Hand Modeling, Analysis, and Recognition

For Vision-Based Human Computer Interaction

Ying Wu and Thomas S. Huang

In the evolution of user interfaces, keyboards were the primary devices in text-based user interfaces, and then the invention of the mouse brought us the graphical user interface. What is the counterpart of the mouse when we are trying to explore three-dimensional (3-D) virtual environments (VEs)?

In many current VE applications, keyboards, mice, wands, and joysticks are the common controlling and navigating devices. However, to some extent, such mechanical devices are inconvenient and unsuitable for natural and direct interaction, because it is difficult for these devices to supply 3-D and high degree of freedom inputs. Although magnetic trackers are being used as sensors for 3-D inputs in some of these devices, they are prone to magnetic interference, and they only supply global motion information.

A more convenient and natural device is desirable to achieve more immersive interaction. The use of hand gestures has become an important part of human computer interaction (HCI) in recent years [1], [24]. To use human hands as a natural interface device, some glove-based devices have been employed to capture human hand motion by attaching sensors to measure the joint angles and spatial positions of hands directly. Unfortunately, such devices are expensive and cumbersome.

Since rich visual information provides a strong cue to infer the inner states of an object, vision-based techniques provide promising alternatives to capture human hand motion. At the same time, vision systems could be very cost efficient and noninvasive. These facts serve as the motivating forces for research in the modeling, analysis, animation, and recognition of hand gestures.

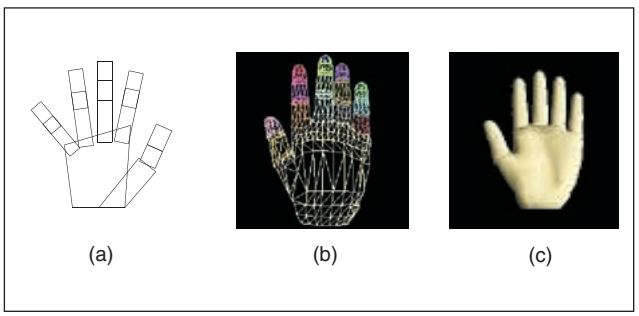
According to different application scenarios, hand gestures can be classified into several categories: conversational gestures, controlling gestures, manipulative gestures and communicative gestures. Sign language is an important case of communicative gestures. Because sign languages are highly structured [33], [37], they are very suitable as a test-bed for vision algorithms [33], [37]. Controlling gestures are the focus of current research in vision-based interfaces [6], [17], [23], [26], [35], [45]. Virtual objects can be located by analyzing pointing gestures [24]. Some display-control applications demon-

1053-5888/01/$10.00©2001IEEE

strate the potential of pointing gestures in HCI [6]. Another controlling gesture is the navigating gesture. Instead of using wands, the orientation of hands can be captured as a 3-D directional input to navigate in VEs [23]. The manipulative gestures can serve as a natural way to interact with virtual objects [32]. Tele-operation and virtual assembly are good examples of such applications. Conversational gestures are subtle in human interaction, which requires careful psychological studies. Vision-based motion capturing techniques can help those studies [5], [22].

There have been many implemented application systems in such domains as VEs, smart surveillance, HCI, teleconferencing, and sign language translation. Zeller et al. [45] presented a VE for a very large scale biomolecular modeling application. This system permits interactive modeling of biopolymers by linking a 3-D molecular graphics and molecular dynamics simulation program. Hand gestures serve as the input and controlling device of the virtual environment. Pavlovic and Berry [23] integrated controlling gestures into the VE BattleField, in which hand gestures are used not only for navigating the VE, but also as an interactive device to select and move the virtual objects in the BattleField. Ju et al. [17] developed an automatic system for analyzing and annotating video sequences of technical talks. Speakers' gestures such as pointing or writing are automatically tracked and recognized to provide a rich annotation of the sequence that can be used to access a condensed version of the talk. Quek [26] presented a FingerMouse application to recognize two-dimensional (2-D) finger movements, which are the input to the desktop. Crowley and Coutaz [6] also developed an application called FingerPaint to use fingers as input devices for augmented reality. Triesch and Maslburg [35] developed a person-independent gesture interface on a real robot that allows the user to give simple commands such as how to grasp an object and where to put it. Imagawa et al. [14] implemented a bidirectional translation system between Japanese Sign Language and Japanese in order to help the hearing impaired communicate with normal speaking people through sign language.

Analyzing hand gestures is a comprehensive task involving motion modeling, motion analysis, pattern recognition, machine learning, and even psycholinguistic studies. There are already several good review papers on human motion analysis [12] and interpretation [24]. However, a comprehensive review of various techniques in hand modeling, analysis, and recognition is needed. Due to the multidisciplinary nature of this research topic, we cannot include all the works in the literature. Rather than function as a thorough review paper, this article serves as a short tutorial to this research topic. In this article, we study 3-D hand models, various articulated motion analysis methods, and gesture recognition techniques employed in current research. We conclude with some thoughts about future research directions. We also include some of our own research results, some of which are shown as examples.



▲ *1. Hand models. (a) Cardboard model, (b) wireframe model, and (c) polygon-mesh model.*

Hand Modeling

Human hand motion is highly articulate, because the hand consists of many connected parts leading to complex kinematics. At the same time, hand motion is also highly constrained, which makes it difficult to model. Usually, the hand can be modeled in several aspects such as shape, kinematical structure, dynamics, and semantics.

