Toward Content-aware Multimodal Tagging of Personal Photo Collections

Paulo Barthelmess, Edward Kaiser and David McGee
Adapx
Seattle, WA - USA

ABSTRACT

A growing number of tools is becoming available, that make use of existing tags to help organize and retrieve photos, facilitating the management and use of photo sets. The tagging on which these techniques rely remains a time consuming, labor intensive task that discourages many users. To address this problem, we aim to leverage the multimodal content of naturally occurring photo discussions among friends and families to automatically extract tags from a combination of conversational speech, handwriting, and photo content analysis. While naturally occurring discussions are rich sources of information about photos, methods need to be developed to reliably extract a set of discriminative tags from this noisy, unconstrained group discourse. To this end, this paper contributes an analysis of pilot data identifying robust multimodal features examining the interplay between photo content and other modalities such as speech and handwriting. Our analysis is motivated by a search for design implications leading to the effective incorporation of automated location and person identification (e.g. based on GPS and facial recognition technologies) into a system able to extract tags from natural multimodal conversations.

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

General Terms

Design; Experimentation; Human Factors

Keywords

Photo Annotation; Automatic Label Extraction; Collaborative Interaction; Multimodal processing; Intelligent interfaces

1. INTRODUCTION

A variety of techniques that take advantage of meta-data associated to multimedia materials such as videos and photos are becoming available. Besides the facilitation of retrieval of photos from large volume datasets, meta-data has been used in a variety of ways, for example to automatically organize photos via clustering and the creation of collages [6], or the automatic creation of videos from selected photos [10]. While the desirability of having annotated materials is recognized, the annotation task itself remains a burdensome, labor intensive endeavor that discourages wider adoption [9].

Social tagging has surfaced as way to promote the creation of meta-data by leveraging the work of multiple people, and the visibility of materials shared by a community. Services such as Flickr have been able to attract a large number of users and to successfully promote the tagging of photos. Such systems offer users the social incentive to tag, and at the same time may facilitate tagging by allowing for some degree of sharing of the associated tagging workload among more than one person.

We have been exploring a technique that is also based on the exploitation of a social phenomenon, but on a much smaller scale than community-based ones. More specifically, we aim to leverage the conversations that take place naturally among groups of friends of families while they discuss photos, e.g. of recent trips, or events such as weddings, birthdays and so on. Our ultimate goal is to exploit the richness of multimodal language - the combination of speech, handwriting, gestures, that takes place during these occasions to automatically extract tags adapted to specific *idiolects* and *ecolects* - tags that are meaningful to specific individuals and families respectively [11]. The challenge becomes identifying discriminative tags able to capture the idiosyncrasies of these specialized vocabularies in a "cheap" way.

Research that has examined social aspects of tagging has focused on characterizations of tagging behavior of large sets of users and tagged elements (e.g. in Marlow et al. [11]), or has proposed mechanisms to further structure tags, for instance by inducing an ontology (e.g. in Schmitz [13]. The social interaction we exploit here is of a different nature. Rather than relying on asynchronous tagging performed by a community, we focus on the discussion of smaller groups of people. Our objective is to understand interactions and the language they entail, looking for opportunities to facilitate tagging by semi-automatic extraction of labels from speech, handwriting and photo content.

Others have looked into exploiting group discussions to extract tags (e.g. Fleck [7] and Qian and Feijs [12]). Our work is distinguished by our focus on the use of multiple modalities to extract and propagate robust labels from unconstrained, large vocabulary conversation streams, supporting natural practices.

In previous work, we identified the role of redundantly delivered handwritten and spoken terms, finding that these terms have significantly higher average word frequency than terms that are just spoken [3]. This high frequency reflects in turn in higher tf.idf (term frequency, inverse document frequency) weights [2], indicating their high discriminative information retrieval power [8].

ICMI'07, November 12-15, 2007, Nagoya, Aichi, Japan.

The present paper extends this previous foundation by examining the relationship of spoken and handwritten multimodal tags to photo content related to location and the identity of people appearing in a photo. The new analysis contributed here is oriented by multimodal processability goals, looking for robust multimodal features to leverage within systems that rely on natural language recognition, with the goal of automating the extraction of photo tags.

1.1 Collection, transcription and annotation

Pilot data was collected using a system that captures speech, handwriting, and sketching performed on photos printed over digital paper. The conditions of the collection and the technology it employed, which takes advantage of digital paper media is described elsewhere [3, 4].

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. Sessions lasted about 38 minutes in average, with a standard deviation of 7.6 minutes. The narrators provided their own pictures (9 or 10), most of which had been taken abroad during vacation or conferences.

