Exploring True Multi-User Multimodal Interaction over a Digital Table

Edward Tse[1], Saul Greenberg[1], Chia Shen[2], Clifton Forlines[2], Ryo Kodama[2]
[1]Dept of Computer Science, University of Calgary, Calgary, Alberta, Canada T2N 1N4
[2]Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, Massachusetts, U.S.A. 02139
+1 403 220 6087
tsee@cpsc.ucalgary.ca, saul.greenberg@ucalgary.ca, {shen, forlines, kodama}@merl.com

ABSTRACT
True multi-user, multimodal interaction over a digital table lets co-located people simultaneously gesture and speak commands to control an application. We explore this design space through a case study, where we implemented an application that supports the KJ creativity method as used by industrial designers. Four key design issues emerged that have a significant impact on how people would use such a multi-user multimodal system. First, **parallel work** is affected by the design of multimodal commands. Second, individual **mode switches** can be confusing to collaborators, especially if speech commands are used. Third, establishing **personal and group territories** can hinder particular tasks that require artefact neutrality. Finally, timing needs to be considered when designing **joint multimodal commands**. We also describe our model view controller architecture for true multi-user multimodal interaction.

Categories and Subject Descriptors
H5.2 [**Information interfaces and presentation**]: User Interfaces. – *Interaction Styles.*

General Terms
Design, Human Factors

Keywords
Tabletop interaction, multimodal speech and gesture interfaces, computer supported cooperative work.

1. INTRODUCTION
Previous systems have explored how existing off-the-shelf single user applications can be wrapped in a way that allows multiple people to interact with it over a digital table via speech and gesture commands [20]. Studies revealed that these speech and gesture commands can be beneficial as they serve as both commands to the computer and as communication to other collaborators [21]. Yet these wrappers are limited by the one user per computer assumption of the underlying application and operating system. Multiple people cannot work in parallel because the computer can only accept a single stream of input.

Cite as:
Tse, E., Greenberg, S., Shen, C., Forlines, C. and Kodama, R. (2007) **Exploring True Multi-User Multimodal Interaction over a Digital Table.** Report 2007-877-29, Department of Computer Science, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada. T2N 1N4. August.

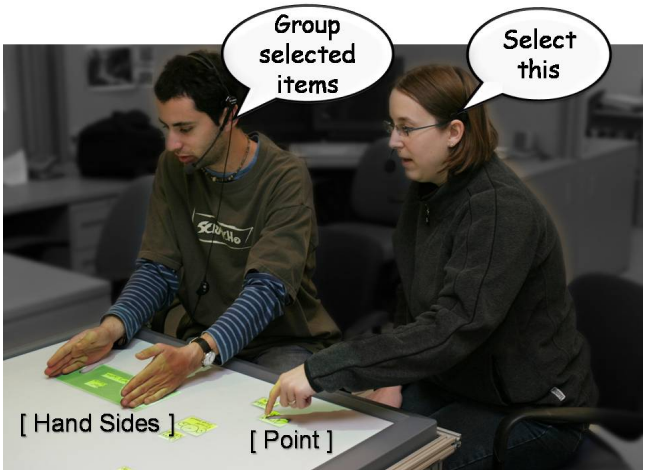

Figure 1. A two person grouping hand gesture.

In contrast, *true* multi-user multimodal interaction recognizes that multiple people are interacting with it, and allows co-located people to simultaneously gesture and speak commands to the groupware application (Figure 1). We explore the design space of such systems through a case study, where we implement an application – the Designers' Environment – that supports the KJ creativity method as used by industrial designers.

To preview what is to come, the KJ creativity method has four basic steps: creating notes, grouping notes, labelling notes, and relating notes. The Designers' Environment supports these four basic steps by **idea sketching** onto digital notes, **grouping** using hand gestures, voice selections, and multimodal selection, **annotation** using handwriting, and **relating** by using sizing gestures or by linking notes with a multimodal command (Table 1). Four key design issues arose during system development.

1. **Parallel work:** The design of multimodal commands can greatly influence collaborators' propensity to engage in parallel work. For example, if the majority of commands were via the speech channel, people may be unwilling to talk over each other, which in turn would favour sequential work.
2. **Mode switching:** While a true multi-user multimodal system provides the potential for independent modes of action (e.g., one person is annotating while another is moving artefacts) confusion arises when people forget what mode they are in. This problem is exacerbated by publicly seen and heard multimodal commands that can give others the false impression that they are all in the same mode.

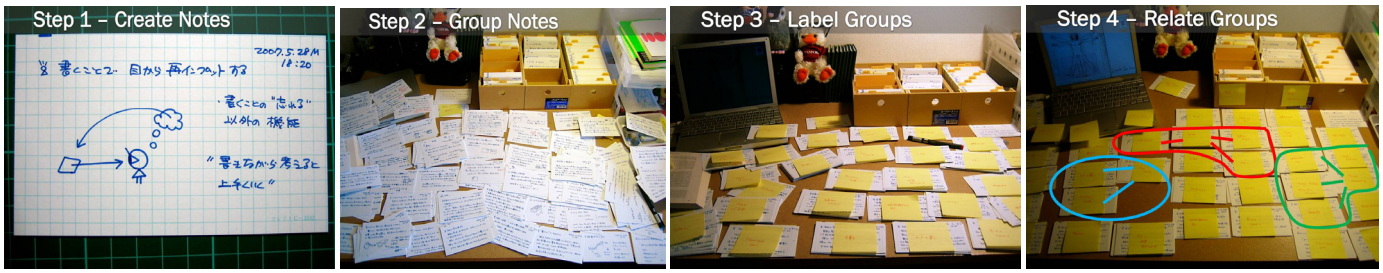

Figure 2. The four steps of the KJ creativity method, images from http://www.flickr.com/photos/hawkexpress

3. **Personal and group territories:** The design of multimodal commands can significantly influence the establishment of personal and group territories in the collaborative workspace [12][15]. By paying attention to how speech and gesture commands are used, designers can develop systems to support or hinder the establishment of personal and group territories.
4. **Joint multimodal commands:** The multimodal interactions of collaborators can be combined into a single joint action (e.g., the joint actions of Figure 1, where one person is extending another person's selection). These commands need to carefully consider the time window for joint inputs to be recognized, or erroneous command additions and omissions may result.

The next section briefly introduces the KJ creativity method. We then describe the Designers' Environment system and detail the four design issues described above. We conclude by describing our implementation and examining related work.

