EURASIP Journal on Applied Signal Processing 2004:11, 1619–1636

The Catchment Feature Model: A Device for Multimodal Fusion and a Bridge between Signal and Sense

Francis Quek

Vision Interfaces and Systems Laboratory, Center for Human Computer Interaction, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
Email: quek@cs.vt.edu

Received 24 October 2002; Revised 16 February 2004

The catchment feature model addresses two questions in the field of multimodal interaction: how we bridge video and audio processing with the realities of human multimodal communication, and how information from the different modes may be fused. We argue from a detailed literature review that gestural research has clustered around manipulative and semaphoric use of the hands, motivate the catchment feature model psycholinguistic research, and present the model. In contrast to "whole gesture" recognition, the catchment feature model applies a feature decomposition approach that facilitates cross-modal fusion at the level of discourse planning and conceptualization. We present our experimental framework for catchment feature-based research, cite three concrete examples of catchment features, and propose new directions of multimodal research based on the model.

Keywords and phrases: multimodal interaction, gesture interaction, multimodal communications, motion symmetries, gesture space use.

1. INTRODUCTION

The importance of gestures of hand, head, face, eyebrows, eye, and body posture in human communication in conjunction with speech is self-evident. This paper advances a device known as the "catchment" [1, 2, 3] and the concept of a "catchment feature" that unifies what can reasonably be extracted from video imagery with human discourse. The catchment feature model also serves as the basis for multimodal fusion at this level of discourse conceptualization. This represents a new direction for gesture and speech analysis that makes each indispensable to the other. To this end, this paper will contextualize the engineering research in human gestures by a detailed literature analysis, advance the catchment feature model that facilitates a decomposed feature approach, present an experimental framework for catchment feature-based research, list examples that demonstrate the effectiveness of the concept, and propose directions for the field to realize the broader vision of computational multimodal discourse understanding.

2. OF MANIPULATION AND SEMAPHORES

In [4], we argue that with respect to human computer interaction (HCI), the bulk of the engineering-based gesture research may be classified as either manipulative or semaphoric. The former follows the tradition of Bolt's "Put-That-There" system [5, 6] which permits the direct manipulation of entities in a system. 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. Semaphore gesture systems predefine some universe of "whole" gestures $g_i \in \mathcal{G}$. Taking a categorial approach, "gesture recognition" boils down to determining if some presentation p_j is a manifestation of some g_i. Such semaphores may be either static gesture poses or predefined stylized movements. The feature decomposition approach based on the catchment feature model advanced in this paper is a significant departure from both of these models.

2.1. Gestures for manipulation

Research employing the manipulative gesture paradigm may be thought of as following the seminal Put-That-There work by Bolt [5, 6]. Since then, there has been a plethora of systems that implement finger tracking/pointing [7, 8, 9, 10, 11, 12], a variety of finger-flying style navigation in virtual spaces or direct-manipulation interfaces [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25], control of appliances [26], in computer games [27, 28, 29], and robot control [30, 31, 32, 33]. Other manipulative applications include interaction with

wind-tunnel simulations [34, 35], voice synthesizers [36, 37, 38], and an optical flow-based system that estimates one of 6 gross full-body gestures (jumping, waving, clapping, drumming, flapping, and marching) for controlling a musical instrument [39]. Some of these approaches (e.g., [30, 36, 37, 40, 41]) 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.

Gestures used in communication/conversation differ from manipulative gestures in several significant ways [42, 43]. 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 possibly only the feature that both may utilize the same body parts.

2.2. Semaphoric gestures

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 semaphoric 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 [44], Zernike moments-based hand pose estimation [45], the application of orientation histograms (histograms of directional edges) for hand shape recognition [46], graph-labeling approaches where labeled edge segments are matched against a predefined hand graph [47] (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") [48], the application of "global" features of the extracted hand (using color processing) such as moments, aspect ratio, and so forth to determine the shape of the hand out of 6 predefined hand shapes [49], model-based recognition using 3D model prediction [50], and neural net approaches [51].

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 $g_i \in \mathcal{G}$ is being performed. Approaches include finite-state machines for recognition of a set of editing gestures for an "augmented whiteboard" [52], trajectory-based recognition of gestures for "spatial structuring" [42, 43, 53, 54, 55, 56], recognition of gestures as a sequence of state measurements [57], recognition of oscillatory gestures for robot control [58], and "space-time" gestures that treat time as a physical 3D [59, 60].

One of the most common approaches for the recognition of dynamic semaphoric gestures is based on the hidden Markov model (HMM) [61]. First applied by Yamato et al. [62] for 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. An 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 [63, 64].

A parametric extension to the standard HMM (a PHMM) to recognize degrees (or parameters) of motion is described in [41, 65]. 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 et al. apply a recursive recognition scheme based on HMMs and utilize a set of rotationally invariant Zernike moments in the hand shape description vector [66, 67]. 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 [68] describe a system using HMMs to recognize a set of 24 dynamic gestures employing an HMM to model each gesture. The recognition rate (92.9%) is high, but it was obtained for *isolated* gestures, that is, 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 [64] describe an 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) [69], recognition of 10 gestures for presentation control [70], music conducting [57, 71], recognition of unistroke-like finger spelling performed in the air [72], and communication with a molecular biology workstation [11].

There is a class of systems that applies a combination of semaphoric and manipulative gestures within a single system. This class is typified by [11] 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). The 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 [73] emphasizes the need for development of systems able to recognize natural, nonposed, and nondiscrete 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.

2.3. Other paradigms

There is a class of gestures that sits between pure manipulation and natural gesticulation. This class of gestures, broadly termed deictics or pointing gestures, has 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 [5, 6], 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. [74, 75] 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. The authors in [76] describe similar work employing mouse pointing for deixis and spoken and typed speech in a system for querying geographical databases. Oviatt et al. [77, 78, 79] 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 [80], and this cotemporality was employed in [77, 78, 79] to support mutual disambiguation of the multimodal channels and the issuing of spatial commands to a map interface. Koons et al. [81] 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. This body of research differs from that reported in this paper in that we address more free-flowing gestures accompanying speech, and are not constrained to the 2D reference to screen or pen-tablet artifacts of pen or mouse gestures.

Wilson et al. [82] proposed a triphasic 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 triphasic 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.

Kahn et al. [12] 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, [83] describe an extension of this work to direct and interact with a mobile robot.

Sowa and Wachsmuth [84, 85] 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[1] 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 [86]. 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

[1] See www.ascension-tech.com.

(hard-to-predict) attributes of each of the 5 target objects were produced by the subjects.

Wexelblat [23] describes a 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 that subjects have viewed. These experiments were rather naive since there is a large body of literature on narration of video episodes [87]. 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 that "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).

3. DISCOURSE AND GESTURE

The theoretical underpinnings of the *catchment feature model* lies in the psycholinguistic theories of language production itself. In natural conversation between humans, gesture and speech function together as a coexpressive whole, providing one's interlocutor access to semantic content of the speech act. Psycholinguistic evidence has established the complementary nature of the verbal and nonverbal aspects of human expression. Gesture and speech are not subservient to each other, as though one were an afterthought to enrich or augment the other. Instead, they proceed together from the same "idea units," and at some point bifurcate to the different motor systems that control movement and speech. For this reason, human multimodal communication coheres topically at a level beyond the local syntax structure. While the visual form (the kinds of hand shapes, etc.), magnitude (distance of hand excursions), and trajectories (paths along which hands move) may change across cultures and individual styles, underlying governing principles that exist for the study of gesture and speech in discourse. Chief among these is the timing relation between the prosodic speech pulse and the gesture [87, 88, 89, 90].

3.1. Growth point theory

"Growth point" (gp) theory [1, 2, 91] assigns the rationale for the temporal coherence across modalities to correspond at the level of communicative intent. This temporal coherence is governed by the constants of the underlying neuronal processing that proceeds from the nascent "idea unit" or "gp." We believe that an understanding of the constants and principles of such speech-gesture-gaze cohesion is essential to their application in multimodal HCI.

While it is beyond the scope of this paper to provide a full discussion of language production and gp theory, we will provide a summary of the theory germane to the development of our model. In [1, 2, 91], McNeill advanced the gp concept that serves as the underlying bridge between thought and multimodal utterance. The gp is the initiating idea unit of speech production, and is the minimal unit of the image-language dialectic [92].

As the initial form of a "thinking-for-speaking" unit [1, 2, 91], the gp relates thought and speech in that it emerges as the newsworthy element in the immediate context of speaking. In this way, the gp is a product of differentiation that (1) marks a significant departure in the immediate context and (2) implies this context as a background. We have in this relationship the seeds for a model of real-time utterance and coherent text formation. The "newsworthiness" aspect of the gp is similar to the rheme-theme model [93, 94] that was employed in [95, 96] for generating speech and gesture and facial expressions, respectively.

3.2. Catchments

An important corollary to gp theory is the concept of the "catchment." The catchment is a unifying concept that associates various discourse components [1, 2, 3, 4, 97]. As a psycholinguistic device, it permits the inference of the existence of a gp as a recurrence of gesture features across two or more (not necessarily consecutive) gestures. The logic for the catchment is that coherent discourse themes corresponding to recurring imagery in the speaker's thinking produce such recurring gesture features. It is analogous to series of peaks in a mountain range that inform us that they were formed by a common underlying process because they share some geological characteristic (even if there are peaks of heterogeneous origins that punctuate the range).

An important distinction needs to be made here with respect to intentionality and wittingness. The speaker always intends to produce a particular catchment although she may be unwitting of its production. This is similar to the particular muscular activations necessary for vocal utterance. While the speaker intends to say the words uttered, she is unwitting of her laryngeal motions, respiratory apparatus, or even prosodic patterning. Nonetheless, both gesture and speech contain rich regularities and characteristics that support modeling and analyses to reveal the points of conceptual coherences and breakpoints in the discourse content.

3.3. The catchment feature model

Note that unlike the "whole gesture" formulation in the gesture recognition literature overviewed earlier, catchments involve only the recurrence of component gesture features. This suggests that one may approach gesture analysis by way

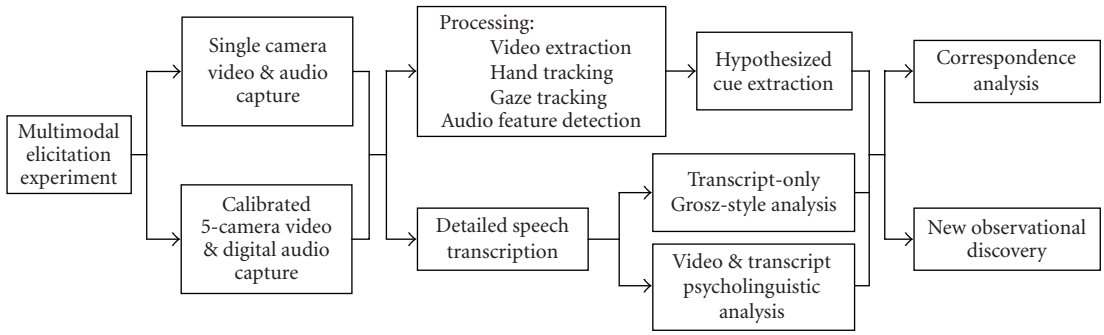

Figure 1: Block diagram of the typical experimental procedure employed.

of decomposing gestures into constituent features and studying their cohesion, segmentation, and recurrence. This is the essence of the catchment feature model proposed here.

