Acting on a visual world: the role of perception in multimodal HCI

Frédéric Wolff, Antonella De Angeli*, Laurent Romary
Laboratoire Loria
BP 239
54506 Vandoeuvre-Les-Nancy
{wolff,deangeli,romary}@loria.fr

Abstract

Following the ecological approach to visual perception, this paper investigates multimodal referring acts in Human-Computer Interaction. Preliminary results from a simulation experiment allow to: (a) clarify the effect of perceptual organization on multimodal communication; (b) provide guidelines for designing effective multimodal interfaces. Demonstrating that both the verbal and the gestural part of a referential act are influenced by perception, we confirm the need and the utility of taking into account perceptual organization to analyze referential expressions in a more robust way. As a conclusion, we show how these results can lead to an actual implementation of a gesture interpretation module directed by the reference analysis process.

Introduction

An efficient way to allow a better use of the constantly growing number of software capabilities can be to provide users with a more natural interaction form to express their communicative intentions. Although the direct-manipulation paradigm reproduces an ecological way of acting upon objects, it still implies a learning phase (to master the rules of the new interaction style) as well as a constant translation activity (to transform conceptual intentions into available elementary actions). Both these activities increase the cognitive workload required to perform computer-supported task (Norman and Draper 1986). Simplifying the translation process from intention to actions, requires interface designers to devise multimodal systems capable of handling spontaneous speech and gesture.

Perception-action

To study user behavior in a multimodal environment, we choose the perception-action cycle as the appropriate unit of analysis (Neisser, 1976). This theoretical framework explains how action planning and execution is controlled by perception and how perception is modified by active exploration. Moreover, the study is based on ecological psychology, an approach to perception, cognition, and action emphasizing the mutuality of organism-environment relationships (Gibson 1979). It is based on the 'validity' of information provided to perception under normal condition, implying, as a corollary, that laboratory studies must be carefully designed to preserve ecological validity. According to this view, perception and action are linked by *affordances*. Optic information about objects conveys their functional properties providing clues about the actions they can support. Functional properties can thus be considered as *affordances* of users' possible actions, as if the object suggested its functionality. For example, a hammer usually induces us to take it by the handle and not by the head, because the handle is visually more graspable.

In this paper, we attempt to extend the concept of affordances to explain referring acts; i.e., communication acts composed by a verbal part (referring NP) and a gestural part (referring gesture). The basic question tackled is the following: *Can different perceptual organizations afford different multimodal actions*?

To answer that, a simulation experiment was run. Perceptual organization was manipulated according to the principles of the *Gestalttheorie* (Wertheimer 1922; Kanizsa 1979). They state that individuals spontaneously organize the perceptual field into groups of percepts. Grouping allows the observer to reduce the original complexity of the stimulus. This is necessary because human capabilities to process separate units are limited. Gestalt laws of perceptual organization describe the principles underlying grouping. The main principle, *prägnanz law*, states that the elements of the visual field tend to be segregated into forms that are the most stable and create a minimal stress. The other principles describe how stability is achieved. In this paper, we focuse on *similarity* (objects are grouped on the basis of their salient physical attributes, such as shape and color), *proximity* (elements are grouped on the basis of their relative proximity), and *good continuation* (shapes presenting continuous outlines have a better configuration than those with discontinuous ones).

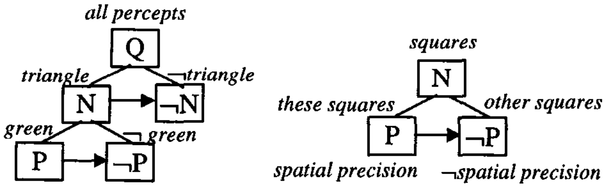
* Also at Cognitive Technology Laboratory. Department of Psychology - University of Trieste. Via dell'Università 7, I-34123 Trieste - Italy. deangeli@univ.trieste.it

Referring acts

The grouping process of perception can be compared to the process of identifying referents. To convey their communicative intentions, speakers have to drive the attention of listeners towards a reduced area of the visual world containing the group of mentioned referents. Referring NPs, possibly in combination with referring gestures, can accomplish this task with a great flexibility in the way features are shared out between language and gesture (Schang and Romary 1994; Bellalem et al. 1995).

