A PURPOSE TAXONOMY OF GESTURAL INTERACTION

FRANCIS QUEK
VISION INTERFACES AND SYSTEMS LABORATORY
CENTER FOR HUMAN COMPUTER INTERACTION
VIRGINIA TECH
(Written: October 22, 2003)

INTRODUCTION

The use of gesture, particularly hand gesture, as a means of communicating with computers and machines is attractive for several reasons. First, many researchers observe that humans possess great facility in performing gestures and appear to do so spontaneously. Second, the hands and arms always 'comes attached' to the human end of the human-computer interaction exchange. There is no need to hunt for the missing remote control or to equip the human with a communicative device if the computer could observe the user and react accordingly. Third, as the space in which we interact extends from one screen to many, from small screen to large, and from the confines of the two-dimensional panel surface into the three-dimensional space beyond it, gestures present the promise of natural interaction that is able to match both the added expanse and dimensionality.

SENSING MODALITY

One may think of the human-end of the interactive chain as being able to produce three key interactive signals: things that can be heard, things that can be seen, and things that can be felt (ignoring taste and smell as currently far-fetched for HCI application). In this sense, the computer's input devices can be thought of as the sensory organs to detect the signals sent by its human partner. Under this formulation, speech interfaces require auditory computer input, and the plethora of input devices by which the user moves a mouse, joystick or depresses keys would constitute the computer's tactile sense. The sensory receptor for gesture is vision. One might relax this vision requirement to allow the use of various glove and magnetic, acoustic or marker-based tracking technologies. For this discussion we shall include these approaches with the caveat that these are intended as waypoints toward the goal of vision-based gesture understanding.

AN ORGANIZING TAXONOMY

To move beyond promise to practice, one needs to understand what the space of gestures is and what it can afford in interaction. We shall organize our discussion around a purpose taxonomy. Interactive 'gesture' systems may be divided into three classes: 1. Manipulative; 2. Semaphoric; and, 3. Conversational. The observation is that the human hands and arms are the ultimate 'multi-purpose tools'. We use them to modify objects around us (moving, shaping, hitting etc.), to signal one another, and in the general service of language. While the psychology and psycholinguistics of gesture is a very involved field (see, for example, (McNeill 1992)) our tri-partite segmentation adequately covers the use of gesture in human computer interaction. These distinctions are not perfunctory – they have great significance for both the vision-based processing strategy employed as well as the design of the interactive system that utilizes the gesture.
MANIPULATIVE GESTURE SYSTEMS

Such systems follow the tradition of Richard Bolt’s “Put-That-There” system (Bolt 1980; Bolt 1982) which permits the direct manipulation of entities in a system. In this work, the user interacted with a large wall-sized display moving objects around the screen. The hand was tracked using an electromagnetic tracker. As will be seen later, this work may also be classified as conversational since co-temporal speech is utilized for in the object manipulation.

We extend the concept to cover all systems of direct control such as ‘finger flying’ to navigate virtual spaces, control of appliances and games, and robot control in this category. The essential characteristic of manipulative systems is the tight feedback between the gesture and the entity being controlled.