2. THE KJ CREATIVITY METHOD

We were approached by a group of industrial designers from a consumer electronics company to develop a system to improve their initial brainstorming activity via the KJ Method. Designers and marketers typically use the initial phases of the KJ Method to collaboratively brainstorm ideas and concepts for new products, to establish customer needs and to explore potential product features. The output of the KJ Method is a list of core needs and features that will be later used and refined by designers in their product sketches.

The paper version of the KJ Method is composed of four basic steps as illustrated in Figure 2. First, multiple people write customer needs, product ideas and comments onto 4x5" cards (Step 1). Each card can be brief with only a title, or it can provide additional details such as a description and illustration. Second, each card is randomly distributed by either shuffling the cards to each collaborator or by shuffling the cards around a table and having each person work on the cards that are closest to them. They then group together similar ideas into piles (Step 2). Third, piles are labelled according to the need/idea that they represent (Step 3). Finally, collaborators relate the notes by drawing links between groups and creating meta-groups representing common themes (Step 4).

3. THE DESIGNERS' ENVIRONMENT

As a case study of a true multi-user multimodal system, we designed and implemented a groupware system for the KJ method that lets co-located people work together over a digital table and personal tablet PCs. Our system is called The Designers' Environment, and we see two people using it in Figure 1. Multiple people create, group, label and relate digital notes using speech and gesture commands; these are summarized in Table 1. In this section, we describe the physical form factor and interactions for each step in this process, and how we leveraged the capabilities of true multi-user multimodal interaction.

3.1 Creating Notes

Notes are the basic unit of the KJ Method. As shown in Figure 2 (Step 1), people create paper notes on 4x5" cards. Similarly, the Designer's Environment supports the creation of basic digital note. However, it also allows people to use a note's contents to search the web for related images and information, and to create new notes based on search results. Multiple people can do all these actions in parallel.

The basic note: Each participant independently creates digital 4x5" cards through a pen-based Tablet PC running our note writing application (Figure 3). They can quickly sketch and hand-write ideas, needs, descriptions and illustrations onto this card. When complete, they send the card to the digital table by tapping a 'Send to Table' button (Figure 3, top right). This automatically places the note in a random location on the digital table, thus mimicking the shuffling of cards of the KJ Method.

Searching and importing images and web pages: People can also import their digital images or snapshots of web pages into a note; while not part of the KJ method, we believe this extra information could help people's discussions. Images and web page content capture experiences, emotions, and concepts that may be hard to express through words or illustration [13].

People can use two methods to achieve this. First, if the person already has a saved image handy, he or she can drag a previously saved image into a 4x5" card, preview and optionally resize it, and then send it to the table as before (Figure 3, the note in the background). Second, a person can search for appropriate web content for related images or web pages and import those. One hand writes the search terms on a note, and then taps a 'Web' button (Figure 3, centre) to begin the search. Handwriting recognition translates the writing into machine-readable text, and feeds the result into Google Image Search. The web page of results is displayed. At

Figure 3. Tablet Note Writing

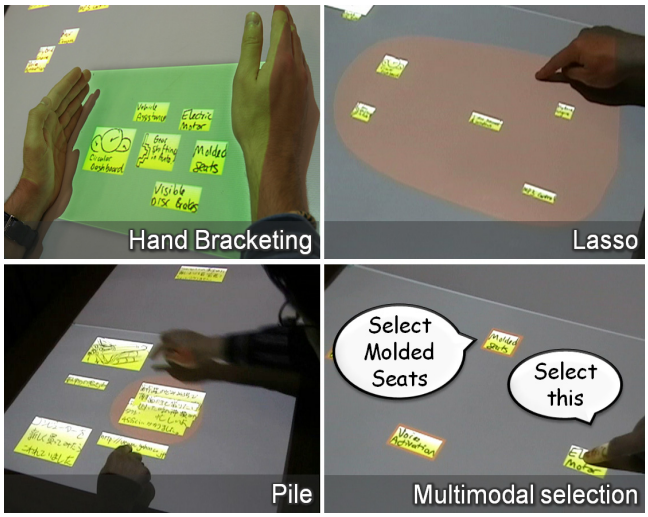

Figure 4. Grouping interactions

this point, one can continue to navigate the web to a particular web page or image by navigating links. Clicking the 'Web' button a second time captures the image or web page on the 4x5" card, which can then be moved to the table. Once on the table, all images can be resized as needed.

Recognizing note contents and using it for searches: After a note is added to the table, its text is automatically processed by a handwriting recognizer and the result is stored as meta-data along with each note. People can reveal this data by several means. Pointing to a note and saying "show recognition" temporarily raises a popup containing the recognized text. Alternately, one

Table 1. The Designers' Environment speech/gesture interface

Speech/Multimodal commands		Gesture commands	
[point to note] **Show recognition**	Reveals the text recognition result for the selected note	**One finger (on a note/group)**	Moves note or group
Recognize this *[note]* / **all notes**	Converts a note into text	**One finger (Empty space)**	Creates a lasso group
[point to note] **Find related images**	Opens a web browser with the note text as a search query	**One finger (annotation mode)**	Draws labels and links
[point to note] **Convert to note**	Converts a web items into an image	**Two fingers**	Zoom note/group
Annotation Mode	Allows one to draw links/create labels	**Five fingers**	Moves group or workspace
[select group] **delete this**	Deletes the selected group	**One hand**	Erases annotations
Select *<say note text>* / **this** *[note]*	Selects a note	**Two hand sides**	Create group between hands
Group selected items	Converts selected notes into a group	**link this** *[note]* **to this** *[note]*	Creates a link for two notes
[select group] **Arrange / Sort Alphabetically**	Tiles items within a group (sorted alphabetically)	**zoom this** *[note]* / **all**	Zooms the camera to a note / all notes
Restore *[group]*	Returns items in a group to their original position	**Make this** *[group]* *<say a colour>*	Changes the group colour

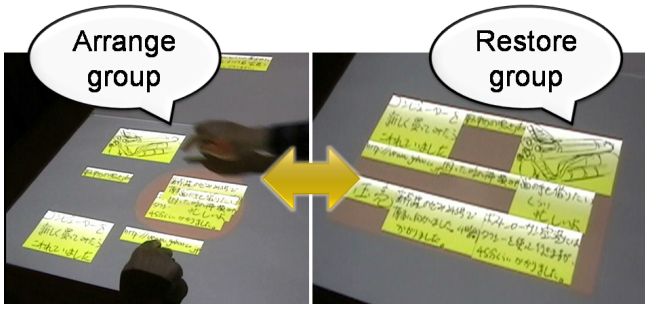

Figure 5. Expanding and collapsing groups

can transform one or more handwritten note into text by saying "[point to note] recognize this", or saying "recognize all notes".

