Behavior Research Methods, Instruments, & Computers
2003, 35 (3), 408-419

SLOT: A research platform for investigating multimodal communication

JAN PETER DE RUITER
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

STÉPHANE ROSSIGNOL and LOUIS VUURPIJL
Catholic University of Nijmegen/NICI, Nijmegen, The Netherlands

DOUGLAS W. CUNNINGHAM
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

and

WILLEM J.M. LEVELT
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

In this article, we present the spatial logistics task (SLOT) platform for investigating multimodal communication between 2 human participants. Presented are the SLOT communication task and the software and hardware that has been developed to run SLOT experiments and record the participants' multimodal behavior. SLOT offers a high level of flexibility in varying the context of the communication and is particularly useful in studies of the relationship between pen gestures and speech. We illustrate the use of the SLOT platform by discussing the results of some early experiments. The first is an experiment on negotiation with a one-way mirror between the participants, and the second is an exploratory study of automatic recognition of spontaneous pen gestures. The results of these studies demonstrate the usefulness of the SLOT platform for conducting multimodal communication research in both human– human and human–computer interactions.

When two humans communicate with each other in a face-to-face situation, they do not only exchange speech. Other so-called *channels* or *modalities*, such as gaze (e.g., eye contact), facial expression, intonation, voice quality, and gestures (e.g., pointing) also play an important role in both the semantic and the socio-emotional aspects of communication (see, e.g., Ellsworth & Ludwig, 1972). Some authors have even claimed that nonverbal communication contributes more to the interpretation of communicative acts than does verbal information (Archer & Akert, 1977; Mehrabian & Wiener, 1967). Archer and Akert (1977) used recorded stimuli in which both the nonverbal information (voice quality) and the verbal information (semantic content) were systematically varied in emotional quality (hostile, neutral, or friendly). In judging the emotional quality of the stimuli in which the nonverbal and verbal qualities did not match, participants indeed attributed more importance to the nonverbal signal. Brown (1986) argued that when judges of emotional quality are presented with incongruous information, the nonverbal aspects of the communication are usually assumed to be under less conscious control by the speaker than the verbal content is, and, therefore, the nonverbal information is perceived to be a more reliable source than the verbal content (p. 498).

Some authors have put forth the weaker claim that nonverbal communication is more suited for expressing feelings and affective information, whereas the content of speech is more often employed to refer to speaker-external events (Argyle, Salter, Nicolson, Williams, & Burgess, 1970). Later, however, Krauss, Apple, Morency, Wenzel, and Winton (1981), using naturalistic stimulus material, demonstrated that in judgments of the emotional quality of an exchange, judgments of the verbal content (presented in the form of transcripts) were a much more reliable source of information than those of video-only or content-filtered speech (presenting voice quality only) when each was compared with judgments of the full video and audio data, which served as a baseline.

Independent of the contribution of nonverbal communication to the overall interpretation of communicative acts, there is no doubt that for the transmission of certain types of abstract information (e.g., conditionals, counter-

This work was supported by Grant IST-2001-32311 from the European Union Project *Conversational Multimodal Interaction with Computers* (COMIC). The authors thank Jolanda Bource, Paula Dings, and Hanneke Ribberink for their meticulous transcription work. Correspondence concerning this article should be addressed to J. P. de Ruiter, Max Planck Institute for Psycholinguistics, P. O. Box 310, NL-6500 AH, Nijmegen, The Netherlands (e-mail: janpeter.deruiter@mpi.nl).

factuals, and tense information), linguistic communication has a clear advantage over nonverbal communication.

The different channels through which people communicate often complement each other. For instance, a spoken statement accompanied by raised eyebrows might be interpreted as a question by the interlocutor. A spoken statement containing the phrase "over there" is usually accompanied by a pointing gesture to indicate where "there" is. In sum, face-to-face interaction is a form of multimodal communication.

The word *multimodality* suggests that the concept refers to more than one perceptual modality (e.g., auditory perception and vision). However, this interpretation of the concept is too narrow to capture the phenomena in which multimodal communication researchers are generally interested. Although both hand gestures and facial expressions are perceived in the visual modality, they clearly represent distinct aspects of communication. The same holds for voice quality and content of speech, both of which rely on the auditory modality. To remedy this definitional problem, we propose the notion of a *semiotic channel.* A semiotic channel is a set of identifiable behavioral units that (1) cannot be performed simultaneously with each other and (2) can be performed simultaneously with (almost) all behavioral elements in other semiotic channels. For example, facial expression and voice quality constitute two different semiotic channels, for it is not possible to have two different facial expressions or two different voice qualities at the same time, whereas all voice qualities can, in principle, be combined with all facial expressions. The proposed definition is, admittedly, not entirely airtight: Although we believe eye gaze and facial expression, to be separate semiotic channels, certain facial expressions, such as the "thinking" expression (Cunningham, Breidt, Kleiner, Wallraven, & Bülthoff, 2003), often contain specific eye-gaze patterns as integral parts. However, given that eye gaze and facial expression can in most cases operate independently of each other, we nevertheless consider them to be separate semiotic channels. Due to the widespread use in the literature of the term *multimodal communication*, we will continue to use this term to refer to the general field of study, but what we mean by it is *communication involving the simultaneous use of more than one semiotic channel.* We will reserve the word *modality* to refer to the perceptual modality (e.g., visual, auditory) and will use the word *channel* to refer to semiotic channels, as defined above.

In order to gain systematic knowledge about how multimodal communication between human participants works, the collection of behavioral data is essential. A straightforward way to obtain data is to record interactions between people and analyze them carefully (see, e.g., Goodwin, 1981; Kendon, 1972). A disadvantage of this naturalistic approach is that one has very little control over the content of the communication, which makes it rather hard to predict what participants are going to communicate about and how they are going to do it. This, in turn, leads to an unacceptable level of variance in both the content of the communication and the use of the different modalities in which the human–human interaction researcher is interested.

A higher level of control over the content of the communication is achieved by Clark and Krych (in press). In their task, one participant had to instruct another in building a prespecified Lego model. An important advantage of this task is that it provides an objective measure of both the efficiency and the quality of the communication (as demonstrated by the duration of the interaction and the number of errors in the resulting Lego model, respectively). Although these tasks yield results that are important for the field of multimodal communication, they are limited to the study of *collaborative* communication. Also, the two participants have different predefined roles: One is "director" and the other is "builder."