Since Bolt’s seminal work, there has been a plethora of systems that implement finger tracking/pointing (Fukumoto, Mase et al. 1992; Kahn, Swain et al. 1994; Crowley, Berard et al. 1995; Quek, Mysliwiec et al. 1995; Kjeldsen and Kender 1996; Pavlovic, Sharma et al. 1996), a variety of ‘finger flying’ style navigation in virtual spaces or direct-manipulation interfaces (Hauptmann 1989; Azarbayejani, Starner et al. 1993; Baudel and Beaudouin-Lafon 1993; Rehg and Kanade 1993; Segen 1993; Wellner 1993; Wexelblat 1994; Kadobayashi, Nishimoto et al. 1998; Segen and Kumar 1998; Kumar and Segen 1999; O’Hagan and Zelinsky 2000), control of appliances (Freeman and Weissman 1995), in computer games (Freeman, Tanaka et al. 1996; Freeman, Anderson et al. 1998; Freeman 1999), and robot control (Friedman 1986; Katkere, Hunter et al. 1994; Triesch and von der Malsburg 1997; Triesch and von der Malsburg 1998). Other manipulative applications include interaction with wind-tunnel simulations (Bryson and Levit 1991; Bryson and Levit 1992), voice synthesizers (Pausch and Williams 1992; Fels and Hinton 1993; Fels 1994), and an optical flow-based system that detects one of six gross full-body gestures (jumping waving, clapping, drumming, flapping, marching) for controlling a musical instrument (Cutler and Turk 1998). Some of these approaches cited (e.g. (Friedman 1986; Fels and Hinton 1993; Fels 1994; Wilson and Bobick 1998; Wilson and Bobick 1999)) use special gloves or trackers, while others employ only camera-based visual tracking. Such manipulative gesture systems typically use the shape of the hand to determine the mode of action (e.g. to navigate, pick something up, point etc.) while the hand motion indicates the path or extent of the controlled motion.

When used in a manipulative fashion, gesture interfaces have a lot in common with other direct manipulation interfaces, the only distinction being the ‘device’ that is used for the interaction. As such many of the same design principles one might apply in building manipulative gesture interfaces. These include ensuring rapid enough visual feedback for the control, the size of and distance to targets of manipulation (see Fitt’s Law), and the considerations for fatigue and repetitive stress order (as when one has to maintain hand positions, poses and attitudes by maintaining muscle tension).

Gestures used in communication/conversation differ from manipulative gestures in several significant ways (Quek 1995; Quek 1996). First, because the intent of the latter is for manipulation, there is no guarantee that the salient features of the hands are visible. Second, the dynamics of hand movement in manipulative gestures differ significantly from conversational gestures. Third, manipulative gestures may typically be aided by visual, tactile or force feedback from the object (virtual or real) being manipulated, while conversational gestures are typically performed without such constraints. Gesture and manipulation are clearly different entities sharing between them possibly only the feature that both may utilize the same body parts.
Semaphoric Gesture Systems

Semaphores are signaling systems in which each body poses and movements are precisely defined to designate specific symbols within some alphabet. Traditionally semaphores may involve the use of the human body and limbs, light flashes, flags etc. Although semaphore use inhabits a miniscule portion of the space of human gestures, it has attracted a large portion for vision-based gesture research and systems. Semaphore gesture systems predefine some universe of ‘whole’ gestures \( g \in G \). Taking a categorial approach, ‘gesture recognition’ boils down to determining if some presentation \( p \) is a manifestation of some \( g \). Such semaphores may be either static gesture poses or predefined stylized movements. Note that such systems are patently not sign-language recognition systems in that only isolated symbols are entertained. Sign languages include syntax, grammar and all the dynamics of spoken language systems. Some attempts have been made to recognize isolated sign-language symbols (e.g. finger spelling), but the distance between this and sign language understanding is as far as that between optical character recognition of individual characters and natural language understanding.

Semaphoric approaches may be termed as ‘communicative’ in that gestures serve as a universe of symbols to be communicated to the machine. A pragmatic distinction between semaphoric gestures and manipulative ones is that the former does not require the feedback control (e.g. hand-eye, force-feedback, or haptic) necessitated for manipulation. Semaphoric gestures may be further categorized as being static or dynamic. Static semaphore gesture systems interpret the pose of a static hand to communicate the intended symbol. Examples of such systems include color-based recognition of the stretched-open palm where flexing specific fingers indicate menu selection (Zhu, Yang et al. 2000), Zernike moments-based hand pose estimation (Hunter, Schlenzig et al. 1995), the application of orientation histograms (histograms of directional edges) for hand shape recognition (Freeman and Roth 1995; Triesch and von der Malsburg 1996), graph-labeling approaches where labeled edge segments are matched against a predefined hand graph (Triesch and von der Malsburg 1996) (they show recognition of American Sign Language, ASL-like, finger spelling poses), a ‘flexible-modeling’ system in which the feature-average of a set of hand poses is computed and each individual hand pose is recognized as a deviation from this mean (principal component analysis, PCA, of the feature covariance matrix is used to determine the main modes of deviation from the ‘average hand pose’) (Lanitis, Taylor et al. 1995), the application of ‘global’ features of the extracted hand (using color processing) such as moments, aspect ratio, etc. to determine the shape of the hand out of 6 predefined hand shapes (Maggioni 1995), model-based recognition using 3D model prediction (Munk and Granum 1997), and neural net approaches (Boehme, Braumann et al. 1998).

