Challenges in the Analysis of Multimodal Messaging

Amy Voida & Elizabeth D. Mynatt
GVU Center, College of Computing
Georgia Institute of Technology
85 5th Street NW, Atlanta, GA 30332
{amyvoida, mynatt}@cc.gatech.edu

ABSTRACT

New forms of computer-mediated communication are increasingly multimodal, providing capabilities for communicating with some combination of text, image, audio, and video. In this paper, we point to the need to develop better methods for studying multimodal communication — more specifically, for studying the communicative role of and relationships among different modalities within their increasingly complex, multimodal semiotic landscapes. We present two challenges in the analysis of multimodal communication, point of view and unit of analysis, both encountered in the context of our study of the use of photo-enhanced instant messaging.

Categories and Subject Descriptors

H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems; H.4.3 [Information Systems Applications]: Communications Applications; H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces — *Collaborative Computing*

General Terms

Human Factors, Design

Keywords

Multimodal analysis, computer-mediated communication, instant messaging, IM, photo-enhanced instant messaging, multimedia messaging, MMS

1. INTRODUCTION

New forms of computer-mediated communication are increasingly multimodal, providing capabilities for communicating with some combination of text, image, audio, and video. In our research, for example, we have explored the use of photo-enhanced instant messaging, a multimodal messaging medium in which photographs can be interleaved with text [16]. Other instant messaging systems provide similar multimodal communicative functionality, including Buddy Vision's Visual Instant Messaging[1] and Picasa's Hello[2], both of which also allow the interleaving of photographs and text in the communicative exchange. A sketch-based research system named Amigo has also supported multimodal messaging, enabling communication with combinations of handwritten text and hand-drawn images [5].

In addition, much recent research in computer-mediated communication has focused on the use of cameraphones and multimedia messaging service (MMS), which allows some combination of text, image, audio, and video to be shared (e.g., [1, 10]). As Kindberg et al. note, however, the cameraphone may also be used for multimodal communication without using MMS, for example by sharing a photograph face-to-face using the cameraphone's display [10]. In that case, the digital photograph is used in conjunction with all the modal richness of the face-to-face communicative context. Other variants of networked digital photography have supported the multimodal exchange of digital photographs with audio [14] or small amounts of text [11].

Most existing research in multimodal computer-mediated communication, including our own, has adopted a primary analytic focus on one modality. In most recent research, the focus has been on the visual modality — in our case because it was the novel addition to an otherwise well-studied domain. In some cases, particularly with respect to studies of MMS and networked digital photography, the image and text are treated as a disambiguated unit of analysis without differentiating between the communicative role of the image and text. To our knowledge, only Koskinen et al. have mentioned the relationship between text and image, noting that text may complement an image and vice versa [11].

As Kress and van Leeuwen argue, however, "particular modes of communication should be seen in *their* environment, in the environment of all the other modes of communication which surround them" (authors' original emphasis) [12]. Kress and van Leeuwen refer to this as the *semiotic landscape*. One important next step for research in computer-mediated communication is to explore more systematically and with greater detail the multimodal semiotic landscape — to explore questions about the role of and relationships among the different modalities used in these communication technologies.

These questions should be of fundamental interest in the design of new computer-mediated communication systems. The relationships among communicative modalities should inform the ways in which the various multimodal features are designed into these technologies. Is there a difference in the relationship between text and image in a medium such as photo-enhanced instant messaging, in which text and images can be interleaved semi-synchronously, and a medium such as MMS, in which the text and image are sent asynchronously as one unit? If there are differences in the relationships among modalities, then the technical and design implications for supporting different types of multimodal communication are significant.

In this paper, we point to the need to develop better methods for studying multimodal computer-mediated communication. We

1. http://www.buddyvision.com/
2. http://www.hello.com/

CSCW'06, November 4–8, 2006, Banff, Alberta, Canada.

provide an overview of existing methods in multimodal analysis and then focus on two challenges we encountered in the analysis of our own multimodal data from our study of photo-enhanced instant messaging [16]. In order to provide an additional concrete basis for our discussion, we also draw from an analytic framework developed by Scott McCloud to describe the various text-image relationships found in another multimodal medium: comics.

2. MULTIMODAL ANALYSIS

> The specialized theoretical and critical disciplines which developed to speak of these arts became equally monomodal: one language to speak about language (linguistics), another to speak about art (art history), yet another to speak about music (musicology), and so on, each with its own methods, its own assumptions, its own technical vocabulary, its own strengths and its own blind spots [13].