Another interesting multimodal task is the *map task* (Anderson et al., 1991). In this task, the 2 participants each have a map in front of them, and one of them (the *route giver*) instructs the other (the *route follower*) to follow a certain route through the map. The maps used by route giver and route follower are not identical, which makes the task more difficult and, therefore, potentially more interesting. As with the director/builder task of Clark and Krych (in press), this task assigns different roles to the 2 participants.

An important reason for us to develop a new research platform instead of using the map task or the director/builder task is that we wanted to use a communication task that is highly structured but does not assign different roles to the 2 participants. Also, since an important goal of the human–human data that we want to generate is to provide information that could produce guidelines for (natural-looking) human–computer interaction (HCI), we wanted our task to involve the use of both electronic pen data and speech—modalities often used in multimodal HCI (Oviatt, 1999). Finally, we wanted to be able to have control over the competitiveness of the interaction, to be able to distinguish between multimodal communication at different levels of competitiveness of the task.

THE COMMUNICATIVE TASK

Criteria

An earlier version of the communicative task in the spatial logistics task (SLOT) was originally developed (see Levelt, 2001) with the following criteria in mind:

1. The experimenter should have a high degree of control over the communicative goals of the participants during the experiment. This should be the case for general (overall) goals as well as for possible subgoals at any time during the course of the experiment.

2. It should be possible to assess objectively the quality of the end result of the communication.

3. Although any task involving naturalistic, unconstrained communication will result in data with high inter- and intrasubject variability, this variability should be kept as low as possible to facilitate analyses with sufficient statistical power.

4. In the task, it should be natural for the participants to use all the channels that are available to them, especially the electronic pen.

5. The task should allow for the manipulation of the presence or absence of certain channels (with the exception of speech, which will always be available) without changing the essence of the communication task or making the task impossible to fulfill.

6. The task should be the same for both participants.

7. The level of competitiveness of the task should be variable.

Task Definition

With these criteria in mind, the SLOT communication task has been defined as a *route negotiation task*. The general idea is as follows: Two participants are shown a map (formally corresponding to a mathematical graph) consisting of cities (vertices) that are connected to each other by roads (arcs). Some of the cities are marked as targets. The participants have to negotiate a route that begins and ends at a specified *start city* and passes through all of the specified targets, while minimizing personal and/or global travel costs. The subjects are given the opportunity to use an electronic pen to draw on the presented maps (implementing a *shared whiteboard* metaphor) to facilitate the negotiation process.

There are two variants of this task—a *cooperative* variant and a *competitive* variant—which will be described separately below. The two modes of SLOT enable the study of potential differences in communicational behavior in cooperative versus competitive social contexts.

Cooperative Mode

In the cooperative mode, all the specified targets are presented in the same color (purple; we have used gray in Figure 1). See Figure 1 for an example of a cooperative SLOT map. The participants have to negotiate a route that passes through all of the targets while keeping the travel costs at a minimum (see below). The participants have to figure out collectively which route will result in the lowest cost. This is a variant of the well-known traveling salesman problem (TSP; see Burkard, 1979, for an overview), the difference from the canonical version of the TSP being that the travel costs for a certain step are not independent of the route that has been traveled up to that point. After a route has been negotiated, one of the participants is asked to submit it to the computer, using the electronic pen, to have it checked for legality and to compute and store the resulting costs, which are also displayed on the tablets.

Competitive Mode

In the competitive mode, there are different targets for each of the participants, marked by the use of a different color for each participant (red and blue; light gray and dark gray in Figure 2). See Figure 2 for an example of a competitive SLOT map. In addition to having his or her own personal targets, each participant also has his or her own cost accounting, and these personal costs are a function of the number of personal targets that have been reached at any point during the route. Note that by different placements of the red and blue target nodes, the degree of competitiveness can be varied. For maps on which the optimal route for the red targets is identical to that for the blue ones, a situation approximately equal to that of the cooperative mode is realized. The larger the difference between the respective optimal routes, the higher the degree of competition.

The Cost Rule

The general cost rule in SLOT is that every step from node to node over a single arc has a cost of 1 + the number of designated targets that still must be reached. As

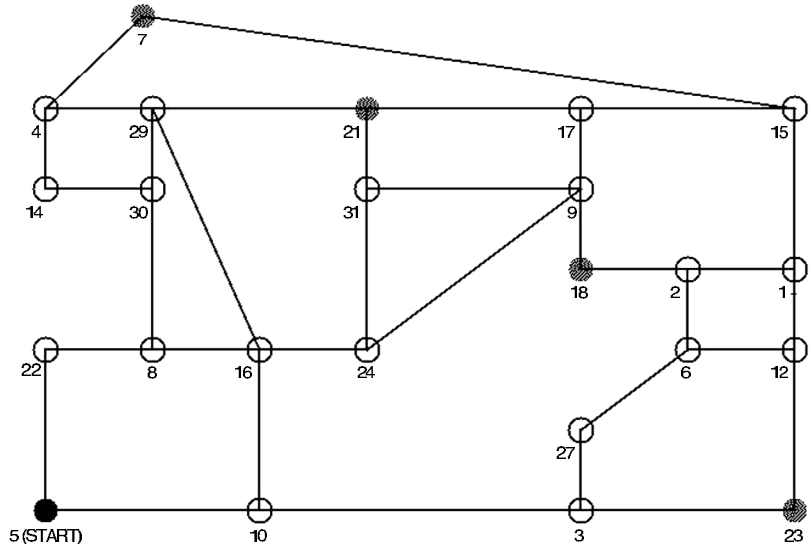

Figure 1. A cooperative SLOT map.

