MAMI: Multimodal Annotations on a Camera Phone

Xavier Anguera
Telefónica Research
Via Augusta 177, 08021 Barcelona, Spain
xanguera@tid.es

Nuria Oliver
Telefónica Research
Via Augusta 177, 08021 Barcelona, Spain
nuriao@tid.es

ABSTRACT

We present MAMI (*i.e.* Multimodal Automatic Mobile Indexing), a mobile-phone prototype that allows users to annotate and search for digital photos on their camera phone via speech input. MAMI is implemented as a mobile application that runs in real-time on the phone. Users can add speech annotations at the time of capturing photos or at a later time. Additional metadata is also stored with the photos, such as location, user identification, date and time of capture and image-based features. Users can search for photos in their personal repository by means of speech. MAMI does not need connectivity to a server. Hence, instead of full-fledged speech recognition, we propose using a Dynamic Time Warping-based metric to determine the distance between the speech input and all other existing speech annotations. We present our preliminary results with the MAMI prototype and outline our future directions of research, including the integration of additional metadata in the search.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; H.5.1 [**Information Interfaces and Presentation**]: Multimedia; H.3.m [**Information Storage and Retrieval**]: Information Search and Retrieval; I.5.4 [**Pattern Recognition**]: Applications; K.8.m [**Personal Computing**]: Miscellaneous

General Terms

Mobile Camera Phones, Speech Annotations, Multimedia Retrieval, User Experience, Digital Image Management

1. INTRODUCTION

Mobile phones have become multimedia devices. It is not uncommon to observe users capturing photos and videos on their mobile phones, instead of using digital cameras or camcorders. As consumers generate an increasing number of digital multimedia content, finding a specific image, audio clip or video becomes a non-trivial task. Often, this rich multimedia content is lost in the users personal repositories due to the lack of efficient and effective tools for tagging and searching the content. One solution to this multimedia data management problem is the addition of annotations or metadata to the content [1, 2, 6, 11], therefore allowing users to search for multimedia information using keywords related to their annotations. However, the vast majority of prior work on personal image management has assumed that the annotation of multimedia content occurs at a *later time* and in a *desktop* computer. Time lag and device and context change significantly reduce the likelihood that users will perform the task, as well as their accurate recall of the context in which a particular photo or video was taken.

Mobile devices are designed to take into account the users' physical environment and situation and can ultimately allow the inference of the image or video content from the context. In recent years there has been some interesting and relevant research directed towards real-time multimodal annotations on mobile phones. Related work takes advantage of GPS-derived location metadata [8] or content-based image retrieval and user verification to achieve high-level metadata [9].

In this area, there are two pieces of previous work that are particularly related to ours. On the one hand, Wilhelm *et al.* [10] have developed a prototype that allows users to annotate digital photos at the time of capture using location (cellID), user name, date and time. The images and their automatic annotations are sent via GSM/GPRS to a server, where the new content and annotations are matched against a repository of annotated images. The server sends annotation *guesses* back to the user as a drop down list, who can confirm the suggested annotation or input new textual annotations. In their user studies, they encountered that limited connectivity and network unpredictability and errors were among the most important problems to address. On the other hand, Hazen *et al.* [2] describe a speech-based annotation and retrieval system for digital photographs. Their system is also implemented as a light-weight client connected to a server that stores the digital images and their audio annotations, together with a speech recognizer for recognizing and parsing query carrier phrases and metadata phrases. Preliminary experiments demonstrate successful retrieval of photographs using purely speech-based annotation and retrieval.

Given all previous work, the main contributions of this paper are: (1) The development of a mobile application, named MAMI (*i.e.* Multimodal Automatic Mobile Indexing), to add multimodal (location, date and time, user identification, audio and image features) metadata to photographs at the time of capture; (2) The implementation of a Dynamic Time Warping-based distance function for comparing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MobileHCI 2008 September 2-5, 2008, Amsterdam, the Netherlands.
Copyright 2008 ACM 978-1-59593-952-4/08/09 ...$5.00.

speech annotations; (3) The development of an application to search and retrieve content from the users personal photo repository by means of speech; (4) The evaluation of the proposed distance function in the context of a photo search task in a small user study.