Modeling the Shape

Hand shape models can be classified into several groups such as geometrical models, physical models, and statistical models. Geometrical models are suitable for 3-D rendering and hand animation applications. Moreover, they could be employed to analyze hand motion using the approach of analysis-by-synthesis [11], [19], [32]. Both physical models and statistical models emphasize hand deformation. The difference is that physical models aim for an explicit representation of deformation, while statistical models characterize hand deformation implicitly by learning from a set of examples.

Spline-based geometrical surface models represent a surface with splines to approximate arbitrarily complicated geometrical surfaces. These spline-based surface models can be made as realistic as possible, but many parameters and control points need to be specified [19]. An alternative is to approximate the homogeneous body parts by simpler parameterized geometric shapes such as generalized cylinders or super-quadrics. The advantage of this method is that it can achieve equally good surface approximation with less complexity [32], [11]. Other than parametric models, free-form hand models are defined on a set of 3-D points [13]. Polygon meshes that are formed by those 3-D points approximate the hand shape, which is computationally efficient. For computational efficiency, cardboard models could be used for visual motion capturing. Each piece in a cardboard model is a 2-D plane, but the joint angles could be adjusted. Examples of different hand models are shown in Fig. 1. Cyber Scanner, MRI techniques, or other space digitizers may be used to obtain the range data directly [13]. Another way is to reconstruct the hand model from multiple images of different views.

Physical hand shape models emphasize the deformation of the hand shape under the action of various forces [37]. The motion of the model is governed by Newtonian dynamics. The internal forces are applied to hold the shape of the model, and the external forces are used to fit the model to the image data. Examples are the simplex mesh model [13] and the finite element method model [36].

Statistical hand shape models [13] learn the deformation of hand shape through a set of training examples that can be 2-D images or range images. Mean shape and modes of variation are found using principal component analysis (PCA). A hand shape is generated by adding a linear combination of some significant modes of variation to the mean shape.

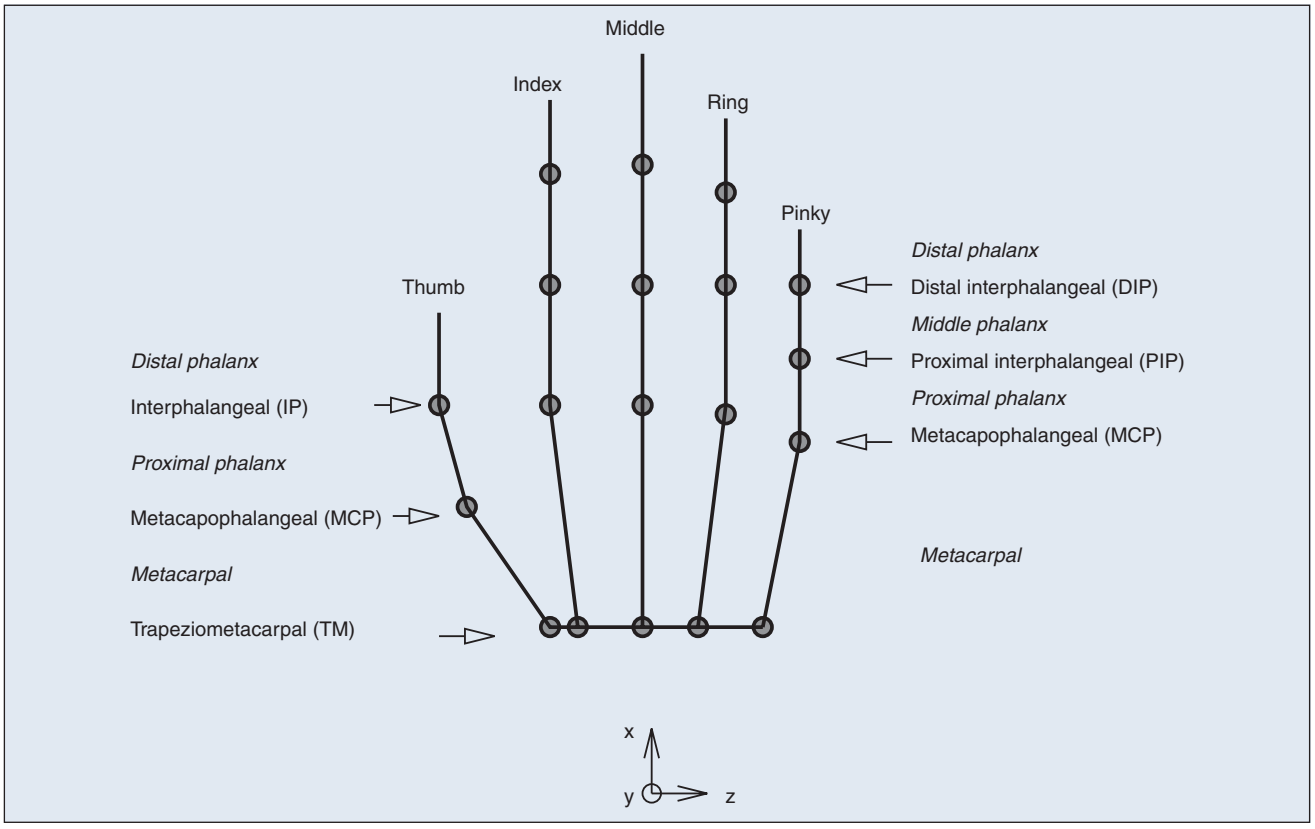
Modeling the Kinematical Structure

Figure 2 shows the skeleton of a hand. Each finger consists of three joints whose names are indicated in the figure. Except for the thumb, there are 2 degrees of freedom (DOF) for metacapophalangeal (MCP) joints, and 1 DOF for proximal interphalangeal (PIP) joints and distal interphalangeal (DIP) joints. For simplicity, the thumb could be modeled by a 5 DOF kinematic chain, with 2 DOF for the trapeziometacarpal (TM) and MCP joint and 1 DOF for the interphalangeal (IP) joint. Considering global hand pose, human hand motion has roughly 27 DOF. The challenge of hand motion analysis lies in the fact that hand motion is highly articulate.

