (Zurich), pp. 135–140, June 1995.

[34] J. L. Crowley, F. Berard, and J. Coutaz, "Finger tacking as an input device for augmented reality," in *Proc. of IWAFGR'95*, (Zurich), pp. 195–200, June 1995.

[35] K. Cho and S. M. Dunn, "Learning shape classes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 882–888, Sept. 1994.

[36] J. Segen, "Controlling computers with gloveless gestures," in *Proceedings of virtual reality systems*, April 1993.

[37] U. Bröckl-Fox, "Real-time 3-d interaction with up to 16 degrees of freedom from monocular image flows," in *Proc. of IWAFGR'95*, (Zurich), pp. 172–178, June 1995.

[38] Y. Cui and J. Weng, "Learning-based hand sign recognition," in *Proc. of IWAFGR'95*, (Zurich), pp. 201–206, June 1995.

[39] B. Moghaddam and A. Pentland, "Maximum likelihood detection of faces and hands," in *Proc. of IWAFGR'95*, (Zurich), pp. 122–128, June 1995.

[40] J. Schlenzig, E. Hunter, and R. Jain, "Recursive identification of gesture inputs using hidden markov models," in *Proceedings of the Second IEEE Workshop on Applications of Computer Vision*, (Sarasota, FL), pp. 187–194, December 5-7 1994.

[41] T. E. Starner and A. Pentland, "Visual recognition of american sign language using hidden markov models," in *Proc. IWAFGR'95*, (Zurich), pp. 189–194, June 1995.

[42] J. Schlenzig, E. Hunter, and R. Jain, "Vision based hand gesture interpretation using recursive estimation," *Proceedings of the 28th Asilomar conference on signals, systems, and computer*, 1994.

[43] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *International Workshop on Automatic Face- and Gesture-Recognition IWAFGR95*, June 1995.

[44] S. Ahmad and V. Tresp, "Classification with missing and uncertain inputs," in *Proceedings of 1993 International Conference on Neural Networks*, vol. 3, pp. 1949–1954, 1993.

[45] J. Davis and M. Shah, "Gesture recognition," Tech. Rep. CS-TR-93-11, Department of Computer Science, University of Central Florida, 1993.

[46] Y. Kuno, M. Sakamoto, K. Sakata, and Y. Shirai, "Vision-based human computer interface with user centered frame," in *Proceedings of IROS'94*, pp. 2923–2029, 1994.

[47] 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.

[48] F. K. H. Quek, T. Mysliwiec, and M. Zhao, "Finger mouse: A freehand pointing interface," in *Proc. of IWFGRA'95*, (Zurich), pp. 372–377, June 1995.

[49] R. Kjeldsen and J. Kender, "Visual hand gesture recognition for window system control," in *Proc. of IWAFGR'95*, (Zurich), pp. 184–188, June 1995.

[50] R. Cipolla, Y. Okamoto, and Y. Kuno, "Robust structure from motion using motion parallax," in *Proceedings of International Conference on Computer Vision*, pp. 374–382, IEEE, 1993.

[51] C. Maggioni, "A novel gestural input device for virtual reality," in *1993 IEEE Annual Virtual Reality International Symposium*, pp. 118–124, IEEE, 1993.

[52] M. W. Krueger, "Environmental technology: Making the real world virtual," *Communications of the ACM*, vol. 36, pp. 36–37, July 1993.

[53] C. Uras and A. Verri, "Hand gesture recognition from edge maps," in *Proc. of IWAFGR'95*, (Zurich), pp. 116–121, June 1995.

[54] J. Davis and M. Shah, "Visual gesture recognition," *IEE Proc. - Vis. Image and Signal Process.*, vol. 141, pp. 101–10, April 1994.

[55] Y. A. Tijerino, K. Mochizuki, and F. Kishino, "Interactive 3-d computer graphics driven through verbal instructions: Previous and current activities at atr," *Computers and Graphics*, vol. 18, no. 5, pp. 621–631, 1994.

[56] J. Davis and M. Shah, "Determining 3-d hand motion," *Proceedings of the 28th Asilomar conference on signals, systems, and computer*, 1994.

[57] C. Maggioni, "Gesturecomputer - new ways of operating a computer," in *Proc. of IWAFGR'95*, (Zurich), pp. 166–171, June 1995.

[58] S. B. Kang and K. Ikeuchi, "Toward automatic robot instruction for perception - recognizing a grasp from observation," *IEEE Transactions on Robotics and Automation*, vol. 9, pp. 432–443, Aug. 1993.

[59] E. Hunter, J. Schlenzig, and R. Jain, "Posture estimation in reduced-model gesture input systems," in *International Workshop on Automatic Face- and Gesture-Recognition IWAFGR95*, June 1995.

[60] J. A. Adam, "Virtual reality," *IEEE Spectrum*, vol. 30, no. 10, pp. 22–29, 1993.

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

[62] K. Ishibuchi, H. Takemura, and F. Kishino, "Real time hand gesture recognition using 3d prediction model," in *Proceedings of 1993 International Conference on Systems, Man, and Cybernetics*, (Le Touquet, France), pp. 324–328, October 17-20 1993.

[63] A. Torige and T. Kono, "Human-interface by recognition of human gestures with image processing. recognition of gesture to specify moving directions," in *IEEE International Workshop on Robot and Human Communication*, pp. 105–110, 1992.

[64] J. F. Abramatic, P. Letellier, and M. Nadler, "A narow-band video communication system for the transmission of sign language over ordinary telephone lines," in *Image sequences processing and dynamic scene analysis* (T. S. Huang, ed.), pp. 314–336, Springer-Verlag Berlin Heidelberg, 1983.

[65] H. Harashima and F. Kishino, "Intelligent image coding and communications with realistic sensations - recent trends," *IEICE Transactions*, vol. E 74, pp. 1582–1592, June 1991.

[66] A. Blake and A. Yuille, *Active Vision.* MIT Press, Cambridge, MA, 1992.

[67] R. Sharma, "Active vision for visual servoing: A review," in *IEEE Workshop on Visual Servoing: Achievements, Applications and Open Problems*, May 1994.

[68] E. T. Levy and D. McNeill, "Speech, gesture, and discourse," *Discourse Processes*, no. 15, pp. 277–301, 1992.

[69] J. Streeck, "Gesture as communication i: its coordination with gaze and speech," *Communication monographs*, vol. 60, pp. 275–299, December 1993.

[70] R. Sharma, T. S. Huang, and V. I. Pavlović, "A multimodal framework for interacting with virtual environments," in *Proc. Symp. on Human Interaction with Complex Systems*, (Greensboro, North Carolina), September 17-20 1995.

[71] M. T. Vo and A. Waibel, "A multi-modal human-computer interface: combination of gesture and speech recognition," in *Adjunct Proceedings of InterCHI'93*, April 26-29 1993.

[72] M. T. Vo, R. Houghton, J. Yang, U. Bub, U. Meier, A. Waibel, and P. Duchnowski, "Multimodal learning interfaces," in *ARPA Spoken Language Technology Workshop 1995*, January 1995.

[73] K. Watanuki, K. Sakamoto, and F. Togawa, "Multimodal inetraction in human communication," *IEICE Transactions on Information and Systems*, vol. E78-D, pp. 609–614, June 1995.