People can also use the note's meta-data of recognized text to search for information related to a note's contents. One searches for related images through the multimodal command "[point to note] find related images"; this triggers a web search on Google Images using the terms in the recognized text. That page is projected onto the digital table. As was done on the tablet, the person can then follow links until a desired web page or specific image is found, and then convert that into a new note by saying "convert to note".

3.2 Grouping Notes

The second step of the KJ method is to group or pile related notes on the table (Figure 2, Step 2). Grouping is supported through several gesture and speech actions on the digital table, as illustrated in Figure 4 and described below. To encourage discussion and coordination between collaborators, all grouping is done on the digital table rather than through the Tablet PCs. All grouping actions can be done simultaneously by multiple people.

Groups are visually represented using a light highlight color (red, green, blue and tan). Each highlighted colour represents the individual that created the grouping. This makes it clear who created each group, which can help facilitate later discussion.

Hand bracketing and lasso grouping: People naturally explore item grouping by moving related notes next to each another. They do this on our table by using a single finger to move either single notes or previously established groups. Participants can then explicitly group each note by either using two hands to bracket an area (Figure 4, top left), or by using a single finger to draw a lasso around the desired notes (Figure 4, top right). In both cases, notes within the contained area are automatically included in the group selection. As an aside, notes can overlap so they look more like piles (Figure 4, bottom left). These groups can then be moved around the table by using five fingers (a grabbing gesture) or by using a single finger on an area within the group that does not contain a note.

Alternately, an empty group or pile can be created by lassoing or hand bracketing an area containing no notes. Notes can be later dragged into this area, which automatically includes them in that group. Empty groups can be deleted using the "[select group] delete this" multimodal command.

Searching notes by speech: Sometimes, people may want to find and select a note that is out of reach, covered by other notes, or lost in the clutter. To help in these cases, one can find notes using speech. Recall that a note's handwritten contents are recognized

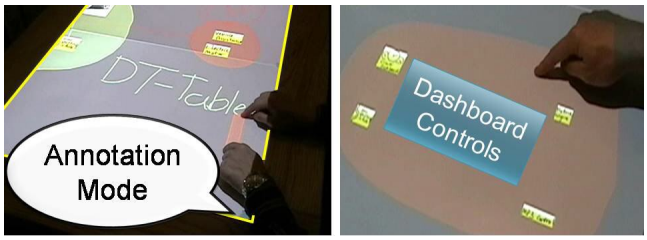

Figure 6. Group Annotation Methods: table annotation (left), adding a group title (right)

and stored as meta-data (§3.1). Under the covers, this meta-data is also automatically added to a speech recognizer. When a person says "select <*say note text*>", the note that best matches that text is highlighted with the users default colour (Figure 4, bottom right). Currently, our recognition system works best if the entire note contents are spoken verbally. We realize that this is hard to do for notes containing large descriptions, and that it does not work over images and illustrations. A future system would benefit from a better speech search system.

Multimodal selection and grouping: Sometimes, people may want to select and group distant notes that are scattered around the table. They do this by first selecting one or more notes with speech (described earlier) or by doing a multimodal selection: "[point to note] select this", and then saying "Group selected items". All selected notes are then moved into a single neatly arranged group. People can collaboratively extend each other's selection, as illustrated in Figure 1.

Rearranging groups: Next, people can rearrange notes in a groups through multimodal speech commands, either to reveal the hidden content of overlapping notes (as in Figure 5, left) or to bring distant notes within a group closer together (as in Figure 4, top right). When one says "Arrange group", a smooth animation sequence re-arranges these notes as non-overlapping adjacent tiles as seen in Figure 5 (right). Arranged notes can be returned to their original location by using the "Restore group" speech command (Figure 5, right).

Sorting notes within a group: The "Sort alphabetically" speech command rearranges a group's notes into alphabetical order. As before, the "Restore group" command returns the group to its original position. Non-textual illustrations appear at the top left corner of a sorted group.

Panning / zooming: As an aside, there may be insufficient space on the table to arrange all notes and groups. Consequently, we allow the entire workspace to be panned using a 5-finger gesture, and zoomed with a 2-finger gesture in an empty area (§5.3).

3.3 Labelling Groups

The third step in the KJ Method is to label groups and areas with a descriptive name (Figure 2, Step 3).

Labelling groups is supported in two ways. The first method is to write the label directly on the digital table surface using a finger (Figure 6, left), and the second is to create a special title note through the Tablet PC (Figure 6, right) and add that to the table.

Tabletop labelling: A person creates a textual label (and other marks) on the digital table by saying "Annotation Mode". A yellow border appears around the digital table indicating that a mode change has occurred for all users. From that point on any

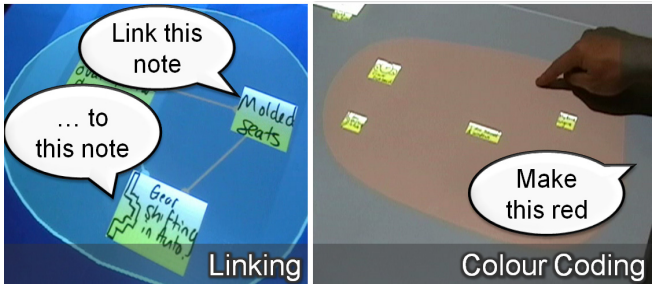

Figure 7. Two methods for relating notes and groups

person can write and draw on the table with their finger, or erase through a whole hand erasing gesture. Saying "Annotation Mode" again exits that mode. Marks always appear atop other table artefacts, i.e., notes and groups. The restriction is that tabletop annotations cannot be moved. If a group is repositioned, any corresponding tabletop annotations must be erased and drawn again. While restrictive, tabletop labelling is at its best when regions of the table space can be split into theme areas.

Tablet labelling: Alternately, people can use the note writing application on the Tablet PC to create special 'title notes' that can become incorporated within a group. One switches to title notes by pressing a title note toggle switch (Figure 3, lower middle); the color of the note changes to dark aqua with white sketching, as with normal notes the person creates the label and presses the "Send to Table" button. Once on the table, a title note can be placed in a group and moved along with it. However, title notes have special features. They are always placed on top of other notes (Figure 6, right). When groups are re-arranged, either spatially or alphabetically, title notes are always placed as the first item in the top left corner.