Each finger can be modeled by a kinematic chain, in which the palm is its base reference frame and the fingertip is the end-effector. When fixing the joint length, hand kinematics can be characterized by its joint angles. The inverse kinematics problem is often involved to calculate joint angles when analyzing finger motion. Generally, gradient-based methods can be used to solve this problem by deriving the kinematical Jacobian [29]. There are other alternatives in the literature such as genetic algorithm [40]. However, such inverse kinematics problem is ill-posed such that a unique solution cannot be guaranteed, which makes the analysis formidable.

Fortunately, natural hand motion is also highly constrained. One type of constraints, usually referred to as static constraints, are the limits of the range of finger motions as a result of hand anatomy, such as $0° \le \theta_{MCP} \le 90°$. These constraints limit hand articulation within a boundary. Another type of constraints describes the correlations among different joints and thus reduces the dimensionality of hand articulation. For example, the motions of the DIP joint and PIP joint are generally not independent, and they could be described as $\theta_{DIP} = (2/3)\theta_{PIP}$ from the study of biomechanics [19], [20]. Although this constraint could be intentionally made invalid, it is a good approximation of natural finger motion.

Unfortunately, not all of such constraints could be quantified in closed forms. There are few studies of finger motion constraints in the literature. A preliminary investigation could be found in [21], in which learning tech-



▲ *2. Hand skeleton structure. Generally, we can assume 2 DOF for the MCP and TM joint, and 1 DOF for all the other joints. Thus, the hand roughly has 21 DOF for its local finger motion.*

Human hand motion is highly articulate, but also highly constrained, which makes it difficult to model.

niques are employed to model the hand configurations space directly by collecting a large set of hand motion data [21]. The computational complexity of finger motion analysis could be reduced significantly when considering such motion constraints.

Modeling the Dynamics

To capture complex hand motion and recognize continuous hand gestures, the dynamics and semantics of hand motion should also be modeled.

Kalman filtering and extended Kalman filtering (EKF) techniques are widely adopted to model the dynamics [32]. EKF works well for some tracking tasks. However, it is based on small motion assumption that often fails to hold in hand motion.

Simple hand gestures can be modeled by a finite state machine [10], but it is insufficient to represent complex hand dynamics. Rule-based approaches can be applied to model complex hand movements [26]. However, many heuristics are needed to construct the rules. Considering the similarities between sign languages and spoken languages, the hidden Markov model (HMM) and its variants are also used to model the hand dynamics [33], [38]. As a generalization of HMM, dynamic Bayesian net [23] is another promising approach to model the hand dynamics. These methods are essentially learning methods that learn the intrinsic dynamics from a set of training data. The knowledge of dynamics and semantics is not explicitly expressed in these methods but implicitly stored in the structures of the learning models.

The learning results of these methods depend on the training data set, structures of learning models, and training methods. One of the common problems of the learning approaches is that generalization of the learning results largely depends on the training data. However, obtaining the training samples is not a trivial problem. Currently, learning dynamics (i.e., behaviors and semantics) of human motion has drawn much attention from researchers in HCI, computer vision, computer graphics, and psychology.

Capturing Hand Motion

Hand motion capturing is finding the global and local motion of hand movements. Several different model-based approaches will be discussed in this section.

Formulating Hand Motion

Highly articulate human hand motion consists of the global hand motion and local finger motion, which can be expressed as $\mathbf{M}=[\mathbf{M}_G,\mathbf{M}_L]$, where $\mathbf{M}$ is the hand motion, $\mathbf{M}_G$ is the global motion, and $\mathbf{M}_L$ is the local motion. Global hand motion that presents large rotation and translation can be written as $\mathbf{M}_G=[\mathbf{R},\mathbf{t}]$, where $\mathbf{R}$ and $\mathbf{t}$ are rotation and translation, respectively. One important issue is how to track reliably the global motion in image sequences.

Local hand motion is articulate, and self-occlusion makes the detection and tracking local hand motion challenging. Local hand motion can be parameterized with the set of joint angles (or hand state), $\mathbf{M}_L=[\Theta]$ where Θ is the joint angle set. Consequentially, hand motion can be expressed as $\mathbf{M}=[\mathbf{R},\mathbf{t},\Theta]$.

One possible way to analyze hand motion is the appearance-based approach, which emphasizes the analysis of hand shapes in images [24]. However, local hand motion is very hard to estimate by this means. Another possible way is the model-based approach [11], [13], [19], [20], [29], [32], [37], [40]. With a single calibrated camera, local hand motion parameters can be estimated by fitting the 3-D model to the observation images. Multiple camera settings are helpful to deal with occlusion [20], [29], [37]. The use of a 3-D model can largely alleviate the problem of depth ambiguity since the structure of the hand is included in the model.

Hand Localization

Hand localization is locating hand regions in image sequences. Skin color offers an effective and efficient way to fulfill this goal. The core of color tracking is color-based segmentation. According to the representation of color distribution in certain color spaces, current techniques of color tracking can be classified into two general approaches: nonparametric [16], [18], [43] and parametric [28], [39]. Figure 3 gives an example of segmentation-based hand localization, in which the input image is segmented by color, and hand blob is localized by grouping skin-color pixels.