In dynamic semaphore gesture systems, some or all of the symbols represented in the semaphore library involve predefined motion of the hands or arms. Such systems typically require that gestures be performed from a predefined viewpoint to determine which \( g \in G \) is being performed. Approaches include finite state machines for recognition of a set of editing gestures for an ‘augmented whiteboard’ (Black and Jepson 1998), trajectory-based recognition of gestures for ‘spatial structuring’ (Quek 1993; Quek 1994; Quek 1995; Wilson and Bobick 1995; Quek 1996; Quek and Zhao 1996; Zhao, Quek et al. 1998), recognition of gestures as a sequence of state measurements (Wilson and Bobick 1995), recognition of oscillatory gestures for robot control (Cohen, Conway et al. 1996), and ‘space-time’ gestures that treat time as a physical third dimension (Darrell and Pentland 1993; Darrell, Essa et al. 1995).

One of the most common approaches for the recognition of dynamic semaphoric gestures is based on the Hidden Markov Model (HMM) (Rabiner 1989). First applied by Yamato, Ohya, and Ishii (Yamato, Ohya et al. 1992) to the recognition of tennis strokes, it has been applied in a myriad
of semaphoric gesture recognition systems. The power of the HMM lies in its statistical rigor and ability to learn semaphore vocabularies from examples. A HMM may be applied in any situation in which one has a stream of input observations formulated as a sequence of feature vectors and a finite set of known classifications for the observed sequences. HMM models comprise state sequences. The transitions between states are probabilistically determined by the observation sequence. HMMs are ‘hidden’ in that one does not know which state the system is in at any time. Recognition is achieved by determining the likelihood that any particular HMM model may account for the sequence of input observations. Typically, HMM models for different gestures within a semaphoric library are rank-ordered by likelihood, and the one with the greatest likelihood is selected. Good technical discussions on the application of the HMM to semaphoric gesture recognition (and isolated sign language symbol recognition) are given in (Assan and Grobel 1997; Hofmann, Heyer et al. 1997).

A parametric extension to the standard HMM (a PHMM) to recognize degrees (or parameters) of motion is described in (Wilson and Bobick 1998; Wilson and Bobick 1999). For example, the authors describe a ‘fish size’ gesture with inward opposing open palms that indicate the size of the fish. Their system encodes the degree of motion in which the output densities are a function of the gesture parameter in question (e.g. separation of the hands in the ‘fish size’ gesture). Schlenzig, Hunter and Jain apply a recursive recognition scheme based on HMMs and utilize a set of rotationally invariant Zernike moments in the hand shape description vector (Schlenzig, Hunter et al. 1994; Schlenzig, Hunter et al. 1994) Their system recognized a vocabulary of 6 semaphoric gestures for communication with a robot gopher. Their work was unique in that they used a single HMM in conjunction with a finite state estimator for sequence recognition. The hand shape in each state was recognized by a neural net. The authors of (Rigoll, Kosmala et al. 1997) describe a system using HMMs to recognize a set of 24 dynamic gestures employing a HMM to model each gesture. The recognition rate (92.9%) is high, but it was obtained for ISOLATED gestures, i.e. gesture sequences were segmented by hand. The problem, however, is in filtering out the gestures that do not belong to the gesture vocabulary (folding arms, scratching head). The authors trained several "garbage" HMM models to recognize and filter out such gestures, but the experiments performed were limited to the gesture vocabulary and only a few transitional garbage gestures. Assan and Grobel (Assan and Grobel 1997) describe a HMM system for video-based sign language recognition. The system recognizes 262 different gestures from the Sign Language of the Netherlands. The authors present both results for recognition of isolated signs and for reduced vocabulary of connected signs. Colored gloves are used to aid in recognition of hands and specific fingers. The colored regions are extracted for each frame to obtain hand positions and shapes, which form the feature vector. For connected signs, the authors use additional HMMs to model the transitions between signs. The experiments were done in a controlled environment and only a small set of connected signs was recognized with 73% of recognition versus 94% for isolated signs.

