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 paper we introduce an approach for semi-automated labeling based on extraction of metadata from naturally occurring conversations of groups of people discussing pictures among themselves.

As the burden for structuring and extracting metadata is shifted from users to the system, new recognition challenges arise. We explore how multimodal language can help in 1) detecting a concise set of meaningful labels to be associated with each photo, 2) achieving robust recognition of these key semantic terms, and 3) facilitating label propagation via multimodal shortcuts. Analysis of the data of a preliminary pilot collection suggests that handwritten labels may be highly indicative of the semantics of each photo, as indicated by the correlation of handwritten terms with high frequency spoken ones. We point to initial directions exploring a multimodal fusion technique to recover robust spelling and pronunciation of these high-value terms from redundant speech and handwriting.

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

Photo Annotation; Automatic Label Extractgion; 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.

1.1 Social opportunities for label elicitation

In this paper we explore a solution that leverages social aspects of photo usage to facilitate label elicitation. We take advantage of the rich communication that takes place when groups of people go over sets of pictures sharing their experiences. We propose a framework for automatically detecting and propagating labels that are produced as part of a natural group conversation using voice, handwriting and sketching.

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. [10, 23, 27]) pointing to strong user preference for sharing of printed photos rather than electronic ones. We exploit Anoto's digital paper technology [2] to provide tangible physical photographs, in support of this natural practice.

1.2 Observant systems

Central to the approach that is detailed here is the notion that groups of people naturally generate a semantically rich multimodal discourse as they discuss content and events associated to photos [23, 27, 10]. We therefore emphasize an interface that provides support for the task while avoiding getting in the way of the interaction. This led us to employ a mostly observant system [14], a system that "eavesdrops on the storytelling" [9].

In observant systems, most of the operations are performed behind the scenes. Minimal interface manifestations take place, mostly to support diffusion of contextual information to other participants that may be co-located or remote. This collaborative interface aims at creating favorable conditions for label elicitation based on which automatic extraction of labels from group conversations takes place.

1.3 The role of multimodal data

We examine the role of multimodal data in helping solve three basic problems associated with the extraction of semantics from unconstrained natural language:

- *Detecting a concise set of meaningful labels* to be associated with each photo as representative of its semantics. This is a particularly acute problem given the variety and richness of

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

the overall language employed by users while narrating their photos.

- *Achieving robust recognition of key semantic terms*. While group interactions may elicit extremely rich semantics, the unconstrained nature of such interactions has traditionally posed problems for natural language recognition systems.
- *Facilitating label propagation*. One of the primary goals of an annotation support tool is clearly to provide easy ways for users to perform the task. We identify multimodal behavior that may lead to lower cost labeling compatible with natural practices.

We show via analysis of a preliminary set of collected data that users choice of handwritten labels are highly indicative of the intended semantics of each associated photograph. The high level of cross-modal redundancy between handwriting and speech that was observed also indicates that multimodal analysis can provide promising cues for disambiguation leading to more robust recognition.

In this paper we describe a multimodal observant system that we implemented, that explores these directions to provide robust automated extraction of labels from natural language elicited during collaborative sharing sessions.

1.4 Paper organization

We start by reviewing the related literature (Section 2). In Section 3 we discuss the interface issues in further detail, and then overview the architecture of the implemented prototype (Section 4). In Section 5 we discuss aspects of the analysis of pilot data providing initial evidence supporting our interface design and processing strategy.

2. RELATED LITERATURE

Commercial photo management tools such as Adobe Photoshop Album, iPhoto, and ACDSee support manual user annotation. While available, these annotation mechanisms are recognized to be tedious and not to offer users enough incentives (e.g. [15, 21, 7]).

Here we review some of the related literature, emphasizing work that support collaboration and/or the use of other media besides typed text.

2.1 Multiuser natural language label extraction

Fleck [9] explores use scenarios for observant systems that "eavesdrop on conversations" and extract metadata from naturally occurring interactions. Recordings collected while subjects browsed photos using a web-based interface and photo albums were transcribed using a Sphinx 2-based speech recognizer trained on transcribed broadcast news data. A 60% error rate is reported.

Qian and Feijs [21] present a tool - Talkim - based on an instant-messaging environment that facilitates photo sharing among a group of distributed participants. Textual annotation is supported with the help of interface buttons that add some structural hints. These structured annotations can be freely positioned over regions of the displayed picture, serving a similar purpose as our localized annotations in promoting collaboration by allowing for attention to be focused onto specific regions of the shared photo artifact. The text of instant messages exchanged by (remote) participants is used as a source of additional labeling information. Messages are classified according to location, time, event and person categories by performing parsing based on a static grammar (expressed as a state machine). Once classified, the full text of the message is attached to the photo, without further extraction of keywords. Some degree of dialog understanding is achieved by analyzing potential question/answer pairs for indications of category (e.g. by determining that the answer to a "where was this taken?" question would provide location information.