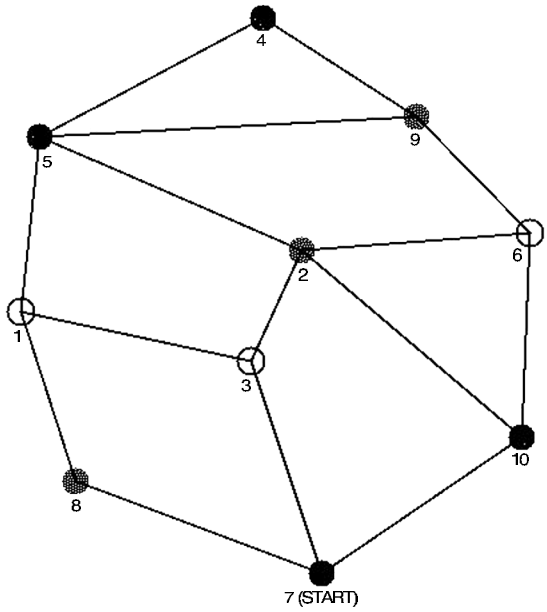

Figure 2. A competitive SLOT map.

soon as a target has been reached, the cost per step drops by 1. After all the targets have been reached, the cost of any subsequent step (i.e., returning to the start node) is equal to 1. In the cooperative mode, this holds for any target that has been reached, whereas in the competitive mode it holds only for those targets assigned to the participant whose score is being calculated. For example, in the competitive map presented in Figure 2, a first step from the start node (7) to Node 8 would have a cost of 4 for both participants, because no target has been reached yet (1 + number targets not yet reached [3] = 4). However, a subsequent step from Node 8 to Node 1 would have cost 3 for the participant assigned the red targets (because Node 8 was a red target) and cost 4 for the participant assigned the blue targets (because no blue target has been reached yet). The motivation for the use of this specific cost rule is to give participants an incentive to try to reach their targets as soon as possible in the competitive mode, and to reach any target as soon as possible in the cooperative mode.

Summary of Task Motivation

The use of a route negotiation task enables the manipulation of the communicational context (competitive vs. cooperative) and the tracking of the overall goal and subgoals at any time during the negotiation. The highly visual nature of the task makes the use of the pen natural (like drawing with a pencil on a real map) but not essential for the participants, and the quality of the result of the negotiation (i.e., a certain route) can be computed objectively.

One critical feature of SLOT is that it allows for *asymmetric* manipulation of the available channels (with the exception of the speech channel), in the sense that a certain channel or modality can be available to one participant but not to the other. This possibility is interesting in its own right, but it is also important for research supporting work in human–machine interaction. The available channels in human–machine interaction are likely to be asymmetrically distributed, due to the simple fact that the automatic (computer-based) recognition of most relevant channels is far less reliable than the corresponding recognition capabilities of humans.

THE HARDWARE PLATFORM

General Description of the Current Laboratory Setup

The 2 participants in a SLOT experiment sit at a small table (with a surface of 1 × 1 m), opposite and facing each other. In front of each of them is a graphical tablet that is positioned almost vertically, at an 80° angle relative to the table surface. The height of the tablets does not prevent either participant from seeing the face of his or her interlocutor. The graphical tablets are both controlled by a single high-performance Pentium-IV personal computer. This computer is operated by the experimenter, who has access to the keyboard, mouse, and screen. The computer and the experimenter operating it are located in the corner of the laboratory, at a distance of at least 2.5 m from the 2 participants, in order to minimize the influence they exert on the ongoing interaction between the participants.

Three video cameras are used to record the participants' behavior. Two cameras are positioned opposite the participants (behind and 0.5 m above their respective interlocutors) to record their facial behavior, whereas a third camera is located perpendicular to the participants to provide an overview of both participants simultaneously and from the side. An omnidirectional stand-alone microphone on a tripod is positioned to the side of the participants' desk to record the speech of both participants.

The signals of the three video cameras and the information represented on the graphical tablets are captured by a video processor and compressed into a 4-fold (2 × 2) split screen image. The split screen image and the signal from the microphone are recorded using a professional quality S-VHS video recorder. This setup allows for the simultaneous and time-locked recording of the speech, nonverbal behavior, and pen behavior of both subjects on one video tape. For additional fine-grained analysis of the pen gestures or handwriting behavior, the pen data stored on the hard disk can be used.

Graphical Tablets

A central feature of the hardware platform for SLOT is the use of WACOM PL-550-02B0 graphical tablets, colloquially known as the "Wacom Cintiq 15X." These are flat desktop LCD devices that function as a replacement for a standard color VGA computer screen. In addition, they allow for drawing and erasing directly on the screen surface with an electronic pen. Tablet coordinates are sampled at 200 Hz with a spatial resolution of 50 points per mm. The pen is pressure sensitive with a resolution of

512 quantification levels. The display allows for viewing angles of up to 160º. The shared whiteboard is implemented by sending an identical VGA signal from the SLOT computer to both tablets (and to the computer monitor of the experimenter) at the same time, using a VGA splitter. However, the pen data are recorded separately for each participant, using two independent USB connections.

An overview of the laboratory setup is given in Figure 3, and a snapshot from a video recording of a SLOT experiment is shown in Figure 4.

THE SOFTWARE PLATFORM

General Environment

The general software environment for which SLOT has been developed is the GNU/Linux operating system, which is a noncommercial variant of the Unix operating system. The particular distribution of Linux that is used is the SuSE 7.3 distribution, although SLOT should run on any GNU/Linux distribution (we know of one SLOT environment running without problems on a RedHat 7.2 Linux system). The programming languages used in SLOT are C (for the SLOT program), C++ (for the analysis tools), and Icon (for the graphical map editor). SLOT is programmed in plain Xlib using the Athena widget set. In order to capture simultaneous data streams from the two USB tablets with minimal delays from operating system overhead, we patched the USB drivers and rebuilt the Linux kernel.

The SLOT Application

In programming SLOT, special care was taken to provide a writing environment that is as similar as possible to a real whiteboard. The rendered ink traces take into account pressure data: If a participant exerts more pressure, the ink trace becomes thicker. Furthermore, the participants can use the back of the pen to erase their (and the other participant's) virtual ink, but not the displayed map.

The SLOT application performs the following steps:

1. It draws a SLOT map that is specified in an experimental design file (made with the editor described earlier) on both WACOM tablets and enters *negotiation mode,* in which it (1) registers the participants' pen gestures during the negotiation phase and displays them in real time on both tablets as virtual ink in blue for 1 participant and in red for the other; (2) stores all pen activity on disk, each sample containing the *x*,*y* coordinates and the corresponding pressure values and time stamp, in milliseconds; (3) registers the participants' erasure activity and processes the virtual ink accordingly; and (4) stores all erasure activity (time stamped) on disk.

2. After the negotiation phase has been completed, it allows the experimenter to put the program into *submission mode*, in which it (1) allows one of the participants (which one is designated in the experimental design file) to enter the negotiated route using the pen; (2) stores the entered route (time stamped) on disk; and (3) evaluates the entered route for legality. If the route is illegal, it provides feedback to the participant about the nature of the error(s) and restarts Step 2.