Other HMM-based systems include the recognition of a set of 6 'musical instrument' symbols (e.g. playing the guitar) (Iwai, Shimizu et al. 1999), recognition of 10 gestures for presentation control (Lee and Kim Sept. 1999), music conducting (Wilson and Bobick 1995; Bobick and Ivanov 1998), recognition of unistroke-like finger spelling performed in the air (Pavlovic, Sharma et al. 1996; Martin and Durand 2000), and communication with a molecular biology workstation (Pavlovic, Sharma et al. 1996).

There is a class of systems that applies a combination of semaphoric and manipulative gestures within a single system. This class is typified by (Pavlovic, Sharma et al. 1996) that combines HMM-based gesture semaphores (move forward, backward), static hand poses (grasp, release, drop etc.) and pointing gestures (finger tip tracking using 2 orthogonally oriented cameras - top and side). System is used to manipulate graphical DNA models.
Semaphores represent a miniscule portion of the use of the hands in natural human communication. In reviewing the challenges to automatic gesture recognition, Wexelblat (Wexelblat 1997) emphasizes the need for development of systems able to recognize natural, non-posed and non-discrete gestures. Wexelblat disqualifies systems recognizing artificial, posed and discrete gestures as unnecessary and superficial. He asks rhetorically what such systems provide that a simple system with key presses for each categorical selection cannot.

CONVERSATIONAL GESTURES

Conversational gestures are those gestures performed naturally in the course of human multimodal communication. This has been variously termed ‘gesticulation’ or ‘co-verbal’ gestures. Such gestures are part of the language and proceeds somewhat unwittingly (humans are aware of their gestures in that they are available to subjective description after they are performed, but they are often not consciously constructed) from the mental processes of language production itself. The forms of these gestures are determined by personal style, culture, social makeup of the interlocutors, discourse context etc. There is a large body of literature in psychology, psycholinguistics, neurosciences, linguistics, semiotics and anthropology in gesture studies that lies beyond the scope of this article. We will list just two important aspects of gestures here. First, hand and arm gestures are made up of up to five phases: preparation, pre-stroke hold, stroke, post-stroke hold, and retraction. Of these, only the stroke that bears the key semiotic content is obligatory. Depending on timing, there may or may not be the pre- and post-stroke holds. Preparations and retractions may be elided depending on the starting and termination points of strokes (a preparation may merge with the retraction of the previous gesture ‘phrase’). Second, there is a temporal synchrony between gesture and speech such that the gestural stroke and the ‘peak of the tonal phrase’ are synchronized (Kendon 1972; Kendon 1980).