As an illustration of this concept, we construct the following multimodal discourse segment (gesture described in brackets): "We will need *speakers* for the talk (two-handed gesture with each hand cupped with fingers extended, palms directed away from the speaker, coinciding with the word "speakers")…. We will set them up at the *right* and *left* of the podium (hands cupped as before, but this time with palm toward the speaker's torso; the left hand moves to a left distal point from the speaker holding the same hand shape, with palm directed at the speaker coinciding with the word "right" and the right hand moving similarly to a right distal point coinciding to the word "left" (with the left hand holding its distal position))…. When the speaker comes up on the *left* of the podium … (right hand in a pointing ASL "G" hand with index finger extended, indicating the path up the podium at the same right distal point as before coinciding with the word "left") …."

In this construction, the speaker established the cupped hand shape as an iconic representation of the speakers in the first utterance. She then establishes the spatial layout of the podium facing her where she places the speakers. Later in the discourse, she reuses the location of the left of the podium to indicate the ascent of the (human) speaker. In this case, we can recognize two catchments. The first, anchored by the iconic hand representations of the audio speakers, registers the coherence of the first two utterances. The second, based on the spatial layout established by the speaker, links the second and third utterances in the narrative (the left of the podium). These utterances may be separated by other utterances represented by the "…"s. In this illustration alone, we can see other features that may be salient in other analyses. For example, the direction of the palms in the iconic representation of the audio speakers establishes the orientation of the podium.

Clearly the number of features one may consider is myriad. The question then becomes what kinds of gestural features are more likely to anchor catchments. One may assume, for example, that the abduction angle of the little finger is probably of minor importance. The key question, then, to bridge the psycholinguistics of discourse production with image and signal processing, is the identification of the set of gestural feature dimensions that have the potential of subtending catchments. This paper presents an approach to answer this question, presents a set of catchment features that have been computationally accessed, proposes a set of metrics to evaluate these features, and proposes directions for our field to further advance our understanding and application of the catchment feature model.

4. EXAMPLES OF CATCHMENT FEATURES

A gesture is typically defined as having three to five phases: preparation, (prestroke hold), *stroke*, (poststroke hold), and retraction [87]. Of these only the stroke is obligatory. It carries the imagistic content and is the pulse that times with the prosodic pulse of speech phrases [87, 90]. The preparation and retraction can be thought of as being pragmatic movements to bring the hand into position for the stroke and to return the hand to rest after the stroke. Often, the retraction of a gesture unit will merge with the preparation of the next one. The prestroke and poststroke holds, if they are present, often serve as a timing function to synchronize the stroke with its speech affiliate.

The catchment features example cited here has been extracted computationally from either monocular or stereo video datasets of human subject experiments. We are in the process of collecting corpora of such data to support this scientific endeavor (see http://vislab.cs.wright.edu). As will be shown, some catchment features relate to individual gestures while others group runs of gestural activity.

4.1. Experimental methodology

To put our body of work in perspective, we will outline the general experimental methodology and the tools we have developed to support the science. Figure 1 lays out a typical experiment based on our methodology.

Figure 1 may be thought of as a general framework for research on the multimodal discourse analysis. The data is first obtained through a multimodal elicitation experiment. Bearing in mind that the makeup of the multimodal performance depends on discourse content (e.g., describing space, planning, narration), social context (i.e., speaking to an intimate, to a group, to a superior, etc.), physical arrangement

(e.g., seated, standing, arrangement of the interlocutor(s)), culture, personal style, and condition of health (among other factors), the elicitation experiment must be carefully designed. In our work, we have collected data on subjects describing their living quarters, making physical group plans, narrating the contents of a cartoon from memory, and trying to convince an interlocutor to take a blood pressure examination. Our data includes "normals" (typically American and foreign students), right- and left-handers, and individuals with Parkinson disease at various stages of disease and treatment.

Video/audio are captured using either single or multiple camera setups. The multiple camera setups[2] involve two stereo-calibrated cameras directed at each of the subject and interlocutor (to date, we have dealt only with one-on-one discourse). We employ standard consumer mini-DV video cameras (previously, our data have come from VHS and Hi-8 as well). The audio comes through boom microphones.

The video is captured to disk and processed using a variety of tools. The hands are tracked using a motion field extractor that is biased to skin color [98, 99, 100, 101] and head orientation is tracked [102]. From the hand motion data, we extract the timing and location of holds of each hand [103]. We also perform a detailed linguistic text transcription of the discourse that includes the presence of breath and other pauses, disfluencies, and interactions between the speakers. The speech transcript is aligned with the audio signal using the Entropic's word/syllable aligner.[3] We also extract both the F_0 and RMS of the speech. The output of the Entropic's aligner is manually checked and edited using the Praat phonetics analysis tool [104] to ensure accurate time tags. This step makes our work immune to any misalignment in the auto-aligner step. This process yields a time-aligned set of traces of the hand motion with holds, head orientations, and precise locations of the start and end points of every speech syllable and pause. The time base of the entire dataset is also aligned to the experiment video. In some of our data, we employ the Grosz "purpose hierarchy" method [105] to obtain a discourse segmentation. The choice of discourse segmentation methodology may vary. Any analysis that determines topical cohesion and segmentation will suffice. The question to which we seek answer is whether the catchment feature approach will yield a discourse segmentation that matches a reasonably intelligent human-produced segmentation.

To support the stringent timing analysis needed for our studies, we developed the Visualization for Situated Temporal Analysis (VisSTA) system for synchronous analysis of video, speech audio, time-tagged speech transcription, and derived signal data [106, 107, 108].

To demonstrate the efficacy of the catchment feature concept, both as a device for language access and as a bridge to signal and image/video processing, we will visit three catchment feature examples.

4.2. Holds and handedness

In the process of discourse, speakers often employ their hands and the space in front of them as conversational resources to embody the mental imagery. Hand use, therefore, is a common catchment feature [3, 109, 110]. In [4, 97, 111], we investigated the detection of hand holds and hand use in the analysis of video data from a living space description. This 32-second (961 frames) data was obtained from a single camera, and so we hand only x (horizontal) and y (vertical) motion data on the hands.

Gesturing may involve one hand (1H), that could either be right (RH) or left (LH), or two hands (2H). The dual of hand use is, of course, resting hand holds (detected LH-only holds indicate RH use and vice versa). In real data, the detection of holds is not trivial. In [103], we describe our RMS motion-energy approach to detect holds while ignoring slight nongestural motions.

Figures 2 and 3 are a synopsis of the result of the catchment analysis. The horizontal dimension of the graphs is time or frame number. From top to bottom, each chart shows the x and y hand motion traces, the marking of the hand-hold durations, the F_0 of the speech audio, and the words spoken. The key discourse segments are labeled (A) through (E). The vertical columns of shading indicate time spans where both hands are stationary.

The subject in the experiment systematically assigned the description of the rear of her dwelling to her LH in sections (A) and (D) (this includes a kitchen area and a spiral staircase). She assigned the front staircase of her home that is on the right-hand side to her RH in section (C), and, whenever she talked about the front of her house, she used symmetric 2H gestures (section (B)). This order was consistently held even though the description included a major discourse repair at the end of (A) where she says "Oh! I forgot to say ..." (RH withdraws sharply from the gesture space (as can be seen in the top x graph labeled (K.1.))). The same hand use configuration marks her returns to the back staircase again in section (D) 5 to 6 phrases later. In this latter section, the holding LH moves slightly as the RH makes very large movements. Since nonsymmetrical 2H movements are unlikely, (see next section) the "dominant motion rule" that attenuates the small movements in one hand in the presence of large nonsymmetric movements in the other hand helped to label the LH as holding (the intuition is that since the body is interconnected, there will always be small movements in other extremities in conjunction with large movements of one arm).

The 2H section labeled (B) may be further subdivided based on the motion symmetry characteristics of the hands. We will discuss this in Section 4.3. At the end of section (B) (F_0 numbers 28–30), we see the final motion of the RH going to rest. This is a retraction signalling the end of the 2H portion (B) and the beginning of the LH portion (C). The retraction suggests that the discourse portions encapsulated by (B) has already ended, placing the words corresponding

[2] Described in http://vislab.cs.vt.edu/KDI/Homepage/equipment.html.

[3] Entropic was acquired by Microsoft that has discontinued support for the xwaves products. The version we are using is a pre-Microsoft acquisition version.

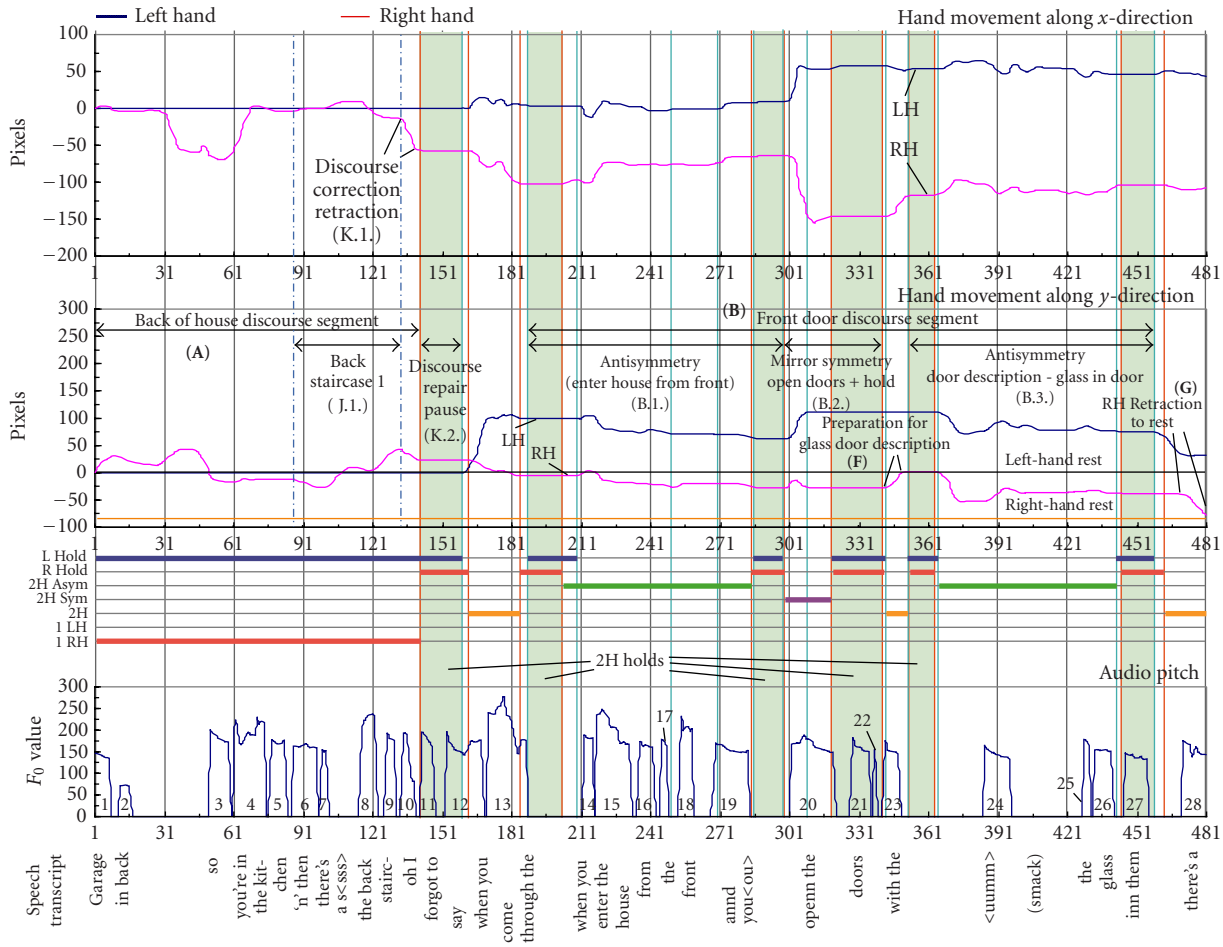

Figure 2: Hand position, handedness analysis, and F_0 graphs for the frames 1–481.

to F_0 units 28–30: "there's a . . . the front . . ." to the following utterance. This correctly preserves the text of the front staircase description. This structure preservation is robust even though the preceding final phrase of (B) is highly disfluent (exhibiting a fair amount of word search behavior).