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.

The pilot data was then hand annotated. Transcriptions were created for speech, digital ink and photo content. Timing information (at a sentence granularity) was associated with each term.

The digital ink was examined to extract handwritten labels and sketched information. Sketches were classified primarily as *regions* (ellipsoids or rectangles marking parts of the scene) and *pointers* (lines and arrows, usually connecting regions of the scene to handwritten labels). Other elements were classified as *decorations*. These included a variety of highlights applied to photos to show for instance the outline of a person, or the skyline of a city, or stalagmites in a cave. Other decorations depicted parts of the scene that laid outside of the photo itself, such as additional furniture not visible in a picture, or elements in a park around a water fountain that were not shown (statistics of the graphical aspects is presented elsewhere [3]).

Photo content annotation identified the people present in each photo (if any). Such identification was performed either directly, based on explicit handwritten labels provided by the participants, or in cases in which no such label was provided, on the visual similarity with previously identified faces. It is worth mentioning that in most cases the annotators did not have trouble identifying the persons depicted in each photo, which points to the fact that the contextual information naturally provided by the participants via speech and handwriting was thorough enough to permit easy identification.

Besides identifying the people present in each photo, annotators added location information, usually corresponding to a short sentence (e.g. "Jenolan Caves", or "Lake Ruby"). Again this was easily obtainable from the larger multimodal context (and can be automated by exploiting GPS or cell id information to search for locations within a geographic information system, as proposed e.g. by Davis et al. [5] and Tuffield et al. [1]).

The transcripts were then segmented by photo discussion. A combined collection of multimodal terms was then built by concatenating information from speech, handwriting and content transcripts. This list was filtered to remove stop-words (those common words that are not deemed to carry significant semantic information). This composite filtered list was then subjected to the analysis as described in the next section.

2. ANALYSIS

Throughout this section we will refer to the words that are produced via speech, handwriting and identified photo content elements (e.g. a person's name) interchangeably as *tags* or *terms*; the discussion of individual photos will be referred to as a *segment*; the set of all segments for a subject is called a *session*. A *collection* is a set of all sessions.

2.1 Content-related tag weights

A first question we address is how salient or important content-related terms are, and how they relate to terms delivered via other modalities. For the purposes of this analysis, we adopt a salience criterium based on tf.idf (term frequency, inverse document frequency) weights [2].

Tf.idf is a standard statistical measure employed in the context of information retrieval and document content analysis. The tf.idf weight of a word in a document within a collection is proportional to the number of occurrences of the word within a document (*tf*) and inversely proportional to the number of occurrences of the word across different documents (*idf*).

For the purposes of this analysis, we equate a document to a segment during which a photo is discussed. A high tf.idf weight of a term used during the discussion of a photo will therefore result from a high frequency of occurrence within a photo discussion, and a relative low occurrence during the discussion of other photos in the collection. The highest ranked terms are then taken to be the most representative tags to be associated to a photo.

We find that content terms referenced during the multimodal discussion have significantly higher tf.idf weights in average (5.92), when compared to the average weight of all terms (2.79) - a 212% increase (Figure 1). This confirms that the salience of the content is indeed reflected in the frequency of references via speech, handwriting or both, a primary pre-condition for constructing a multimodal mechanism exploiting content recognition.

This finding also points to the potential high payoff in recovering references to content, as these constitute a class with the highest discriminative power as measured by the tf.idf weights.

2.2 Tag distribution

Figure 2 shows the distribution of content-related tags according to the modality employed to reference them. Of the content referenced during a discussion, 74% of the references were provided redundantly via speech and handwriting. Furthermore, almost 60% of all redundantly handwritten and spoken tags are content-related. In 23% of the cases references to content were just spoken. Cases in which the references were just handwritten were rare (3%), and corresponded furthermore to a style adopted by just one of the subjects.

The dominance of redundantly spoken and handwritten content references indicates that techniques exploiting the redundancy between hypotheses generated by a content analyzer, a speech recognizer and a handwriting recognizer can be profitably applied.

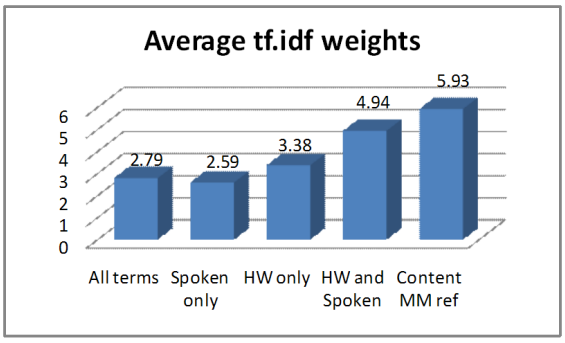

Figure 1: Tf.idf weights for classes of tags. From left to right a) all terms; b) terms that were only spoken; c) terms that were only handwritten; d) terms that were redundantly handwritten and spoken; e) content referenced via speech, handwriting or both.