There is a class of gestures that sits between pure manipulation and natural gesticulation. This class of gestures, broadly termed deictics or pointing gestures, have some of the flavor of manipulation in its capacity of immediate spatial reference. Deictics also facilitate the ‘concretization’ of abstract or distant entities in discourse, and so are the subject of much study in psychology and linguistics. Following Bolt (Bolt 1980; Bolt 1982), work done in the area of integrating direct manipulation with natural language and speech has shown some promise in such combination. Earlier work by Cohen et al (Cohen, Dalrymple et al. 1989; Cohen, Dalrymple et al. 1998) involved the combination of the use of a pointing device and typed natural language to resolve anaphoric references. By constraining the space of possible referents by menu enumeration, the deictic component of direct manipulation was used to augment the natural language interpretation. (Neal, Thielman et al. 1989; Neal, Thielman et al. 1998), describe similar work employing mouse pointing for deixis and spoken and typed speech in a system for querying geographical databases. Oviatt et al (Oviatt 1999; Oviatt, DeAngeli et al. 1999; Oviatt and Cohen 2000) extended this research direction by combining speech and natural language processing and pen-based gestures. We have argued that pen-based gestures retain some of the temporal coherence with speech as with natural gesticulation (Quek, Yarger et al. 2000), and this co-temporality was employed in (Oviatt 1999; Oviatt, DeAngeli et al. 1999; Oviatt and Cohen 2000) to support mutual disambiguation of the multimodal channels and the issuing of spatial commands to a map interface. Koons et al (Koons, Sparrell et al. 1993; Koons, Sparrell et al. 1998) describe a system for integrating deictic gestures, speech and eye gaze to manipulate spatial objects on a map. Employing a tracked glove, they extracted the gross motions of the hand to determine such elements as ‘attack’ (motion toward the gesture space over the map), ‘sweep’ (side-to-side motion), and ‘end reference space’ (the terminal position of the hand motion). They relate these spatial gestural references to the gaze direction on the display, and to speech to perform a series of ‘pick-and-place’ operations. In Kendon’s parlance, such conventionalized gestures that may or may not accompany speech are termed emblems (Kendon 1988). Chai et al (Chai, Pan et
al. 2002) describes a MIND system that fuses speech with deictic gestures and some emblems (e.g. hand draws a question mark while speaker says “What is this?”). The system models the speaker's intention and attention to fuse the different modalities.

Wilson, Bobick and Cassell (Wilson, Bobick et al. 1996) proposed a tri-phasic gesture segmenter that expects all gestures to be a rest-transition-stroke-transition-rest sequence. They use an image-difference approach along with a finite state machine to detect these motion sequences. Natural gestures are, however, seldom clearly tri-phasic in the sense of this paper. Speakers do not normally terminate each gesture sequence with the hands in their rest positions. Instead, retractions from the preceding gesture often merge with the preparation of the next.

Kettebekov et al (Kettebekov, Yeasin et al. 2002; Kettebekov, Yeasin et al. 2003) describe a system that combines speech prosody and gesture to recognize two kinds of gestures used in television weather reporting. The reporter speaks and gestures in front of a green screen over which the weather map is superimposed for transmission. The two gesture types are ‘point’ and ‘contour’ (controlled motion where the reporter tracks a region or weather front). Combining speech prosody features and motion velocities and accelerations, the system uses a HMM to recognize the two gestures along with their preparation and retraction phases. The system expects each gesture to have a separate preparation and retraction, and one might argue that performances in front of a green-screen while one watches oneself on a monitor contributes to motion dynamics that are somewhat unnatural (even pointing gestures will become ‘visually tracked’ motions somewhat similar to manipulations).

Kahn et al (Kahn, Swain et al. 1994) describe their Perseus architecture that recognizes a standing human form pointing at various predefined artifacts (e.g. coke cans). They use an object-oriented representation scheme that uses a ‘feature map’ comprising intensity, edge, motion, disparity, and color features to describe objects (standing person and pointing targets) in the scene. Their system reasons with these objects to determine the object being pointed at. Extending Perseus, (Franklin, Kahn et al. 1996) describe an extension of this work to direct and interact with a mobile robot.