Indeed, the way referential acts are produced depends on the complexity of extracting the percepts which are to be referred to. To do so, both communication modes (speech and gesture) often convey complementary information that helps *categorizing* and *localizing* referents. On the one hand, the referring NP allows the listener to determine the category of referred percepts. On the other hand a localization is carried out by using either spatial referring NP or referring gesture. As regards categorization, the role of a referential expression is to filter a perceptual category according to the most salient features of the object. This process depends in particular on the semantic characteristics of the expression (e.g. demonstrative vs. definite usages, see below). As regards localization, the gesture determines the spatial features of the referred percepts. Such referential extraction allows the listener to isolate the focused percepts from the shared visual context.

Categorization can be performed according to various criteria which are intrinsic to the perceptual features. From a linguistic point of view, this filtering action builds an opposition, also called *axiology*, between the different kinds of percepts occurring in the scene (Gaiffe, Reboul and Romary 1997). For instance a definite expression such as «the green triangles» organizes the referential space into opposite categories as shown in figure 1a. At a first level, the use of the definite expression operates a contrast between N-objects (N being the category) and non-N-objects, which is then refined, in our example, into two opposites sub-categories : the green and non-green triangles. A demonstrative expression such as «those squares», also defines a similar contrast as shown in figure 1b. The initial context, where the contrast is to be performed, results from the filtering of the more global context to keep only N-object. The contrasting process then relies on the ability to isolate an element –or in the case of plural, group of elements– which is specifically in focus.



a «The green triangles» b «these squares»

Figure 1: Examples of axiology

A deictic gesture, on an otherwise verbally limited context, also leads to a reduction of the focus of the listener. The perceptual criteria used then rely on the features which are extrinsic to the percept, and rather belong to properties of perceived group, or percepts, the user refers to. Features that define a group can be found in its localization, its topology, as well as in its general shape.

However, the rate of each reduction seems to be related to the heterogeneity of the elements, of their properties, as well as of the involved groups and their salience. In some cases, when the salience of the referents corresponds to their intrinsic features, this type of discrimination can be more easily done verbally, by a referential NP. For instance, a group of red triangles spread among gray squares, provides the speaker with an easy linguistic access because of its strong salience. In other cases, when the scene displays groups with a sufficiently contrasted topology, the reduction can be more efficiently expressed by a deictic gesture. For instance, a circling gesture can easily isolate a group that is well separated from other percepts. These reductions reflect the contrastive effort which has to be produced by the speaker so that the listener will be able to isolate the referenced percepts.

If the features of referential access seem to be determined by the complexity of the scene and of the perceived groups, as we will show in the second part of this paper, the difficulty of extracting referents belonging to different groups also seems to influence the characteristics of the referential effort. Indeed, it is easier to reference all the elements of a same group in a scene, than several percepts spread in different groups. In the first case, the salience of each group will influence the verbal and gestural accessibility. In the second case, the possible spreading of referents among different groups requires to build another group in an explicit way.

As a whole, these results are strongly enkeeping with what has been observed for referring acts in a pure linguistically based interaction. We deal here with an extension of the concept of relevance as defined by (Sperber and Wilson 1986) and which refines the classical Gricean maxims of brevity and efficiency. The only difference is that the evaluation of the ratio between the cognitive load to compute a given utterance and the number of inferences it fires must comprise the perceptual features available to the adressee combined with the gestural information. Besides, our observation that referential interpretation is heavily based upon localized perceptual groups complements the notion of contrastive sets advocated by (Dale and Reiter 1992), not to say that they correspond to the same kind of representations, from a cognitive point of view.