The robustness of the hand use feature illustrated here bears out its utility as a catchment feature.

4.3. *Symmetry classification*

The portion of the living space description of Figure 2 labeled (B) is further segmented into three pieces labeled (B.1) to (B.3). These are separated by columns of vertical shading that mark periods when both hands are holding. The x-(lateral) symmetry characteristic marks (B.1) and (B.3) as generally positive x-symmetric (both hands moving in same x-direction) and (B.2) as negative x-symmetric. This divides the "front of the house" description into three pieces—describing the frontage, entering through the front doors, and description of the doors, respectively.

This brings us to our second catchment feature of motion symmetry of 2H gestures. Concerning symmetry in sign language and gesture, Kita writes, "When two strokes by two hands coincide in sign language, the movements obey the well-known Symmetry Condition, which states that the movement trajectory, the hand orientation, the hand shape, and the hand-internal movement have to be either the same or symmetrical . . . the Symmetry Condition also holds for gestures." [112, 113]. In fact, it appears that when both hands are engaged in gesticulation with speech, there is almost always a motion symmetry (either lateral, vertical, or near-far with respect to the torso), or one hand serves as a platform hand for the other moving hand. To test the veracity of this claim, one needs only perform the simple experiment attempting to violate this condition while both hands are engaged in gesticulation. This tyranny of symmetry for two moving hands during speech seems to be lifted when one hand is performing a pragmatic task (e.g., driving while talking and gesturing with the other hand). Such pragmatic movements also include points of retraction of one hand (to transition to a one-handed (1H) gesture) and preparation of one hand (to join the other for a two-handed (2H) gesture or to change the symmetry type).

In [114, 115], we investigated a finer-grain analysis of this motion symmetry using a signal correlation approach. We

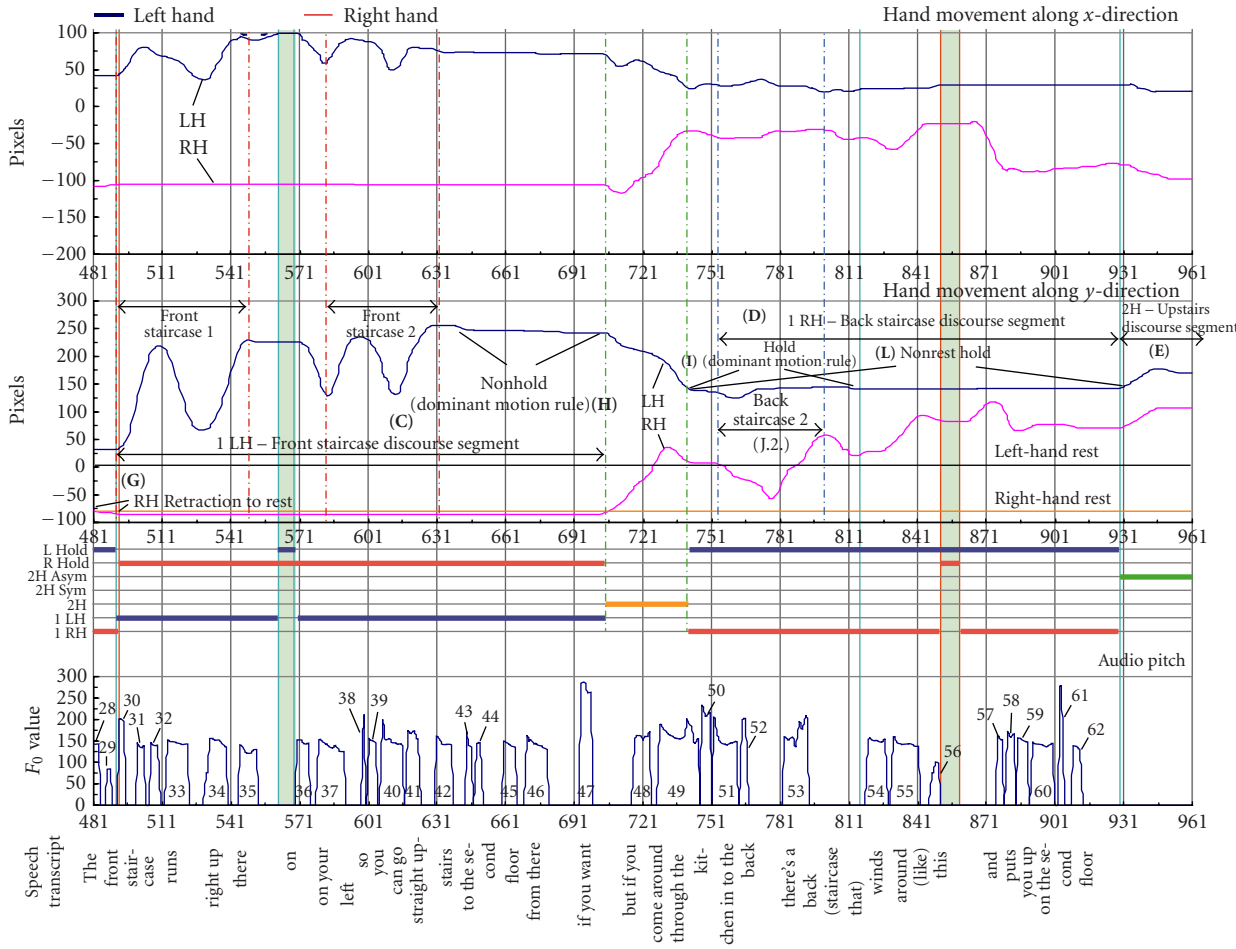

Figure 3: Hand position, handedness analysis, and F_0 graphs for the frames 481–961.

apply the correlation relationship

$$r_u = \frac{\sum_F \sum_u (S_L - \overline{S_L})(S_R - \overline{S_R})}{\sqrt{\sum_F \sum_u (S_L - \overline{S_L})^2 \sum_F \sum_u (S_R - \overline{S_R})^2}}, \quad (1)$$

where S_L and S_R are LH and RH motion trajectories, respectively, $\overline{S_L}$ and $\overline{S_R}$ are the mean values of S_L and S_R, F denotes the frame number, and u denotes the positional value (if u is the x value of the hand position, we are computing lateral symmetry).

Equation (1) yields the global property between left-hand signal and right-hand signal. To obtain local symmetry information, we employ a windowing approach: $S_w^L = W \star S_L$; $S_w^R = W \star S_R$, where W is the selected window and $\star$ denotes convolution.

Hence, the local symmetry of the two signals may be computed with a suitable window:

$$r_{u_w} = \frac{\sum_{F_w} \sum_{u_w} (S_w^L - \overline{S_w^L})(S_w^R - \overline{S_w^R})}{\sqrt{\sum_{F_w} \sum_{u_w} (S_w^L - \overline{S_w^L})^2 \sum_{F_w} \sum_{u_w} (S_w^R - \overline{S_w^R})^2}}, \quad (2)$$

where $\overline{S_w^L}$ and $\overline{S_w^R}$ are the mean values of S_w^L and S_w^R, respectively, and w defines the window size.

Taking

$$\mathbf{P}_L(t) = [x_L(t) y_L(t) z_L(t)]^T,$$
$$\mathbf{P}_R(t) = [x_R(t) y_R(t) z_R(t)]^T \quad (3)$$

as the LH and RH motion traces, respectively (x is lateral, y is vertical, and z is front-back with respect to the subject's torso), we can compute the correlation vector $\mathbf{R}_w(t) = [r_{x_w}(t) r_{y_w}(t) r_{z_w}(t)]^T$.

The size of the convolving window is critical since too large a window will lead to oversmoothing and temporal inaccuracies of the detected symmetries. Too small a window will lead to instability and susceptibility to noise. We chose a window size of 1 second (30 frames). This gave us reasonable noise immunity for our data while maintaining temporal resolution. The drawback was that the resulting symmetry profiles detected were fragmented (i.e., there were "dropouts" in profiles). Instead of increasing the window size to obtain a smoother output, we applied a rule that a dropout below a certain duration between two detected symmetries of the same polarity (e.g., a dropout between two runs of positive symmetry) is deemed to be part of that symmetry. We chose a period of 0.6 second for the dropout threshold.

Table 1: x-symmetry.

Number	Beginning time	Duration	Correlation coefficients	Time from previous feature	Speech and comments
1	5.44	0.17	0.65	0.00	when [you come]
2	5.91	0.17	−0.63	0.30	thro[ugh the]
3	6.91	0.63	0.84	0.83	[when you enter the hou]se
4	7.81	0.13	−0.65	0.27	[from the] front
5	8.34	0.13	−0.43	0.40	from the fr[ont]
6	8.94	0.33	0.67	0.47	a[nd you]
7	9.84	0.40	−0.89	0.57	open the
8	10.78	0.13	−0.72	0.53	[doors] with …
9	11.38	0.27	−0.83	0.47	[the] … <um> … the glass
10	12.08	0.37	0.85	0.43	the … […] … <um> …
11	12.75	0.30	0.85	0.30	the … [<um>] … the glass
12	13.15	0.17	0.65	0.10	the … <um> […] … the glass
13	14.01	0.20	0.71	0.70	the … <um> … [the g]lass

Table 2: y-symmetry.

Number	Beginning time	Duration	Correlation coefficients	Time from previous feature	Speech and comments
1	5.44	0.63	−0.92		when yo[u come through the] …
2	6.91	0.63	0.94	0.83	wh[en you enter the house]
3	7.81	0.13	0.65	0.27	house [from the] front
4	8.64	0.13	−0.52	0.70	front … [and] you … open
5	9.84	0.40	−0.91	1.07	[open the] … doors with
6	11.38	0.20	0.67	1.13	doors wi[th the] …
7	12.08	0.37	0.91	0.50	doors with the … […] … <um> …
8	12.75	0.30	0.78	0.30	with the … [um] … the
9	14.01	0.43	0.88	0.97	with the … [um] … [the gla]ss

This adequately filled in the holes without introducing oversmoothing (given inertia, the hands could not transition from a symmetry to nonsymmetry and back in 0.6 second).

4.3.1. 2D living space description symmetries

Tables 1 and 2 present the start time, duration, correlation coefficient, time from previous symmetric feature, and the words uttered (marked in brackets). We summarize these tables in Figure 4. The two lines above each text line represent positive symmetries, and the two lines beneath represent negative symmetries.The lines closer to the text represent x-symmetries, and the lines farther from the text represent y-symmetries. The line segments are numbered as per Tables 1 and 2, showing the contiguous runs of symmetry.

