Gesture is not just pointing

Lisa D. Harper, Daniel P. Loehr, Anthony J. Bigbee
The MITRE Corporation
1820 Dolley Madison Boulevard
McLean, Virginia, USA, 22102
{lisah, loehr, abigbee}@mitre.org

Abstract

In a preliminary study of 3D multi-player wargame interactions, we observed that gesture is used for much more than the simple deixis handled by most current multimodal systems. Multimodal referring expressions are found to convey both semantic and pragmatic information. Bi-directional multimodal systems concerned with referential coherence in discourse must take into account the role and function of gesture in multimodal communication.

Introduction

Most systems today that enable multimodal (language and gesture) input emphasize commands and requests in a very restricted domain of application. Furthermore, the types of gestures that these systems interpret are, primarily, deictic pointing gestures (e.g., Neal *et al.* 1989, Binot 1992, Huls *et al.* 1995). The usefulness of deictic pointing in natural language generation has also received considerable attention (e.g., Kobsa 1986, Neal *et al.* 1989, Wahlster 1991).

More recently, some multimodal systems have extended the notion of gesture beyond deictic pointing. Koons *et al.* (1993) describe a system that has the ability to interpret *iconic* gestures. Iconic gestures represent both an object and some attribute of shape or motion. By using relative hand position and orientation users can position and orient objects in a graphical scene such as ßlt the planet like *this.*ÓJohnston *et al.* (1997) describe the QuickSet multimodal system This system integrates speech and pen gestures in dynamic interaction with maps and other visual displays. QuickSet allows a series of gestures of different types such as occurs when a user circles an entity on the map, and utters "follow this routeÓ and draws an arrow indicating the route to be followed. Also, complex gestures are possible that communicate not only an entity type but also to specific temporal or spatial properties of an entity. For example, users can create entities with orientation or movement properties such as start and stop times. Both of these systems extend the notion of entity reference by enabling a richer notion of gesture communication than had been realized in earlier systems. However, observation of human-human communication in richly visual environments indicates that gesture interacts with language in even more complex ways (for example, McNeill 1992, Kendon 1972). Cassell (in press) distinguishes spontaneous body movements associated with spoken language from more conscious gestures used in the manipulation of devices in human-computer interfaces. We are interested in parallels between these two uses of gesture. The larger question this research is engaged with is how we may use gesture to convey information in visual output.

1 Current Study

This paper describes research into how gesture and speech are used together. One goal of this program is to generate automated visualizations that will improve a military commanderÕs situation awareness and help her convey her situation understanding and intent more effectively. Current visual displays are static and describe discrete moments in time. They do not effectively communicate the fluidity of events in a battlespace. We are engaged in a process of examining how people interpret and

understand events in collaborative war game tasks; war games refers to role playing military decisionmakers while executing a battle using a common physical or simulation model. In an ongoing series of learning and data collection events, human participants play cooperative roles side-by-side around a large 10Õx10Õx2Õ three-dimensional model city. The game playing is much like ÒdungeonsÓ where a facilitator interactively describes what is happing to individual gamers. The gamers are responsible for making decisions and acting on them both individually and collaboratively. In order to manage the entire game, the commander in charge requests frequent situation updates from the subordinate players. From examining our video footage weÕve observed that this type of role-playing involves story-telling from the perspective of multiple roles. Players play their own role but are also asked to relay their perceptions from the point of view of the enemy they are engaging.

Our ultimate goal is to identify elements of the shared situation that can be represented and communicated in automatically generated visualizations. These visualizations should enable human participants not in the actual environment to visualize and understand events as described by participants in that battlespace. In after-action discussions, exercise participants have noted how difficult it is to communicate what is happening to those participants not working around the 3D model. They relied very heavily on the use of definite descriptions in combination with gesture to refer to entities and events. The question we are addressing is what techniques do people use in face-to-face collaborative problem-solving that can be employed in a dialogue system that produces primarily visual output. What we are doing differently from the embodied agent community is attempting to see how we can enable a system to communicate using a combination of speech, gesture, and visual presentation without necessarily using embodied agents.