The need for empirical research

The major problem in developing multimodal systems is connected to communication variability. Such a variability is as much present in the verbal part than in the gestural one, so that the communication protocol can not be reduced to stereotypic shapes. Even though a lot of studies have aimed at improving the understanding and also the computation of verbal utterances (see, for instance, the proceedings of last ISSD 96 in Philadelphia, or of the ESCA Workshop on Spoken Dialogue Systems Visgo, 1995) providing a less artificial linguistic protocol, only a few works have dealt with gesture variability (Oviatt, De Angeli and Khun 1997) and flexibility (Streit 1997). This bias has led to some kind of weakness in our understanding and thus in our ability to compute automatically complex gestures, which has led to a standard execution form: pointing has to be included in the visual referent.

This current lack could be explained by the features of most gestural devices and by the relatively few accurate data related to gesture parameters available to build and validate new models. Contrary to human-human communication, gesturing in HCI often requires to manipulate artificial mediators. A lot of traditional gesture devices can be categorized as mediators: mouse, trackball, joystick, pad or VR glove. In spite of this, devices exist which allow to limit this *artificial limb effect* at different levels. For instance, the touch screen reduces the presence of the intermediate layer sensitive to actions, but still requires some physical contacts. Other devices allow to reduce the latter constraint by, for instance, the location of an ultrasonic transmitter ring, or the disruption of a weak electric field produced by the hand position (Zimmerman *et al.* 1995). However, such gesture devices cannot fullly convey the richness of natural gesture. In human-human communication, locutors often use multipolar gesture, which cannot be reduced to one point, as in the case of a one-finger gesture. Multipolar gestures change the way of capturing action features, because it becomes possible to perform multi-point gestures composed by non-sequential events. Limiting gestural events to one point also reduces the potential lexical diversity of actions, and thus the ratio of semantic that the gestural movement can bring into multimodal utterances. It becomes then difficult for the user to produce utterances like «Draw a line that long» with an appropriate two-finger gesture. New generation devices can handle multi-point gestures, like the multipolar pad, or also the use of video recognizers which are able to locate hand positions (Littman, Drees and Ritter 1996).

The fact that gestural interaction has for long been limited to simple pointing gesture can explain that no real urge has emerged to collect accurate data for this communication mode. Here, we present a simulation experiment where users were free to perform different kinds of gestures. These data have been collected in order to build an empirical reference interpretation model based on users' spontaneous behavior. Results will give insights about the effect of perceptual groups on verbal and gestural accessibility. In the last part of the paper, a general discussion will lead to evaluate the consequences of these influences in the modelization of multimodal analysis principles and to define some new experimental works.

Woz Simulation

In this section, we present a pilot study of an ongoing research project aimed at developing empirical predictive models that account for communication behavior in a multimodal HCI. As previously stressed, such a knowledge is essential to the design of future systems and to the development of interfaces capable of overcoming the technical constraints of the system without diminishing the intrinsic naturalness of multimodal communication. The pilot study was designed to test the reliability of the simulation environment (including system, task, procedure) and to provide preliminary results about the role of the perceptual field organization in gesture and speech production.

Method

Participants. Seven students from the University of Nancy participated in the simulation as volunteers. All were French native speakers.

Procedure. After reading hard-copy instructions describing system functions and task requirements, the participant engaged a dialogue with the simulated system to perform a typical computer-supported task, moving objects into folders. Interaction was based on speech and gestures, mediated by a microphone and an electronic pen.

Thirty different scenes were presented to each participant. The user's screen displayed a collection of objects and 8 boxes. To inhibit pure verbal references, objects were abstract-shape figures (De Angeli, Petrelli and Gerbino 1996). They could be targets or distractors. Targets were collections of two or three same-shape stimuli that have to be moved into the box displaying their figure. Distractors were exclusively used to manipulate perceptual field organization and did not have to be moved. At the end of the session, each participant filled in a user's satisfaction questionnaire and was debriefed.