One of the nonparametric approaches is based on color histograms [16], [18]. Because color space is quantized by the structure of the histogram, this technique shares the same problem with nonparametric density estimation, in which the level of quantization will affect the estimation. How to select a good quantization level of the color histogram is not trivial. Although nonuniform quantization would perform better than uniform quantization, it is much more complicated. Another nonparametric approach is proposed in [43] based on the self-organizing map, an unsupervised clustering algorithm to approximate color distribution. Generally, these nonparametric approaches work effectively when the quantization level is set properly and there are sufficient data.

Parametric approaches model the color density in parametric forms such as Gaussian distribution or Gaussi-

an mixture models [28], [39]. Expectation-maximization (EM) offers a way to fit probabilistic models to the observation data. The difficulty of model order selection could be handled by heuristics [28] or cross validation.

To lead to a robust and efficient localization, besides the color cue, hand shape and motion could also be employed for localization. One important research problem is the integration or fusion of multiple cues [2], [33].

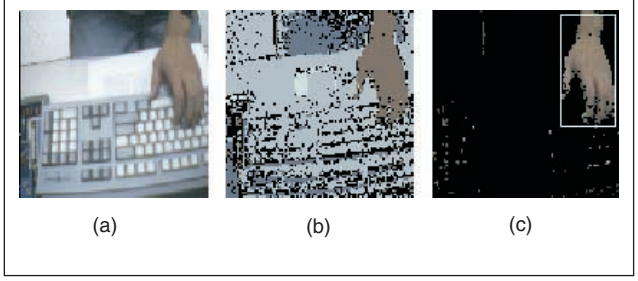
Selecting Image Features

To estimate the parameters of the model, some image features should be extracted and tracked to serve as the observation of the estimators. Hand image features can be geometric features such as points, lines, contours, and silhouettes [19]. Fingertip is one of the frequently used features, because the positions of fingertips are almost sufficient to recognize some gestures due to the highly constrained hand motion [20]. Color markers are often used to help track the 3-D positions of fingertips [20], [11]. Some researchers estimate the positions and orientations of fingertips by fitting a 3-D cylinder to the images [11]. Line fitting is also a frequently used technique to detect the fingertips [29].

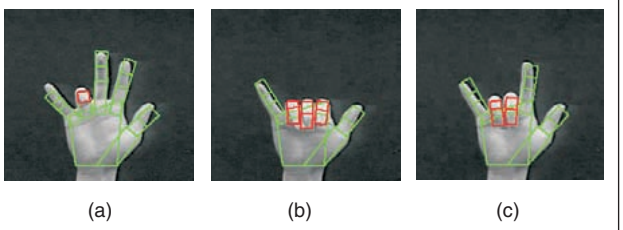
Capturing Hand Motion in Full DOF

To capture articulate hand motion in full DOF, both global hand motion and local finger motion should be determined from video sequences. It is a challenging problem to analyze and capture hand motion, because the hand is highly articulate. Different methods have been taken to approach this problem. One possible method is the appearance-based approaches, in which 2-D deformable hand shape templates are used to track a moving hand in 2-D. However, this method is insufficient to recover full articulations, because it is difficult to infer finger joint angles based on appearances only.

Another possible way is the 3-D model-based approach, which takes the advantages of *a priori* knowledge built in the 3-D models. This approach aligns a 3-D model to images or even range data by estimating the parameters of the model. In 3-D model-based methods, image features could be looked as the image evidence or image observation of a 3-D model that is projected to the image plane. A 3-D model with different parameters will produce different image evidence. Model-based methods recover the joint angles by minimizing the discrepancy between the image feature observations and projected 3-D model hypotheses [11], [13], [19], [20], [29], [32], [40], which is a challenging optimization problem. Two important tasks in the model-based approach are determining the match and searching the hand joint angles space. Some examples are shown in Fig. 4, in which the parameters of a cardboard hand model are adjusted to match three input images. Generally, due to the huge search space of hand articulation, the optimization involved is difficult and computationally intensive.



▲ *3. Hand localization. (a) Input image, (b) segmentation result, and (c) hand blob located by analyzing the segmented image pixels.*



▲ *4. Capturing articulate hand motion using a cardboard hand model. Hand pose and finger joint angles could be recovered by fitting the model to the images. The fitting minimizes the discrepancy between image feature observations and projected models.*

Many methods tend to estimate the global and local hand motion simultaneously. In [29], the hand was modeled as an articulate stick figure, and point and line image features were used for the registration. Hand motion capturing was formulated as a constrained nonlinear programming problem. The drawback of this approach is that the optimization is often trapped in local minima. Another idea is to model the surface of the hand [11], [19], [32], and then hand configurations can be estimated using the analysis-by-synthesis approach, in which candidate 3-D models are projected to the image plane and the best match is found with respect to some similarity measurements. If the surface model is very fine, an accurate estimation can be obtained. However, those hand models are user dependent. Rough models can only give approximate estimations [32].

To ease the optimization, a decomposition method can be adopted to analyze articulate hand motion by decoupling hand motion to its global motion and local finger motion. Global motion is parameterized as the pose of the palm, and local motion is parameterized as the set of joint angles. A two-step iterative algorithm could be used to find an accurate estimation [40]. Given an initial estimation, hand pose is estimated using least median of squares with joint angles fixed. Then the joint angles are recovered by a genetic algorithm with the global hand pose fixed. Those two steps are alternately iterated until the solution converges [40].

Gesture Recognition

Meaningful gestures could be represented by both temporal hand movements and static hand postures. Hand postures express certain concepts through hand configurations, while temporal hand gestures represent certain actions by hand movements. Sometimes, hand postures act as special transition states in temporal gestures and supply a cue to segment and recognize temporal hand gestures. Although hand gestures are complicated to model because the meanings of hand gestures depend on people and cultures, a set of specific hand gesture vocabulary can always be predefined in many applications, such as VE applications, so that the ambiguity can be limited.