2. SYSTEM DESCRIPTION

MAMI's two modes of use are depicted in Figures 1 and 2. Figure 1 depicts an example of MAMI's capture and annotate functionality. This mode implements the picture taking functionality and addition of metadata at the time of capture: time and date, location, user identity, speech annotation and image-based features. When the user takes a digital photo or video with their mobile phone (step 1 on the Figure), this component automatically gathers available contextual metadata (step 2) and allows users to enter an audio annotation at the point of capture (step 3) via a push-to-talk method. The digital content and annotations are stored in MAMI's picture and metadata databases, which are locally stored on the phone (step 4).

The second mode of use, illustrated in Figure 2, allows users to search and retrieve photos from their personal repository by means of speech. The user provides an input utterance via a push-to-talk method (step 1). MAMI computes a Dynamic Time Warping distance measure between the input speech and all existing audio annotations in the user's digital media database (steps 2 and 3). MAMI returns the N (where N is typically 4) photos whose speech annotations are the closest to the input speech (step 4), and presents them to the user (Step 5). All the processing is carried out on the phone.

If the user has high-bandwidth connectivity and desires to do so, the user's pictures and associated metadata can be uploaded to a remote server.

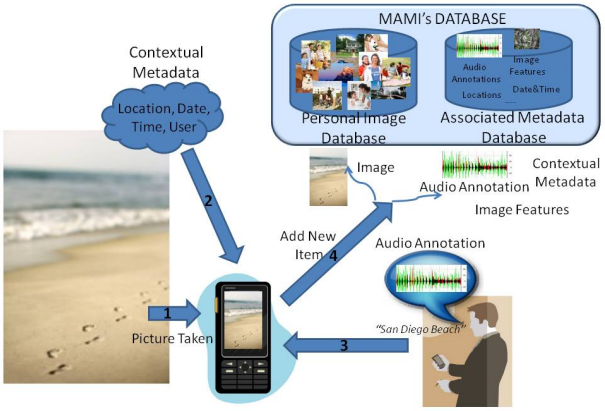


Figure 1: MAMI in capture and annotation mode.

Note that in this paper we focus on presenting the speech annotation and search aspects of MAMI. Other aspects of the MAMI prototype, such as the image-based features used as metadata, multimodal search and its user interface evaluation are still ongoing work and will appear in later publications.

MAMI is implemented as a Windows Mobile application. In our experiments, we have used the HTC Touch^{TM} phone running Windows Mobile 6.0. Figure 3 displays MAMI's interfaces. Figure 3 (a) shows MAMI in capture-and-annotate mode. Figure 3 (b) displays MAMI's search and retrieval

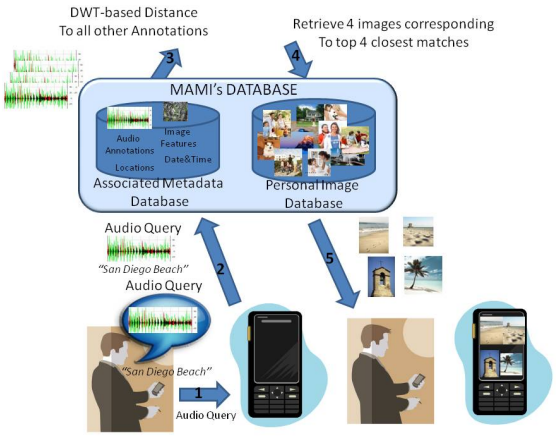
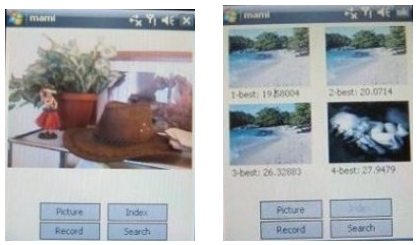


Figure 2: MAMI in search mode.

mode, where the four closest pictures to the input query *beach* are displayed.