Design. The original study was based on a complex design manipulating several perceptual factors. In accordance to the aims of this paper, only a sub-sample of the corpus has been considered. It comprises data collected under two stimulus segregation conditions: High vs. Low Salience of group. In the high-salience condition, targets were easily perceived as a group, clearly separated from distractors. In this case, proximity and good continuation supported similarity. In the low-salience condition, targets were spontaneously perceived as elements of a broader heterogeneous group that included distractors. In this case, proximity and good continuation acted in opposition to similarity.

Semi-Automatic Simulation. The system was simulated by the Wizard-of-Oz (Fraser and Gilbert 1991). An ad hoc system was developed in order to support a semi-automatic simulation on two connected SUN SPARC workstations.

The Wizard could observe user's actions on a graphical interface, where he also composed system answers (Figure 2). Wizard was supported by interface constraints and several prefixed answers. These strategies have been found to increase simulation reliability by reducing response delays and lessening the attention demanded of wizards (Oviatt *et al.* 1992).

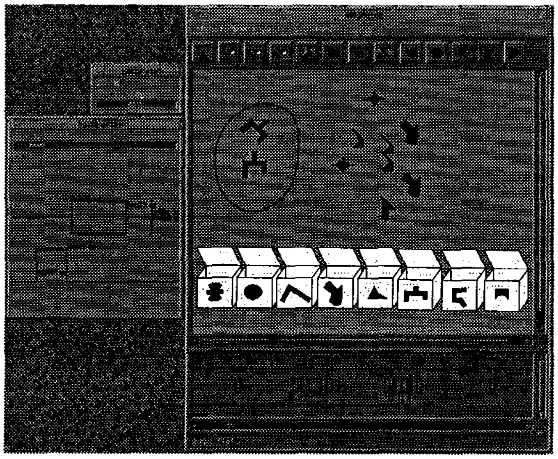


Figure 2. The Wizard Screen in the High Salience condition

All interaction data were recorded in such a way that it is possible to replay the entire experiments with precise information. Each record is actually made up of an audio file, a task evolution file as well as a gesture file, providing accurate numerical data.

Data coding. The corpus was analyzed with respect to *moving commands*: i.e., communication actions aimed at moving targets. Individual commands were defined as single conversational turns produced by users. According to the number of objects displaced by it, each moving command was classified as *group-oriented* or *element-oriented,*

The gestural and the verbal part of each group-oriented command were further tabulated according to the strategy adopted to convey the concept of group. Two general strategies were found in the corpus: *group-reference* and *individual-reference.* As far as gestures are concerned, group-reference was achieved by showing the perimeter or area of the group; individual-reference by indicating each element one by one. As regards language, group-reference was achieved by means of plural deictic anchor or target description (i.e., «these objects», «the two isolated objects, the two forms»; in French «ces objets», «les deux objets isolés, les deux formes»); and individual-reference by singular linguistic anchors (i.e., «this item and this item», «this object and also this one » in French «cette pièce et cette pièce», «cet objet ainsi que celui-ci»). At a multimodal level commands were classified as *group-reference, mixed-reference,* and *individual-reference,* according to the strategies adopted in each modality (speech and gesture).

Gestures were defined *as trajectories in certain parameter space* and classified in one of the following categories: 0-d (pointing); 1-d (targetting); 2-d (circling, free-form, scribbling). A first scoring was performed by two independent judges watching the audio-video logging files. It appeared that pointing could have different degrees of precision. In some cases, it reproduced a very precise point with no movement at all. In others, it resembled a small straight lines or a small spot. To test if these differences were intentional or exclusively due to the gestural interface, a small experiment was run. Using the same simulation procedure that has been previously described, five persons were explicitly required to reproduce the three gestures. A corpus of 285 gestures allowed to discriminate between technical error and intentionality, as well as to set category boundaries. Then 0-d gestures were further classified as: dot, spot or line. Each pointing gesture was a dot when the range of the movement on the x-y axis was inferior to 4 pixels. Otherwise, it was classified as a spot or a line according to the presence or absence of backward movements (x+y/distance< .75 for spots).

Double scoring was conducted for 20% of the reported variables. All measures had a reliability of .95 or above.