Hand Posture Recognition

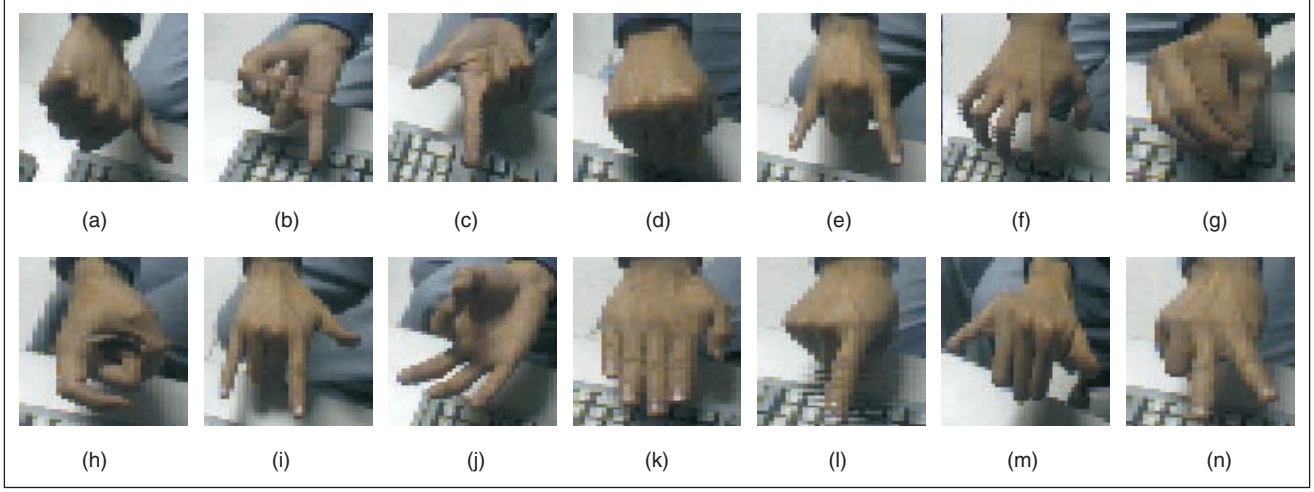
Different from sign languages, the gesture vocabulary in VE applications is structured and disambiguated. Some simple controlling, commanding, and manipulative gestures are defined to fulfill natural interaction such as pointing, navigating, moving, rotating, stopping, starting, and selecting. These gesture commands can be simple in the sense of motion; however, many different hand postures are used to differentiate and switch among the commanding modes. For example, only if we know a gesture is a pointing gesture would it make sense to estimate its pointing direction. View-independent hand posture recognition is a natural requirement in many VE applications. In most cases, because users do not know where the cameras are, the naturalness and immersiveness will be ruined if users are obliged to issue commands to an unknown direction.

One approach is the 3-D model-based approach, in which the hand configuration is estimated by taking advantage of 3-D hand models [11], [13], [19], [20], [29], [32], [40]. Because hand configurations are independent of view directions, these methods could directly achieve view-independent recognition. Different models use different image features to construct feature-model correspondences. Joint angles can be estimated by minimizing a projected surface model and some image evidences such as silhouettes in the light of analysis-by-synthesis [11], [19], [20]. However, this approach needs good surface models and the process of projection and comparison is expensive. Alternatively, point and line features are employed in kinematical hand models to recover joint angles [29], [32], [40]. Hand postures could be estimated accurately if the correspondences between the 3-D model and the observed image features are well established. Physical models and statistical models [13] were also employed to estimate hand configurations. However, the ill-posed problem of estimating hand configuration is not trivial. Many current methods require reliable feature detection, which is plagued by self-occlusion. Another drawback is that it is not trivial to achieve user independence, because 3-D models should be calibrated for each user.

Because the estimation of hand joint angles is difficult, an alternative approach is the appearance-based approach [7], [31], [35], [42], which aims to characterize the mapping from the image feature space to the possible hand configuration space directly from a set of training data. This approach often involves learning techniques. Because image data are generally high dimensional, and manually labeling a large data set will be very time consuming and tedious, there are two major difficulties for this approach: automatic feature selection and training data collection. The research of the first problem has been investigated widely, and there have been many discussions about feature extraction [35] and selection [7]. However, little has been addressed on how to collect the training data automatically. In [42], a hybrid learning approach was proposed to employ a large set of unlabeled images in training. Images for different hand postures are shown in Fig. 5.

Temporal Gesture Recognition

Some temporal gestures are specific or simple and could be captured by low-level dynamic models. However, many high-level activities have to be represented by more complex gesture semantics, so modeling the low-level dy-



▲ *5. Recognizing different hand postures.*

namics is insufficient. The HMM technique and its variations are often employed in modeling, learning, and recognition of temporal signals. Because many temporal gestures involve motion trajectories and hand postures, they are more complex than speech signals. Finding a suitable approach to model hand gestures is still an open research problem. Practical large-vocabulary gesture recognition systems by HMM are yet to be developed. A similar problem is the recognition of human motion.

Recognizing Low-Level Motion

Modeling the low-level dynamics of human motion is important not only for human tracking but also for human motion recognition. It serves as a quantitative representation of simple movements so that those simple movements can be recognized in a reduced space by the trajectories of motion parameters.