Similarly to these systems, the photo annotator we describe in this paper explores social situations to extract labels from natural language. These systems as a rule explore single modalities (text or speech), and do not take advantage of handwriting and gestures to promote annotation. Our work is distinguished by our use of multiple modalities to extract and propagate robust labels from unconstrained, large vocabulary conversation streams, and by our use of tangible materials to support natural work practices.

2.2 Single-user natural language processing

Chen et al. [4] describes a single-user system that can be used to process on-the-spot spoken annotations recorded by the digital camera itself while a picture is being taken. Spoken utterances follow a simple prescribed grammar that contains a category keyword (*people, location, taken_on, event*) followed by a corresponding label sequence. Recorded speech is transcribed offline via a speech recognizer. Robustness is achieved via a query expansion mechanism, which extends recognized words with related ones extracted from a thesaurus. To cope with mis-recognitions particularly in the presence of out-of-vocabulary words that are common in photo annotation that include names of places and people, a statistical model of word replacements is built on top of n-best recognizer results. This allows for instance the system to predict that the label "Niagara" might be commonly mis-recognized as "Naipaul", "manical", "neither", "sidebar" and "Sniper". By extending queries with this extended set, Chen et al. [4] show that better recall and precision can be achieved.

Pinzon and Singh [29] describs a system - eVITAe - that supports structured annotations via a display for location and time information, implemented over a map and a timeline panel respectively. Voice annotations are transcribed using a Sphinx speech recognizer. Thresholding is applied to prune out mis-recognitions.

In Shoebox [18] up to five (compound) nouns are extracted from transcriptions of spoken notes and from textual notes. Noun phrases are first identified by a Brill tagger; heuristics are then applied to favor high-frequency proper nouns, aiming at selecting names of people and places. Audio annotations are played back as users hover over pictures in the system's photo browser. Shoebox also allows users to search for pictures based on image region similarities.

In Aria [16], a system that observes user textual email messages automatically extracts query terms from the spatial vicinity of the cursor as users type. a photo browser is then continuously updated to display pictures from a database that match the current recovered query items. Labeling takes place when users choose specific pictures from the selected set to incorporate into their email. The system is therefore able to learn labels incrementally based on indirect user supervision. Other work (e.g. [28]) similarly explores retrieval feedback as a source of labeling information.

Unlike the systems just described, our approach is based primarily on collaboration scenarios, during which label descriptions are produced as a matter of fact by participants. Unlike these systems, that focus primarily on textual and spoken unimodal input, we explore multiple modalities. An exception is Show&Tell [26], which supports single-user multimodal language by integrating (mouse) gestures with speech. Show&Tell integrates photo analysis techniques with multimodal annotation, aiming at identifying objects of interest in a photo, e.g. buildings and roads in an aerial photograph, and attaching descriptions to these objects. The additional linguistic context provided by users is used to perform more robust image interpretation. Vocabulary and syntax of accepted utterances are constrained to achieve better recognition results.

In contrast, our approach to natural language processing is based on a learning mechanism that leverages redundant speech and handwriting for mutual disambiguation [20]. We support extraction of labels from unconstrained, large vocabulary, natural multi-party dialogues.

2.3 Other multi-user annotation systems

Some work also provides support for annotation using alternative media, and may support collaboration, but is distinguished from the previous by the fact that no automatic recognition is attempted.

Debaty et al. [7] overview a few systems developed at the HP Laboratory in Palo Alto that take advantage of collaboration to support photo annotation. Their MemoryViewer is a peer-to-peer system for distributing and displaying photos. This system allows users to input audio annotations, besides textual ones. This system is similar to Picasa's Hello [11] in the sense that it facilitates synchronous photo distribution for a group of remote participants. StoryMail provides mobile-phone based support for building a sequence of pictures that can then be narrated via audio. The sequence can then be distributed for asynchronous viewing to multiple recipients.

Collaborative construction and organization of digital photo collections in face-to-face story sharing sessions was addressed by the Personal Digital Historian (PDH) project at MERL [24]. An interface built using the DiamondSpin toolkit is projected on an interactive horizontal surface such as the DiamondTouch multiuser touch-tabletop surface. Material can be organized along categories related to "who", "when", "where" and "what" questions. The system provides for rich and fluid selection of objects and their arrangement into meaningful structures.