In discourse analysis, Kress and van Leeuwen present one of the more comprehensive treatments of multimodal analysis [13]. Their goal is to identify common principles behind multimodal communication; their focus is on the practice of meaning-making in communication. Kress and van Leeuwen note four domains in which meaning-making occurs: discourse, design, production, and distribution. From this perspective, modes are resources that "allow the realisation of discourses" when combined through design and produced and distributed in any number of media [13]. Kress and van Leeuwen apply their framework for discourse analysis to various multimodal artifacts including magazine layouts and multimedia CD-ROMs. Their framework has not yet been extended to apply to conversational interactions.

In conversation analysis, researchers have taken up issues of non-verbal in addition to verbal modes of communication. Its practitioners have appropriated emerging technologies for recording visual data (e.g., video); in doing so, they have had to derive extensions to conversation analysis to account for additional data about non-verbal communication practices [4]. Researchers such as Goodwin (e.g., [8]) and Heath and Luff (e.g., [9]) have extended conversation analysis to study the visual aspects of talk-in-interaction, particularly aspects such as posture, gaze, and gesture. To the extent that the subject of a digital photograph in computer-mediated communication is intended and interpreted as conveying a particular aspect of non-verbal communication, specific research on posture, gaze, or a particular gesture (e.g., pointing) may be productively brought to bear on specific instances of the digital photograph in computer-mediated communication.

These embodied extensions of conversation analysis often reflect a situated extension as well, noting that verbal and non-verbal communication are inextricably situated in the context of interaction. The talk and gesture of archeologists, for example, must be construed with respect to the traces of color in the dirt that they are examining [8]. The talk and gesture of workers in the London Underground must be construed with respect to the computer displays in front of which they are seated [9]. This situated extension to conversation analysis has more recently been applied by Crabtree et al. to conversation about photos — photo-talk [3]. Here, conversation analytic techniques have been applied in situations where the photo is the object of conversation and where the photograph grounds the conversation. We are, however, aware of no research that utilizes this fine-grained an analysis and that addresses the use of images as a first-class communicative object, where an image may carry the entire weight of a conversational turn.

3. CONTEXT OF RESEARCH

Broadly, our research agenda is to understand the impact of digital photography (via webcams, cameraphones, or networked digital cameras) on informal communication practices in both work and social settings. This research plays out across multiple research endeavors of interest to the computer-supported cooperative work community, from communication among mobile workgroups to multigenerational family communication.

3.1. Lascaux

In the remainder of this paper, we draw from one study of the use of Lascaux, a photo-enhanced instant messaging client. Elsewhere, we have described the study and the data set in more detail and have proposed six themes of the communicative appropriation of the images from Lascaux's use [16]. Here, we begin to wrestle with the meaning of those images within their more complex multimodal semiotic landscape.

With Lascaux, users are able to take still photos from a live webcam feed and insert them inline into an instant message as easily as they are able to insert text. Lascaux users see their own live webcam feed at the bottom of the chat window and can click a "Send Photo" button at any time to capture and send the current image.

Lascaux was used by 8 self-selected participants over the course of 4 months. 202 Lascaux transcripts were gathered, including 806 images. In general, a Lascaux encounter emphasized both text and image as first-class communicative objects. In the context of instant messaging, the transcripts showed an experimental, fluid, coparticipatory interleaving of text and image.

3.2. Relationships Between Text and Image

While our aim is to study the use of multimodal messaging to identify naturally-occurring relationships between text and image,

Table 1. McCloud's seven categories of word and picture combinations [15].

Combination	Description
Word Specific	"where pictures *illustrate*, but don't significantly *add* to a largely *complete* text"
Picture Specific	"where words do little more than add a *soundtrack* to a visually told sequence"
Duo-Specific	"in which words and pictures send essentially the *same message"*
Additive	"where words *amplify* or *elaborate* on an image or *vice versa"*
Parallel	"words and pictures seem to follow very different courses — without *intersecting"*
Montage	"where words are treated as integral *parts* of the picture"
Interdependent	"where words and pictures go *hand in hand* to convey an idea that neither could convey *alone"*

the complexity of this analysis caused us to search out existing frameworks that might be used for initial scaffolding and inspiration. In thinking about the interplay between text and image, we turned to a framework originally developed to describe the various relationships between text and image in the comic genre [15]. McCloud suggests that there are 7 categories of word-picture combinations: word specific, picture specific, duo-specific, additive, parallel, montage, and interdependent (Table 1).

We in no way mean to imply that this is the "right" framework for analysis. We use this framework here to provide the reader with a concrete sense of the kind of text-image relationships we find interesting as well as to provide concrete examples for the following discussion of two analytic challenges.