Some low-level motions can be represented by simple dynamic processes, in which a Kalman filter is often employed to estimate, interpolate, and predict the motion parameters. As the extension of the Kalman filtering technique in the case of non-Guassian noises, the CONDENSATION algorithm could also be used to recognize temporal trajectories [30]. However, those low-level dynamics models are not sufficient to represent more complicated human motions.

Some human activities could be represented as a complex, multistate model in [25], in which several alternative models were employed to represent human dynamics, one for each class of response. Model switching is based on the observation of the state of the dynamics. This approach produces a generalized maximum likelihood estimate of the current and future values of the state variables. Recognition is achieved by determining which model best fits the observation.

Recognizing High-Level Motion

Many applications need to recognize more complex gestures that include semantic meaning in the movements. Modeling the low-level dynamics alone is not sufficient in such tasks.

An approach to this problem is rule-based modeling [26], [8], in which the high-level motion can be explicitly represented by a set of rules, and the recognition is achieved by rule-based induction. One of the difficulties lies in the fact that constructing the rule system needs quite a lot of heuristics. Such rules could also be represented using the finite state machine technique [11], [15]. Temporal events are represented by a state transition diagram, in which each state indicates possible gesture states at the next moment. By using a rest state, all unintentional actions can be ignored. Although such methods are simple, they lack the ability to model the large variation in the temporal gestures, since the same gesture may have different temporal characteristics.

One possible way to analyze hand motion is the appearance-based approach, which emphasizes the analysis of hand shapes in images.

To model the large variation in the gestures, the HMM technique seems a promising approach. A more detailed review will be given in the next section. As a generalization of HMM, another promising approach to modeling the semantics of temporal gestures is the dynamic Bayesian network [23], which provides a more flexible structure of a graphical model to represent temporal signals.

Gesture Recognition by HMM

The HMM is a type of statistical model widely used in speech recognition. Due to the similarity between speech recognition and temporal gesture recognition, HMM is also employed to recognize human motion in recent years, and is used to model the state transition among a set of dynamic models [4], [25]. HMM has the capacity for modeling not only the low-level dynamics but also some high-level motion [34].

There are also many variations of HMM. In [44], gestures were modeled by a multidimensional HMM, which contains more than one observation symbol at each time. This approach is able to model multipath gestures and provide a means to integrate multiple modalities to increase the recognition rate. Because the output probability of feature vectors of each state in HMM is unique, HMM can handle only piecewise stationary processes that are not adequate in gesture modeling. The partially observable Markov decision process [9] was introduced for temporal matching. Standard HMM was extended to include a global parametric variation in the output probabilities of the HMM to handle parameterized movements such as musical conducting and driving by EM algorithm [38].

When the Markov condition is violated, conventional HMMs fails. HMMs are ill suited to systems that have compositional states. The coupled HMM technique [3] was presented for coupling and training HMMs to model interactions between processes that may have different state structures and degrees of influence on each other. Coupled HMMs are well suited for applications requiring sensor fusion across modalities.

Future Research Directions

Although much progress has been made in recent years, there are still many issues related to gesture analysis and recognition that need to be adequately addressed in the future.

Robust Hand Localization

Although the idea of localizing the hand by tracking skin color is straightforward, in practice, there are some challenging problems of color tracking. Many color tracking techniques assume controlled lighting. However, due to the dynamic scenes and changing lighting conditions, the color distribution over time is nonstationary. If a color classifier is trained under a specific condition, it may not work well in other scenarios. Besides the large variation in skin colors, in some VE applications, because the graphics rendered in the display keep changing, the reflected lights would probably change the skin color as well. This color consistency problem is not trivial in tracking skin color.

Recently, some researchers have begun to look into the nonstationary color distribution problem in color tracking. Several color model updating methods have been proposed to solve this problem [28], [41], [43]. However, handling the nonstationary color is still an open research problem. In the meantime, to achieve a robust hand localization system, multiple cues should be integrated. Better approaches for integration should be studied in the future research.

Modeling the Constraints

Although the hand is highly articulate, the natural finger motion is also highly constrained. These constraints largely reduce the possible hand configuration space. Consequently, the search space would be significantly reduced in hand posture estimation, and the articulate motion capturing would be more efficient. Unfortunately, most of such constraints are impossible to be represented explicitly, partly due to the large variation in finger motion.

However, to achieve robust and efficient estimation of hand configuration and realistic hand animation, such constraints have to be modeled. Instead of explicit modeling, learning techniques could be taken to characterize the hand configuration space. A more profound investigation should be conducted.

Motion Editing for Animation

Realistic articulate hand animation should be considered in the future. The human animation produced by many current animation systems looks very unrealistic and still looks like robots, due to the fact that the current motion model is too simplified and largely dependent on kinematics. To achieve realistic animation, the natural motion constraints should be integrated with animation systems. Another research direction is to achieve personalized animation, in which different styles of motion can be produced with low costs. Schemes of avoiding the violation of body constraints and collision detection should be built into animation systems.

Recognizing Temporal Patterns

Although HMM is used widely in speech recognition, and many researchers are applying HMM to temporal gesture recognition, current examples of gesture recognition by HMM are still with very limited vocabularies. Compared to HMM in speech recognition, data collection for HMM training in temporal gesture recognition is very difficult, which is part of the reason that large vocabularies are prevented. A crucial issue of training data collection is motion capturing. Due to its lower cost and noninvasive nature, the vision-based motion capturing would be one of the ideal approaches to collect motion training data. However, there are many challenging and unsolved problems in vision-based motion capturing techniques. Another issue is gesture co-articulation, which makes the extraction and segmentation of gesture commands even harder in continuous hand movements.