3. It computes the costs incurred by the participants and displays them.

4. It goes back to Step 1 until there are no more maps to be displayed.

Additional Computational Tools

In addition to the SLOT program, to run experiments, some other tools have been developed for the preparation and evaluation of SLOT maps.

Graphical map editor. In order to interactively design SLOT maps, an interactive graphical map editor was written. The program was written in the Icon programming

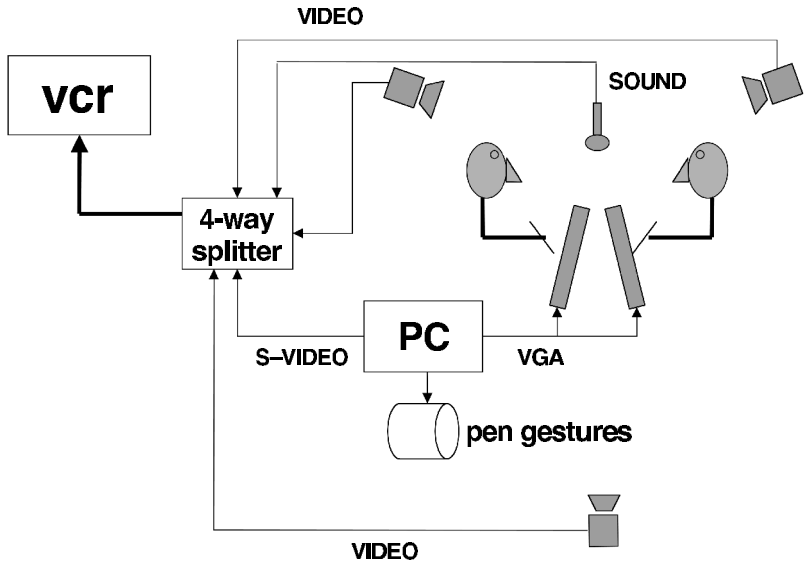

Figure 3. Schematic diagram of interconnected hardware in the SLOT laboratory.

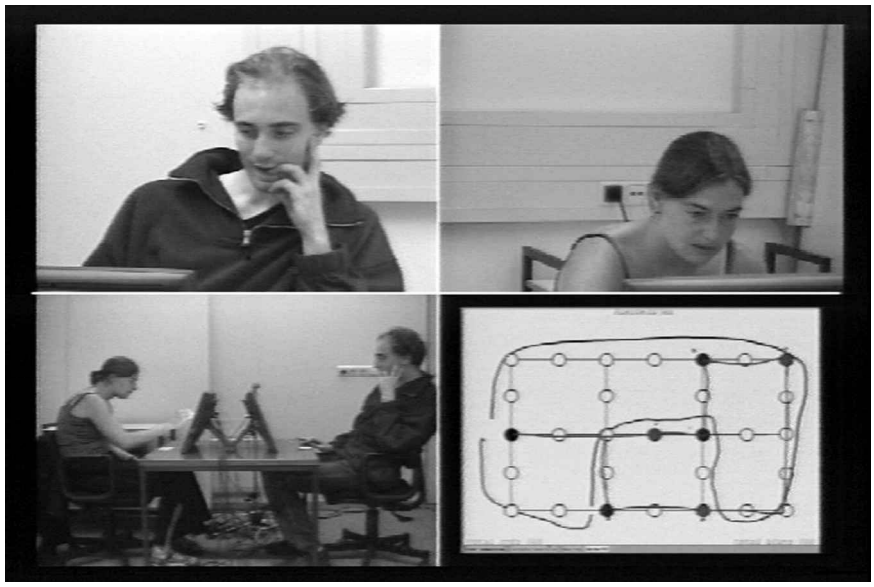

Figure 4. Snapshot from SLOT video recording.

language, which is portable among almost all existing computer platforms (e.g., Mac, Linux, and Windows NT). The program allows for the creation of maps by using the mouse for selection, placement, and movement around nodes, vertices, and the keyboard for elementary operations such as storing maps and assigning labels (e.g., "target," "start") to nodes. Map definitions are stored in xml format for easy editing and compatibility with existing multimodal annotation tools.

Map analyzer. The map analyzer is a program written in C++ for the Linux platform. It performs several calculations on SLOT maps that are relevant to the evaluation of certain dependent and independent variables in SLOT experiments. For any given map, the program computes the following properties: (1) the optimal (i.e., lowest cost) route for reaching all targets,[1] (2) the optimal route for the the participant assigned the red targets (competitive maps) and the associated cost, (3) the cost of the above route for the participants assigned the blue targets, (4) the optimal route for the the participant assigned the blue targets (competitive maps) and the associated cost, (5) the cost for the above route for the participants assigned the red targets, (6) a measure of how non-Euclidian a map is, defined as the standard deviation (*SD*) of the physical (screen) distances between every two connected nodes after these distances are standardized to an average of 1 (if all of the actual screen distances are the same, this measure is equal to zero), and (7) a measure of the intrinsic unfairness of a competitive map. This is defined as the difference between the costs of the optimal route to reach all targets for the participant assigned the red targets and the one assigned the blue targets, respectively.

SLOT pen data analyzer and transformer. The pen data collected through SLOT is stored in an efficient data format in order to avoid significant disk access delays. A program was developed to examine these data.

The program displays the collected pen data in real time and can be used to analyze pen movements produced by the participants during a SLOT experiment. In addition, the stored pen data can be transformed to the UNIPEN format (see http://www.unipen.org), which is the de facto standard for storing dynamic handwriting data. For this data format, many tools are available that can be used to annotate and analyze the data.

Preprocessing SLOT Data

Before the data generated in SLOT experiments can be used to address specific research questions, they need to be preprocessed by human coders. The participants' speech needs to be transcribed into a textual format, whereas the other channels under investigation are to be coded into discrete categories with their corresponding time intervals. For this, a multimodal transcription tool such as EUDICO (Brugman, Russel, Broeder, & Wittenburg, 2000) can be used on the digitized video recordings. See Table 1 for an overview of a number of channels that can be annotated in SLOT data.

METHOD

In this section, we present some initial results generated in our SLOT laboratory, to illustrate the type of research question that can be investigated using SLOT. First, we will describe the experiment (more information can be obtained from de Ruiter, 2003), and then we will present and discuss two illustrative results from this experiment.