4. ANALYTIC CHALLENGES

While individuals communicatively appropriate the multiple modalities of text and image in what appears to be an intuitive manner, the analysis of the two modalities is less than intuitive. In the following section, we describe two challenges in the analysis of multimodal instant messaging: challenges of point of view and unit of analysis.

4.1. Point of View: Intent and Interpretation

The challenge of point of view is one that is certainly present in monomodal communication, but is compounded in multimodal communication — particularly multimodal communication that involves images. The following instant messaging transcript is one of the shortest and simplest in our data. The entire communicative interaction consisted of one line of text and one image:[3]

Scott [Fri May 23 15:09:53 EDT 2003]:
Hello!
Scott [Fri May 23 15:09:59 EDT 2003]:

We asked three individuals to examine this text-image pair and code the relationship as one of McCloud's seven categories. We also asked the three individuals for a justification for their categorization. One individual coded the exchange as *additive* because the image elaborated on the text — providing the additional context that someone else was also watching the conversation unfold and that communication should be guarded in the presence of a third party. A second individual coded the exchange as *picture specific* as she believed the text did no more than provide a soundtrack to the image (the two individuals saying "Hello"). The third individual coded the exchange as *duo-specific*; this individual interpreted the photograph as being of one relevant individual waving, conveying essentially the same message as the textual "Hello!" and of one conversationally-irrelevant individual occupying the background.

3. In all transcripts, identifying information in the text has been anonymized but idiosyncrasies of language have been preserved. All images are presented unaltered, with the participants' consent.

The challenge of point of view here is reflected in the fact that the first of these individuals was the sender and producer of the instant message. The second individual was the recipient and consumer. The third individual was an independent researcher who was not privy to the context in which the conversation was carried out.

Kress and van Leeuwen point out that there are two kinds of participants in communication: represented participants (the subjects of the communication) and interactive participants (the sender/producer and the recipient/consumer of the communication) [12]. Significant relationships among these participants include those between the interactive participants as well as those between the interactive participants and the represented participants. To these relationships, we might add the researcher and his or her relationships with both the represented and interactive participants.

In our example, the sender/producer coded the excerpt differently because the communicative intent of his image was more nuanced than was the recipient/consumer's interpretation of it. However, both interactive participants had relationships with the represented participants (the represented participant in the foreground was the sender/producer; the represented participant in the background was a colleague known to both interactive participants), such that they understood their significance in the meaning of the conversation. The researcher, however, while able to infer that the represented participant in the foreground was the sender/producer, did not know the other represented participant and was not able to infer that his presence carried any meaning.

As art historian Ernst Gombrich contends, "the innocent eye is a myth" [6]. Each of these individuals' analytic eyes were privy to different communicative contexts and social relationships. The challenge is in answering the question: when analyzing conversation, which "eye" does the analysis rely upon? From the standpoint of a discipline in which it is often the eventual goal to inform system design, one might argue that the modal choice is in the hands of the sender/producer. Thus, the argument might hold that the intentionality of the sender/producer is the "eye" that is most relevant for making design choices about multimodal features. A counterargument comes from aestheticians and philosophers who talk about "the intentional fallacy" [2]. This argument holds that too much emphasis is placed on what the sender/producer thinks and not enough on what meaning the recipient/consumer takes away. As Minor White, photographer, argues, "photographers frequently photograph better than they know" [17]. One might also make a pragmatic third argument that it will be methodologically impossible to always know the intent of the sender/producer or the interpretation of the recipient/consumer and so the responsibility for the analytic "eye" must fall on an external researcher.

4.2. Unit of Analysis: Communicative Context

A challenge perhaps even more permutationally problematic than the first is the question of the unit of analysis. How does one know what text pertains to what image or vice versa? We illustrate this point with a concrete example — a complete instant messaging transcript only one turn longer than our previous example but still significantly shorter than the average transcript in our data. And yet, by increasing the length of the transcript by only one turn, we significantly increase its complexity:

mokona [Fri Apr 11 20:09:00 EDT 2003]:
just holler when you do...i'm gonna go make some porridge
mokona [Fri Apr 11 20:09:08 EDT 2003]:

mokona [Fri Apr 11 20:09:16 EDT 2003]:
[the cat] is watching traffic again

Returning to McCloud's seven categories of text-image relationships, we asked the sender/producer of this transcript to categorize the following: (a) the relationship between the first textual turn and the image, (b) the relationship between the image and the second textual turn, and (c) the relationship among the image and text as a whole. Here, "mokona" returned with three completely different responses. She felt that the first textual turn (the sender/producer making porridge) and the image (of a bed and a cat in the window) were happening in *parallel,* following different courses without intersecting. The relationship between the image and the second textual turn was categorized as *duo-specifi*c, that is they sent essentially the same message (her cat watching traffic). Finally, when asked to analyze the entire exchange as a whole, "mokona" felt that the relationship was *word-specific*. She felt that the image did not add significantly to a largely complete text but did illustrate the mood, which she characterized as being "laid-back and uneventful."