3.4 Relating Groups

The final step of the KJ Method is to relate groups (Figure 2, Step 4). Typically, people draw links between related notes, and spatially create meta-groups that represent common themes. In the Designers' Environment, people can perform these actions through public speech and gesture commands over the digital table. First, related groups can be moved closer together. Second, meta-groups can be created and links between contained groups can be drawn. Third, groups and notes can be emphasized by resizing them or by colour coding them to highlight common group themes.

Arranging and emphasizing groups within the workspace: As mentioned in §3.2, groups can be moved and resized to affect their visual prominence. Because this usually runs into spatial constraints, people can also pan the workspace and zoom into areas to see more details. To view the contents of a single group in high detail one can select the group and say "zoom this" to see a smooth animation enlarging the group to fit the entire screen. To restore the view so that all notes are visible on the table users can use a "zoom all" speech command. Thus one can zoom in to inspect a group in detail, and then zoom out to move it to other related groups. We should mention that all workspace pan and zoom actions are global, and thus affect the entire table area.

Creating explicit meta-groups: Explicit meta-groups can be created in much the same way as groups are created. That is, people can create meta-groups by using two hand sides around existing groups (left) or they can lasso around existing groups

using a single finger (right). Meta-groups behave as regular groups and can be easily moved with a single finger on an empty area or with five fingers over the entire group.

Linking notes and groups: Notes and groups can be linked using the annotation functionality of the Designers' Environment. A person says "annotation mode", and then draws lines linking related items; these lines will dynamically follow the notes and groups they are connected to. A person can also use a special multimodal command "[point to note/group] link this [point to another note/group] to this" where the note is selected using a single finger (as illustrated in Figure 7, left). A directional arrow is drawn on the destination node or group to reinforce the order of the hierarchy.

Colour coding groups: Finally, groups with similar themes can share a common colour. By default, groups are given the default color of the individual that created it. To modify the colour of a particular group, a person selects it with a single finger and says "make this <*say a colour*>" (Figure 7, right).

4. DESIGN ISSUES

We developed the Designers' Environment to help us understand the design space of true multi-user multimodal interaction groupware for co-located interaction. Along the way, we encountered a number of design issues that we believe generalize to other groupware applications of this type. In this section, we detail these issues to provide insights for future designers.

4.1 Parallel work

As mentioned in the introduction, the primary benefit of a true multi-user multimodal groupware system is that they recognize that multiple people are interacting with it, and they can allow co-located people to *simultaneously* gesture and speak commands to the groupware application. However, we found that the design of multimodal commands can greatly influence collaborators' propensity to simultaneously interact with the system. Some of the factors that affect parallel work include the effect of using of voice commands, the work area size, and the gesture size.

The effect of voice commands: The computer is able to isolate and recognize the speech of multiple people because each person has their own microphone connected to separate speech recognizers. This means that multiple people can simultaneously issue voice commands to the system. The problem is that voice commands are also a very public action: all voice commands are audible by others in the collaborative process regardless of their current activity. Yet in practice social protocols discourage collaborators from speaking simultaneously over each other. Thus while the system allows parallel activity, people may choose to work sequentially instead. To mitigate this effect, we argue that designers of multimodal systems should avoid using voice commands for those actions that are likely to be done simultaneously by multiple people, but that they should use voice commands when people are likely to interleave their utterances [21]. For example, the KJ Method encourages people to jot down notes simultaneously. While we could have used voice recognition for rapid entry of notes, this may have inhibited simultaneous entry. Instead, we chose to let people enter notes using a pen on the tablet as this method lets people easily engage in parallel individual work.

The effect of small work areas and large gestures on individual work: People often work on highly individual tasks, even when working together towards a common goal. A gestural interface, especially one that requires large gestures over a small table, may affect people's ability to do their individual work in parallel with others. Other's gestures may be distracting or simply get in the way. On large tables, we know that people create personal work areas, which makes it easier to pursue individual wrok. [10][14][15]. In our case, the table was somewhat small. Thus we gave people individual Tablet PCs that serve as personal work areas, where people used it to create and publish notes. More generally, we argue that designers can increase the amount of parallel work by ensuring that each collaborator has enough space to serve as a personal work area. This could be part of the table, or it could be a separate device as we have done.

The effect of gesture size: Generally speaking, manipulating artefacts on larger work surfaces require larger gestures, e.g., as people move items, or when they group a large set of notes. This is beneficial as large gestures are more visible to other collaborators; they involve more motion and often involve the movement of the entire arm. This produces consequential communication that others can use for awareness and coordination [9]. However within the context of parallel work, this awareness can distract from individual concentration. Large gestures – in addition to requiring more time to complete – can shift the attention of collaborators away from their current task. Conversely a smaller gesture could result in actions that are less visible to others. A balance must be struck between the two. Designers wishing to increase parallel individual work could benefit from using smaller gestures. Conversely, if designers wish to increase collaboration and communication, they could benefit from using larger gestures For example, consider the difference between the two methods of creating labels in the Designers' Environment. The annotation method encourages collaboration, as one has to write in large letters directly atop the table. The tablet method encourages individual work, as it uses a small gesture area for writing on the tablet that is generally hidden from others. As another example, the large five finger gesture for moving groups is more visible than the one-finger gesture for moving a single note. This makes sense, as the group movement of Step 4 should inspire more conversation than note placement of Step 2 (Figure 2).

4.2 Mode switching

Since true multi-user multimodal systems recognize multiple people, each person can potentially be working and switching between separate individual modes. For example, one person could be in an 'annotation mode' while another is in a 'moving mode'. While seemingly beneficial, we encountered three key issues that made individual modes difficult for multiple people to understand; these are described below.

Public voice commands: Some voice commands trigger individual mode switching, while others could trigger global mode switching. The problem is that one person's voice command is heard by others, and this may mistakenly give others the false impression that they are in the same mode. For example, they may believe that the mode switch is global when it is in fact individual, or that the mode switch does not affect them when it is in fact global. One solution uses clearly stated voice commands that result in global mode switches (e.g., the 'Annotation Mode'

or 'Zoom this' speech command), while favouring gestures for commands that result in individual mode switches (e.g., moving and/or scaling a note).

Mode Visualization: If modes cannot be avoided in the design of a true multi-user multimodal system, some mode awareness is crucial to avoiding confusion regarding individual modes. This is usually done by altering the appearance of the objects affected. In distributed groupware, modes are often suggested by overlaying mode information on the telepointer. Yet in direct touch environments such as a digital table – where a pen or finger directly interacts with the digital display – information in the immediate vicinity of the touch point can be occluded by a person's hands. One partial solution slightly offsets the mode visualization so that it is not occluded by the hand [24]. Another solution could highlight what objects or areas are affected by a user's touch. For example, we already saw how a yellow border around the table's perimeter is used to mark the global annotation mode (Figure 6, left).