Participants

Sixteen Dutch-speaking participants from the participant pool of the Max Planck Institute for Psycholinguistics were paid to take part

Table 1
Channels Typically Annotated in SLOT Data

Channel	Registration Medium	Focus of Analysis
Speech	Video soundtrack	Speech acts (content analysis) Turn-taking behavior
Intonation	Video soundtrack	Intonational contours/phrases
Gaze	Video	Eye contact between participants
Posture	Video	Body position shifts
Head orientation	Video	Use of head-tilt signals
Facial expression	Video	Finding set of used facial expressions
Pen gesture	Video + digital registration	Classes of pen gesture, relation of pen gesture to speech (timing, content)
Handwriting	Video + digital registration	Use of labels, words, exclamation marks, etc.

in the experiment. Eight participants were female and 8 were male. The participants were randomly paired to create eight mixed-gender dyads. None of the participants knew his or her experimental partner.

Materials

For the experiment, 10 competitive SLOT maps were generated with the map editor described earlier. Two of these maps were used for practice purposes, and the other 8 were designed for the actual experiment. Each map contained three red targets and three blue ones. Using the map analyzer, four of the experimental maps were created such that the intrinsic unfairness was high (ranging from 10 to 24, $M = 17$), whereas for the other four it was low (each of them rating 1). The order of presentation of the maps was randomized and was the same for all dyads.

Design

The principal between-dyads manipulation in this experiment is that for four of the dyads, there was a one-way (half-silvered) mirror standing on the table between the two tablets. Lighting conditions were arranged so that 1 of the participants could clearly see his or her partner through the glass, whereas the participant on the other side could not see his or her partner, due to the mirror effect. The aim of this between-dyads manipulation was to study the effect of an asymmetric distribution of the visual modality between the participants in a dyad. The principal within-dyad manipulation was the intrinsic unfairness of the used maps, which was high for four experimental maps and low for the other four. The independent variables of gender, color played in SLOT (red or blue), and side of the mirror on which the participants in the mirror condition sat were counterbalanced.

Procedure

The participants were handed written instructions in which the rules of the SLOT task were explained. They were encouraged to minimize both their personal costs and the global (summed) costs. There was no mention of a strict time limit, but the participants were told that the experiment was expected to last about 45 min to 1 h. During a session with the two practice maps, the participants could get used to the shared whiteboard and the electronic pen and to submitting the negotiated route into the computer using the electronic pen. After the practice session, the participants had an opportunity to ask questions, after which the eight experimental maps were negotiated. During negotiation, the participants were allowed to freely draw on the tablets as much as they wanted. After reaching an agreement about the route to follow, the participants would signal this to the experiment leader, who would put SLOT in route-submission mode. The participants took turns entering the negotiated routes into the computer.

Transcription

For detailed information about the transcription procedure and the used maps, we refer the reader to de Ruiter (2003). For the purposes of the analysis below, we will mention a few relevant aspects of the transcription process. Time markers were created that allowed us to temporally locate the beginning and end of the negotiation for every map. Pen gestures were transcribed initially by marking their beginnings and ends in time. The beginning, end, and transcript of each speech turn produced by each of the participants were coded. The cost scores were stored automatically by the SLOT computer.

Results and Discussion

In this section, we will present and discuss two analyses of data collected in the SLOT experiment. The first focuses on the effect of blocking the visual modality for only one of the participants in a SLOT session on efficiency and performance in the negotiation task. The second analysis concerns the use of the pen-gesture data generated in SLOT to investigate the possible use of multimodal information to improve automatic pen-gesture recognition and interpretation.

Modality Effects on Performance and Efficiency

Since the main focus of this study is the possible effects of asymmetrical availability of the visual modality on negotiation performance and efficiency, the main dependent variables of interest in this study are the duration of the negotiation and the cost of the negotiated route. A general linear model analysis was performed to investigate a possible main effect of the presence of the mirror (the fixed factor *mirror*) and of the fairness of the maps (the fixed factor *fairness*). The data entered into the analysis were the durations of the eight experimental maps for each of the eight dyads, leading to a total of 64 data points. See Table 2 for a summary of the results. As can be seen, both the factors mirror and fairness have a substantial effect on the negotiation times.

The main effect of fairness was significant [$F(1,63) = 6.13, p = .016$], as was the main effect of mirror [$F(1,63) = 4.30, p = .042$]. The interaction of fairness and mirror was not significant [$F(1,63) < 1$]. Both the presence of the mirror and the (un)fairness of the maps led to an increase of

Table 2
Mean Negotiation Durations (in Seconds) by Condition

Factor	No Mirror	Mirror
Fair	81.8	150.8
Unfair	163.2	221.9

the duration of the negotiation. This increased duration could have been caused either by the participants' producing more speech or by their being silent for a larger proportion of the time. Additional analyses of both total speech time (the sum of the duration of the speech of both participants in the negotiation) and silence (the time during which neither of the participants spoke) revealed that the prolonged duration for the factor mirror was due to longer silences, whereas that for the factor fairness was due to both more speech *and* more silence. That there was more speech and more silence with the unfair maps is not surprising, because these maps are harder to negotiate, which could be expected to lead to more discussion and also to longer periods of pondering over proposals. See Table 3 for the average total speech time and silence time per negotiation in the four conditions. Note that speech time and silence time do not add up to negotiation time, due to the fact that the participants often spoke simultaneously.

The effect of mirror on speech was not significant [$F(1,63) < 1$], whereas the effect of mirror on silence was significant [$F(1,63) = 7.3$, $p = .009$], indicating that the effect of the mirror was indeed to increase not the total duration of speech, but rather the total duration of silence. The main effects of fairness on both speech and silence were significant [speech: $F(1,63) = 7.23$, $p = .009$; silence: $F(1,63) = 5.0$, $p = .03$].

We now analyze the effect of the presence of the mirror on the cost scores of the negotiation. This analysis will be performed not on dyads, as the previous analysis was, but on individual participants. Three groups of participants can be identified: those who had no mirror in front of them (*duplex*), those who were on the side of the mirror where they could not see their partners (*nonvisual*), and those who were on the side of the mirror where they could see their partners (*visual*). The average cost over all eight maps was 46.56 ($SD = 6.97$) for the duplex participants, 45.41 ($SD = 10.03$) for the nonvisual participants, and 45.66 ($SD = 8.99$) for the visual participants. None of these differences reached conventional levels of significance (all Fs < 1), supporting the conclusion that the presence of the mirror did not have an effect on the final outcome of the negotiation.