When analyzing the use of digital photographs in computer-mediated communication, the question of what unit of analysis to take up is a challenging one. One could argue that if the end goal is to design computer-mediated communication technologies in which the sender/producer is making modal choices, then one could argue for an analysis of each image as it flows in the conversation — how the image relates to the conversation that exists up to that point in time. One might also argue that both text and image should be analyzed with respect to the most closely related text or image, whatever is the most conceptually coherent unit. In our example, this unit might be the image and the textual turn immediately following it. Finally, one might argue for a holistic perspective; that is, we must understand broadly what role both text and image play in the conversation as a whole, a similar approach as that taken by Gombrich using the divisions of language proposed by Karl Bühler [7].

5. CONCLUSION

In this paper, we have pointed to an important area of future research in studying multimodal computer-mediated communication: understanding the use of different modalities within their increasingly complex semiotic landscape. We have also described the analytic challenges of point of view and unit of analysis, encountered in our study of multimodal messaging.

We hope that this paper will inspire others to help undertake the future research required to address the complexity of the analytic challenges present in the study of multimodal computer-mediated communication.

6. REFERENCES

[1] Battarbee, K. (2003). Defining co-experience. In *Proceedings of the Conference on Designing Pleasurable Products and Interfaces*. New York: ACM Press, pp. 109-113.

[2] Wimsatt, W.K. & Beardsley, M.C. (1987). The intentional fallacy. Reprinted in J. Margolis (Ed.), P*hilosophy looks at the arts*, 3rd ed. Philadelphia: Temple University Press.

[3] Crabtree, A., Rodden, T., & Mariani, J. (2004). Collaborating around collections: Informing the continued development of photoware. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. New York: ACM Press, pp. 396-405.

[4] Erickson, R. (2004). Origins: A brief intellectual and technological history of the emergence of multimodal discourse analysis. In P. LeVine & R. Scollon (Eds.), *Discourse and technology: Multimodal discourse analysis.* Washington, D.C.: Georgetown University Press, pp.196-207.

[5] Fabersjö, H., Windt, E., Wridell, Y., & Sanneblad, J. (2003). Amigo - wireless image based instant messaging for handheld computers. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*. New York: ACM Press, pp. 910-911.

[6] Gombrich, E.H. (1960). *Art and illusion: A study of the psychology of pictorial representation*. Princeton, NJ: Princeton University Press.

[7] Gombrich, E.H. (1982). The visual image: Its place in communication. In E.H. Gombrich (Ed.), *The image and the eye*. Oxford, UK: Phaidon Press, pp. 137-161.

[8] Goodwin, C. (2003). The semiotic body in its environment. In J. Coupland & R. Gwyn (Eds.), *Discourses of the body*. New York: Palgrave/Macmillan, pp. 19-42.

[9] Heath, C. & Luff, P. (1992). Explicating face-to-face interaction. In G.N. Gilbert (Ed.), *Researching social life*. London: Sage, pp. 306-327.

[10] Kindberg, T., Spasojevic, M., Fleck, R., & Sellen, A. (2005). The ubiquitous camera: An in-depth study of camera phone use. *IEEE Pervasive Computing*, 4(2), 42-50.

[11] Koskinen, I., Kurvinen, E., & Lehtonen, T. (2002). *Mobile Image*. Edita, Finalnd: IT Press.

[12] Kress, G. & van Leeuwen, T. (1996). *Reading images: The grammar of visual design.* London: Routledge.

[13] Kress, G. & van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. New York: Oxford University Press.

[14] Mäkelä, A., Giller, V., Tscheligi, M., & Sefelin, R. (2000). Joking, storytelling, artsharing, expressing affection: A field trial of how children and their social network communicate with digital images in leisure time. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York: ACM Press, pp. 548-555.

[15] McCloud, S. (1993). *Understanding comics: The invisible art.* New York: Harper Collins.

[16] Voida, A. & Mynatt, E.D. (2005). Six themes of the communicative appropriation of photographic images. In *Proceedings of the ACM Conference on Human Factors in Computing Systems.* New York: ACM Press, pp. 171-180.

[17] White, M. (1957). Criticism. *Aperture,* 2(2), 29-30.