By our rule, we have the x-symmetries yielding the following 12 longer segments: "you come," "through the," "When you enter the house," "from the front," "And you," "open the doors with the," and "<ummm> <smack> the glass."

Taking the superset of these segmentations (i.e., if a y segment contains an x segment, we take the longer segment and vice versa), we have the following segmentation: (1) "When you come through," (2) "… when you enter the house from the front," (3) "and you …," (4) "open the doors *with the*," (5) "*with the* … <ummm> <smack> … the glass," (overlapping segments are in italics).

This analysis preserves the essence of the (B.1)–(B.3) segmentation with some extra detail. The utterance (3) "and you …" between (B.2) and (B.3) is set apart from the latter and is essentially the retraction for the "open the doors" gesture (both open palms begin facing the speaker and fingers meeting in the center, mid-torso and swings out in an iconic representation of a set of double doors) and the preparation of the "glass in the doors" representation (the subject moves both hands synchronously in front of her with a relaxed open palm as though feeling the glass in the door). Also, the correlation-based algorithm correctly extracted the segment (1) "When you come through" that was missed by the earlier analysis (and by the human coders). This utterance was, in fact, an aborted attempt at organizing the description. The subject had begun talking about going through the double doors. She began and aborted the same "opening the doors" (we know these are double doors that open inward only from the gesticular imagery, it was never said) gesture as she later

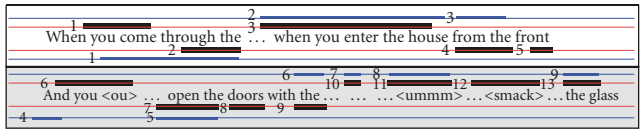

Figure 4: Symmetry-labeled transcript.

completed in (4) "open the doors." She realized that she had not yet introduced the front of the house and did so in (2). This demonstrates the catchment feature that represents the mental imagery of the corresponding gp.

4.3.2. 3D spatial planning data symmetries

A second experiment captured by two stereo-calibrated cameras demonstrates the symmetry catchment feature in 3D [115]. In this experiment, a subject is made privy to a plan to capture a family of intelligent wombats that have taken over the town theater in a ficticious town for which there is a physical model. She is then video-taped discussing the plan and fleshing it out with an interlocutor.

The dataset comprised 4,669 video frames (155.79 seconds). In the x-symmetry data, there were 32 runs of symmetries. Of these, 7 occurred during the interlocutor's turn where the subject clearly pantomimed her interlocutor, most likely to show interest and assent. This leaves 25 detected symmetry runs cotemporal with the subject's speech. In the y-symmetry data, 37 runs were extracted. Of these, one was erroneous, owing to occlusion of the hands in the video, and 6 took place during the interlocutor's turn. This leaves 30 detected y-symmetry runs accompanying speech.

For this dataset, we compared the start and end of each run of symmetry to the Grosz purpose-hierarchy-based analysis of the discourse text. We would expect the symmetry transitions to correspond to discourse shifts.

Combining both x- and y-symmetries, we have a total of 56 runs of symmetry. This gives 112 opportunities for finding discourse transitions. The purpose hierarchy yielded 6 level-1 discourse segments, 18 level-2 segments, 18 level-3 segments, and 8 level-4 segments. There were 59 unit transitions and 71 speaker-interlocutor turn changes.

Of the 112 symmetry-run starts and ends, 63 coincided with purpose-hierarchy discourse unit (DU) transitions. Of these, 25 transitions coincided with x-symmetry terminals, and 28 transitions coincided with y-symmetry terminals. Note that it is possible for two terminals to detect the same transition (i.e., if both x- and y-symmetries detect the same transition or when the end of one symmetry run coincides with the end of a discourse segment, and the next symmetry run begins with the next discourse hierarchy segment).

We introduce another concept that is becoming evident in our analysis of the 3D symmetry data—that of directional dominance. We noticed that the symmetry coefficients along different axes were more chaotic for some runs as compared with others. For example, in a particular discourse region, we have a run of positive correlations in x but not in y, and in other discourse regions, the reverse is the case. Upon investigation of the discourse video, we noticed that at these junctures, we perceived the speaker's gestures to be dominantly symmetrical in the direction indicated by the coherent correlations. There are, nonetheless, equally strong correlations (in terms of absolute correlation value) in the more fragmented dimensions. The reason is that while the speaker "intends" a particular symmetry (say moving the hands outward laterally in an "opening gesture"), the biometrics of arm movement dictate some collateral symmetry in the y and z dimensions as well. In this case, the absolute distance traversed in x dominates the y and z movements. We cannot simply filter out small movements since some motion symmetries are intentionally small. We can, however, detect the dominant direction in a symmetric run in terms of the relative total traversals and select the corresponding symmetries as the "true" ones.

4.4. Space use analysis

The final catchment feature example we will visit is that of space use (SU). Space and imagery are inseparable. Obviously, one expects gesture to access space, where space is the immediate "subject matter," but speakers recruit spatial metaphors in gesture even when not speaking about space (as formalized by the "mental spaces" concept [116, 117]). A lateral differentiation of gestures across the midline of the gesture space, for example, reflects the lateral arrangement of objects in the reference space even when the content of speech does not mention space [118]. A related concept is that of the "origo" (see [87, page 155], [119]). In a sense, all language can be thought of as referential. References comprises three components: the thing referenced (and its location), the act of referencing, and the viewpoint (or origo) from which the reference is made. In a pointing gesture, by analogy, these correspond to the thing and location pointed to, the pointing finger configuration and motion, and the origin from which the gesture is made.

In [120, 121], we investigate the application of SU patterns as a catchment feature. For some DU, $D(i)$, the corresponding pattern of SU may be captured by a hand occupancy histogram (HOH) $\mathcal{H}(i)$. $D(i)$ is any DU (e.g., a phrase, sentence, or "paragraph"). The gesture space in front of the speaker is divided into a $K \times K$ (we use 50×50) occupancy grid. At each time interval (we use the camera frame rate of 30 fps), within $D(i)$, we increment each cell in $\mathcal{H}(i)$ by a weighted distance function:

$$\triangle H_t(u,v) = f_w\Big(\big|[u,v]^T,[x_t,y_t]^T\big|\Big), \tag{4}$$

$$f_w\Big(\big|[u,v]^T,[x_t,y_t]^T\big|\Big) = \frac{\mathcal{G}\Big(\big|[u,v]^T,[x_t,y_t]^T\big|\Big)}{\sum_{u,v}\mathcal{G}\Big(\big|[u,v]^T,[x_t,y_t]^T\big|\Big)}. \tag{5}$$

Equation (5) is a normalized sigmoidal function, where

$$\mathcal{G}(d) = \begin{cases} \dfrac{1-\epsilon-\mathcal{F}(k,d)}{1-2\epsilon} & \text{for } d < k, \\ 0 & \text{for } d \geq k, \end{cases} \tag{6}$$

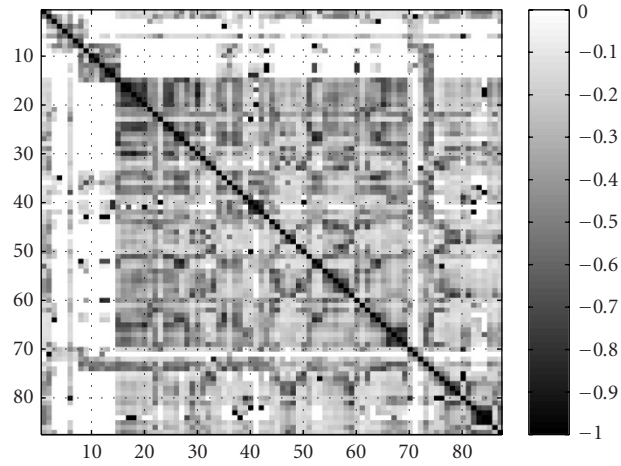

Figure 5: Sentence SCM.

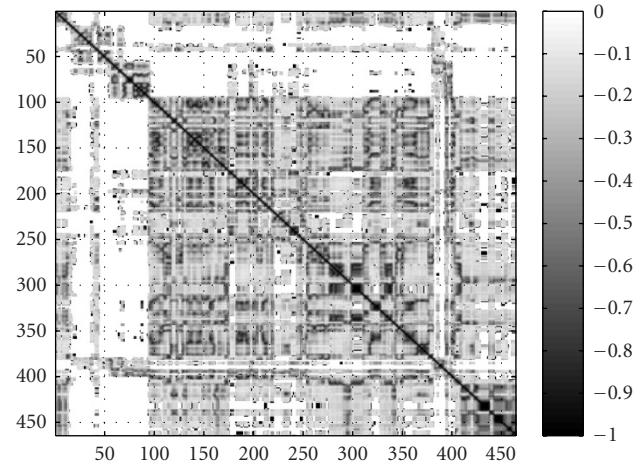

Figure 6: Discrete time origo correlation matrix.

where $0 < \epsilon \ll 1$ (we use 0.01) and $\mathcal{F}(k,d) = 1/(1 + e^{-a(d-k/2)})$, where $a = -(2/k)\ln(\epsilon/(1-\epsilon))$. The parameter k controls the size of the sigmoid function. Empirically, we found quarter of the size of the occupancy grid to be an adequate value for k.

Equation (4) determines the likelihood that the hand is in any particular cell at time t. It serves two purposes: it avoids the discretization problem where the hand is judged to be in a specific grid location when it is near a grid boundary; and it allows us to use a much finer-grain grid with the attendant advantage of smoothing out uncertainties in the location of the hand. Hence, for each computed hand location above our physical town model, (4) produces a "location likelihood" distribution at each time slice. For each DU, $D(i)$, we compile a discourse-specific "SU histogram":

$$\mathcal{H}(i) = \sum_{t=t_S(D(i))}^{t_E(D(i))} \triangle H_t(u,v), \tag{7}$$

where $t_S(D(i))$ and $t_E(D(i))$ are the start and end times of $D(i)$.

If DUs, $D(i)$, and $D(j)$ share a common SU nexus (SUN), this may be discovered by correlating $\mathcal{H}(i)$ against $\mathcal{H}(j)$. The problem is that we do not know a priori where the SUNs will be, what shapes they may take, and if there may be more than one SUN in a particular DU. Take the example where $D(i)$ encompasses SUN o_m and o_n and $D(j)$ contains o_m and o_p. A simple sum of least squares correlation may penalize the two as different when they in fact share o_m. We devised a fuzzy-AND correlation function that examines only the normalized intersection between two HOH's. We define the cellwise masking of $\mathcal{H}(i)$ by $\mathcal{H}(j)$ as follows:

$$\left(C^i_{(u,v)} \,|\, C^j_{(u,v)} > 0\right) = \begin{cases} C^i_{(u,v)} & \text{for } C^j_{(u,v)} > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

$C^i_{(u,v)}$ and $C^j_{(u,v)}$ being (u,v) cells of $\mathcal{H}(i)$ and $\mathcal{H}(j)$, respectively.