Discussion of Modality Effects on Performance and Efficiency

To summarize, the previous analysis revealed the following:

1. Neither the presence of the one-way mirror nor the side on which the participants in the mirror condition were seated had any effect on the outcome of the negotiation.

Table 3
Mean Duration of Speech and Silence (in Seconds) by Condition

Factor		No Mirror	Mirror
Fair	Speech	48.2	60.6
	Silence	57.7	120.6
Unfair	Speech	108.4	88.9
	Silence	109.0	177.4

2. The presence of the one-way mirror made negotiations last longer, and this was due to the participants' being silent for longer periods of time.

3. The more intrinsically unfair the map, the more speech was produced and the longer was the silence time.

To start with the last finding, it shows that one important criterion of SLOT—namely, the possibility of varying the level of competitiveness of the negotiation—was met. With respect to the first two findings, it is interesting to compare them with the results by Drolet and Morris (2000), which are remarkably similar. In their study, they completely blocked visual contact between the negotiators and compared negotiation times and results with those of another group, for which visual contact was possible. They found that negotiation times were 25% longer in the condition without visual contact (in our study, they were about 50% longer in the mirror condition) and that the negotiation results were unaffected. They attribute their findings to a higher level of rapport (see Tickle-Degnen & Rosenthal, 1990) between participants who could see each other, and we have no reason to disagree with their explanation. The study reported here, in combination with the study by Drolet and Morris, suggests that the availability of the visual modality affects negotiation times but not the outcome, and that these effects on negotiation time and outcome are independent of whether the visual modality was blocked for both or just one of the negotiators. Apparently, in order for the visual modality to facilitate the development of rapport, the participants' possibility of exchanging visual signals must be *mutually manifest* (Sperber & Wilson, 1995). This finding also has important consequences for HCI research in which 3-D animations of a human-like head are used, but in which the computer cannot perceive the visual signals of the human user. If the goal is to increase efficiency in HCI by creating rapport, this will work only if users have at least the illusion that the computer can see them as well. How this illusion can be created without giving the computer access to the visible behavior of the user is an important question that will be addressed in future research.

Analysis and Automated Recognition of Multimodal Pen Gestures

The combination of pen and speech input is widely used in multimodal human–computer interaction (Oviatt, 1999). In general, the use of two or more modalities is required in natural interaction dialogs with the computer to make it possible to disambiguate user utterances in either modality. In this section, we describe how the manual analysis and annotation of the data acquired through SLOT is used to develop automated systems for the recognition of pen gestures produced in the SLOT task.

In any pattern recognition task such as the present example, knowing about the participants' speech and gesture repertoires and the availability of properly annotated data is of paramount importance. These are needed to provide insight into how users generate and use speech and gesture, to train and test the required pattern recognition

algorithms, and also for the design of user interfaces (Oviatt et al., 2000; Potamianos, Kuo, Pargellis, Saad, & Zhou, 1999; Rossignol et al., 2003). We have annotated the pen gestures and speech utterances in order to identify possible *classes* of pen gestures. It was found that the interpretation of pen gestures is generally not possible without the accompanying speech and the visual context in which they occurred (i.e., the SLOT map on the whiteboard). This is a phenomenon that has also been observed in manual gestures and speech (de Ruiter, 2000; McNeill, 1992). Note that the gestures recorded in the SLOT environment were produced for the benefit of the understanding of another human, and that the recipient of the pen gestures understood them without any apparent problems.

Two annotators marked and classified a total of 454 pen gestures in the SLOT data, using the recorded image of the pen gesture and the accompanying speech of the producer of the pen gesture. Four main classes of pen gestures where identified. In the data, the most frequent type was the *trajectory* ($N = 313$). A trajectory is an uninterrupted line connecting two or more cities, corresponding to a (partial) route through the negotiated map. The second class that we identified was the *marker* class ($N = 113$). A marker is used to indicate a certain point on the map, usually a city. The third class is called *directional pointing* ($N = 18$), which consists of a line or arrow that indicates a certain direction on the map. The final class is called *area* ($N = 10$). This is an encircling type of pen gesture, indicating an area on the map that contains more than one city. We refer to Figure 5 for examples of recorded pen gestures from each of the four classes.

For the development of automated pen-gesture recognition systems, we used the even-numbered gestures in the data set for the training of a classification method, and we used the odd-numbered gestures for testing. For the subsequent computational analysis, the high-resolution pen data from the tablets that are stored on the SLOT computer were used. As is apparent from the second row in Figure 5, samples from different classes can have highly similar shapes.

Two well-known classification methods were used for the automatic recognition: the *multilayered perceptron* (MLP; see Rumelhart, Hinton, & Williams, 1986) and the *k-nearest neighbor method* (KNN; see Duda & Hart, 1973). From each pen gesture, seven features were extracted, to be used as feature vectors for training and testing the classifiers. The features we used were: (1) the number of points with high curvature (*curvature* being defined as angular velocity divided by absolute velocity); (2) the total number of samples (sampling rate = 200 Hz); (3) the spatial distance between the first and last samples; (4) the spatial distance between the first sample and the sample occurring after half the duration of the pen gesture (which we call the *halfway sample*); (5) the spatial distance between the halfway sample and the last sample; (6) the surface area of the bounding box enclosing the entire gesture; and (7) the ratio of the width and the height of the bounding box enclosing the gesture.

For details about the type of features used in pen gesture recognition, see Rubine (1991).

Both the KNN and the MLP used the same training data and testing data. The recognition result of the MLP was

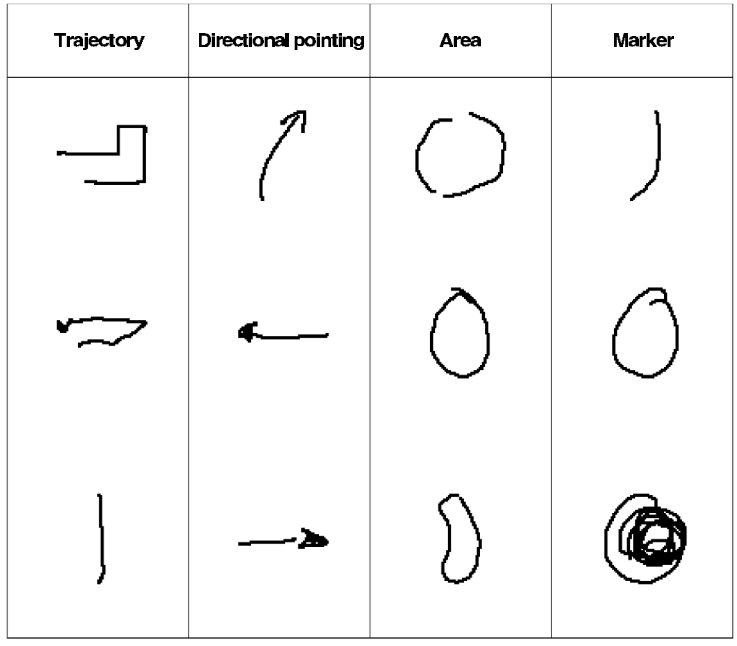

Figure 5. Examples of recorded pen gestures of the four different classes.