In a word, a good representation of temporal gestures needs to be found in future research. It could be a very different representation from speech signals. Motion interpretation is a quite ill-posed problem, in which cognitive science and psychological studies may be combined. In the near future, it may be very possible to develop task-specific gesture systems, but we are still far from a general purpose temporal gesture recognition and understanding system.

Other Open Questions

To achieve an immersive interaction, multimodality by integrating hand gesture and speech should also be addressed adequately in the future. Recent research shows that there is a complementary among different modalities such as hand gesture and speech [27]. More profound research should be conducted.

Current research focuses on single hand gestures for simplicity. However, two-handed gestures should also be studied in the future, since they are more expressive and allow more natural interaction.

Conclusions

In this article, we reported the past development on the research of human hand modeling, analysis, and recognition in the context of HCI. Several aspects of the hand can be modeled such as shape, kinematical structure, and dynamics. Different hand models are used in different applications. Three-dimensional hand models offer a rich description to fully capture hand motion. Static hand posture recognition and temporal gesture recognition are the two main parts of gesture recognition. However, we are far from building a general-purpose gesture recognition system.

Overall, at the current state of the art, vision-based gesture tracking and recognition are still in their infancy. In order to develop a natural and reliable hand gesture inter-

face, substantial research efforts in computer vision, graphics, machine learning, and psychology should be made.

Acknowledgment

This work was supported in part by National Science Foundation Grants CDA-96-24396 and IRI-96-34618 and NSF Alliance Program.

Ying Wu received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 1994 and the M.S. degree from Tsinghua University, Beijing, China, in 1997. Currently, he is pursuing his Ph.D. degree in electrical and computer engineering in the University of Illinois at Urbana-Champaign (UIUC). Since 1997, he has been a Graduate Research Assistant at the Image Formation and Processing Group of the Beckman Institute for Advanced Science and Technology at UIUC. During summer 1999 and 2000, he was a Research Intern with the Vision Technology Group, Microsoft Research, Redmond, WA. His current research interests include computer vision, computer graphics, machine learning, human-computer intelligent interaction, image/video processing, multimedia, and virtual environments. He received the Robert T. Chien Award at the University of Illinois at Urbana-Champaign in 2001.

Thomas S. Huang received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, China, and the M.S. and Sc.D. degrees in electrical engineering from Massachusetts Institute of Technology (MIT), Cambridge, MA. He was on the Faculty of 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, he has worked at the MIT Lincoln Laboratory, the IBM Thomas J. Watson Research Center, and Rheinishes Landes Museum in Bonn, West Germany, and he held Visiting Professor positions at the Swiss Institutes of Technology in Zurich and Lausanne; the University of Hannover, Germany; INRS-Telecommunications of the University of Quebec, Montreal, Canada; and the University of Tokyo, Japan. He has served as a consultant to numerous industrial firms and government agencies both in the United States and abroad. His professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. He has published 12 books and over 400 papers in network theory, digital filtering, image processing, and computer vision. He is a Founding Editor of the *International Journal of Computer Vision*, *Graphics and Image Processing* and Editor of the "Springer Series in Information Sciences," published by Springer Verlag.

Dr. Huang is a Fellow of the International Association of Pattern Recognition, and the Optical Society of American 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 received the IEEE Signal Processing Society's Technical Achievement Award in 1987 and the Society Award in 1991. He was awarded the IEEE Third Millennium Medal in 2000. He received the Honda Lifetime Achievement Award for "contributions to motion analysis" in 2000. He received the IEEE Jack S. Kilby Medal in 2001. Dr. Huang is a member of the National Academy of Engineering.

References

[1] J. Aggarwal and Q. Cai, "Human motion analysis: A review," in *Proc. IEEE Nonrigid and Articulated Motion Workshop*, 1997, pp. 90-102.

[2] Y. Azoz, L. Devi, and R. Sharma, "Reliable tracking of human arm dynamics by multiple cue integration and constraint fusion," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Santa Barbara, CA, 1998, pp. 905-910.

[3] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 994-999.

[4] C. Bregler, "Learning and recognition human dynamics in video sequences," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 568-574.

[5] J. Cassell, *A Framework for Gesture Generation and Interpretation*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[6] J. Crowley, F. Berard, and J. Coutaz, "Finger tracking as an input device for augmented reality," in *Proc. Int. Workshop Automatic Face and Gesture Recognition,* Zurich, Switzerland, 1995, pp. 195-200.

[7] Y. Cui and J. Weng, "Hand sign recognition from intensity image sequences with complex background," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1996, pp. 88-93.

[8] R. Cutler and M. Turk, "View-based interpretation of real-time optical flow for gesture recognition," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, Japan, 1998, pp. 416-421.

[9] T. Darrell and A. Pentland, "Active gesture recognition using partially observable Markov decision processes," in *Proc. IEEE Int. Conf. Pattern Recognition*, 1996, vol. 3, pp. 984-988.

[10] J. Davis and A. Bobick, "The representation and recognition of action using temporal templates," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 928-934.

[11] J. Davis and M. Shah, "Visual gesture recognition," *Vision, Image, and Signal Processing*, vol. 141, pp. 101-106, Apr. 1994.

[12] D.M. Gavrila, "The visual analysis of human movement: A survey," *Computer Vision and Image Understanding*, vol. 73, pp. 82-98, Jan. 1999.

[13] T. Heap and D. Hogg, "Towards 3D hand tracking using a deformable model," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, Killington, VT, 1996, pp. 140-145.

[14] K. Imagawa, S. Lu, and S. Igi, "Color-based hands tracking system for sign language recognition," in *Proc. Int. Conf. Face and Gesture Recognition*, 1998, pp. 462-467.