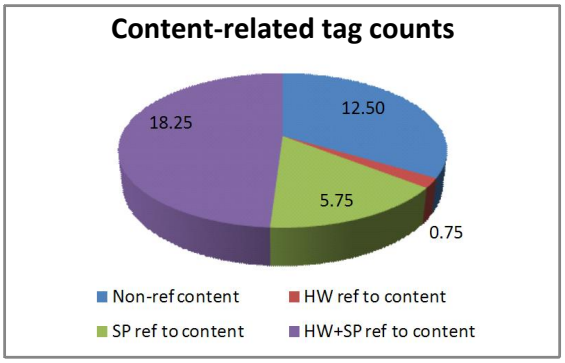

Figure 2: Average number of content-related tags per session.

2.3 Content spanning multiple photos

In a considerably large number of instances - 34% - of all photo content, the content present in a photo was not referenced, either via speech or via handwriting.

Content that is not reference emerges more commonly in situations in which the same places or people appear repeatedly in multiple photos. Typically only a portion these repeated instances are referenced via speech or handwriting. We found that 61% of the content-related tags appear in more than one photo; in average, repeated terms appear in about 2.5 photos.

From the perspective of an automated tag extractor, the relatively large amount of content that is not directly referenced in other modalities points to the advantage of cross-photo tag propagation mechanisms. Automatic content analysis can then be bootstrapped by existing multimodal references, providing tags to be associated with faces and locations that are identified solely from content analysis.

A high degree of cross-photo repetitions is also found for tags in general - 46% of all the tags appear in more than one photo; the average repetition for all tags is close to 3. The bulk of the repeated tags correspond to spoken tags that have relatively low tf.idf weights (given their repetition across multiple documents).

2.4 Relative contribution of content-related tags

Even though content-related tags account for the highest overall tf.idf weights, there are relative few tags of this kind, and their overall contribution is relatively small (Table 1. Content-related tags are just 5% of the overall number of tags. When we sum the tf.idf weights of all content-related tags and divide that by the sum of the tf.idf weight for all tags, we see that this class of tags contributes 8.6% of the overall mass of tf.idf weights for all terms.

Class	Count	% Count	Avg weight	% Weight
All terms	825.75	100.0%	2.79	100.0%
Non ref content	12.50	2.0%	2.72	1.6%
HW ref to content	0.75	0.1%	1.14	0.1%
SP ref to content	5.75	0.6%	4.60	1.1%
HW+SP ref to content	18.25	2.5%	6.33	5.8%
Total content tags	37.25	5%		8.6%

Table 1: Tag distribution showing relative contribution of content-related classes. *Count* is the average number of terms per class; *% Count* gives class count divided by the total number of terms for all classes; *Avg weight* lists the average tf.idf for the class; *% Weight* lists class contribution against the overall mass of tf.idf for all classes.

2.5 Time of reference

An implication in terms of system design is that these few but valuable tags may be hard to recover from a much larger set of tags, most of which will be primarily spoken.

To provide additional insights that may contribute to the development of a technique to identify these tags, we examined the time of introduction of new tags, focusing primarily on content-related ones. We divided the time span during which a photo was discussed in five sub-segments of equal duration and computed the number of tags of a class introduced during each of these sub-segments. Figure 3 shows that content-related tags are introduced predominantly at the beginning of the discussion - almost 90% of the content-related tags referenced via speech, handwriting or both appear at the beginning of a photo discussion segment. In contrast, the rate of introduction of new tags in general remains fairly stable and evenly distributed throughout the discussion.

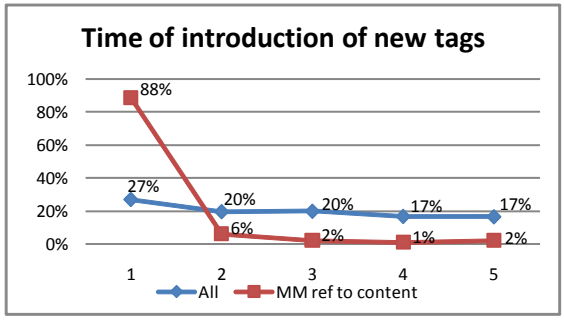

Figure 3: Percent of new tags introduced within five temporal sub-segments. References to content via speech, handwriting or both are introduced primarily at the first fifth of a photo discussion segment.