While the systems we just described offer interesting collaboration features and point to the richness of group annotation and organization (e.g. in PDH [24]), they don't provide the interpretation of the rich multimodal language for automatic extraction and propagation of labels as we propose in this paper.

2.4 Annotation over digital paper pictures

Norrie and her group at ETH Zürich explore annotation over documents printed on digital paper. Of particular relevance to what we present here is the work on annotation of mammography paper reproductions [8]. Regions of a mammogram can be marked by circling them with the digital pen. Handwritten annotations can then be associated with these regions. Regions can overlap, thus allowing for hierarchical annotations, e.g. generic ones over the whole mammogram complemented by detailed annotations over specific sub-regions. PaperPoint [25] is a digital paper application that allows users to control a PowerPoint presentation by tapping on printed images of the slides. Collaborative annotation and presentation control is possible when multiple pens are used.

ButterflyNet [30] focuses of supporting notes takes by botanists in the field. Notes taken on digital paper notebooks using Anoto's technology can be enhanced by the inclusion of digital photos placeholders using a gestural language. Subsequent browsing integrates the handwritten notes with associated photos and other sensor data e.g. GPS, video and audio. Browsing can be directed by pen taps on the physical paper notes, creating an effect that is similar to the "slide show" functionality we adopted in our work. Some functionality for photo annotation appears to be supported via handwriting on a portable device lcd screen, rather than on paper.

While we adopt digital paper in support of natural work practices as well, our photo annotator has a different focus, providing services for annotation based on multimodal language analysis.

The system presented here adds to a long lineage of systems produced by our group in the past decades such as Quickset [5], NIS Map [6], and Charter [3]. These systems, as the one presented here, explore ways to take advantage of natural user practices and naturally occurring multimodal language as the means to support users in non-disruptive, cognitive-load conscious ways. Of particular relevance for our present discussion is Rasa [17, 6], which pioneered the use of paper-based interfaces within our group.

3. ANNOTATION INTERFACE

In this section we overview how we leverage multimodal group behavior to provide an interface that supports users' natural practices, and that provides multiple opportunities for automatic label extraction and propagation. 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.

Our goal in designing the interface we describe here was to provide support to a variety of collaboration scenarios in which photo annotation could take place. Both co-located and distributed collaboration are supported. Various levels of formality can be accommodated, including groups of friends getting together to exchange personal experiences on the one hand, and more formal work e.g. by a group of analysts on the other hand.

3.1 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. [10], 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. [27] highlights the importance of the materiallity of prints, and the affordances of the material to promote the fluid, non-sequential kinds of interaction that make photo sharing enjoyable.

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 [2]. 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.

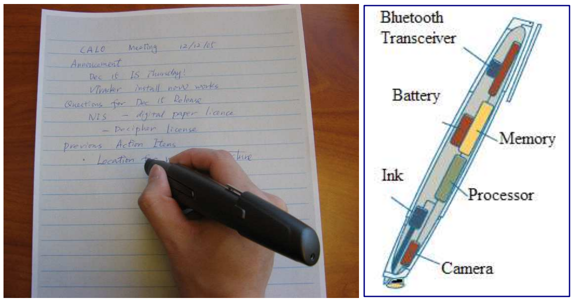

Figure 1: (a) Digital pen and paper; (b) Digital pen system design.

Multiple participants can write over shared or replicated printed sheets with their own individual pens. These multiple sources of electronic ink are then consolidated by the system, as detailed in Section 4.

3.2 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.

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 may remain engaged and "on-task".

3.3 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 speech and handwriting (Section 5. 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 system's vocabulary "on-the-fly" by bootstrapping the redundancy across modalities to enroll new terms that will bias future recognition towards high-relevance terms.

3.4 Propagation via abbreviations

Given the goal of facilitating the labeling task, we chose to support label propagation via more concise abbreviated labels. In the scenario we have in mind users would first establish the spelling and pronunciation of a term by handwriting and speaking, and would subsequently refer to the previously introduced term via an abbreviation whose semantics is provided by speech.

The assumption upon which this mechanism is based is that a relatively large subset of people and places depicted in photos presented during a sharing session will appear in more than one photo. The data collected so far supports this assumption at least partially. We observe that even when the photos bear little relationship to one another either by time or location, still certain key players (usually family members) will tend to re-occur with higher frequency.