[15] K. Jo, Y. Kuno, and Y. Shirai, "Manipulative hand gestures recognition using task knowledge for human computer interaction," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, Japan, 1998, pp. 468-473.

[16] M. Jones and J. Rehg, "Statistical color models with application to skin detection," Compaq Cambridge Res. Lab. Tech. Rep. CRL 98/11, 1998.

[17] S. Ju, M. Black, S. Minneman, and D. Kimber, "Analysis of gesture and action in technical talks for video indexing," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 595-601.

[18] R. Kjeldsen and J. Kender, "Finding skin in color images," in *Proc. 2nd Int. Conf. Automatic Face and Gesture Recognition*, 1996, pp. 312-317.

[19] J.J. Kuch and T.S. Huang, "Vision-based hand modeling and tracking for virtual teleconferencing and telecollaboration," in *Proc. IEEE Int. Conf. Computer Vision*, Cambridge, MA, June 1995, pp. 666-671.

[20] J. Lee and T. Kunii, "Model-based analysis of hand posture," *IEEE Comput. Graph. Appl*., vol. 15, no. 5, pp. 77-86, Sept. 1995.

[21] J. Lin, Y. Wu, and T.S. Huang, "Modeling human hand constraints," in *Proc. Workshop on Human Motion*, Dec. 2000, pp. 121-126.

[22] D. McNeill, *Hand and Mind*. Chicago, IL: Univ. Chicago Press, 1992.

[23] V. Pavlovic, "Dynamic Bayesian networks for information fusion with applications to human–computer interfaces," Ph.D. dissertation, Univ. Illinois at Urbana-Champaign, 1999.

[24] V. Pavlovic, R. Sharma, and T.S. Huang, "Visual interpretation of hand gestures for human computer interaction: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 677-695, July 1997.

[25] A. Pentland and A. Liu, "Modeling and prediction of human behavior," *IEEE Intell. Veh.*, pp. 350-355, 1995.

[26] F. Quek, "Unencumbered gesture interaction," *IEEE Multimedia*, vol. 3. no. 3, pp. 36-47, 1996.

[27] F. Quek *et al*., "Gesture, speech, and gaze cues for discourse segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. II, 2000, pp. 247-254.

[28] Y. Raja, S. McKenna, and S. Gong, "Colour model selection and adaptation in dynamic scenes," in *Proc. European Conf. Computer Vision*, 1998, pp. 460-475.

[29] J. Rehg and T. Kanade, "Model-based tracking of self-occluding articulated objects," in *Proc. IEEE Int. Conf. Computer Vision*, 1995, pp. 612-617.

[30] J. Rittscher and A. Blake, "Classification of human body motion," in *IEEE Int. Conf. Computer Vision*, Corfu, Greece, 1999, vol. I, pp. 634-639.

[31] R. Rosales and S. Sclaroff, "Inferring body pose without tracking body parts," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2000, vol. II, pp. 721-727.

[32] N. Shimada et al., "Hand gesture estimation and model refinement using monocular camera—Ambiguity limitation by inequality constraints," in *Proc. 3rd Conf. Face and Gesture Recognition*, 1998, pp. 268-273.

[33] T. Starner et al., "A wearable computer based American sign language recognizer," in *Proc. IEEE Int. Symp. Wearable Computing*, Oct. 1997, pp. 130-137.

[34] P. Stoll and J. Ohya, "Application of HMM modeling to recognizing human gestures in image sequences for a man–machine interface," in *Proc. IEEE Int. Workshop on Robot and Human Communication*, 1995, pp. 129-134.

[35] J. Triesch and C. von de Malsburg, "Robust classification of hand postures against complex background," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, 1996, pp. 170-175.

[36] L. Tsap et al., "Human skin and hand motion analysis from range image sequences using nonlinear FEM," in *Proc. IEEE Nonrigid and Articulated Motion Workshop*, 1997, pp. 80-88.

[37] C. Vogler and D. Metaxas, "ASL recognition based on a coupling between HMMs and 3D motion analysis," in *Proc. IEEE Int. Conf. Computer Vision*, Mumbai, India, Jan. 1998, pp. 363-369.

[38] A. Wilson and A. Bobick, "Recognition and interpretation of parametric gesture," in *Proc. IEEE Int. Conf. Computer Vision*, 1998, pp. 329-336.

[39] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, no. 7, pp. 780-785, July 1997.

[40] Y. Wu and T.S. Huang, "Capturing articulated human hand motion: A divide-and-conquer approach," in *Proc. IEEE Int. Conf. Computer Vision*, Corfu, Greece, Sept. 1999, pp. 606-611.

[41] Y. Wu and T.S. Huang, "Color tracking by transductive learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. I, Hilton Head Island, SC, June 2000, pp. 133-138.

[42] Y. Wu and T.S. Huang, "View-independent recognition of hand postures," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. II, Hilton Head Island, SC, June 2000, pp. 88-94.

[43] Y. Wu, Q. Liu, and T.S. Huang, "An adaptive self-organizing color segmentation algorithm with application to robust real-time human hand localization," in *Proc. Asian Conf. Computer Vision*, Taipei, Taiwan, Jan. 2000, pp. 1106-1111.

[44] J. Yang, Y. Xu, and C. Chen, "Gesture interface: Modeling and learning," in *Proc. IEEE Int. Conf. Robotics and Automation*, vol. 2, 1994, pp. 1747-1752.

[45] M. Zeller et al., "A visual computing environment for very large scale biomolecular modeling," in *Proc. IEEE Int. Conf. Application-Specific Systems, Architectures and Processors*, Zurich, Switzerland, 1997, pp. 3-12.