After cellwise masking, we normalize the resulting histogram. We denote this histogram by $\|\mathcal{H}(i)|\mathcal{H}(j) > 0\|$. Each cell in this histogram represents the probability that the hand was in that cell during DU, $D(i)$, if it shares a SUN with $D(j)$. We denote this by $P(C^i_{(u,v)}|C^j_{(u,v)} > 0)$. With this set up, we can perform the correlation of $\|\mathcal{H}(i)|\mathcal{H}(j) > 0\|$ with $\|\mathcal{H}(j)|\mathcal{H}(i) > 0\|$ by taking the cellwise fuzzy-AND $\mathcal{H}(i) \otimes \mathcal{H}(j)$:

$$\sum_{u,v} \min\left(P(C^i_{(u,v)} \,|\, C^j_{(u,v)} > 0), P(C^j_{(u,v)}|C^i_{(u,v)} > 0)\right). \tag{9}$$

Note that $\mathcal{H}(i) \otimes \mathcal{H}(j) = 1$ if $i = j$.

Applying $\mathcal{H}(i) \otimes \mathcal{H}(j)$ to all i, j, we obtain a $N \times N$ SU correlation matrix (SCM), where N is the number of DUs. Examples of such matrices are shown in Figures 5 and 6.

Contiguous DUs linked semantically by SU should yield blocks of high correlation cells along the diagonal of the SCM. Consequently, semantic discourse shifts should manifest themselves as gaps between such blocks. These would correspond to minima in the diagonal projections in the correlation matrix normal to the (i,i) diagonal. We compute this SU coherence projection vector (SCPV) for each diagonal cell (i,i) as the following sum:

$$P_0(i) = \sum_{k=-d}^{d} \mathrm{SCM}(i+k, i-k) - 1.0, \tag{10}$$

where d is the range of cells over which the projection is taken. Since the (i,i)th cell is always 1.0, we subtract 1.0 from each vector element. The parameter d controls the range of the DU neighborhood that exerts effect on a vector element. The value of d obviously depends on the granularity of the DUs we use.

To improve the sensitivity of the SCPV, we include the "between" diagonal projections from the $(i, i+1)$ cells:

$$P_1(i) = \sum_{k=-d}^{d-1} \mathrm{SCM}(i+k, i+1-k). \tag{11}$$

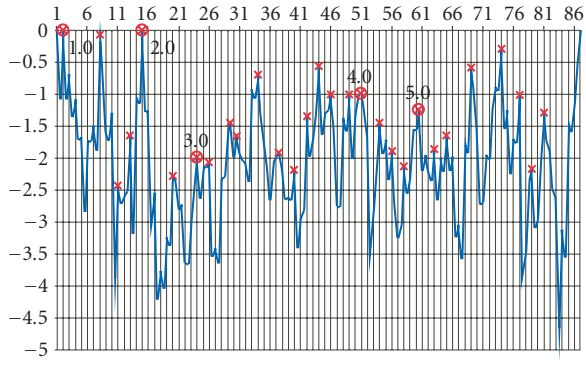

Figure 7: Sentence SCPV.

Combining P_0 and P_1, we obtain the $2N - 1$ projection vector $\mathcal{P}$. An example of SCPV is shown in Figure 7 (the plot is inverted so that the SU transitions are peaks).

We tested the SU catchment feature on the same spatial planning data introduced in Section 4.3.2. For interactants making plans with the aid of a terrain map, the space in the plane of the map often serves as "address space." Hence, we mapped the SU HOHs of the subject's dominant hand in the x-z plane above the village model (in discourse that does not use a horizontally oriented artifact, the dominant SU plane will be the x-y plane (the vertical plane in front of the speaker's torso)).

Sentential DU

We performed an independent sentential parse using only the grammatical syntax for segmentation that yielded 87 discourse units. Using the start and end times of these units, we computed the 87 HOHs and obtained the SCM shown in Figure 5. The 87 sentences are numbered on the 2 axes. The larger dark rectangles along the 1.0 autocorrelation diagonal correspond to higher contiguous SU cohesion. Figure 7 shows the SCPV for these sentence units where the peaks correspond to SU transitions. In this case, since the sentences are large DUs, the value of d in (10) and (11) was set to 3. 31 transitions were detected. Of these, only 3 did not correspond to valid purpose-hierarchy transitions. All 5 level-1 transitions were correctly extracted. These are numbered in Figure 7. Six SCPV peaks were detected at the start of the speaker's turn, and 6 were associated with the withdrawal of the speaker's hands at the end of her turn. Of the 3 nonpurpose-hierarchy transitions, 2 were at the start of the speaker's turn when she reintroduced her hand to the terrain map. Only one detected SCPV peak did not correspond to either a turn-exchange or a purpose-hierarchy DU transition. This took place in a rather complex situation when the speaker and interlocutor were speaking simultaneously.

Discrete-time DU

We performed a second analysis where the discourse was segmented into a series of overlapping one-second long DUs at a uniform interval of 0.333 seconds (every tenth video frame).

Table 3: Discrete time SCPV peaks correspondences.

Event	No.	Event	No.
Transition	45	Repair	3
Interlocutor	9	Action stroke	1
Start turn	8	New transition	1
End turn	7	Unaccounted	5
New place	3		

This produced 465 units and 465 HOHs. The 465×465 SCM is displayed in Figure 6. It should be noted that Figures 6 and 5 are remarkably similar although the latter was generated from sentences of varying time durations. Both SCMs depict the same information about the flow of the discourse. A 931-element SCPV was derived from the discrete-time unit SCM in Figure 6. The value for d was set to 15 (or 5 seconds).

A total of 75 peaks were found in the SCPV. Table 3 summarizes the discourse events that correspond to the SCPV peaks. Note that the event counts sum up to more than 75 because an SCPV peak may coincide with more than one event (e.g., at a speaker turn change that coincides with a discourse transition).

The beginnings of all 6 level-1 purpose-hierarchy units were correctly detected (among a total of 45 transitions found). Of the 15 turn exchanges detected as SCPV peaks, 6 did not coincide with a hierarchy transition. There were 9 SCPV peaks when the subject was silent and the interlocutor was speaking. Most of these occurred because the subject imitated the gestures of her interlocutor or pantomimed what she was describing (most probably to show that she was following the discussion). There was one pragmatic hand movement when she moved her hands onto her hips while her interlocutor was speaking, and a couple of times the subject retracted her hands to rest when it became clear that the interlocutor turn would be extended. The new-place events occurred when a new location was introduced in the middle of a DU and the hand location moved from its origo role to the deictic target. In one of the three instances, the speaker says, "*we are gonna go over to [breath pause]|| 35 'cause*" (the double vertical bars represent the SCPV peak point). In this case, the hand moves after the breath pause to the location of "house 35."

In certain speech repairs, there is a tendency for a speaker to withdraw her hand from the gesture space to reintroduce it [3, 4, 97, 111, 122]. This accounts for the 3 repair instances detected as SCPV peaks. The action-stroke event occurred when the subject said, "... scare the wombats || out through the front." In this case, the hand indicates the path along which the wombats will be chased.

The new-transition event was missed in the original manual coding. The subject actually introduced the ideas of "wombats" and the town "theater" for the first time with the utterance: "and see the thing is || is there are *wombats* in the *theater*...." The SCPV peak flags a large withdrawal of the hand backward terminating in a downward beat at the word "wombats." At this point, the nondominant hand enters the scene and both hands join forces assuming the

G-hand pose and pointing emphatically at the theater coinciding with the utterance "in the theater." The hands then stay around the theater while she describes the events to take place there. Hence, we most likely have the introduction of a new SUN centered around the theater. There is always debate as to when a new purpose phrase begins, and this may be an example where the SU shift may provide better guidance for the coder. In any case, since the purpose hierarchy as coded did not flag a discourse shift, we did not count this SCPV peak as a discourse transition in Table 3. There were 5 SCPV peaks for which we could not determine a cause.

One might argue that the sentential coding experiment is not completely independent of the purpose hierarchy because sentential structure is related to semantic discourse content. In the discrete-time experiment, discounting the SCPV peaks that took place during the interlocutor's turn and the 6 nontransition turn changes, 45 out of 60 detected peaks corresponded to semantic discourse transitions. This is more significant in that there is no other reason that a 0.333-second interval graph should adhere to the purpose-hierarchy structure other than gestural structuring of discourse content. Apart from 5 cases, the other 10 SCPV peaks correspond to other non-SU discourse phenomena (such as speech repairs).

5. DISCUSSION AND FUTURE DIRECTIONS

Thus far, we have laid out a perspective of multimodal communication based on sound psycholinguistic theory. Beginning from the relation between mental imagery and the gp, we motivated the concepts of the catchment and consequently the device of the catchment feature model along with the corollary concept of feature decomposition approach for gesture analysis. To lend concreteness to these concepts, we presented three catchment features along with the analyses of how they facilitate analysis of the entire multimodal communicative performance. The model has been applied to study multimodal gesture-speech disfluency phenomena [4, 97, 122, 123, 124], timing of prosody and gesture as discourse focal points [102, 125, 126], and the communicative deficits attendant to Parkinson disease [127, 128, 129]. Other catchment features we have investigated include oscillatory gestures and hand shape [55, 56, 130].

The catchment feature model provides a locus for multimodal fusion at the level of mental imagery and discourse planning. As such, it suggests several future directions for the field of multimodal communication research beyond the obvious research in identifying, extracting, and testing new catchment features.

First, there is need for measures of catchment feature efficacy. If the question is whether a particular catchment feature detector is accurate, paradigms of classifier performance evaluation such as those employing false positives and negatives and receiver operating characteristics [131] would suffice. This does not, however, address the question of the efficacy of a particular catchment feature. Given particular discourse and social contexts, subject matter, personal styles, and so forth, a specific catchment feature could be perfectly extracted, but of limited efficacy. We propose a *power/penalty* evaluation that applies to particular contexts. In the "SU" example cited above, there were 59 points of discourse topic/level transitions in the expertly coded transcription. The discrete-time SU detector extracted 75 peaks of which 45 corresponded to coded transitions. This indicates that in the context of spatial plan conveyance over a terrain representation between the 2 subjects, we properly extracted 45 transitions out of 59 opportunities, yielding a power of 76.27%. The penalty of applying this catchment feature model is 30 nontransition SU peaks out of 75 peaks or 40%. This power/penalty analysis bears out the intuition that in conveying a spatial/temporal plan with access to a model of the terrain, a speaker may organize her discourse plan around the physical artifact.

The second element needed to advance the field is the availability of coded video corpora. The power/penalty analysis highlights two requirements in this context: (1) the need for sufficient coded data; (2) the need for corpora around a taxonomy of discourse conditions. It is obvious that the usefulness of a power/penalty analysis for a single dataset is of limited utility (apart from showing the potential of a particular catchment feature). Given behavioral variances due to personal styles, cultural contexts, and social situations, we have to either randomize these distributions or specify the conditions to constitute a single class (e.g., spatial/temporal planning for American English-speaking military personnel with equally ranked individuals). This permits the computation of power/penalty statistics across multiple datasets (e.g., across college students of varying national origins and personal styles engaged in the "wombat" planning discourse). This requires carefully planned experiments, coding schemes, and the identification of classes of discourse contexts. While an exhaustive taxonomy discourse contexts may not be practical, the identification and classification of certain "useful" contexts (e.g., American trained teachers tutoring Latin-American third graders in English as a second language) is essential.

Third, the development of standardized tools such as VisSTA [106, 107, 108] and Anvil [132] to visualize and analyze temporally situated multimodal discourse is essential.[4] Since these datasets are necessarily multimedia (time-tagged transcriptions, audio, video, motion traces, etc.), the field will be impeded if every researcher has to develop their own set of these tools.