Global action awareness and interruption: Global actions on the table, while often necessary, can be problematic. They can be extremely distracting to others, or can lead to changes that others do not want. We suggest two ways of mitigating these problems. The first method is to increase people's awareness of another's global acts. We do this by leveraging public speech commands and by using large gestures. Both produce consequential communication that others can use for coordination [9]. *Continuous* global workspace manipulations leverage large whole handed gestures, and take time to do. *Discrete* commands unfold over time rather than done all at once by invoking an animation sequence so the action takes time to unfold. This increases their visibility. Second, we allow others to interrupt a global action when they do not agree with it or if they want to discuss it further. They can stop a person in the middle of a continuous manipulation, or touch the digital table to stop the animation of a discrete manipulation.

Artefact modes: The behavioural characteristics of an artefact should try to match people's expectations. Confusion can arise when people expect modes to extend only to the artefact. For example, people using the Designers' Environment expected to be able to write annotations on groups and have them move with the group. Instead, our system provided a global "Annotation Mode" where annotations would reside in a layer above other artefacts (such as notes and groups). Thus, annotations would stay in place even if a group was moved. An alternative may have been to provide a gesture or toggle switch to allow a single finger to annotate over a group.

In summary, we already know that modes in traditional interfaces should be avoided, although this is hard to do in practice. The same is true in multi-user multimodal systems. Designers should try to avoid introducing modes if they can. If modes are necessary, they should carefully consider people's mode expectations, how they understand each other's multimodal mode switching actions, and how modes are visualized on the surface.

4.3 Personal and group territories

We already know that people naturally establish personal and group territories in the collaborative workspace [12][15]. Consequently, the design of items, item containers and interactions on the workspace can have a significant impact on how people establish these territories. In this section we take a closer look at our artefacts and how they affect territories on the digital surface.

Items such as individual images, notes, or illustrations can be rotated to establish personal and group territories. Studies have shown that in collaborative work, people often rotate items towards themselves in their own personal territories, while items inside a group territory are typically rotated in a compromised orientation that all collaborators can read [12]. In some cases, the designer may want to promote the table as a group territory by hindering the establishment of personal item territories. This could be done by limiting the rotation of artefacts to a single orientation. For example, since the goal of the KJ method is to treat each idea as equal, collaborators sit along a single table side and organize notes using a common orientation (Figure 2). This practice is replicated in the Designers' Environment.

Item size on a digital table also has an impact, where its visual prominence influences the amount of attention it receives from collaborators [15]. For example, since we had limited screen real estate in our implementation, individual notes were cropped to remove any blank space not occupied by ink strokes. Thus each note would consume a minimal amount of screen space on the digital table. In practice however, this made notes with lots of text more visually prominent. Ideas would have been treated more equally if all notes were the same size as they are in the KJ Method.

Item containers are areas of the digital surface used to hold items, and can include tools for sorting, labelling, organizing and relating the contained items. To encourage discussion, we identified the person who made the group by coloring the container with that person's colour. Yet this can mistakenly give the impression that these containers are a personal territory, and that others should not manipulate its contents. In practice, container marking and location on the table can profoundly affect how they are viewed as personal vs. group territories.

4.4 Joint multimodal commands

Joint multimodal commands are commands issued by multiple people that overlap (or interleave) with one another in time. There are two types of joint commands: Independent joint commands and dependent joint commands.

Independent joint commands happen when people interact simultaneously (but independently) to achieve a joint action that might otherwise require several sequential steps by one person working alone. For example, in the Designers' Environment it is possible to move groups and to pan the workspace at the same time. Two people can work together to move a group to an off-screen location, e.g., one person could pan the workspace to the left to reveal an unused area as the other person moves the group to that spot. The result is that two people can move a group faster than a single individual could on their own.

For independent joint multimodal commands to work, people have to time and coordinate their actions closely. For this to work, the system must be very responsive. It has to animate changes in what seems like real time (so people get appropriate feedthrough of the other person's actions and resulting state). It must also carry out commands with no delay. Using the example above, if a person tries to drop a group as the other is panning, lags in either

the panning or in the drop action could result in the group being misplaced.

Dependent joint commands explicitly leverage the (speech and gesture) inputs of multiple people as a single command to the system. For example, Figure 1 shows an instance of joint grouping in the Designers' Environment. Two people are selecting notes simultaneously: the woman uses the command "select this [note]" while the man uses a hand bracketing gesture and says "group selected items." The end result is that the items selected by the man with the hand bracketing gesture and the single item selected by the woman will be combined into a single group. Under the covers, the interleaving multimodal grouping command searches for selection actions made by other collaborators within a 5 second threshold and adds them to the current grouping command.

When dependent joint commands are used, it can be difficult to achieve appropriate timing for closure. Usually joint multimodal commands accept input from collaborators within a time window of several seconds before and after the joint command has been issued. A *short joint command window* could result in collaborator's inputs being missed in the joint command. Consider Figure 1, if the man said "group selected items" before the woman had finished adding the last item to the group a new group would be created with an item missing, and the woman's selection would still be active. To correct this action, one would need to undo the selection and move the note to the newly created group. A *long joint command window* could result in a collaborator's input being included by mistake. If the woman in Figure 1 wanted to create a separate group with her multimodal selection this input might be erroneously included into the man's multimodal grouping command. To correct this action one would need to find the note within the newly created group, remove it and reselect it so that it can be added to a new group. The timing of such joint multimodal commands will vary depending on the nature of the commands and the kinds of interactions seen in practice.

5. ARCHITECTURE

While our primary goal is to consider design issues in true multi-user multimodal groupware systems, we also explain how we implemented such a system to benefit future system designers.

A common approach used to support multiple people simultaneously interacting over artefacts in a shared workspace in distributed systems is to use a model-view-controller (MVC) architecture. Items on the shared workspace are stored on a single machine (model) and can be manipulated by multiple people working over separate computers (controllers) and they can view the shared workspace on their own screens (view). As illustrated in Figure 8, we leverage the model-view-controller approach to support multiple co-located people using speech and gestures in the Designers' Environment. All of the individual notes and their respective groups/links are stored in a dictionary of key/value pairs: this is the model (Figure 8, row 3). Notes can be created using multiple Tablet PCs and manipulated using gesture and speech input, these are the controllers (Figure 8, rows 1&2). Collaborators can collaborate over a shared view on the digital table or they can view notes and manipulate notes after the meeting on a standard desktop computer, this is the view (Figure 8, row 4).