In order to be able to successfully support abbreviations, it is necessary for the system to be able to recover the semantics of (the full term) the abbreviation refers to. Our strategy is based once again on exploiting multimodal language characteristics. Once a term has been introduced in full (via handwriting and speech), the system is able to determine both its spelling and pronunciation. Further spoken references in the temporal vicinity of a an abbreviated handwritten term can then be recovered via the application of heuristics (see [13] for details of this mechanism).

While the data we have collected so far does not provide enough evidence of this multimodal behavior (speech plus abbreviated handwriting), given that users are not explicitly advised of this capability of the system, there is evidence in other domains (e.g. lecturing [1]) that substantiate the assumption that people will indeed speak in full the terms they write down in abbreviated form, which is what our approach requires for attempting a semantic recovery.

4. SYSTEM DESCRIPTION

We now turn our attention to the technical aspects of the architecture of the implemented photo annotator. We concentrate on the description of the multimodal backend components.

Server-side processing provides recognition services for speech, gesture, and handwriting. These streams are then fused within a temporal chart parsing framework [12]. Since the system is not currently able to process incoming audio streams in real-time, we employ a replay infrastructure to process captured data offline.

The observant nature of the system makes this offline processing acceptable, given that no immediate extensive feedback is required. Awareness mechanisms (via ink propagation) and unimodal processing of ink are real-time and suffice to support most aspects of collaboration even when shared displays are used.

Here we briefly overview the processing mechanism - details of the integration mechanism for label extraction and learning can be found e.g. in [3, 13].

- An ensemble of *Speech recognizers* based on Carnegie Mellon University's Sphinx recognizer [22] was employed, including four phone recognizers, and a word / phrase-spotter.

- The *Ink Segmenter* performs a first pass segmentation of ink points, grouping strokes into glyphs based on temporal and spatial relationships. This initial segmentation is then refined based on contextual information by NIS Sketch and Handwritten recognizers described below.

- *NIS Sketch* is a sketch recognizer that handles the classification of electronic ink glyphs into tokens corresponding to what will be interpreted as handwriting or graphical annotation elements such as lines and arrows. NIS Sketch outputs a list of hypotheses ranked by likelihood. The decision of how to interpret each of these symbols is made by Multiparser - the multimodal fusion engine.

- The *Handwriting agent* interacts with NIS Sketch to separate handwriting from other drawings such as lines and arrows that might be included in photo annotations. The underlying handwritten recognition is provided by Microsoft's handwriting recognizer.

- *Multiparser* is a temporal chart parser that handles multimodal fusion. Fusion in Multiparser is based on a temporal chart parsing technique and is guided by grammar rules that specify operations and constraints expressed in terms of unification over feature structures [12].

5. ANALYSIS OF THE PILOT CORPUS

The multimodal collaborative system we developed was used to collect some initial pilot data to verify assumptions and inform further development particularly of our multimodal processing techniques. We report on four data collection sessions.

5.1 Participants and scenario

Four groups of three participants, recruited amongst lab members, took part in each of the pilot collection. For each session, a different participant was the *narrator* - the one telling the stories and performing the annotations. The other two participants' function was to provide an *audience*, thus replicating to some extent the social photo sharing situation we envision as being potentially conducive to label elicitation. Of the four narrators (which we refer to as "N1-4"), one was female and three were male. Two of the participants were not native English speakers.

Participants sat around a small meeting table upon which the stack of photos was placed. Two of the participants (narrator and one member of the audience) sat side-by-side, and the third participant sat across the table (Figure 2).

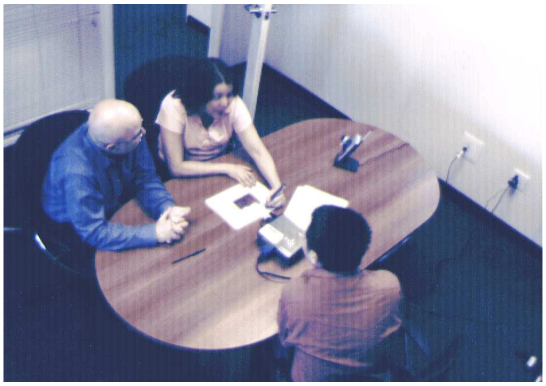

Figure 2: Collection scenario, with the three participants around a meeting table.

A projector displayed the image of the photo currently being annotated. No direct interaction with the computers took place during the annotation sessions. As the system detected which image was being written on, the display was automatically updated, avoiding the need for any explicit slide transitioning user interface operation, as was our design goal.