Table 4
Confusion Matrix for Pen Gesture Classification Using MLP

	Gesture Is Recognized as:				
Gesture Is of Type:	Marker	Area	Trajectory	Directional pointing	*N*
Marker	79.37%	0%	17.46%	3.17%	63
Area	100.00%	0%	0%	0%	6
Trajectory	2.01%	0%	97.99%	0%	149
Directional pointing	0%	0%	55.56%	44.44%	9

88.1%, which is much better than the 71.8% we obtained with the KNN (using $k = 5$). The MLP used seven input nodes, two hidden layers with eight nodes each, and four output nodes. In Table 4, the confusion matrix for the MLP analysis is shown.

As can be seen from the confusion matrix, the MLP is unable to distinguish between area and marker: Every area is recognized as a marker. Also, directional pointings are more often than not recognized as trajectories. This is because the pen gestures of markers and areas, and also those of trajectories and directional pointings, often have similar shapes (see the middle row of Figure 5). Furthermore, the total number of area gestures and directional pointings in our data set is low, making correct recognition more difficult due to limited training data.

Clearly, more information is required if similar gestures of different classes are to be distinguished. Within the SLOT domain, another potential source of information is the location of the cities on the map. This information could be used to disambiguate between circular markers (which enclose only one city) and areas (which enclose two or more cities). The distinction between trajectories and directional pointings would be improved by use of the information that trajectories always pass through cities on the map, whereas directional pointings rarely do.

The approach that is explored here is to consider distinctive fragments of the accompanying speech as new, orthogonal features. For each pen gesture, we located the speech turn of the producer of that pen gesture that was temporally closest to the occurrence of the pen gesture. From the transcription of that turn, we counted the occurrences in speech of five words that often accompany pen gestures. The relative frequencies of these words could perhaps be used to disambiguate conflicting cases. The words that we counted were the Dutch words *die*, *deze*, *zo*, *hier*, and *daar* [English "that"/"those," "this"/"these," "this way" (manner or path), "here," and "there," respectively]. These words were chosen because they often occur in utterances in which gesture and speech are combined. In Table 5, the relative frequency with which these words co-occurred with the pen gestures from the four different classes is presented.

It can be concluded that the occurrence of the word *hier* (English "here") indicates that it is likely that a marker or area was generated by the user. Furthermore, if the word *zo* (English "this way") is used, it is likely that a trajectory or directional pointing was generated.

The use of multiple classifiers for difficult pattern recognition problems is a well studied technique. The approach followed here is described in Vuurpijl, Schomaker, and Van Erp (in press), in which the comparison of outputs from multiple classifiers is used to resolve conflicting situations. To distinguish between the confusing gesture classes, the following two-staged classification scheme was used. During the first classification stage, the results from both classifiers were considered on the basis of shape information only. In cases in which both the KNN and the MLP yielded the same output, this was considered as the final output of the combined classifier. If each yielded a different output, the speech signal was considered to rule between outcomes. If the speech features could not be used to disambiguate between classes, the outcome of the MLP was used. The latter decision was based on the fact that the MLP yields the highest recognition results. Note that with this method, if both classifiers are wrong, there is no way to correct the output of the classification.

In all cases in which both classifiers agreed during the first classification stage, the resulting classification was correct. In 63 cases, a conflict between the two classifiers was observed. In 13 of these 63 confusions, the two classifiers yielded different outcomes, but both were wrong. In 22 cases, the speech signal did not contain extra information (e.g., there were no occurrences of words). From

Table 5
Probability of Occurrence of Words in Concurrent Speech for Each Pen Gesture Class

	Pen Gesture Class				
Concurrent Speech	Marker	Area	Trajectory	Directional pointing	*N*
die ("that"/"those")	.30	.22	.24	.15	84
deze ("this"/"these")	.59	.78	.24	.08	125
zo ("this way")	.29	.67	1.00	1.00	234
hier ("here")	1.00	1.00	.13	.00	161
daar ("there")	.05	.11	.07	.23	24

these 22 cases, the MLP outcome resulted in four errors. For the remaining 28 cases, the speech features were considered to determine the final outcome. In 16 cases, this resulted in an error, amounting to a total of 33 (13 + 4 + 16) errors, yielding a recognition performance of only 85.5%.

Although our present automatic gesture recognizer does not perform at the desired near-100% level, using the natural SLOT data has given us some valuable insights into the nature of human–human pen gestures. Researchers in multimodal HCI often discuss and try out multichannel integration using pen gestures and speech, but our data (both the qualitative impressions from the manual annotations and the results from automatic recognizers) suggest that the visual context in which pen gestures occur, and possibly also a deeper level of semantic analysis of the speech, are indispensable for accurate automatic gesture recognition. The SLOT platform is well suited for generating accurately recorded natural pen gestures intended for other humans, together with coexpressive speech, which occur in well-defined 2-D contexts.

DISCUSSION

The SLOT route negotiation task is a communication task that imposes some structure on the communication without disrupting the natural flow of conversation. Its structure is symmetrical, in that both participants in an experiment have the same task and role. Nevertheless, the platform allows for the systematic research of asymmetric modality and channel distributions without changing the basic structure of the task. Furthermore, it enables the efficient recording and analysis of speech, pen gesture, and several channels of nonverbal communication.

In this article, we have presented two illustrative research projects using SLOT. First, we showed an example of an asymmetric modality manipulation (in this case the visual modality) that yielded nontrivial results regarding efficiency and performance in negotiation. The finding that efficiency in interaction is served by *mutually manifest* visual communication is relevant both for fundamental communication research and for HCI applications involving an animated head. In our second study, we tried to train automatic recognizers on pen gestures that were not performed for the computer, but used in human–human communication in the same way as two people might use a real whiteboard together. The results of this study serve to focus and guide further research in human pen gestures and their relationship to the accompanying speech and the visual context.