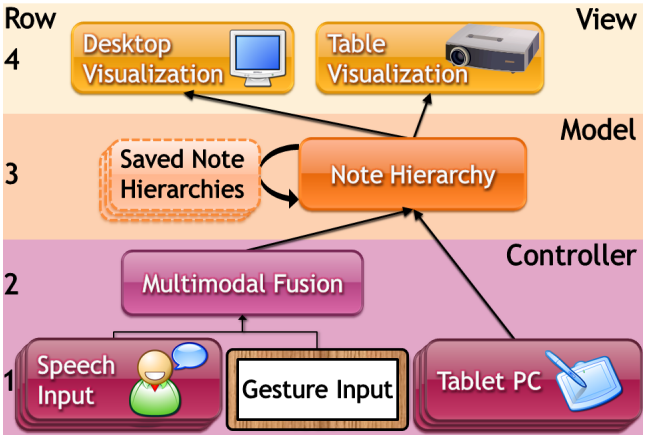

Figure 8. The Designers' Environment architecture

5.1 Model

To allow simultaneous input from multiple people in the co-located setting we first need a common model for manipulating data across multiple input devices and computers. We store all note hierarchy information in a dictionary of key-value pairs using a distributed networking toolkit (the GroupLab Collabrary [3]). This networking toolkit stores all note information on a common shared sever. It allows distributed and local controller applications to modify values within the dictionary and provides notifications of updated values to subscribed view applications. As illustrated below, we store each note and image as a separate key value pair, the key represents the note/image number and the value contains the ink strokes/image data serialized to a byte stream. The relations between notes are stored in the dictionary as a separate key representing an ID and a value containing two object IDs (notes, groups, images, etc) for starting and ending nodes. Groups contain a list of items, item positions and region information describing the bounding region and lasso points.

```
/note/1  = [ Ink Strokes ]
/image/2 = [ DesignSketch.jpg ]
/link/3  = /note/16, /image/7
/group/1/items = /note/1, /image/2, /group/4
/group/1/itemPositions = (X1,Y1), (X2,Y2), (X4,Y4)
/group/1/Region = (X, Y, Width, Height, Lasso)
```

5.2 Controllers

To support public and private actions in a multimodal co-located environment we need to provide collaborators with a variety of interaction options (as illustrated in Figure 8, row 1).

Handwriting recognition opens up new interaction possibilities for those used to pen and paper as it is possible for people to leverage more capabilities of computers. We perform handwriting recognition using the Microsoft Tablet PC SDK to convert each written note to text form. The recognition results are used in speech selections (e.g., "select tinted windows"), note conversion (e.g., "recognize this [note]") and web search queries on the Tablet PC or Table (e.g., "[point to note] find related images"). The most likely result is placed as a search query (to Google Images) into a browser window where the user can then click on links and images as desired. On the Tablet PC when a user clicks on the Send to Table button or when a user says "convert to note" on the digital table, the system checks if the current web address is an image. If it is an image, it sends that image to the server. Otherwise it uses the GroupLab Collabrary [3] to capture an

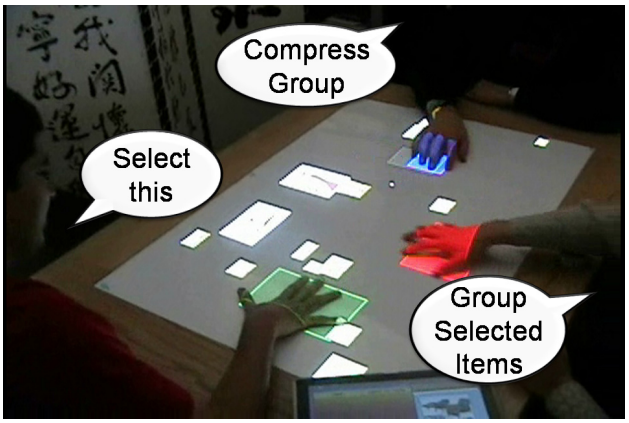

Figure 9. Double Diamond Touch & multi-user identification

image of that web page and sends the image to the server. As well, the ink strokes of a note are serialized for transportation using the Tablet PC SDK and sent to the server.

Speech Commands: The speech controller is tightly coupled to the contents of the model, where the speech vocabulary is extended on the fly to include the recognized text of new notes. For example, if someone uses the Tablet PC to add a note with "halogen headlights" as the contents, this will appear in the model and the speech command vocabulary will be expanded to include a "select halogen headlights" command. In our implementation, speech commands are recognized from multiple people using noise cancelling Labtec LVA 7330 headsets connected to individual Tablet PCs. It is also possible to recognize the speech of multiple people using multiple sound cards. From the software perspective, we use GSI Demo [22] a toolkit for gesture and speech interaction in co-located environments to send the speech recognition results from multiple tablet PCs to the server hosting the note hierarchy (our model).

Gesture Commands: Three basic functions are needed to support simultaneous artefact manipulation by multiple people on a digital table display. First, multiple simultaneous inputs (at least one per user) need to be detected by the digital table. Second as discussed in §4.1, the digital table must provide a reasonable amount of space for people to engage in parallel work. Finally, the digital table must be able to determine the person generating the touch if user identity is to be used in the application (e.g., for the multi-user multimodal fusion described next). In our implementation we originally chose to use a Diamond Touch table to detect multiple simultaneous inputs from multiple people. However, we wanted a larger table size so that three or four people could have their own non overlapping personal space for object manipulation. For this we chose a 148 x 116 cm Double Diamond Touch digital table [23] as illustrated in Figure 9. This one-off table comprises two standard Diamond Touch table, where it is treated as a contiguous surface. Finally, the digital table also provides user level identification, as illustrated by the different coloured boxes in Figure 9.

Multimodal Fusion: The speech and gesture actions of a single person or multiple people can be fused into a single command (Figure 8, row 2). Our system provides two types of multimodal command fusions: multimodal command unions and aggregate multimodal commands.

Multimodal command unions combine only a single speech and gesture input into a command e.g., "select this [note]". Multimodal command unions must include all appropriate speech and gesture components before it will be fused and passed on as a single command. If either the speech or gesture components are missing for a multimodal command union, the actions are ignored. For example, if the system recognizes a "select this" speech command and no one is pointing to a note then the command will be ignored. We extend the multimodal command unions presented by Cohen et al. [6] to a multi-user setting by allowing others to include their speech and gesture commands as input to a multimodal command union. As discussed in §4.4, this can result in recognition errors if people are engaged in parallel work and do not intend to complete the multimodal command of their partner. For example, a partner may be moving a note at around the same time another person says "select this". We mitigate this issue in two ways: first, we allow others to complete a multimodal command only if the originator of the command does not complete it within a reasonable amount of time. Second, we only allow others to complete a multimodal command after it has been made public through a speech command, thus any prior artefact manipulations are treated as parallel work.

