GESTURE AND INTERACTION
FRANCIS QUEK

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, seen, and felt (ignoring taste and smell as currently far-fetched for HCI). 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 organize our discussion around a purpose taxonomy. Interactive ‘gesture’ systems may be divided into three classes: 1. Manipulative; 2. Semaphoric; and, 3. Conversational. 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 our tri-partite segmentation adequately covers the use of gesture in HCI. 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 that permits the direct manipulation. The user interacted with a large wall-sized display moving objects around the
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 object manipulation.

We extend the concept to cover all systems of direct control. 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, a variety of ‘finger flying’ style navigation in virtual spaces or direct-manipulation interfaces, control of appliances, in computer games, and robot control. Other manipulative applications include interaction with wind-tunnel simulations, voice synthesizers, 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. Some of these approaches 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. 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, 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. 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 redefine some universe of ‘whole’ gestures $g \in G$. Taking a categorial approach, ‘gesture recognition’ boils down to determining if some presentation $p_b$ 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 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, the application of orientation histograms (histograms of directional edges) for hand shape recognition, graph-labeling approaches where labeled edge segments are matched against a predefined graph of finger-spelling-like hand 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, the application of ‘global’ features of the extracted hand (using color processing) such as moments, aspect ratio, etc. to determine recognize a set of hand shapes, model-based recognition using 3D model prediction, and neural net approaches.

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 semaphore is being performed. Approaches include finite state machines for recognition of a set of editing gestures for an ‘augmented whiteboard’, trajectory-based recognition of gestures for ‘spatial structuring, recognition of gestures as a sequence of state measurements, recognition of oscillatory gestures for robot control, and ‘space-time’ gestures that treat time as a physical third dimension.

One of the most common approaches for the recognition of dynamic semaphoric gestures is based on the Hidden Markov Model (HMM). First applied by Yamato, Ohya, and Ishii in 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.

In a typical HMM application Rigoll, Kosmala, and Eickeler (1997) were able to train a system to achieve 92.9% accuracy in recognizing 24 dynamic semaphores using manually-segmented isolated semaphores. This illustrates the weakness of such approaches, in that some form of presegmentation or other constraint (e.g. requiring fixed semaphore sequences, the user always ‘home’ their hands for each semaphore) is needed.

Semaphores represent a miniscule portion of the use of the hands in natural human communication. A major reason for their dominance in the literature is that they are the low-hanging fruit.

**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.

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, work done in the area of integrating direct manipulation with natural language and speech has shown some promise in such combination. Earlier work 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. Such systems have, for example, been employed for querying geographic databases. A natural extension of this concept is the combination of speech and natural language processing with pen-based gestures. The effectiveness of such interfaces is that pen-based gestures retain some of the temporal coherence with speech as with natural gesticulation, and this co-temporality was employed to support mutual disambiguation of the multimodal channels and the issuing of spatial commands to a map interface. Others have developed systems that resolve speech with deixes in regular video data.

In Kendon’s parlance, a class conventionalized gestures that may or may not accompany speech are termed emblems. The North American “OK” hand gesture is a typical emblem. While the temporal speech/emblem relationship is different from that of free-flowing gesticulation emblematic gestures in conjunction with speech have been proposed for such applications as map interaction.

Another approach to co-verbal gesticulation is to parse hand movements into gesture phases. Wilson, Bobick, and Cassell (1996), for example, developed a tri-phasic gesture segmenter that expects all gestures to be a rest-transition-stroke-transition-rest sequence (ignoring pre- and post-stroke holds). They required that the hand return to rest after every gesture. In another work, Kettebekov, Yasin and Sharma (2003) fused speech prosody and gesticular motion of a television weather reporter (in front of a green screen) to segment the phases and recognize two classes of gestures (deictics and contours). All gestures are constrained to have separate preparations and retractions. They employed a HMM formalization.

Sowa and Wachsmuth (2000) describe a study based on a system for using co-verbal iconic gestures for describing objects in the performance of an assembly task in a virtual environment. In this work, subjects wearing electromagnetically tracked gloves describe contents of a set of 5 virtual parts (e.g. screws and bars) that are presented to them in wall-size display. 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.
This author and colleagues approach conversational gestures from the perspective of the involvement of mental imagery in language production. 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’ 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. They demonstrate discourse segmentation by analyzing ‘hand use’, kinds of motion symmetries of two-handed gestures, gestural oscillations, and space use distribution.

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.

FURTHER READING