These projects are by no means the only ones that are currently performed with SLOT. Studies are in preparation that compare the pen gestures from SLOT with spontaneously produced 3-D hand gestures (see de Ruiter & Wilkins, 1998). Furthermore, the speech data from SLOT experiments are being used to develop quantitative models of turn-taking and to investigate the use of visual signals to regulate turn-taking behavior. Another study is aimed at comparing efficiency in the SLOT task with and without the use of the pen. Finally, the *negotiation moves* in SLOT are annotated and analyzed in detail for study of the structure of negotiation under different levels of competitiveness and with varying channel availability.

The SLOT platform has not only served us well to address the issues we had in mind during its design, but it has also inspired many new research directions. We invite other researchers in the field of multimodal communication to consider using this platform as well.

Availability

Researchers or institutions interested in using the SLOT platform are advised to contact the first author for further information on how to obtain copies of our software and manuals. All the hardware components used in SLOT are available on the market.

REFERENCES

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., & Thompson, H. S. (1991). The HCRC Map Task Corpus. *Language & Speech*, **34**, 351-366.

Archer, D., & Akert, R. M. (1977). Words and everything else: Verbal and nonverbal cues to social interpretation. *Journal of Personality & Social Psychology*, **35**, 443-449.

Argyle, M., Salter, V., Nicolson, H., Williams, N., & Burgess, P. (1970). The communication of inferior and superior attitudes by verbal and nonverbal signals. *British Journal of Social & Clinical Psychology*, **9**, 222-231.

Brown, R. (1986). *Social psychology* (2nd ed.). New York: Free Press.

Brugman, H., Russel, A., Broeder, D., & Wittenburg, P. (2000, May). *EUDICO. Annotation and exploitation of multi media corpora.* Paper presented at the International Conference on Language Resources and Evaluation, Athens.

Burkard, R. E. (1979). Traveling salesman and assignment problem: A survey. In P. L. Hammer, E. L. Johnson, & B. H. Korte (Eds.), *Discrete optimization* (Vol. 1, pp. 193-215). Amsterdam: North-Holland.

Clark, H. H., & Krych, M. A. (in press). Speaking while monitoring addressees for understanding. *Journal of Memory & Language.*

Cunningham, D., Breidt, M., Kleiner, M., Wallraven, C., & Bülthoff, H. (2003, May). *How believable are real faces? Towards a perceptual basis for conversational animation.* Paper presented at the Conference on Computer Animation and Social Agents (CASA), New Brunswick, NJ.

de Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 284-311). Cambridge: Cambridge University Press.

de Ruiter, J. P. (2003). Research on modality effects on performance quality and efficiency (IST COMIC Public Report). Available at http://www.hcrc.ed.ac.uk/comic/documents.

de Ruiter, J. P., & Wilkins, D. P. (1998, December). *The synchronisation of gesture and speech in Dutch and Arrernte (an Australian Aboriginal language): A cross-cultural comparison.* Paper presented at the Conférence Oralité et Gestualité, Besançon, France.

Drolet, A. L., & Morris, M. W. (2000). Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Journal of Experimental Social Psychology*, **36**, 26-50.

Duda, R., & Hart, P. (1973). *Pattern classification and scene analysis.* New York: Wiley.

Ellsworth, P. C., & Ludwig, L. M. (1972). Visual behavior in social interaction. *Journal of Communications*, **22**, 375-403.

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers.* New York: Academic Press.

Kendon, A. (1972). Some relationships between body motion and speech. In A. W. Sigman & B. Pope (Eds.), *Studies in dyadic communication.* New York: Pergamon.

Krauss, R. M., Apple, W., Morency, N., Wenzel, C., & Winton, W.

(1981). Verbal, vocal, and visible factors in judgments of another's affect. *Journal of Personality & Social Psychology*, **40**, 312-319.

LEVELT, W. J. M. (2001). *SLOT: A spatial logistics planning task for COMIC* (Unpublished report). Nijmegen: Max Planck Institute for Psycholinguistics.

MCNEILL, D. (1992). *Hand and mind*. Chicago: Chicago University Press.

MEHRABIAN, A., & WIENER, M. (1967). Decoding of inconsistent communications. *Journal of Personality & Social Psychology*, **6**, 109-114.

OVIATT, S. [L.] (1999). Ten myths of multimodal interaction. *Communications of the ACM*, **42**, 75-81.

OVIATT, S. L., COHEN, P. R., WU, L., VERGO, J., DUNCAN, L., SUHM, B., BERS, J., HOLZMAN, T., WINOGRAD, T., LANDAY, J., LARSON, J., & FERRO, D. (2000). Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions. *Human–Computer Interaction*, **15**, 263-322.

POTAMIANOS, A., KUO, H., PARGELLIS, A., SAAD, A., & ZHOU, Q. (1999, June). *Design principles and tools for multimodal dialog systems*. Paper presented at the ESCA Workshop on Interactive Dialog and Multi-modal Systems, Koster Irsee, Germany.

ROSSIGNOL, S., TEN BOSCH, L., VUURPIJL, L., NEUMANN, A., BOVES, L., DEN OS, E., & DE RUITER, J. P. (2003, June). *Multi-modal interaction for bathroom design: The role of human factors*. Paper presented at the HCI International 2003, Crete, Greece.

RUBINE, D. (1991). Specifying gestures by example. *Journal of Computer Graphics*, **25**, 329-337.

RUMELHART, D. E., HINTON, G. E., & WILLIAMS, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318-362). Cambridge, MA: MIT Press.

SPERBER, D., & WILSON, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Oxford: Blackwell.

TICKLE-DEGNEN, L., & ROSENTHAL, R. (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, **1**, 285-293.

VUURPIJL, L., SCHOMAKER, L., & VAN ERP, M. (in press). Architectures for detecting and solving conflicts: Two-stage classification and support vector classifiers. *International Journal of Document Analysis and Recognition.*

NOTE

1. The complexity of this computation increases exponentially with the number of target nodes, which means that in practice it is feasible only for maps having a maximum of 10 targets.

(Manuscript received September 19, 2002;
revision accepted for publication May 14, 2003.)