Results

A corpus of 98 moving commands has been analyzed. Independently from group salience, a strong preference towards the more economic procedure emerged: 92% of the commands were group-oriented actions. Moreover, with only 3 exceptions, commands were performed multimodally.

As regards reference strategies at the multimodal level, a strong consistency between modalities was found. Only 1 out of 3 commands was based on a mixed strategy. It is worth noticing, that all mixed inputs were composed by verbal group-references amplified by gestural individual-references; whereas all gestures providing group-references were always accompanied by verbal group-references. Most of the commands (40%) followed an individual-reference strategy; 28% a group reference strategy.

The distribution of cases in the three reference categories differs according to group-salience (χ^2= 18.38, d.f.= 2, *p*< .001). In the High-Salience condition, group-reference was the most frequent strategy, and individual-references occurred as frequently as mixed-references (Table 1). On the contrary, in the Low-Salience condition, individual reference was the most frequent strategy, whereas group-reference was very rare.

	Group	Mixed	Individual
High	46	27	27
Low	5	41	54

Table 1. Percentages of the three reference strategies as a function of experimental conditions.

To summarize, reference at the multimodal level is affected by perceptual field organization. However, the effect is stronger with respect to the gestural part of the input (χ^2=14.96, d.f.= 1 p< .0001) than with respect to the verbal one (χ^2= 6.68, d.f.= 1, p< .01).

A corpus of 184 gestures has been analyzed. As regards gesture dimensions, 0-d was the most frequently used (79%), followed by 2-d (14%), and 1-d (7%). Typically, pointing reproduced a spot (74%), otherwise it was a line (17%), or a spot (9%). To test whether pointing precision was different when users referred to a group or to a single percept, the sample was considerably increased by scoring other experimental scenes requiring to refer only to one element. Moreover, given the higher percentages of dot, only two pointing categories were considered: High-precision (dot) and Low-precision (spot and line). A crosstabs analysis comparing Access Type (Group vs. Element) to Gesture Precision (High vs. Low) showed a significant difference on frequency distribution in the four cells (χ^2= 5.04, d.f.= 1, p< .05). Dots occurred more frequently in the case of group access (78%) than in the case of percept access (60%). Moreover, when two percepts were identified by a group access mediated by pointing, low-precision gesture occurred more frequently on the first percept (30%) than on the second (3%), χ^2= 9.95, d.f.= 1, p< .01. In this case, we found also a clear preference towards the upper-down direction (χ^2= 31.16, d.f.= 2, p< .001). When percepts where displayed on a vertical line, the upper one was pointed out first 82% of the time.

Surprisingly, 2-d gestures were not associated to group reference since they occurred with the same percentages when people referred to group or to individual target. All the circling followed an anti-clockwise direction.

Discussion

Our findings confirm previous results (Oviatt, De Angeli and Khun 1997; De Angeli, *et. al.* 1998) showing that in a visual/spatial domain multimodal communication is highly preferred to the unimodal one (only 3 commands out of 98 were performed unimodally).

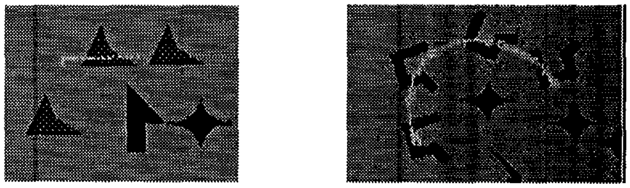
As regards commands at a multimodal level, it is worth noticing that gestures were always accompanied by a verbal anchor, in the form of a deictic expression or of a description. This contrasts with previous results (Oviatt, De Angeli and Khun 1997), showing that most multimodal constructions (59%) produced during a multimodal simulation did not contain any spoken deictic. In our opinion, such a difference can be explained considering not only the difficulty for naming the abstract shapes and the interaction language (American vs. French) but also a fundamental distinction in the two interfaces, i.e., the system reaction to user's gestures. Indeed, in the previous simulation the pen provided a detectable feedback to the users' gestures, while in the present simulation no feedback was provided. As showed elsewhere (De Angeli *et al.* 1998), the presence of a feedback appears to favor the elision of verbal anchors leading to pure referring gesture (Petrelli et. al. 1997). To further investigate the effect of feedback on multimodal communication an experiment, based on the same simulation paradigm described in this paper, is contemplated.