5.2 Datasets

The narrators provided their own pictures (9 or 10 - see Table 1), most of which had been taken abroad during vacation or conferences. Digital photos were sent in advance by the narrators, so that they could be prepared for the collection. Processing consisted of printing on digital paper, so that the photos would be overlayed on top of the Anoto pattern required for pen operation.

Photos were printed placed at the center of the page, with wide margins at the sides. The main reason we adopted this style is that it is more flexible in terms of organization. Since the printing preparation is performed by the experimenters, we preferred to avoid assuming a specific organization, e.g. by printing multiple pictures per page, that might conflict with the one preferred by the subjects. We acknowledge that the generous empty margins around each photo might influence the annotation style, and intend to verify the differences in language when larger portions of the page are taken by the photos.

5.3 Procedure

At the beginning of the session, narrators were told to use the pen and digital paper as they would any conventional pen and paper. The instruction provided indicated that, as they went over each picture, they could make annotations using the pen. They were told that the annotations they made would be interpreted by a system that would then provide them with labeled photos, which would make it easier for them to e.g. search or browse pictures based on their contents.

No references were made during the instruction to the system's capabilities or limitations. In particular, narrators were not informed they could use abbreviations, pointing, or what the role of speech and handwriting would be in the interpretation. The intention at this point was to collect data that would be naturally produced given the circumstances.

Sessions lasted about 38 minutes in average, with a standard deviation of 7.6 minutes (Table 1).

Session	N1	N2	N3	N4	Avg	Std
Duration (min)	27	43	43	40	38.3	7.6
Photos	9	10	10	9	9.5	0.6
Time / photo	3.0	4.3	4.3	4.4	4.0	0.7

Table 1: Session summary.

5.4 Collection infrastructure

Collection was achieved via the collaborative multimodal system infrastructure described in Section 4. Since no backend processing was required at collection time, processing was lightweight enough to be run on laptop computers, making it conveniently mobile, an important aspect considering the variety of collection scenarios we want to address.

Two laptop computers were used to run the required software. One of the machines ran the ink collection and display software, as well as the unimodal ink recognition components, while the other was used to collect the audio recordings from the close talking microphone.

Shared display was acomplished via a projection onto a screen. Annotation was performed over pre-printed digital paper photos using Nokia's SU-1B pens streaming over Bluetooth. Printing of photo digital paper were performed on a high quality color laser printer (an OKI C5400).

The main narrator's audio was collected via a close-talking microphone. Audio from additional subjects (the "audience") was collected via an open microphone attached to a digital room camera.

5.5 Preliminary analysis

As expected, participants were highly engaged in the task, a fact that had a clear positive impact in the amount and variety of the multimodal language elicited. This was informally verified by examining the video footage. The discussions could be seen to be lively and participants appeared to be immersed in the discussions.

We report the results of this analysis in the next sections. We start by examining the characteristics of unimodal handwriting and sketching (Section 5.5.1). The analysis of multimodal speech and ink (presented in Section 5.5.2) suggest that handwritten labels may provide a concise set of terms that are indicative of the topics of discussion surrounding each photo, corroborating our hypothesis. In Section 5.6 we explore how our multimodal technique can be used to rosbustely recover the semantics of these high-value terms.

5.5.1 Handwriting and sketching

Participants reported during debriefing being at ease adding handwritten information to photos. This is consistent with observations reported in the literature (e.g. [10, 23, 27]) indicating the preference for sharing printed photos, as well as the practice of handwriting info on photos - even though on regular photos annotations are added in general to the back side.

Table 2 presents the results of the handwritten recognition evaluation. Recognition ranged from 26% to 56%, with a 42% average. Lowest recognition results could be observed when labels were not written horizontally, but at an angle (as is the case for N3). Additional graphic decorations also seemed to have served as confounders - text around or embedded in more elaborate sketched elements were often not recognized.

Sketched gestures were found to be common in the collected data (Table 3). Arrows and lines were the most common graphical elements sketched on digital paper (29% and 28% of the sketched non-handwritten information respectively); circles of various kinds represented 20% of the elements. More elaborate elements (such as maps, or representations of furniture) accounted for 16% of the sketches. These in some cases illustrated parts of the environment depicted by the photo that could not be seen, or highlighted features of a photo (e.g. the outline of a person barely visible in a photo). Lines were the most successfully recognized elements (73% recognition rate); only 10% of the arrows were recognized. The sketch recognizer was not trained to recognized other elements.