Fourth, along with the investigation of individual catchment features, there needs to be research in combining ensembles of catchment features and speech. Even within a specific discourse context, the imagistic content of different discourse segments may be represented by different catchments. Different catchment features may properly mark a topical unit or not (leading to a penalty). Research into temporal fusion of multiple features will be essential to the advancement of the field.

[4]The Linguistic Data Consortium has begun the task of cataloging such tools in http://www.ldc.upenn.edu/annotation/gesture/.

6. CONCLUSION

Our list of requirements is not intended to be exhaustive. The field computational multimodal discourse analysis is young and many voices and perspectives are necessary to realize its potential. This paper seeks only to present the basic catchment feature model that permits the fusion of different communicative modes, and bridges what may be reasonably extracted by signal and video processing with the realities of how humans communicate multimodally.

We have argued from a detailed analysis of the literature that the body of engineering research on gesture has heretofore clustered around manipulative and semaphoric gestures. We have motivated our catchment feature model from a sound psycholinguistic basis. The ideas of the growth point and catchment suggest that the locus of fusion across communicative mode may be accomplished at the level of the underlying discourse plan and imagistic conceptualization. The focal points of newsworthy items of an utterance that carries the discourse plan along are precisely where both gesticular activity and prosodic emphases emerge (and merge). Imagery is not conveyed by "whole gestures" but by particular features of the gestures that represent the imagery. Although these gestural canvases are sketched out and painted from discourse to discourse at the moment of conceptualization, they create consistent feature spaces within each discourse. The catchment feature model, hence, serves as a bridge between the discourse conceptualization and the entities that may be extracted from discourse video.

We presented our experimental framework for catchment feature-based research, and visited three examples of catchment features: hand use, symmetry characteristics, and space use to demonstrate the efficacy of the catchment feature model.

Finally, this paper lays out some of the needs of the new domain of computational multimodal discourse analysis. We believe it is in the understanding of how humans communicate multimodally that we can approach multimodal human-computer interaction in a cogent way.

The resulting science has the potential for such breakthrough applications as improved speech recognition by accessing segmentation information in the gestural stream, multimodal transcription and enrichment of multimedia meeting records, study of communicative deficits in such diseases as schizophrenia and Parkinson's disease, and advanced communicative human-computer interfaces.

ACKNOWLEDGMENTS

This research and the preparation of this paper has been supported by the US National Science Foundation (NSF) STIMULATE Program, Grant no. IRI-9618887, "Gesture, Speech, and Gaze in Discourse Segmentation"; NSF KDI Program, Grant no. BCS-9980054, "Cross-Modal Analysis of Signal and Sense: Multimedia Corpora and Tools for Gesture, Speech, and Gaze Research"; NSF ITR program, Grant no. ITR-0219875, "Beyond the Talking Head and Animated Icon: Behaviorally Situated Avatars for Tutoring"; and the Advanced Research and Development Activity (ARDA) VACEII Grant no. 665661, "From Video to Information: Cross-Model Analysis of Planning Meetings." Finally, much appreciation goes to our extended research team, especially David McNeill, a friend and colleague, upon whose psycholinguistic research this work is based.

REFERENCES

[1] D. McNeill, "Growth points, catchments, and contexts," *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, vol. 7, no. 1, pp. 22–36, 2000.

[2] D. McNeill, "Catchments and context: non-modular factors in speech and gesture," in *Language and Gesture*, D. McNeill, Ed., chapter 15, pp. 312–328, Cambridge University Press, Cambridge, UK, 2000.

[3] D. McNeill, F. Quek, K.-E. McCullough, et al., "Catchments, prosody and discourse," in *Oralité et Gestualité, ORAGE 2001 (Speech and Gesture 2001)*, C. Cavé, I. Guaïtella, and S. Santi, Eds., pp. 474–481, Aix-en-Provence, France, June 2001.

[4] F. Quek, D. McNeill, R. Ansari, et al., "Multimodal human discourse: gesture and speech," *ACM Transactions on Computer-Human Interaction*, vol. 9, no. 3, pp. 171–193, 2002.

[5] R. A. Bolt, "Put-that-there," *Computer Graphics*, vol. 14, no. 3, pp. 262–270, 1980.

[6] R. A. Bolt, "Eyes at the interface," in *Proc. ACM CHI Human Factors in Computer Systems Conference*, pp. 360–362, Gaithersburg, Md, USA, March 1982.

[7] J. Crowley, F. Berard, and J. Coutaz, "Finger tracking as an input device for augmented reality," in *Proc. IEEE International Workshop on Automatic Face and Gesture Recognition*, pp. 195–200, IEEE Computer Society, Zurich, Switzerland, June 1995.

[8] R. Kjeldsen and J. Kender, "Toward the use of gesture in traditional user interfaces," in *Proc. 2nd International Conference on Automatic Face and Gesture Recognition (FG '96)*, pp. 151–156, IEEE Computer Society, Killington, Vt, USA, October 1996.

[9] F. Quek, T. Mysliwiec, and M. Zhao, "FingerMouse: a freehand pointing interface," in *Proc. IEEE International Workshop on Automatic Face and Gesture Recognition*, pp. 372–377, IEEE Computer Society, Zurich, Switzerland, June 1995.

[10] M. Fukumoto, K. Mase, and Y. Suenaga, "Real-time detection of pointing actions for a glove-free interface," in *Proc. IAPR Workshop on Machine Vision Applications*, pp. 473–476, Tokyo, Japan, December 1992.

[11] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Gestural interface to a visual computing environment for molecular biologists," in *Proc. 2nd International Conference on Automatic Face and Gesture Recognition (FG '96)*, pp. 30–35, IEEE Computer Society, Killington, Vt, USA, October 1996.

[12] R. Kahn, M. Swain, P. Prokopowicz, and R. Firby, "Gesture recognition using the perseus architecture," in *Proc. 2nd IEEE Conference on Computer Vision and Pattern Recognition*, pp. 734–741, IEEE Computer Society, San Francisco, Calif, USA, June 1996.

[13] J. Segen, "Controlling computers with gloveless gestures," in *Proc. Virtual Reality Systems Conference*, pp. 2–6, New York, NY, USA, March 1993.

[14] P. Wellner, "Interacting with paper on the DigitalDesk," *Communications of the ACM*, vol. 36, no. 7, pp. 87–96, 1993.

[15] J. Segen and S. Kumar, "Fast and accurate 3D gesture recognition interface," in *Proc. 14th International Conference on Pattern Recognition*, vol. 1, pp. 86–91, IEEE Computer Society, Brisbane, Australia, August 1998.

[16] S. Kumar and J. Segen, "Gesture based 3d man-machine interaction using a single camera," in *IEEE International Conference on Multimedia Computing and Systems*, vol. 1, pp. 630–635, IEEE Computer Society, Florence, Italy, June 1999.

[17] J. Segen and S. Kumar, "Shadow gestures: 3D hand pose estimation using a single camera," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 479–485, IEEE Computer Society, Fort Collins, Colo, USA, June 1999.

[18] J. Rehg and T. Kanade, "DigitEyes: vision-based human hand tracking," Tech. Rep. CMU-CS-93-220, Carnegie Mellon University, Pittsburgh, Pa, USA, December 1993.

[19] P. Nesi and D. A. Bimbo, "Hand pose tracking for 3-D mouse emulation," in *Proc. IEEE International Workshop on Automatic Face and Gesture Recognition*, pp. 302–307, IEEE Computer Society, Zurich, Switzerland, June 1995.

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

[21] R. Kadobayashi, K. Nishimoto, and K. Mase, "Design and evaluation of gesture interface of an immersive walk-through application for exploring cyberspace," in *Proc. 3rd International Conference on Automatic Face and Gesture Recognition (FG '98)*, pp. 534–539, IEEE Computer Society, Nara, Japan, April 1998.

[22] R. O'Hagan and A. Zelinsky, "Visual gesture interfaces for virtual environments," in *1st Australasian User Interface Conference*, pp. 73–80, IEEE Computer Society, Canberra, Australia, January–February 2000.

[23] A. Wexelblat, "An approach to natural gesture in virtual environments," *ACM Transactions on Computer-Human Interaction*, vol. 2, no. 3, pp. 179–200, 1995.

[24] A. Azarbayejani, T. Starner, B. Horowitz, and A. Pentland, "Visually controlled graphics," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 602–605, 1993.

[25] A. G. Hauptmann, "Speech and gestures for graphic image manipulation," in *Proc. Conference on Human Factors in Computing Systems (CHI '89)*, pp. 241–245, IEEE Computer Society, New York, NY, USA, May 1989.

[26] W. T. Freeman and C. Weissman, "Television control by hand gestures," in *Proc. IEEE International Workshop on Automatic Face and Gesture Recognition*, pp. 179–183, IEEE Computer Society, Zurich, Switzerland, June 1995.

[27] W. T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma, "Computer vision for computer games," in *Proc. 2nd International Conference on Automatic Face and Gesture Recognition (FG '96)*, pp. 100–105, IEEE Computer Society, Killington, Vt, USA, October 1996.

[28] W. T. Freeman, "Computer vision for television and games," in *Proc. International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, p. 118, Corfu, Greece, September 1999.

[29] W. T. Freeman, D. Anderson, P. Beardsley, et al., "Computer vision for interactive computer graphics," *IEEE Computer Graphics and Applications*, vol. 18, no. 3, pp. 42–53, 1998.

[30] M. B. Friedman, "Gestural control of robot end effectors," in *Proc. SPIE Conference on Intelligent Robots and Computer Vision*, vol. 726, pp. 500–502, Bellingham, Wash, USA, January 1986.

[31] A. Katkere, E. Hunter, D. Kuramura, J. Schlenzig, S. Moezzi, and R. Jain, "Robogest: Telepresence using hand gestures," Tech. Rep. VCL-94-104, Visual Computing Laboratory, University of California, San Diego, Calif, USA, December 1994.

[32] J. Triesch and C. von der Malsburg, "Robotic gesture recognition," in *Proc. International Gesture Workshop: Gesture and Sign Language in Human Computer Interaction*, I. Wachsmuth and M. Fröhlich, Eds., pp. 233–244, Springer, Bielefeld, Germany, September 1997.

[33] J. Triesch and C. von der Malsburg, "A gesture interface for human-robot-interaction," in *Proc. 3rd IEEE International Conference on Automatic Face and Gesture Recognition (FG '98)*, pp. 546–551, IEEE Computer Society, Nara, Japan, April 1998.

[34] S. Bryson and C. Levit, "The virtual wind tunnel: An environment for the exploration of three-dimensional unsteady flows," in *Proc. IEEE Visualization '91*, pp. 17–24, San Diego, Calif, USA, October 1991.

[35] S. Bryson and C. Levit, "The virtual wind tunnel," *IEEE Computer Graphics and Applications*, vol. 12, no. 4, pp. 25–34, 1992.

[36] S. S. Fels and G. E. Hinton, "Glove-talk: A neural network interface between a data-glove and a speech synthesizer," *IEEE Transactions on Neural Networks*, vol. 4, no. 1, pp. 2–8, 1993.

[37] S. S. Fels, *Glove-TalkII: mapping hand gestures to speech using neural networks—an approach to building adaptive interfaces*, Ph.D. thesis, University of Toronto, Toronto, Canada, August 1994.

[38] R. Pausch and R. Williams, "Giving candy to children: user-tailored gesture input driving an articulator-based speech synthesizer," *Communications of the ACM*, vol. 35, no. 5, pp. 58–66, 1992.

