Demo: Collaborative Multimodal Photo Annotation over Digital Paper

Paulo Barthelmess, Edward Kaiser, Xiao Huang,
David McGee, Philip Cohen
Natural Interaction Systems
Seattle, WA, USA

ABSTRACT

The availability of metadata annotations over media content such as photos is known to enhance retrieval and organization, particularly for large data sets. The greatest challenge for obtaining annotations remains getting users to perform the large amount of tedious manual work that is required. In this demo we show a system for semi-automated labeling based on extraction of metadata from naturally occurring conversations of groups of people discussing pictures among themselves. The system supports a variety of collaborative label elicitation scenarios mixing co-located and distributed participants, operating primarily via speech, handwriting and sketching over tangible digital paper photo printouts. We demonstrate the real-time capabilities of the system by providing hands-on annotation experience for conference participants. Demo annotations are performed over public domain pictures portraying mainstream themes (e.g. from famous movies).

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Collaborative Computing; Synchronous interaction; H.5.2 [**User Interfaces**]: Natural language; Input devices and strategies; I.2.6 [**Learning**]: Language acquisition

General Terms

Design; Experimentation; Human Factors

Keywords

Demo; Photo Annotation; Collaborative Interaction; Multimodal processing; Intelligent interfaces

1. INTRODUCTION

Retrieval, support and organization of photo collections can be enhanced via annotation, e.g. to support indexing, clustering and automatic generation of albums and collages. While the value of annotations is well recognized, finding ways to motivate users to undertake the tedious annotation work that is required has remained elusive.

In our work we explore a solution that leverages social aspects of photo usage to facilitate label elicitation. Central to the approach that we demonstrate is the notion that groups of people naturally generate a semantically rich multimodal discourse as they discuss content and events associated to photos [5, 6, 3]. We therefore emphasize an interface that provides support for the task while avoiding getting in the way of the interaction. This collaborative interface aims at creating favorable conditions for label elicitation based on which automatic extraction of labels from group conversations takes place.

We provide a simple to use multimodal collaborative interface that aims at supporting a variety of use scenarios, ranging from informal family gatherings to more formal meetings in which analysts might examine and annotate photos. Both co-located and remote collaboration are supported.

Of particular interest is the support provided for annotation over paper-based photo artifacts. There is evidence (e.g. [3, 5, 6]) pointing to strong user preference for sharing of printed photos rather than electronic ones. We exploit Anotos digital paper technology [1] to provide tangible physical photographs, in support of this natural practice.

This is a companion demo for a paper submitted in separate to the ICMI Conference [2].

2. COLLABORATIVE INTERFACE FOR ANNOTATION

Two complementary aspects form the foundation of our approach to facilitation of photo annotation: 1) a collaborative interface that is conducive to the expected labeling behavior and 2) automatic support for label extraction and propagation through the analysis of multimodal language.

2.1 Digital paper annotation

To promote our goal of supporting more fluid and natural types of interaction, we include support for annotation over photos printed on digital paper. Participants of an interaction may perform annotation by handwriting on the photos themselves, as they would on regular paper documents. Users are thus freed from having to directly operate a computer interface, and can concentrate fully on the task.

This design is consistent with findings that point to differences in perception and use between electronic and printed pictures. The tangible nature and ease of manipulation of physical photographs makes prints much preferred in sharing situations. Frohlich et al. [3], for instance, reports that only seven out of one hundred and twenty seven sharing episodes reported in their study involved digital pictures. Van House et al. [6] highlights the importance of the materiallity of prints, and the affordances of the material to pro-

ICMI'06, November 2–4, 2006, Banff, Alberta, Canada.

mote the fluid, non-sequential kinds of interaction that make photo sharing enjoyable. There is also evidence in other domains that paper-based interfaces may reduce the cognitive load associated to collaborative tasks [4].

Digital paper and pen provide a natural interface for users to annotate pictures. The underlying technology we exploit is based on Anoto's Digital Pen and Paper [1]. Anoto-enabled digital paper is plain paper that has been printed with a special pattern, like a watermark. The pattern consists of small dots with a nominal spacing of 0.3 mm (0.01 inch). These dots are slightly displaced from a grid structure to form the proprietary Anoto pattern.