	N1		N2		N3		N4	
Line	2	7%	17	32%	19	50%	1	5%
Arrow	11	39%	3	6%	12	32%	14	70%
Circle	4	14%	17	32%	4	11%	3	15%
Other	11	39%	16	30%	3	8%	2	10%
Total	28	100%	53	100%	38	100%	20	100%

Table 3: Sketching symbols used.

5.5.2 Speech plus handwritting

Not surprisingly handwritten information was considerably more concise than the accompanying speech. In fact, handwriting represents less than 2% of the overall number of term instances produced via speech and handwriting within our pilot data corpus.

To investigate the main question - whether the information users choose to handwrite corresponds to key descriptive terms, we analyzed the correlation between spoken words and handwritten labels. This analysis was based on hand-transcription of handwriting and speech. Here we report on the results for the two narrators that were native English speakers, those for which speech recognition results might be at a level that would allow the application of the cross-modal learning we want to explore.

Table 4 shows the number of terms before and after stop word removal, as well as the amount of redundant handwritten and spoken terms. In order to match handwriting to speech we expanded handwritten abbreviations so that they would match the spelling of how they might appear in the speech transcript. In matching handwritten terms to speech terms we counted as matches terms that differed only in simple number form, as determined by adding or removing a terminal ”s”.

#		N1	N2	Avg	Std
1	All Word Types	1623	1681	1652	41
2	Stop words removed (WT)	899	920	910	15
3	Average WT Frequency	2.16	2.52	2.34	0.25
4	Average WT Rank	196.92	162.76	179.84	24.15
5	Handwritten Words (HW)	89	66	77.5	16.23
6	Redundant Words (RW)	79	66	69	5.66
7	Average RW Frequency	3.97	4.93	4.45	0.68
8	Average RW Rank	83.2	71.74	77.47	8.10
9	Freq. increase of RW	83.8%	95.6%	90%	8%
10	Rank increase of RW	57.7%	55.9%	57%	1%
11	Redundancy RW / HW	90%	100%	95%	7%

Table 4: Speech and handwritten (HW) data for the two native English speakers. Redundant word (RW) are those that were spoken and handwritten.

Table 4, row #1 shows the number of individual word types that occured (either spoken or handwritten). Row #2 shows the number of word types remaining after stop word removal. Row #6, shows the number of word types that were delivered redundantly, both handwritten and spoken. The averages in rows 3,4,6, and 7 are figured after stop word removal, while the percentages of handwritten words that were also spoken redundantly (row #10) includes stop words.

If we compare the average rank and frequency of all occuring word types (rows #3 and #4) to the averages for those word types that were delivered redundantly (rows #7 and #8), we see a striking increase in average rank and frequency. The rank increase is 57.7% and 55.9% (row #10) and the frequency increase is 83.8% and 95.6% (row #9) for N1 and N2 respectively. The percentage of handwritten words that are also spoken redundantly (row #11) is 90% and 100% for N1 and N2 respectively. In other words, we find not only that the handwritten labels are present redundantly in the speech, but also that the handwritten labels correspond to words that occur with high frequency while discussing a photo. This in turn reflects the relative importance the handwritten terms have to the topic under discussion, and supports our claim that redundantly presented terms are dialogue critical.

These preliminary findings suggest potential gains for techniques that concentrate processing resources on recovering these terms from multimodal and redundant terms, as we discuss in the next Section.

5.6 Robust multimodal label recovery

While we expect that the recognition rates of sketch, speech and handwritting recognizers will be considerably enhanced with retraining, the fact remains that recognition of unconstrained group interactions will be error prone for the foreseeable future.

To address this intrinsic limitation, we explore ways to leverage the high rate of redundancy between handwritten labels and high-frequency spoken terms to enhance label recovery. These high-

#	HW Labels	N1		N2		N3		N4		Avg		Std
1	Total	89	100%	66	100%	42	100%	32	100%	57.3	100%	25.5
2	Recognized	32	36%	37	56%	11	26%	16	50%	24.0	42%	12.5
3	Recoverable (in n-best)	14	16%	3	5%	2	5%	4	13%	5.8	10%	5.6
4	Abbreviations	6	7%	1	2%	4	10%	1	3%	3.0	5%	2.4
5	Labels / photo	9.9		6.6		4.2		3.6		6.1		2.9

Table 2: Handwritting evaluation results.

value elements are in turn used to correct handwritting and speech transcripts, potentially boosting the overall recognition.

To illustrate the potential of the approach, we selected a small subset of the data (one photo) to process. The technique currently requires labor intensive parameter tuning, making it infeasible to process larger amounts of data.