Aggregate multimodal commands can accept multiple speech and gesture inputs of multiple people, e.g. Figure 1 shows one person saying "[point to note] select this" while another says "group selected items". The issue of parallel work is exacerbated in the aggregate setting because prior commands also need to be considered as input. For example, the prior selections in Figure 1 could be used as input to the grouping command. Our approach is to have a short (and customizable) time frame where the input of others can be included, also discussed in §4.4.

To implement multimodal fusion we use GSI Demo [22] to collect the speech and gesture actions of multiple people into a single computer. We then fuse the speech and gesture actions of multiple people (based on the rules described above) into commands.

Turn taking policies can be used to avoid conflicts when multiple people try to manipulate the same object simultaneously. For small items we used a first person wins turn taking policy. E.g., when multiple people try to move a note at the same time in the Designers' Environment, the first person to come in contact with the note must release it before others can manipulate it. This policy allows people to move objects around the display without the fear of others manipulating the object under them. Conversely, for global workspace manipulations (like panning described in §3.4) we used a last person wins turn taking policy. E.g., a panning action can be interrupted by placing five fingers on the table. In practice, most conflicts are resolved by social protocol; our turn taking policies are designed to merely to assist the social process already used over physical tables.

5.3 Views

The purpose of the view is to present a visualization of the model (in this case the note hierarchy) for co-located collaborators (Figure 8, row 4). A shared tabletop view is beneficial in the co-located environment because people can see the digital content and the body language of others at the same time. If a shared view is used it should provide sufficient resolution for multiple people to manipulate artefacts in parallel. As illustrated in Figure 9, we use two adjacent 1024x768 LCD projectors aligned along the long

edge produce a total resolution of 1536x1024 on the digital table. Both projectors are connected to the server and the two displays as treated as one large display. This resolution is sufficient for viewing and manipulating around 50 notes on the digital table but would be difficult to manage over 100 notes without scaling. The view also provides a seemingly infinite digital workspace so that people do not feel constrained by the resolution limitations of the digital table. We achieve this in the Designers' Environment by allowing people to zoom individual notes or the entire workspace (§3.4).

Animations: In the co-located environment, the view should provide smooth animations to avoid the jarring effects of artefacts disappearing from a users view unexpectedly. For example, the "group selected items" would be confusing to others who might be manipulating a selected note, thus we smoothly animate the movement all of the selected notes into a new compressed group. Similarly, global transitions such as the "zoom all" command smoothly animate the scaling of the entire workspace.

Desktop reviewing: After the steps of the KJ method have been completed in the co-located setting, collaborators may wish to review groupings and notes at a later time on their own desktops. To support this, we provide a "save note hierarchy" speech command that saves the current note hierarchy to a file (Figure 8, row 3). The model can be later loaded using the "load note hierarchy" speech command or using a keyboard on any desktop computer. Items can still be viewed and manipulated using the mouse if desired.

Piccolo Direct3D: In our implementation the view is achieved using the Piccolo Direct3D toolkit by Bederson et al. [2]. This toolkit provides a high level software application programmer's interface that allows notes to be efficiently rendered using graphics hardware accelerated primitives and textures. This toolkit also provides tools to simplify the animation of notes within the digital workspace and provides camera panning and zooming tools that make it trivial to develop a compelling zoomable workspace. Piccolo Direct3D provides the tools needed to create a smooth, responsive, and visually appealing user experience with the Designer's Environment.

6. RELATED WORK

Computer support for designers: There are many tools designed to support informal brainstorming by designers (e.g., Cognoter [7], Smart Ideas [SmartTech.com], PReSS [5]). Most existing systems are designed for a single person working with a keyboard and a mouse for jotting down ideas, or for a distributed group to work together. For example, PReSS [5] is intended solely for real time distributed interaction. While Cognoter is intended for people located in a meeting room, people contribute ideas by typing them on individual computers which then appear on a large wall display [7]. In studies of this system, users had a tendency to focus on their own display rather than looking at the shared large display [19]. Other brainstorming systems are more oriented to group decision support, and they typically demand a rigid and formal process that must be followed exactly. This has proven ineffective for the informal brainstorming and sharing of ideas used in the early stages of design [19][11]. Indeed, Buxton argued that the informal nature of sketches is crucial to creative design practice [4].

KJ method: The KJ method is an established design practice, and several research systems have been designed to support this practice digitally. GUNGEN by Yuizono, et al. [26] provided support for the KJ method in a distributed environment where distance separated collaborators could still engage in collaborative brainstorming using a keyboard and mouse interface. The implementation of this system is very similar to Cognoter [7].

Hybrid physical and digital interfaces: Several researchers have explored the use of hybrid physical and digital interfaces to support design practice. The Designers Outpost by Klemmer et al. supported a mix of physical and digital interaction as designers' could write on individual sticky notes and then have a camera capture the notes as they were placed on an upright SmartBoard [11]. Lucero et al. extended this interaction for creating mood (or emotion) boards on a digital table by mounting a camera above the table [13]. The camera could capture images, magazine articles, and other physical objects placed on the digital table by momentarily turning the screen a green colour and performing background subtraction on the image.

Multimodal co-located interaction: A handful of systems have also explored how multiple people can interact with speech and gestures, although these were done over existing single user applications rather than over true groupware systems. Tse et al. explored multi-user interaction over geospatial applications such as Google Earth, Warcraft III and The Sims [20]. These systems could not support parallel work as they were fundamentally limited by the one user per computer assumption of current operating systems. To work around this limitation, Tse et al. [23] explored a split view setting where two computer displays would be projected onto a shared digital tabletop. Collaborators could work in parallel because they were working on separate computers. However, they could not engage in joint multimodal commands and interactions across displays.

7. CONCLUSION

In this paper, we presented an initial exploration into the design and implementation of a true multimodal co-located system. We described a real world case study involving the brainstorming practices of industrial designers. Using this case study we explored issues that future designers of multimodal co-located systems should consider. We also described our architecture for managing simultaneous speech, gesture and pen inputs from multi people and multiple computers. Our goal is that this would help to further the design of future multi-user multimodal systems.