In the rest of this paper we describe some initial observations in regard to how people use combinations of language and gesture to talk about entities in an immediate and shared visual context. To this aim, we will provide examples of some *kinds* of multimodal referring expressions that occur in our corpus. We will follow with a presentation of some preliminary results on a small test study concerned with relations between intonation and gesture in regard to narrative perspective and referential coherence. Though we are engaged in a more careful analysis of our video corpus of wargame interactions on other levels, we are not yet ready to present any formal results or analyses at this time.

2 Multimodal Referring Expressions

In multimodal communication, where language communication and direct manipulation paradigms are mixed, it is evident that speakers may use gesture to communicate different types of information to a listener. For example, gesture may be used to:

- **Identify an object, event, or property or group of objects, events, or properties** (e.g., *deictic* gesture[1] as in Òwhat are these red dots?Ó<pointing or encircling gesture>)
- **Introduce an object, event, or property** (e.g., *iconic* gesture as in Òmake the car move like thisÓwhile drawing an arrow gesture pointing the direction)
- **Perform an action** (e.g., directly manipulate an object as in a *pantomimic* gesture: — Òdraw the house like thisÓ; *propositional gestures* as in Òmove this <point to an object> over there <point to a location>; or *symbolic/emblematic* gesture as in ÒdeleteÓwhere a speaker marks a cross over a graphical icon)
- **Perform a focusing function** (e.g., user gestures to select or highlight one ore more graphical elements)
- **Draw attention to something** (e.g., Òdo you know what the green dots are?Ówhile indicating gesturally to a graphical instance)

[1] The italicized gesture types in these examples are drawn from a scheme described Rim and Schiaratura (1991) and McNeill (1992) in a classification of hand gestures.

These activities are not mutually exclusive. A gesture may be part of an ostensive referring act (deictic pointing toward a particular referent), for example, but may also perform a focusing function.

Because we are working in a 3D face-to-face paradigm, we also observe gestures not associated with direct manipulation. For example, as people relay a "story" of what is happening, players use gestures in space in various ways. For example, players frequently use *metaphoric* gestures to describe elements of a plan (e.g., The next phase <accompanied by a vertical metaphoric gesture>). Similarly, we also see the use of *beat* gestures (small baton-like movements in space) that serve as indicators for a shift in narrative point-of-view. This is certainly bears on referential coherence in discourse and will be discussed in greater detail below.

Many multimodal researchers give motivations why multiple modes might be more desirable than a single mode (Cohen 1992, Martin 1998, Hauptmann and MacAvinney 1993, Oviatt 1996). Individually, each mode has different strengths and weaknesses[2]. Potentially, multiple modes allow users to take advantage of the strengths of each mode while providing mechanisms for overcoming the weakness of each. Though multimodal referring expressions may convey complementary semantic information, perceptual constraints affect the form of an utterance and, in turn, affect discourse context.

2.1 Ambiguity and Context

Gesture interpreted in and of itself may be ambiguous. Wahlster (1998) notes that even simple deictic pointing can lead to ambiguity. Since pointing is fundamentally ambiguous without contextual information, a user model and a discourse model are used for the interpretation of ambiguous pointing gestures. For example, graphical objects can offer ambiguity in terms of the level of granularity of reference. Users working in a graphical environment and selecting a point on a particular location on the screen could intend one of the following:

- reference to a particular location on the screen (coordinates of the pixel)
- reference to a particular graphical icon on which the click occurred
- reference to a group of objects if the icon is part of a perceptual grouping
- reference to a particular type exemplified by the selected object (e.g., "the green ones")
- reference to an arbitrary exemplar (ambiguously a definite or indefinite description)

Furthermore, gestural acts may contribute to ambiguity of reference (e.g., the user points successively to two or three green objects and say "the green ones" but, in fact, intends to refer to all of the green ones visible on the display including ones referred to by pointing.) Gesture may be used to simplify verbal referring expressions but at the same time, may introduce ambiguities that can only be resolved by contextual information. For example, if someone says "you go that way" and "I go this way", this person may be indicating a general direction. However, this person may be referring to a plan that indicates specific routes in each direction. The only way to interpret this gesture in context is to have information about the shared plan.