Figure 3 shows a photo annotation corrected by our multimodal SHACER (our Speech and HAndwriting reCognizER) technique [13]. By aligning handwriting hypotheses with phone sequence hypotheses produced by an ensemble of recognizers (see Section 4), SHACER was able to recover the correct interpretation of the label "17", originally misrecognized as "12".

The hovering label displayed below the recognized label in Figure 3) indicates that the system was able to recover the semantics ("seventeen") from the associated speech, and recognize that it referred to this numeral. The hovering bubble attempts to provide a lightweight indication of the recovered semantics for user consumption.

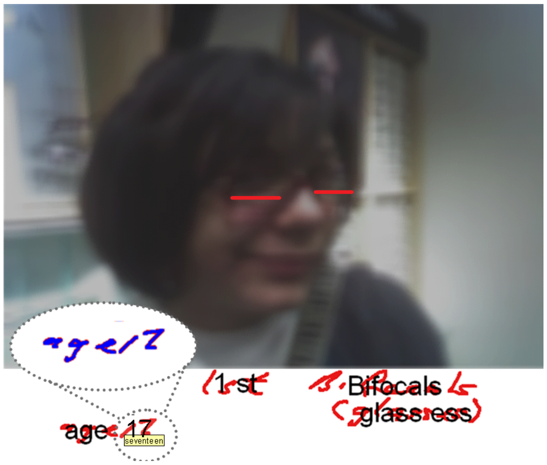

Figure 3: Recognition results corrected via our multimodal re-scoring technique overlayed on original ink (image blurred for privacy).

For the alignment described above to be successful, it is necessary for the correct interpretation to be present in the recognizers' list of alternates, as the approach is based primarily on re-scoring. To verify the potential impact of the technique on the collected corpus, we examined how often the correct term was recoverable from the recognizers' hypotheses (Table 2 - row #3). We found that on average 17% (min 6%, max of 25%) of the handwriting misrecognitions could potentially be corrected via our multimodal alignment technique. More in depth presentations of this technique can be found in [13].

6. SUMMARY AND FUTURE WORK

We presented a system that was implemented to exploit social phenomena to provide an attractive environment for photo annotation. We proposed automatically extracting labels from the multimodal streams produced by groups of people sharing experiences while looking at photographs. The non-intrusive, observant nature of the system, coupled with support for the use of tangible materials - photos printed on digital paper - aims at facilitating current use practices, thus avoiding distractions that might result from a more explicit introduction of technology into the process.

We analyzed pilot data and verified that handwritten labels offer a concise set of label terms, well correlated with high frequency spoken terms. This corroborates our strategy of selecting handwritten labels as the primary indicators of photo semantics. We pointed to recognition challenges and analyzed the potential for exploiting multimodal data to partially address them. Redundancy across speech and handwritten modalities, verified in our analysis of the data, appears to provide opportunities for significant recognition improvements. We illustrated the approach to robust multimodal recognition via the analysis of a small subset of the data. We showed that label correction and semantic recovery, a challenging task, could be achieved using a technique that re-scores recognizers' hypotheses based on the alignment of handwriting and speech.

Much remains to be explored. Future directions include the following:

- We expect to collect additional data under different formality conditions, involving co-located, distributed and mixed scenarios, with and without displays. Besides providing us with additional insights into the domain, we expect the data will be used to retrain our recognizers, boosting recognition rates.
- We want to extend the extraction to contemplate a broader semantic range of identified labels. We want to populate a richer ontology capturing label categories (identifying e.g. location, time, people and other classes identified in the data), and recover higher level connections, for instance relationships among people, and narrative threads that may help in organizing photo collections, both of which were observed in the data.
- The techniques presented here are complementary to work that exploits other contextual information that can be obtained by sensors , e.g. time and location [19], or by image interpretation using vision-based techniques (e.g. [26]). This additional contextual information could be used to offer additional opportunities for label propagation and robust 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).

7. REFERENCES

[1] R. J. Anderson, R. Anderson, C. Hoyer, and S. A. Wolfman. A study of digital ink in lecture presentation. In *CHI 2004: The 2004 Conference on Human Factors in Computing Systems*, Vienna, Austria, 2004.

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

[3] P. Barthelmess, E. Kaiser, X. Huang, and D. Demirdjian. Distributed pointing for multimodal collaboration over sketched diagrams. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, New York, NY, USA, October 2005. ACM Press.

[4] J. Chen, T. Tan, P. Mulhem, and M. Kankanhalli. An improved method for image retrieval using speech annotation. In *Proc. 9th International Conference on Multimedia Modeling (MMM 2003)*, pages 15–32, Taipei, January 2003.