As a system, the Designers' Environment shows promise. While we did not do a formal study, we did present the Designers' Environment to our industrial designers. They commented that they enjoyed being able to touch the table and interact with notes. They reacted positively to the features of the Designers' Environment not available in their physical paper setting (e.g., sorting). Our designers also provided useful suggestions for features that our system could provide, some which were incorporated in the version of the system reported in this paper. Future work includes continuing the participatory design process with our industrial designers, to include note hierarchy logging and playback, support for different file formats, the ability to easily include physical media, and to evaluate the refined system in practice as designers use it to brainstorm actual product ideas.

As a case study, the Designers' Environment helped us discover issues that we believe are valuable to the design of any true multi-user, multimodal tabletop system. Of course, we have barely scratched the surface of such systems. From the multimodal co-located perspective, we would like to explore the use of open speech vocabularies for things like dictation for note writing, searching and web browsing. Other multimodal input devices could be brought to play. There are questions about how we would link such a multi-user, multimodal table to an equivalent distant table and group (called Mixed Presence Groupware [18]). And of course, we need to explore how the design issues mentioned in this paper extend to other true multi-user multimodal system designs and to include technologies such as a large digital wall display for peripheral information.

While both multi-user interaction and multimodal interaction have a long history, the combination of the two is still fairly novel. We recognize that there is much left to do.

8. ACKNOWLEDGMENTS

Removed for blind review.

9. REFERENCES

[1] Aliakseyeu, D., Subramanian, S., Lucero, A., and Gutwin, C. (2006) Interacting with piles of artifacts on digital tables. Proc. AVI '06, ACM Press, 159-162.

[2] Bederson, B. B., Grosjean, J., & Meyer, J. (2004). Toolkit Design for Interactive Structured Graphics, IEEE Transactions on Software Engineering, 30 (8), pp. 535-546.

[3] Boyle, M. and Greenberg, S. (2005) Rapidly Prototyping Multimedia Groupware. *Proc. Distributed Multimedia Systems '05*, Knowledge Systems Institute.

[4] Buxton, W. (2007) Sketching User Experiences: Getting the Design Right and the Right Design, Morgan Kaufmann, ISBN-13: 978-0123740373.

[5] Cox, D. and Greenberg, S. (2000) Supporting Collaborative Interpretation in Distributed Groupware. Proc. CSCW 00, ACM Press, 289-298.

[6] Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L. and Clow, J., QuickSet: Multimodal interaction for distributed applications. *Proc. ACM Multimedia*, 1997, 31-40.

[7] Foster, G. and Stefik, M. (1986) Cognoter: theory and practice of a collaborative tool. Proc. CSCW '86, ACM Press, 7-15.

[8] Greenberg, S., Gutwin, C., and Roseman, M. (1996). Semantic Telepointers for Groupware. *Proc OzCHI '96*, IEEE Computer Society Press. 54-61.

[9] Gutwin, C., and Greenberg, S. (2004) The importance of awareness for team cognition in distributed collaboration. In E. Salas, S. Fiore (Eds) *Team Cognition: Understanding the Factors that Drive Process and Performance*, APA Press, 177-201.

[10] Hall, E. (1966) *The Hidden Dimension.* Anchor Books

[11] Klemmer, S. R., Newman, M. W., Farrell, R., Bilezikjian, M., and Landay, J. A. (2001) The designers' outpost: a tangible interface for collaborative web site. Proc. UIST '01, ACM Press, 1-10.

[12] Kruger, R., Carpendale, M.S.T., Scott, S.D., Greenberg, S.: Roles of Orientation in Tabletop Collaboration: Comprehension, Coordination and Communication. In Journal of Computer Supported Collaborative Work, 13(5-6), 2004, pp. 501–537.

[13] Lucero, A., Martens, J. (2005) Mood Boards: Industrial Designers' Perception of Using Mixed Reality. Proc. SIGCHI.NL Conference 2005, 13-16.

[14] Rogers, Y. and Lindley, S. (2004) Collaborating around vertical and horizontal large interactive displays: which way is best? Interacting with Computers (16), 1133-1152. Elsevier.

[15] Scott, S.D., Carpendale, M.S.T, & Inkpen, K.M. (2004). Territoriality in Collaborative Tabletop Workspaces. Proc. CSCW '04, 294-303.

[16] Somer, R., (1969) Personal Space: The Behavioural Basis of Design, Spectrum, ISBN 0-13-657577-3.

[17] Tang, A., Tory, M., Po, B., Neumann, P., and Carpendale, M. S. T. (2006). Collaborative Coupling over Tabletop Displays. Proc.CHI '06. (April 24-27, Montreal, Quebec). pp: 1181-1190. ACM Press.

[18] Tang, A., Neustaedter, C. and Greenberg, S. (2006) VideoArms: Embodiments for Mixed Presence Groupware. Proc. the 20th BCS-HCI British HCI 2006 Group Conference (Sept 11-15, Queen Mary, University of London, UK).

[19] Tatar, D. G., Foster, G., and Bobrow, D. G. (1991) Design for conversation: lessons from Cognoter. Int. J. Man-Mach. Stud. 34, 2 (Feb. 1991), 185-209.

[20] Tse, E., Shen, C., Greenberg, S. and Forlines, C. (2006) Enabling Interaction with Single User Applications through Speech and Gestures on a Multi-User Tabletop. Proc. AVI'06, 336-343, ACM Press.

[21] Tse, E., Shen, C., Greenberg, S., and Forlines, C. (2007) How pairs interact over a multimodal digital table. Proc. CHI 07, ACM Press, 215-218.

[22] Tse, E., Greenberg, S. and Shen, C. (2006) GSI DEMO: Multiuser Gesture / Speech Interaction over Digital Tables by Wrapping Single User Applications. Proc. ICMI'06, ACM Press.

[23] Tse, E., Greenberg, S., Shen, C., Barnwell, J., Shipman, S. and Leigh, D. (2007) Multimodal Split View Tabletop Interaction Over Existing Applications. Report 2007-869-21, Dept. of Computer Science, University of Calgary, Canada.

[24] Vogel, D. and Baudisch, P. (2007) Shift: A Technique for Operating Pen-Based Interfaces Using Touch. Proc. CHI 2007. p. 657-666.

[25] Wigdor, D., Balakrishnan, R. (2005). Empirical investigation into the effect of orientation on text readability in tabletop displays. *Proc. ECSCW '05*.

[26] Yuizono, T., Munemori, J., and Nagasawa, Y. (1998) GUNGEN: Groupware for a New Idea Generation Consistent Support System. Proc. APCHI '98, IEEE Computer Society, 357-363.