2.2 Discourse Context

There appears to be no simple mapping between the form of a linguistic expression and what the speaker intends to communicate (as characterized as discourse plans). A dialogue system that is able to both interpret and generate multimodal referring expressions needs a discourse model that is informed about saliency and focus of attention. However, current models of discourse focus do not take into account

[2] For example, language facilitates complex queries with the ability to express quantification, attribute and object relations, negation, counterfactuals, categorization, ordering, and aggregate operations. Gesture is more natural for manipulating spatial properties of objects (size, shape, and placement) in graphical environments.

visual attentional processes. As with traditional dialogue systems, multimodal discourse processing must be support two related tasks: 1) identify of a particular referent and 2) realize the form of referring expression. In order to identify a referent by making reference to perceptual features, the discourse model must have access to some model of the display. Realizing the form of a referring expression depends crucially on knowing about affordances and characteristics of the environment (De Angeli *et al.* 1998).

Multimodal communication poses new challenges not traditionally encountered in speech or text processing. We must now be able to recognize, identify and produce well-formed multimodal (or cross-media) referring expressions. This is made complicated in several ways:

- Gestural elements in multimodal referring expressions may participate in multiple goals or contribute information at different abstract levels of interpretation (e.g., semantic, pragmatic);
- Even in regard to demonstrative reference, there is not necessarily a direct mapping between an intended referent and the thing being pointed at;
- Furthermore, visual cues have an influence on how verbal and gestural information is linked in referring expressions.

3 Multimodal Pragmatic Expressions

In addition to the multimodal referring expressions discussed above, language users also produce multimodal *pragmatic* expressions, in which they make use of and change the evolving discourse context. Two modalities previously studied (separately) in light of their contribution to discourse structure are beat gestures and intonation. In our preliminary study, we looked at how these two modalities function *together* in relation to discourse structure. We have discovered that beat gestures and intonation provide complementary information for the listener regarding the interpretation of the utterance in context.

3.1 Beat Gestures and Discourse Structure

Of the gesture types identified by researchers such as McNeill (1992) and Rim and Schiaratura (1991), most carry propositional meaning. An exception is the beat gesture (cf McNeill 1992), so named because it occurs with the "rhythm" of the speech, on stressed syllables. While other, perhaps less spontaneous, gestures typically have three phases (an optional preparation, a main stroke, and an optional retraction), beat gestures are small, two-phase flicks of the wrist or fingers. They have also been called "baton" gestures, due to the impression that they are conducting the rhythmic flow of speech.

Instead of carrying propositional meaning, beat gestures are "extranarrative". As McNeill explains (crediting insights by Cassell and others), speakers can operate on one of at least three levels. The first is "narrative", in which the speaker is relaying simple, chronological events of the story line itself. The second is "metanarrative", in which the speaker explicitly convey information about the structure of the narrative, by talking "about" the story. Examples may include the introduction to the story, or backtracking to fill in previous events. Finally, in the "paranarrative" role, the speaker steps out of the storytelling role altogether, and interacts as fellow conversant with her listeners. The specific contribution of beat gestures, according to McNeill, is to signal that the speaker is shifting from one of these narrative levels to another. The utterance doing the "shifting" may contain several beats, sprinkled throughout its stressed syllables.

The contribution of gestures to discourse structure, then, is the use of beats to signal a "narrative level" shift in relation to the discourse.