A user can write on this paper using a pen with Anoto Functionality (Figure 1), which consists of an ink cartridge, a camera in the pen's tip, and a Bluetooth wireless transceiver sending data to a paired device. When the user writes on the paper, the camera photographs movements across the grid pattern, and can determine where on the paper the pen has traveled. In addition to the Anoto grid, which looks like a light gray shading, the paper itself can have anything printed upon it using inks that do not contain carbon. Multiple participants can write over shared or replicated printed sheets with their own individual pens.

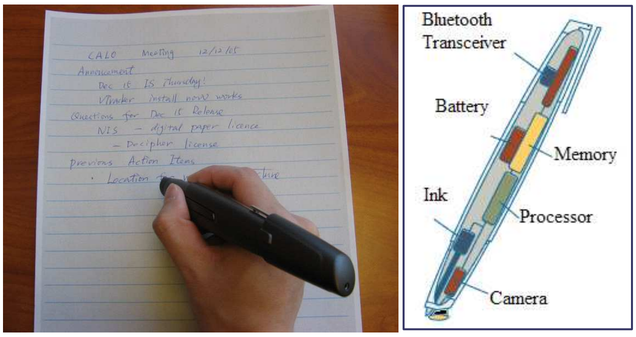

Figure 1: (a) Digital pen and paper; (b) Digital pen system design.

2.2 Shared display and "slide show"

To support potential remote participants and accommodate larger groups, annotations performed on paper can be displayed in (close-to) real-time, overlayed on an electronic version of the image being annotated, replicating the appearance of the annotations performed on paper, e.g. on a large screen TV or projection screen. The basic collaboration infrastructure of the system handles the necessary conversions so that the appearances of the displays are similar independently of differences in screen aspects and resolutions that may exist across distributed displays.

The system is able to detect which photo is being written on by maintaining an association between the specific paper patterns and the corresponding images printed on them. The display is updated automatically as users begin writing on each page, allowing users to control a slide show-like display by touching their pens to the photo printouts.

The automatic updating of the shared display allows for the kind on hands-off operation that we envisioned. Because no direct manipulation of a computer interface is required, users can remain engaged and "on-task". The level of engagement we observed during pilot collection suggests that the interface achieved the desired goal.

2.3 Automated label extraction

One problem faced by systems that take responsibility for analyzing unconstrained group interaction and extracting labels from natural language streams is how to determine which terms within user utterances are relevant descriptions that should be associated to specific photos.

The solution we explore here is based on the hypothesis that the information users choose to handwrite correspond to key descriptive terms. Evidence from the pilot corpus we are collecting using our system supports both the conciseness and discriminatory value of handwritten labels as well as the high degree of redundancy between and across modalities [2]. We therefore select as primary labels those terms that are handwritten on a photo printout.

We concentrate multimodal processing resources on recovering these terms in the most robust fashion possible. This is in turn achieved by exploiting the redundancy within and across modalities, as users handwrite and speak (repeatedly) the high value labels. Our goal is to dynamically adapt the systems vocabulary on-the-fly by bootstrapping the redundancy across modalities to enroll new terms that will bias future recognition towards high-relevance terms.

Multimodal processing currently requires multi-phase offline processing, mainly due to the non-realtime nature of the ensemble of speech recognizers that are used. For the purpose of the demonstration, we restrict the functionality to the unimodal recognition that is achievable in real-time, namely, handwritten recognition and some degree of sketch recognition.

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or the Department of Interior-National Business Center (DOINBC).

3. REFERENCES

[1] Anoto Corporation. Anoto technology - how does it work? `http://www.anotofunctionality.com/cldoc/aof3.htm`, May 2006.

[2] P. Barthelmess, E. Kaiser, X. Huang, D. McGee, and P. Cohen. Collaborative multimodal photo annotation over digital paper. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*. ACM Press, 2006.

[3] D. Frohlich, A. Kuchinsky, C. Pering, A. Don, and S. Ariss. Requirements for photoware. In *CSCW '02: Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 166–175, New York, NY, USA, 2002. ACM Press.

[4] S. Oviatt, A. Arthur, and J. Cohen. Quiet interfaces that help students think. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology (UIST'06)*, 2006.

[5] K. Rodden and K. R. Wood. How do people manage their digital photographs? In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 409–416, New York, NY, USA, 2003. ACM Press.

[6] N. Van House, M. Davis, Y. Takhteyev, N. Good, A. Wilhelm, and M. Finn. From 'what?' to 'why?': The social uses of personal photos. `http://www.sims.berkeley.edu/\~vanhouse/vanhouse\_et\_al\_2004a.pdf`, 2004.