The innovative contribution concerns the influence of perception on multimodal communication. Multimodal commands showed a high correlation between the reference strategies adopted by the two modalities to access groups. However, despite its unfrequent occurrence, the mixed-reference strategy (i.e., a verbal plural-reference accompanied by a gestural individual-reference) may still constitute a problem if multimodal constructions are resolved without considering the visual context. Indeed, the deictic 'these' has to be associated to *n* gestures (n corresponds to the number of elements forming the group), but not to other eventual gestures occurring to indicate different elements (in our case the boxes where elements had to be moved) associated to separate deictic anchors.

We also demonstrated that the effect of perception was stronger on referring gesture rather than on referring NP. This difference underlines the association between perception and physical actions which is higher than between perception and cognitive actions.

Our results lead to the conclusion that group-references occur almost only when the referred group is easily detectable i.e. in the High-Salience condition. Therefore under this condition, it is necessary to extend the pointing-inclusion paradigm in order to allow users to express their communicative intentions in a natural way. Such an extension has to consider the variability of gesture forms and meanings, as well as their possible ambiguity. The last phenomenon is due to the non one-to-one relationship between gesture shape and its meaning. Indeed, we have demonstrated that the same gesture can convey different semantic interpretations, as when a pointing action is performed in order to refer either to an individual element of a group or to the whole group; and when a circling is drawn to refer either to inner objects or to strike objects. Semantic ambiguity can be handled considering either the verbal part or the organization of the perceptual context on which the user is acting. Hence the modelization and the implementation are also required to take care of the graphic layout of the user's interface, in order to build robust multimodal reference interpreters.

Gestural ambiguities are often due to the fact that gestures can be based on perceptual groups. The corpus showed other problematic cases, as the ones presented in Figure 2. The difficulty to resolve reference can be explained by intra-group or inter-group competition of possible candidates.



a «Mets ces 3 objets ...»

b «Deplacez ces objets ...»

Put these three objects.. *Move these objects...*

Figure 3. Inter-group and intra-group ambiguities.

For instance the targeting gesture, performed in figure 3a, generates an intra-group ambiguity in choosing either the percept or the group. This can be resolved by taking into account the verbal categorization of granularity. The example in figure 3b, presents an inter-group ambiguity whether the gesture is considered as circling or targeting. As opposed to the first example, here the choice of the appropriate referential candidate (between the U-shaped group and the star-shaped percept) has to be made according to a measure of gestural relevance to group or percept, i.e. to the perceptual organization. Such a measure requires to calculate the rate of matching of the gestural trajectory and the perceptual candidates.

To treat these ambiguities, the information flow between the linguistic and the gestural modules has to be defined as illustrated in Figure 4. The linguistic module generates the axiologic structure allowing the gestural module to generate, as a return value, referential hypotheses within their contextual frame (dotted objects are those which do not belong to the returned contextual frame).

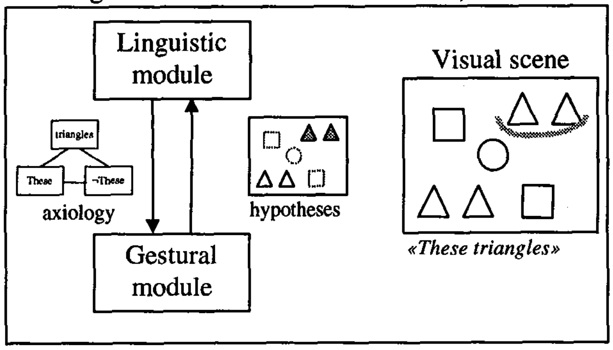


Figure 4. Information flow between the linguistic and the gestural modules.