3.2 Intonation and Discourse Structure

The contribution of intonation to discourse structure has been analyzed by Pierrehumbert and Hirschberg (1994) (hereafter PH94). The

idea is based on Pierrehumbert's description of intonational contours as consisting of high and low tones, combining to form the full intonational melody over an utterance. The full theory will not be described here. Rather, we will note the contribution of two "pitch accents" (tonal movements attached to a stressed syllable), a simple high (H*) and a simple low (L*) (the asterisk denotes that the tone is aligned with the stressed syllable).

PH94 argue that intonation over some entity is used to describe the *relationship* between the entity and some other entity or entities already in the discourse context. To use specific examples relevant to our study, an H* (a simple high pitch accent) indicates that the accented entity is *new*, in relation to what is already in the discourse context, while an L* signifies that the accented entity is *not* new to the discourse. An analogy is the use of the indefinite ("a bus") to introduce a new entity, versus the definite ("the bus") to refer an existing entity.

The contribution of intonation to discourse, then, is to signify the relationship (in our example, *new* or *not new*) of the accented item with the discourse context.

3.3 Beat Gestures, Intonation, and Discourse Structure

Having described the separate contributions of both gesture and intonation to discourse, we now describe our informal investigation as to how they are used together. We analysed a small section of our videotapes to locate the beats used by one of the participants, using the methodology in McNeill (1992). Once the beats were identified, we coded their lexical equivalents for intonational features, according to Pierrehumbert's scheme. We then looked to see what each modality contributed to the discourse structure, according to the relative theories described above.

In every case, we found that the beats did indeed occur on utterances with shifts of narrative level, confirming McNeill's claim. (An example is in a backtracking utterance starting with "first of all, he…" to fill in missing information. Here, the beat was on "he"). However, PH94's theory fared less well. In the above example, the referent of "he" had been mentioned in the preceding utterance (as evidenced by the use of the pronoun). Based on this, PH94 would predict an L* accent, as the entity was obviously not new. Yet in this, as in most such cases in our study, the accent was H*, theoretically reserved for *new* entities.

However, when PH94's theory is extended to use the *levels* of context which McNeill uses, it fares much better. In the above example, the phrase "first of all, he…" signals a shift from the narrative level to a metanarrative level. Now, in this metanarrative level, "he" becomes the *first reference* to the entity in the *new* level. The use of an H* now is felicitous; it signals the introduction of the entity to the *new* level of the discourse structure. McNeill notes independently that often proper names are used, when pronouns would typically be expected, precisely in such situations when a new narrative level has been reached. Pronouns are used for existing entities, while proper names are used for new ones. Similarly, H* is used for entities that may not be new to the overall discourse, but are new to the discourse level just entered.

Thus, the contributions of gesture and intonation to discourse are complementary. Beat gestures signal that a different discourse level has been entered, and intonation signals how to interpret an entity in relation (*new* or *not new*) to the discourse level just entered.

We intend to conduct more investigations along these lines. Yet our preliminary findings may have implications for multimodal analysis and generation systems. In analysis, by tracking the gestures and intonation carefully, we may gain more insight into the discourse structure, which may provide more accuracy in resolving referents. In generation, we may be able to

combine gesture and intonation in ways more natural then current systems allow.

Conclusion

Gesture, even pen gesture in sketch interaction, clearly plays an active role in communication that extends beyond deictic pointing. Clearly, we must consider the role of gesture in the communication of referring and pragmatic acts. In face-to-face dialogue, particularly when referring to entities in an immediate perceivable environment, how do people decide what form of referring expression to use? Principles derived from textual and spoken discourse should provide foundational ground. However, we must begin to acknowledge the role that context and human perception plays in how multimodal referring expressions are planned and understood. Computational models of discourse will need to explore and extend notions of saliency, attention, and discourse structure to account for how multimodal referring and pragmatic expressions are generated in multimedia presentation systems.