A system looking at discovering content-related tags could therefore more profitably concentrate its efforts at the beginning of the discussion, assigning lower confidence to tags recovered later than the first quarter (or even fifth) of the discussion time. Conversely, a system able to determine a period of concentrated references to content might be able to use that as evidence to help segment the discourse by raising the likelihood that a new photo is being discussed. The latter depends naturally on the availability of highly

likely content information, something that is not available in most cases.

3. SUMMARY AND FUTURE WORK

We have presented an analysis of collaborative photo annotation pilot data, focusing particularly on the interplay between multimodal language and photo content, aiming at informing the construction of a multimodal system that integrates automated content analysis into a system that automatically extracts tags from unconstrained group discussions. In summary, the main findings reported here are:

- Content-related tags have high tf.idf weights, 212% higher than the average weight of all tags.
- Content-related references are delivered primarily multimodally - 74% of the references are both handwritten and spoken. These multimodal references represent 60% of all the redundantly handwritten and spoken tags.
- Content-related tags represent a small portion - 5% - of all tags, and correspond in average to an 8.6% of the overall tf.idf weight mass for each segment.
- This small set is introduced over a concentrated period of time, right at the beginning of a new photo discussion. Almost 90% of content-related spoken or handwritten tags are introduced during the first fifth of the time span of a photo discussion segment.
- A sizeable amount of content - 34% - is not referenced via speech or handwriting. That happens primarily in situations in which the same person or place appears in more than one photo.
- There is a relatively high degree of commonality among photo contents. In average, we found that about 60% of the content appears in more than one photo, with an average recurrence of about 2.5.

The work presented here is a preliminary step towards the integration of automated photo content analysis into a multimodal system. Besides collecting more data, future work towards this end includes:

- The analysis presented here is based on manual transcriptions and therefore represents a best-case scenario. We plan therefore to assess the impact of mis-recognitions and to develop strategies for overcoming the uncertainty that is intrinsic in recognition-based approaches. The high degree of redundancy observed, and the recurrent nature of the content lead us to believe that mutual disambiguation techniques might be profitably used to recover content-tags robustly even in presence of mis-recognitions.
- A stronger notion of document relevance is required to permit an analysis based on recall and precision. We are planning to run experiments in which subjects will provide retrieval evaluation, leading to relevance scores, and insights into the nature of queries.

Acknowledgments

The authors benefited from discussions with David Demirdjian. Candice Erdmann performed trancription; Alex Arthur provide audio and video collection software; Xiao Huang contributed to the data collection. The authors thank them, as well as the volunteers that took part in the collection.

4. REFERENCES

[1] M. T. andM. Mischa S. Harris, D. Duplaw, A. Chakravarthy, C. Brewster, N. Gibbins, K. OHara, F. Ciravegna, D. Sleeman, N. Shadbolt, and Y. Wilks. Image annotation with photocopain. In *Proceedings First International Workshop on Semantic Web Annotations for Multimedia (SWAMM)*, 2006.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[3] 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.

[4] P. Barthelmess, D. McGee, and P. Cohen. The emergence of representations in collaborative space planning over digital paper: Preliminary observations. In *CSCW 2006 Workshop on Collaborating over Paper and Digital Documents (CoPADD)*, 2006. Available at `http://www.copadd.ethz.ch/abstracts/11.pdf`.

[5] M. Davis, M. Smith, F. Stentiford, A. Bambidele, J. Canny, N. Good, S. King, and R. Janakiraman. Using context and similarity for face and location identification. In *Proceedings of the IS&T/SPIE 18th Annual Symposium on Electronic Imaging Science and Technology Internet Imaging VII*. IS&T/SPIE Press, 2006.

[6] N. Diakopoulos and I. Essa. Mediating photo collage authoring. In *UIST '05: Proceedings of the 18th annual ACM symposium on User interface software and technology*, pages 183–186, New York, NY, USA, 2005. ACM Press.

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

[8] E. Kaiser, P. Barthelmess, C. Erdmann, and P. Cohen. Multimodal redundancy across handwriting and speech during computer mediated human-human interactions. In *Computer Human Interaction (CHI)*, 2007.

[9] 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.

[10] N. Kuwahara, K. Kuwabara, N. Tetsutani, and K. Yasuda. Using photo annotations to produce a reminiscence video for dementia patients. In *3rd International Semantic Web Conference (ISWC2004)*, 2004. Demo Papers.

[11] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM Press.

[12] 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.

[13] P. Schmitz. Inducing ontology from flickr tags. In *Proc. of the Collaborative Web Tagging Workshop (WWW Š06)*, May 2006.