[39] R. Cutler and M. Turk, "View-based interpretation of real-time optical flow for gesture recognition," in *Proc. 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 416–421, IEEE Computer Society, Nara, Japan, April 1998.

[40] A. D. Wilson and A. F. Bobick, "Nonlinear PHMMs for the interpretation of parameterized gesture," in *Proc. 3rd IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 879–884, IEEE Computer Society, Santa Barbara, Calif, USA, June 1998.

[41] A. D. Wilson and A. F. Bobick, "Parametric hidden Markov models for gesture recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 884–900, 1999.

[42] F. Quek, "Eyes in the interface," *Image and Vision Computing*, vol. 13, no. 6, pp. 511–525, 1995.

[43] F. Quek, "Unencumbered gestural interaction," *IEEE Multimedia*, vol. 4, no. 3, pp. 36–47, 1996.

[44] X. Zhu, J. Yang, and A. Waibel, "Segmenting hands of arbitrary color," in *Proc. 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 446–453, Grenoble, France, March 2000.

[45] E. Hunter, J. Schlenzig, and R. Jain, "Posture estimation in reduced-model gesture input systems," in *Proc. IEEE International Workshop on Automatic Face and Gesture Recognition*, pp. 290–295, IEEE Computer Society, Zurich, Switzerland, June 1995.

[46] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *Proc. IEEE International Workshop on Automatic Face and Gesture Recognition*, pp. 296–301, IEEE Computer Society, Zurich, Switzerland, June 1995.

[47] J. Triesch and C. von der Malsburg, "Robust classification of hand postures against complex backgrounds," in *Proc. 2nd International Conference on Automatic Face and Gesture Recognition (FG '96)*, pp. 170–175, IEEE Computer Society, Killington, Vt, USA, October 1996.

[48] A. Lanitis, C. J. Taylor, T. F. Cootes, and T. Ahmed, "Automatic interpretation of human faces and hand gestures using flexible models," in *Proc. IEEE International Workshop on Automatic Face and Gesture Recognition*, pp. 98–103, IEEE Computer Society, Zurich, Switzerland, June 1995.

[49] C. Maggioni, "Gesturecomputer—new ways of operating a computer," in *Proc. IEEE International Workshop on Automatic Face and Gesture Recognition*, pp. 166–171, IEEE Computer Society, Zurich, Switzerland, June 1995.

[50] K. Munk and E. Granum, "On the use of context and a priori knowledge in motion analysis for visual gesture recognition," in *Proc. International Gesture Workshop: Gesture and Sign Language in Human-Computer Interaction*, I. Wachsmuth and M. Fröhlich, Eds., pp. 123–134, Springer, Bielefeld, Germany, September 1997.

[51] H.-J. Boehme, A. Brakensiek, U.-D. Braumann, M. Krabbes, and H.-M. Gross, "Neural architecture for gesture-based human-machine-interaction," in *Proc. International Gesture Workshop: Gesture and Sign Language in Human Computer Interaction*, I. Wachsmuth and M. Fröhlich, Eds., pp. 219–232, Springer, Bielefeld, Germany, September 1997.

[52] M. J. Black and A. D. Jepson, "Recognizing temporal trajectories using the condensation algorithm," in *Proc. 3rd IEEE International Conference on Face and Gesture Recognition*, pp. 16–21, IEEE Computer Society, Nara, Japan, April 1998.

[53] F. Quek, "Hand gesture interface for human-machine interaction," in *Proc. Virtual Reality Systems Conference*, pp. 13–19, New York, NY, USA, March 1993.

[54] F. Quek, "Toward a vision-based hand gesture interface," in *Proc. Conference on Virtual Reality Software and Technology*, pp. 17–29, Singapore, Singapore, August 1994.

[55] F. Quek and M. Zhao, "Inductive learning in hand pose recognition," in *Proc. 2nd International Conference on Automatic Face and Gesture Recognition (FG '96)*, pp. 78–83, IEEE Computer Society, Killington, Vt, USA, October 1996.

[56] M. Zhao, F. Quek, and X. Wu, "RIEVL: recursive induction learning in hand gesture recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1174–1185, 1998.

[57] A. D. Wilson and A. F. Bobick, "Configuration states for the representation and recognition of gesture," in *Proc. IEEE International Workshop on Automatic Face and Gesture Recognition*, pp. 129–134, IEEE Computer Society, Zurich, Switzerland, June 1995.

[58] C. J. Cohen, L. Conway, and D. Koditschek, "Dynamical system representation, generation, and recognition of basic oscillatory motion gestures," in *Proc. 2nd International Conference on Automatic Face and Gesture Recognition (FG '96)*, pp. 60–65, IEEE Computer Society, Killington, Vt, USA, October 1996.

[59] T. J. Darrell and A. P. Pentland, "Space-time gestures," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR '93)*, pp. 335–340, New York, NY, USA, June 1993.

[60] T. J. Darrell, I. A. Essa, and A. P. Pentland, "Task-specific gesture analysis in real-time using interpolated views," Tech. Rep. 364, MIT Media Laboratory Vision and Modeling Group, Mass, USA, 1995.

[61] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[62] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov models," in *Proc. IEEE Computer Vision and Pattern Recognition*, pp. 379–385, Champaign, Ill, USA, June 1992.

[63] F. G. Hofmann, P. Heyer, and G. Hommel, "Velocity profile based recognition of dynamic gestures with discrete hidden Markov models," in *Proc. International Gesture Workshop: Gesture and Sign Language in Human Computer Interaction*, I. Wachsmuth and M. Fröhlich, Eds., pp. 81–95, Springer, Bielefeld, Germany, September 1997.

[64] M. Assan and K. Grobel, "Video-based sign language recognition using hidden Markov models," in *Proc. International Gesture Workshop: Gesture and Sign Language in Human-Computer Interaction*, I. Wachsmuth and M. Fröhlich, Eds., pp. 97–109, Springer, Bielefeld, Germany, September 1997.

[65] A. D. Wilson and A. F. Bobick, "Recognition and interpretation of parametric gesture," Tech. Rep. 421, MIT Media Laboratory Vision and Modeling Group, Mass, USA, 1998.

[66] J. Schlenzig, E. Hunter, and R. Jain, "Recursive identification of gesture inputs using hidden Markov models," in *Proc. 2nd IEEE Workshop on Applications of Computer Vision*, pp. 187–194, Sarasota, Fla, USA, December 1994.

[67] J. Schlenzig, E. Hunter, and R. Jain, "Vision based hand gesture interpretation using recursive estimation," in *28th Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1267–1271, Pacific Grove, Calif, USA, October-November 1994.

[68] G. Rigoll, A. Kosmala, and S. Eickeler, "High performance real-time gesture recognition using hidden Markov models," in *Proc. International Gesture Workshop: Gesture and Sign Language in Human Computer Interaction*, I. Wachsmuth and M. Fröhlich, Eds., pp. 69–80, Springer, Bielefeld, Germany, September 1997.

[69] Y. Iwai, H. Shimizu, and M. Yachida, "Real-time context-based gesture recognition using HMM and automaton," in *Proc. International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pp. 127–134, Corfu, Greece, September 1999.

[70] H.-K. Lee and J. H. Kim, "An HMM-based threshold model approach for gesture recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 961–973, 1999.

[71] A. F. Bobick and Y. A. Ivanov, "Action recognition using probabilistic parsing," Tech. Rep. 449, MIT Media Laboratory Vision and Modeling Group, Mass, USA, 1998.

[72] J. Martin and J.-B. Durand, "Automatic handwriting gestures recognition using hidden Markov models," in *Proc. 4th IEEE International Automatic Face and Gesture Recognition*, pp. 403–409, Grenoble, France, March 2000.

[73] A. Wexelblat, "Research challenges in gesture: open issues and unsolved problems," in *Proc. International Gesture Workshop: Gesture and Sign Language in Human Computer Interaction*, pp. 1–11, Springer, Bielefeld, Germany, September 1997.

[74] P. Cohen, M. Dalrymple, D. Moran, et al., "Synergistic use of direct manipulation and natural language," in *Human Factors in Computing Systems: CHI'89 Conference Proceedings*, pp. 227–234, ACM, Addison Wesley Publishing, Austin, Tex, USA, April 1989.

[75] P. Cohen, M. Dalrymple, D. Moran, et al., "Synergistic use of direct manipulation and natural language," in *Readings in Intelligent User Interfaces*, M. Maybury and W. Wahlster, Eds., pp. 227–234, Morgan Kaufmann Publishers, New York, NY, USA, April 1998.

[76] J. G. Neal, C. Y. Thielman, Z. Dobes, S. M. Haller, and S. C. Shapiro, "Natural language with integrated deictic and graphic gestures," in *Readings in Intelligent User Interfaces*, M. Maybury and W. Wahlster, Eds., pp. 38–51, Morgan Kaufmann Publishers, New York, NY, USA, July 1998.

[77] S. Oviatt, A. DeAngeli, and K. Kuhn, "Integration and synchronization of input modes during multimodal human-computer interaction," in *Proc. Conference on Human Factors in Computing Systems (CHI '97)*, vol. 1, pp. 415–422, Atlanta, Ga, USA, March 1997.

[78] S. Oviatt and P. Cohen, "Multimodal interfaces that process what comes naturally," *Communications of ACM*, vol. 43, no. 3, pp. 43–53, 2000.

[79] S. Oviatt, "Mutual disambiguation of recognition errors in a multimodal architecture," in *Proc. Conference on Human Factors in Computing Systems (CHI '99)*, vol. 1, pp. 576–583, New York, NY, USA, May 1999.

[80] F. Quek, R. Yarger, Y. Haciahmetoglu, et al., "Bunshin: A believable avatar surrogate for both scripted and on-the-fly pen-based control in a presentation environment," in *Emerging Technologies, SIGGRAPH 2000*, New Orleans, La, USA, July 2000.

[81] D. Koons, C. Sparrell, and K. Thorisson, "Integrating simultaneous input from speech, gaze, and hand gestures," in *Intelligent Multimedia Interfaces*, M. Maybury, Ed., pp. 53–62, American Association for Artificial Intelligence, MIT Press, 1998.

[82] A. D. Wilson, A. F. Bobick, and J. Cassell, "Recovering the temporal structure of natural gesture," in *Proc. 2nd International Conference on Automatic Face and Gesture Recognition (FG '96)*, pp. 66–71, IEEE Computer Society, Killington, Vt, USA, October 1996.

[83] D. Franklin, R. E. Kahn, M. J. Swain, and R. J. Firby, "Happy patrons make better tippers creating a robot waiter using perseus and the animate agent architecture," in *Proc. 2nd International Conference on Automatic Face and Gesture Recognition (FG '96)*, pp. 253–258, IEEE Computer Society, Killington, Vt, USA, October 1996.

[84] T. Sowa and I. Wachsmuth, "Coverbal iconic gestures for object descriptions in virtual environments: an empirical study," in *Post-Proceedings of the Conference of Gestures: Meaning and Use*, Porto, Portugal, April 2000.

[85] T. Sowa and I. Wachsmuth, "Understanding coverbal dimensional gestures in a virtual design environment," in *Proc. ESCA Workshop on Interactive Dialogue in Multi-Modal Systems*, P. Dalsgaard, C.-H. Lee, P. Heisterkamp, and R. Cole, Eds., pp. 117–120, Kloster Irsee, Germany, June 1999.