References

Binot, J. L., L. Debille, et al. (1992) *Multimodal Integration in MMI2: Anaphora resolution and Mode Selection*. Proceedings of WWDU, Berlin.

Cassell, J., T. Bickmore, et al. (to appear) *Conversation as a System Framework: Designing Embodied Conversational Agents*, in Cassell, J. et al. (eds.), *Embodied Conversational Agents.* MIT Press, Cambridge, MA, 1999.

Cassell, J. (in press). *A Framework for Gesture Generation and Interpretation.* In R. Cipolla and A. Pentland (eds.) Computer Vision in Human-Computer Interaction. Cambridge University Press.

Cohen, P. R. (1992) *The Role of Natural Language in a Multimodal Interface*. UIST '92, Proceedings of the ACM Symposium on User Interface Software and Technology, Monterey, CA.

De Angeli, A. D., F. Wolff, et al. (1999) *Relevance and Perceptual Constraints in Multimodal Referring Actions*. Proceedings of the Workshop on Deixis, Demonstration and Deictic Belief at ESSLLI XI

Hauptmann, A. G. and P. McAvinney (1993) *Gestures with Speech for Graphics Manipulation.* Intl. J. Man-Machine Studies 38: 231-249.

Huls, C., E. Bos, et al. (1995) *Automatic Referent Resolution of Deictic and Anaphoric Expressions*. Computational Linguistics 21(1): 59-79.

Johnston, M., P. Cohen, et al. (1997) *Unification Multimodal Integration*. 35th Annual Meeting of the Association for Computational Linguistics.

Kobsa, A. (1986) *Combining Deictic Gestures and Natural Language for Referent Identification*. Proceedings of COLING.

Koons, D. B., C. J. Sparrell, et al. (1993) *Integrating Simultaneous Input from Speech, Gaze, and Hand Gestures*. Intelligent Multimedia Interfaces. M. Maybury. Meno Park, CA, MIT Press**:** 257-276.

Martin, J. C. (1997) *Toward Ò Intelligent Ó Cooperation Between Modalities: The Example of a System Enabling Multimodal Interaction with a Map*. Proceedings of International Joint Conference on Artificial Intelligence (IJCAIÕ97) Workshop on Intelligent Multimodal Systems, Nagoya, Japan.

McNeill, D. (1992) Hand and Mind: What Gestures Reveal about Thought. University of Chicago.

Neal, J. G. and S. C. Shapiro (1991) *Intelligent Multi-Media Interface Technology*. Intelligent User Interfaces. J. W. Sullivan and S. W. Tyler. Reading, Massachusetts, ACM Press, Addison-Wesley Publishing**:** 11-43.

Neal, J. G., C. Y. Thielman, et al. (1989) *Natural Language with Integrated Deictic and Graphic Gestures*. Proceedings of the 1989 DARPA Workshop on Speech and Natural Language.

Oviatt, S. L. (1996) *Multimodal Interfaces for Dynamic Interactive Maps*. Proceedings of Conference on Human Factors in Computing Systems: CHI '96, ACM Press.

Pierrehumbert, J. and J. Hirschberg (1994) *The Meaning of Intonational Contours in the Interpretation of Discourse.* In Cohen, Morgan, and Pollack (eds.) Intentions in Communication.

Rime, B. and L. Schiaratura (1991) *Gesture and Speech*. In Fundamentals of Nonverbal Behaviour. R. S. Feldman and B. Rime, Cambridge University Press**:** 239-281.

Wahlster, W. (1991) *User and Discourse Models for Multimodal Communication*. Intelligent User Interface. J. W. Sullivan and S. W. Tyler. Reading, Massachusetts, ACM Press, Addison-Wesley Publishing**:** 45-68.

Wahlster, W. (1998) *User and Discourse Models for Multimodal Communication*. Readings in Intelligent User Interfaces. M. T. Maybury and W. Wahlster. San Fransisco, Morgan Kaufmann.