Our results are preliminary, but they clearly show the role of perception in multimodal communication. Moreover, they confirm the value of simulation as a tool for building HCI predictive models that provide design guidelines for effectively integrating motor-visual language and verbal language. Further studies are planned to deeper investigate how users distribute their communication intentions across language, vision, and gesture, as well as to clarify how different modalities influence each other. Our next empirical works should attempt to evince universal communication principles by running inter-cultural studies comparing different linguistic and gestural codes.

References

Bellalem, N., Romary, L. and Schang 1995. Which representations for a proper treatment of referring expressions in a man-machine multimodal dialogue. In Proceedings of the ESCA Workshop on Spoken Dialogue Systems, Visgo, Denmark, 61-64.

Dale Robert and Ahud Reiter 1992, Computational Interpretations of Gricean Maxims in the Generation of Referring Expressions, Actes Coling-92.

De Angeli, A., Petrelli, D. and Gerbino, W. 1996. Interface Features Affecting Deixis Production: A Simulation Study. In Proceedings of the Workshop of the Integration of Gesture in Language and Speech, ICSLP-96, 195-204.

Fraser, N. M. and Gilbert, G. N. 1991. Simulating speech systems. *Computer, Speech and Languages*, 5: 81-99.

Gaiffe B., Reboul A. and Romary L. 1997, Les SN définis: anaphore, anaphore associatives et cohérence. In Book Relations anaphoriques et (in)cohérence, Ed. DeMulder, Tasmowski-DeRyck, Vetter, *Rodopi Amsterdam*

Gibson, J.J. 1979. *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.

Kanizsa, G. 1979. *Organization in vision.* New York: Praeger.

Littman E., Drees A. and Ritter H. 1996. Visual Gesture Recognition by a Modular Neural System. In Proceedings of ICANNN'96, Springer

Neisser, U. 1976. Cognition and Reality. San Francisco: Freeman & Co

Norman, D. A. and Draper, S. W. 1986. *User Centered System Design: New Perspectives on Human-Computer Interaction.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Oviatt, S., Cohen, P. R., Fong, M. and Frank, M. 1992. A Rapid Semi-automatic Simulation Technique for Investigating Interactive Speech and Handwriting. In Proceedings of the International Conference on Spoken Language Processing, 2, 1351-1354.

Petrelli, D., De Angeli, A., Gerbino, W. and Cassano, G. 1997. Referring in Multimodal Systems: The Importance of User Expertise and System Features. In Proceedings of the Workshop on Referring Phenomena in a Multimedia Context and Their Computational Treatment, ACL-EACL, Madrid, 14-19.

Schang, D. and Romary, L. 1994. Frames, a unified model for the representation of reference and space in a Man-Machine Dialogue. In Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP-94), Yokohama.

Sperber Dan and Deidre Wilson 1986, Relevance, communication and cognition, Basil Blackwell, Oxford.

Streit, M. 1997. Active and Passive Gestures - Problems with the Resolution of Deictic and Elliptic Expressions in a Multimodal System. In Proceedings of the Workshop on Referring Phenomena in a Multimedia Context and Their Computational Treatment, ACL-EACL, Madrid.

Wertheimer, M. 1922. Untersuchungen zur Lehre von der Gestalt I. *Psychologische Forschung,* 1: 47-58.

Zimmerman T., Smith J., Paradiso J., Allport D. and Gershenfeld N. 1995. Applying Electric Field Sensing to Human-Computer Interfaces. In Proceedings of CHI'95, 280-287

De Angeli A., Gerbino W., Cassano G., Petrelli D., 1998,Visual Display, Pointing, and Natural Language: The power of Multimodal Interaction ,Advanced Visual Interfaces Conference, AVI'98,

Sharon Oviatt, Antonella De Angeli, Karen Kuhn, 1997, Integration and synchronization of input modes during multimodal human-computer interaction, In Conference on Human Factors in Computing Systems: CHI97, New York, ACM Press