[86] S. Prillwitz, R. Leven, H. Zienert, T. Hanke, and J. Henning, *Hamburg Notation System for Sign Languages—An Introductory Guide*, Signum Press, Hamburg, Germany, 1989.

[87] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*, University of Chicago Press, Chicago, Ill, USA, 1992.

[88] A. Kendon, "Some relationships between body motion and speech: An analysis of an example," in *Studies in Dyadic Communication*, A. Siegman and B. Pope, Eds., pp. 177–210, Pergamon Press, Elmsford, NY, USA, 1972.

[89] A. Kendon, "Gesticulation and speech: Two apsects of the process of utterance," in *Relationship Between Verbal and Nonverbal Communication*, M. Key, Ed., pp. 207–227, Mouton Press, The Hague, Paris, France, 1980.

[90] A. Kendon, "Current issues in the study of gesture," in *The Biological Foundations of Gestures: Motor and Semiotic Aspects*, J.-L. Nespoulous, P. Perron, and A. R. Lecours, Eds., pp. 23–47, Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1986.

[91] D. McNeill and S. Duncan, "Growth points in thinking-for-speaking," in *Language and Gesture*, D. McNeill, Ed., chapter 7, pp. 141–161, Cambridge University Press, Cambridge, UK, 2000.

[92] L. Vygotsky, "Thinking and speaking," in *The Collected Works of L. S. Vygotsky, Problems of General Psychology (translated by N. Minnick)*, R. W. Rieber and A. S. Carton, Eds., vol. 1, pp. 39–285, Plenum Press, New York, NY, USA, 1987.

[93] M. Halliday, *Intonation and Grammar in British English*, Mouton Press, The Hague, Paris, France, 1967.

[94] E. Hajicová and P. Sgall, "Topic and focus of a sentence and the patterning of a text," in *Text and Discourse Constitution*, J. S. Petöfi, Ed., pp. 70–96, De Gruyter, Berlin, Germany, 1988.

[95] J. Cassell, M. Stone, B. Douville, et al., "Modeling the interaction between speech and gesture," in *Proc. 16th Annual Conference of the Cognitive Science Society*, A. Ram and K. Eiselt, Eds., pp. 153–158, Lawrence Erlbaum Associates, Atlanta, Ga, USA, 1994.

[96] C. Pelachaud and S. Prevost, "Coordinating vocal and visual parameters for 3d virtual agents," in *Proc. 2nd Eurographics Workshop on Virtual Environments*, M. Göbel, Ed., pp. 99–106, Monte Carlo, Monaco, January 1995.

[97] F. Quek, D. McNeill, R. Ansari, et al., "Multimodal human discourse: gesture and speech," Tech. Rep. VISLab-01-01, VISLab, Wright State University, Dayton, Ohio, USA, September 2002.

[98] F. Quek and R. Bryll, "Vector coherence mapping: a parallelizable approach to image flow computation," in *Proc. Asian Conference on Computer Vision*, vol. 2, pp. 591–598, Hong Kong, China, January 1998.

[99] F. Quek, X.-F. Ma, and R. Bryll, "A parallel algorithm for dynamic gesture tracking," in *Proc. ICCV '99 International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (RATFG-RTS '99)*, pp. 64–69, Corfu, Greece, September 1999.

[100] R. Bryll and F. Quek, "Accurate tracking by vector coherence mapping and vector-centroid fusion," submitted to International Journal of Computer Vision.

[101] B. Zarit, B. Super, and F. Quek, "Comparison of five color models in skin pixel classification," in *Proc. ICCV '99 International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real Time Systems (RATFG-RTS '99)*, pp. 58–63, Corfu, Greece, September 1999.

[102] F. Quek, R. McNeill, D. Bryll, et al., "Gesture, speech, and gaze cues for discourse segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, vol. 2, pp. 247–254, Hilton Head Island, SC, USA, June 2000.

[103] R. Bryll, F. Quek, and A. Esposito, "Automatic hand hold detection in natural conversation," in *IEEE Workshop on Cues in Communication*, Kauai, Hawaii, USA, December 2001.

[104] P. Boersma and D. Weenik, "Praat, a system for doing phonetics by computer," Tech. Rep. 132, version 3.4., Institute of Phonetic Sciences, University of Amsterdam, Amsterdam, The Netherlands, 1996, http://www.fon.hum.uva.nl/praat/.

[105] C. Nakatani, B. Grosz, D. Ahn, and J. Hirschberg, "Instructions for annotating discourse," Tech. Rep. TR-21-95, Center for Research in Computing Technology, Harvard University, Cambridge, Mass, USA, 1995.

[106] F. Quek, R. Bryll, C. Kirbas, H. Arslan, and D. McNeill, "A multimedia system for temporally situated perceptual psycholinguistic analysis," *Multimedia Tools and Applications*, vol. 18, no. 2, pp. 91–114, 2002.

[107] F. Quek and D. McNeill, "A multimedia system for temporally situated perceptual psycholinguistic analysis," in *Proc. 3rd International Conference on Methods and Techniques in Behavioral Research, Measuring Behavior 2000*, p. 257, Nijmegen, The Netherlands, August 2000.

[108] F. Quek, Y. Shi, C. Kirbas, and S. Wu, "VisSTA: A tool for analyzing multimodal discourse data," in *7th International Conference on Spoken Language Processing*, vol. 1, pp. 221–224, Denver, Colo, USA, September 2002.

[109] D. McNeill et al., "Catchments, prosody and discourse," *Gesture*, vol. 1, no. 1, pp. 9–33, 2002.

[110] D. McNeill et al., "Dynamic imagery in speech and gesture," in *Multimodality in Language and Speech Systems*, B. Granström, Ed., pp. 27–44, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

[111] F. Quek, D. McNeill, R. Ansari, et al., "Gesture cues for conversational interaction in monocular video," in *Proc. ICCV '99 International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real Time Systems (RATFG-RTS '99)*, pp. 119–126, Corfu, Greece, September 1999.

[112] S. Kita, I. van Gijn, and H. van der Hulst, "Movement phases in signs and co-speech gestures, and their transcription by human coders," in *Proc. International Gesture Workshop: Gesture and Sign Language in Human Computer Interaction*, I. Wachsmuth and M. Fröhlich, Eds., pp. 23–35, Springer, Bielefeld, Germany, September 1997.

[113] I. van Gijn, S. Kita, and H. van der Hulst, "How phonological is the symmetry condition in sign language," in *Proc. 4th Holland Institute of Linguistics Phonology Conference*, Leiden, The Netherlands, January 1999.

[114] F. Quek, Y. Xiong, and D. McNeill, "Gestural trajectory symmetries and discourse segmentation," in *Proc. 7th International Conference on Spoken Language Processing*, vol. 1, pp. 185–188, Denver, Colo, USA, September 2002.

[115] Y. Xiong, F. Quek, and D. McNeill, "Hand gesture symmetric detection and analysis in natural conversation," in *Proc. 4th IEEE International Conference on Multimodal Interfaces (ICMI '02)*, pp. 179–184, Pittsburgh, Pa, USA, October 2002.

[116] G. Fauconnier, *Mental Spaces: Aspects of Meaning Construction in Natural Language*, MIT Press, Cambridge, Mass, USA, 1985.

[117] E. Sweetser, "Blended spaces and performativity," *Cognitive Linguistics*, vol. 11, no. 3/4, pp. 305–333, 2000.

[118] K.-E. McCullough, "Visual imagery in language and gesture," in *Annual Meeting of the Belgian Linguistic Society*, D. McNeill, Ed., Brussels, Belgium, November 1992.

[119] C. Büehler, "The deictic field of language and deictic words," in *Speech, Place, and Action*, R. J. Jarvella and W. Klein, Eds., pp. 9–30, John Wiley & Sons, London, UK, 1982.

[120] F. Quek, D. McNeill, R. Bryll, and M. Harper, "Gesture spatialization in natural discourse segmentation," in *Proc. 7th International Conference on Spoken Language Processing*, vol. 1, pp. 189–192, Denver, Colo, USA, September 2002.

[121] F. Quek, R. Bryll, D. McNeill, and M. Harper, "Gestural origo and loci-transitions in natural discourse segmentation," in *IEEE Workshop on Cues in Communication*, Kauai, Hawaii, December 2001.

[122] L. Chen, M. Harper, and F. Quek, "Gesture patterns during speech repairs," in *IEEE International Conference on Multimodal Interaction*, pp. 155–160, Pittsburg, Pa, USA, October 2002.

[123] A. Esposito, K.-E. McCullough, and F. Quek, "Disfluencies in gesture: Gestural correlates to filled and unfilled speech pauses," in *Proc. of IEEE Workshop on Cues in Communication*, Kauai, Hawaii, December 2001.

[124] L. Valbonesi, R. Ansari, D. McNeill, et al., "Analysis of speech and gestures: gesture frequency during fluent and hesitant phases in speech," in *6th Multi Conference on Systemics, Cybernetics and Informatics (SCI '02)*, Orlando, Fla, USA, July 2002.

[125] L. Valbonesi, R. Ansari, D. McNeill, et al., "Temporal correlation of speech and gestures focal points," in *Gesture: The Living Medium Conference*, Austin, Tex, USA, June 2002.

[126] L. Valbonesi, R. Ansari, D. McNeill, et al., "Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures," in *Proc. European Signal Processing Conference (EUSIPCO' 02)*, Tolouse, France, September 2002.

[127] F. Quek, M. Harper, Y. Haciahmetoglu, L. Chen, and L. Ramig, "Speech pauses and gestural holds in Parkinson's disease," in *7th International Conference on Spoken Language Processing*, vol. 4, pp. 2485–2488, Denver, Colo, USA, September 2002.

[128] F. Quek, R. Bryll, M. Harper, L. Chen, and L. Ramig, "Audio and vision-based evaluation of Parkinson's disease from discourse video," in *IEEE International Symposium of Bio-Informatics and Bio-Engineering*, pp. 246–253, Bethesda, Mass, USA, November 2001.

[129] F. Quek, R. Bryll, M. Harper, L. Chen, and L. Ramig, "Speech and gesture analysis for evaluation of progress in LSVT treatment in Parkinson's disease," in *11th Biennial Conference on Motor Speech: Motor Speech Disorders*, Williamsburg, Va, USA, March 2002.

[130] R. Bryll, R. Gutierrez-Osuna, and F. Quek, "Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets," *Pattern Recognition*, vol. 36, no. 6, pp. 1291–1302, 2003.

[131] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, no. 4857, pp. 1285–1293, 1988.

[132] M. Kipp, "Anvil: Annotation of video and spoken language," http://www.dfki.de/~kipp/anvil/.

Francis Quek is the Director of the Center for Human Computer Interaction (CHCI), and Professor of computer science at Virginia Polytechnic Institute and State University. He also directs the Vision Interfaces and Systems Laboratory at the CHCI. Francis received both his B.S.E. degree (summa cum laude) in 1984 and M.S.E. degree in 1984 in electrical engineering from the University of Michigan. He completed his Ph.D. in computer science and engineering at the same university in 1990. He also has a Technician's Diploma in electronics and communications engineering from the Singapore Polytechnic in 1978. Francis is a Member of the IEEE and ACM. He performs research in multimodal verbal/nonverbal interaction, vision-based interaction, multimedia databases, medical imaging, collaboration technology, human-computer interaction, computer vision, and computer graphics. He leads several multiple-disciplinary research efforts to understand the communicative realities of multimodal interaction.