[5] P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow. Quickset: Multimodal interaction for distributed applications. In *Proceedings of the Fifth ACM International Multimedia Conference*, 1997.

[6] P. R. Cohen and D. R. McGee. Tangible multimodal interfaces for safety-critical applications. *Communications of the ACM*, 47(1):41–46, 2004.

[7] P. Debary, P. Goddi, R. Gossweiler, R. Rajani, A. Vorbau, and J. Tyler. Enabling informal communication of digital stories. Technical Report HPL-2004-180, HP Laboratories Palo Alto, 2004.

[8] C. Decurtins, M. C. Norrie, and B. Signer. Digital annotation of printed documents. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 552–555, New York, NY, USA, 2003. ACM Press.

[9] M. Fleck. Eavesdropping on storytelling. Technical Report HPL-2004-44, HP Laboratories Palo Alto, 2004.

[10] 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.

[11] Google, Inc. Picasa's hello. `http://www.hello.com`, 2006.

[12] M. Johnston, P. Cohen, D. McGee, S. Oviatt, J. Pittman, and I. Smith. Unification-based multimodal integration. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 1997.

[13] E. Kaiser. Using redundant speech and handwriting for learning new vocabulary and understanding abbreviations. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*. ACM Press, 2006.

[14] E. Kaiser and P. Barthelmess. Edge-splitting in a cumulative multimodal system, for a no-wait temporal threshold on information fusion combined with an under-specified display. In *Proceedings Interspeech 2006 - ICSLP)*, 2006.

[15] J. Kustanowitz and B. Shneiderman. Annotation for personal digital photo libraries: Lowering barriers while raising incentives. Technical Report HCIL-2004-18, Univ. of Maryland, January 2005.

[16] H. Lieberman, E. Rosenzweig, and P. Singh. Aria: an agent for annotating and retrieving images. *Computer*, 34(7):57–62, July 2001.

[17] D. McGee and P. Cohen. Creating tangible interfaces by augmenting physical objects with multimodal language. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI 2001)*, 2001.

[18] T. Mills, D. Pye, D. Sinclair, and K. Wood. Shoebox: A digital photo management system. Technical Report 2000.10, AT&T Laboratories, Cambridge, 2000.

[19] M. Naaman, R. B. Yeh, H. Garcia-Molina, and A. Paepcke. Leveraging context to resolve identity in photo albums. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 178–187, New York, NY, USA, 2005. ACM Press.

[20] S. Oviatt. Mutual disambiguation of recognition errors in a multimodal architecture. In A. Press, editor, *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 576–583, 1999.

[21] Y. Qian and L. M. G. Feijs. Exploring the potentials of combining photo annotating tasks with instant messaging fun. In *MUM '04: Proceedings of the 3rd international conference on Mobile and ubiquitous multimedia*, pages 11–17, New York, NY, USA, 2004. ACM Press.

[22] M. Ravishankar. *Efficient Algorithms for Speech Recognition*. PhD thesis, Carnegie Mellon University, School of Computer Science, Pittsburgh, PA, May 1996. Also published as Technical Report CMU-CS-96-143.

[23] 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.

[24] C. Shen, N. Lesh, and F. Vernier. Personal digital historian: story sharing around the table. *interactions*, 10(2):15–22, 2003.

[25] B. Signer. Fundamental concepts for interactive paper and cross-media information spaces. Dissertation, ETH Zürich, Switzerland, 2005. No. 16218.

[26] R. Srihari and Z. Zhang. Show&Tell: A semi-automated image annotation system. *IEEE Multimedia,*, 7(3):61–71, Jul-Sep 2000.

[27] 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.

[28] L. Wenyin, S. Dumais, Y. Sun, H. Zhang, M. Czerwinski, and B. Field. Semi-automatic image annotation. In M. Hirose, editor, *Human-Computer Interaction–Interact '01*, pages 326–333. IOS Press, 2001.

[29] B. Wu, R. Singh, P. Gupta, and R. Jain. eVitae: An event-based electronic chronicle. In *Proc. International Conference on Extending Database Technology (EDBT)*, 2004. Demonstration Paper.

[30] R. B. Yeh, C. Liao, S. Klemmer, F. Guimbretière, B. Lee, B. Kakaradov, and J. S. A. Paepcke. Butterflynet: A mobile capture and access system for field biology research. In *CHI: ACM Conference on Human Factors in Computing Systems*, Montréal, Québec, Canada, 2006.