Sowa and Wachsmuth (Sowa and Wachsmuth 1999; Sowa and Wachsmuth 2000) describe a study based on a system for using coverbal iconic gestures for describing objects in the performance of an assembly task in a virtual environment. They use a pair of CyberGloves for gesture capture, three Ascension Flock of Birds electromagnetic trackers mounted to the subject’s back for torso tracking and wrists, and a headphone-mounted microphone for speech capture. In this work, subjects describe contents of a set of 5 virtual parts (e.g. screws and bars) that are presented to them in wall-size display. The gestures were annotated using the HAMBURG NOTATION SYSTEM for sign languages (Prillwitz and et.al. 1989). The authors found that “such gestures convey geometric attributes by abstraction from the complete shape. Spatial extensions in different dimensions and roundness constitute the dominant ‘basic’ attributes in [their] corpus … geometrical attributes can be expressed in several ways using combinations of movement trajectories, hand distances, hand apertures, palm orientations, hand-shapes, and index finger direction.” In essence, even with the limited scope of their experiment in which the imagery of the subjects was guided by a wall-size visual display, a panoply of iconics relating to some (hard-to-predict) attributes of each of the 5 target objects were produced by the subjects.

Wexelblat (Wexelblat 1995) describes research whose goal is to “understand and encapsulate gestural interaction in such a way that gesticulation can be treated as a datatype - like graphics and speech - and incorporated into any computerized environment where it is appropriate.” The author does not make any distinction between the communicative aspect of gesture and the manipulative
use of the hand, citing the act of grasping a virtual door knob and twisting as a 'natural' gesture for opening a door in a virtual environment. The paper describes a set of experiments for determining the characteristics of human gesticulation accompanying the description of video clips subjects have viewed. The experiment seeks answers to such questions as whether females produce fewer gestures than males, and whether second language speakers do not produce more gestures than native speakers. While the answers to these questions are clearly beyond the capacity of the experiments, Wexelblat produces a valuable insight that “in general we could not predict WHAT users would gesture about.” Wexelblat also states “there were things in common between subjects that were not being seen at a full-gesture analysis level. Gesture command languages generally operate only at a whole gesture level, usually by matching the user's gesture to a pre-stored template. … [A]ttempting to do gesture recognition solely by template matching would quickly lead to a proliferation of templates and would miss essential commonalities” [of real gestures].

Quek, McNeill and colleagues approach conversational gestures from the perspective of the involvement of mental imagery in language production (see, for example, (McNeill 2000; McNeill, Quek et al. 2001; McNeill, Quek et al. 2002; Quek 2002; Quek, McNeill et al. 2002; Quek 2003)). The idea is that if gesticulation is the embodiment of the mental imagery that, in turn, reflects the ‘pulses’ of language production, then one might be able to access discourse at the semantic-level by gesture-speech analysis. They approach this using the psycholinguistic device of the ‘catchment’ (McNeill 2000) by which related discourse pieces are linked by recurrent gesture features (e.g. index to a physical space and a specific hand shape). The question becomes what computable features have the semantic range to carry the imagistic load. Quek et al have demonstrated discourse segmentation by analyzing ‘hand use’ (Quek, McNeill et al. 1999; Quek, McNeill et al. 2002), kinds of motion symmetries of two-handed gestures (Quek, Xiong et al. 2002; Xiong, Quek et al. 2002), gestural oscillations (Quek and Xiong 2003), and space use (Quek, Bryll et al. 2001; Quek, McNeill et al. 2002).

CONCLUSION

Gesture use in human-computer interaction is a tantalizing proposition because of the human capacity for gesture and because such interfaces permit direct access to large and three-dimensional spaces. The user does not even need to manipulate an input device other than the appendages with which they come. We have laid out a ‘purpose taxonomy’ by which we can group gesture interaction systems, and by which design may be better understood.

REFERENCES


Quek, F., Y. Xiong, et al. (2002). Gestural Trajectory Symmetries and Discourse Segmentation. 7th Int Conf on Spoken Language Proc., Denver, CO.