(a) (b)

Figure 3: MAMI's two modes of use : (a) Capture and annotate. (b) Search.

3. SPEECH ANNOTATIONS

As mobile phones computation capabilities increase, so does the complexity of the applications that can be run on them. However, mobile devices have their own limitations for browsing and entering information, including smaller screen size and slower text input than in desktop systems. At the same time, mobile devices are today multimodal systems and other modalities can be leveraged to improve the user experience. In fact, speech has been claimed to be a natural input modality for mobile phones [3]. Today there are some commercial and research mobile phone-based prototypes that perform speech recognition in real-time. Most of these systems, however, need to operate in a closed domain and with limited vocabulary, making their use difficult in applications beyond what they were designed for.

Moreover, the problem of personal multimedia data annotation typically involves frequent out-of-vocabulary terms, such as proper names of people and locations, in addition to a wide range of domains that the annotations could belong to. Therefore, we propose a pattern matching metric to carry out the search, without the need of recognizing what those annotations *mean*. In addition to being domain independent, this technique is language independent and requires much lower computational cost than any state-of-the-art speech recognition system.

3.1 Silence Detector

As explained above, each speech utterance is obtained via a push-to-talk method: the user presses a button on MAMI's interface to start and stop recording the annotation. This approach generates a variable amount of silence and/or noise at the beginning and ending of each utterance, in addition to other noises caused by the user or pre-existing in the background. As these factors can seriously jeopardize the success of pattern matching algorithms, they need to be accounted for and controlled.

In order to trim the signal to contain only the region where the image annotation is present, we consider the region of interest to be the region with the highest average energy, surrounded by silence and eventually some impulsive noise (for example, due to pressing the phone's button).

First, we find the point in time in the recording, T_{MaxAvE}, with the highest average energy by means of a 200ms smoothing sliding window, computed every 100ms. Such length was empirically estimated so that plosive and voiceless fricative sounds would not trigger false word endings. The start, T_{start}, and end, T_{end}, of the speech signal is computed as the points in the input signal at either side of T_{MaxAvE} where the average energy falls below the 90% of the signal energy range. Therefore, each speech utterance consists of the portion in the input speech between T_{start} and T_{end}.

In the current implementation, the acoustic matching algorithm takes the selected utterance as a single unit. Therefore, it does not handle multiword annotations. A slight modification of the silence detection algorithm would be suitable to segment individual utterances when the user's annotation consists of more than one word.

3.2 Feature Extraction and Parameter Selection

Acoustic features are extracted once the signal has been trimmed. N Mel Frequency Cepstral Coefficients (MFCC) [4] are computed every 10 or 20ms, without adding any delta cepstrums or energy. These acoustic features (AF) are part of the standard processing in speech analysis. They represent the spectral shape of the acoustic signal, with a scale factor to mimic the human auditory system behavior.

Therefore, each image (I_i) has K associated acoustic feature vectors ($AF_i^1, ..., AF_i^K$) containing the MFC coefficients of the annotation recorded by the user for image I_i. These feature vectors are computed on the indexing step and they are stored in MAMI's metadata database for later use. This process represents the most computationally expensive step in the system, and it is entirely carried out in the cellphone. Therefore, it is desirable to use the simplest but yet effective acoustic parameters possible.

Table 1 and Figure 4 summarize our experiments in acoustic feature selection, *i.e.* the selection of the optimum set of features that better represents each acoustic utterance while discriminating it from all other speech annotations. The experiments were carried out in a small digital image library of 47 images, with 5 audio annotations per image. See Section 4 for a detailed description of the library.

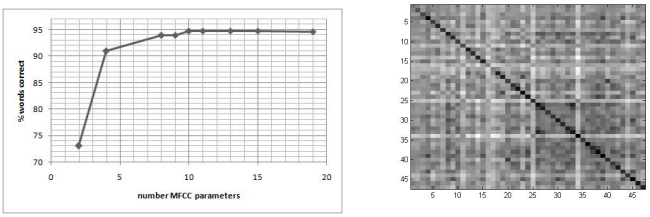
In order to quantitatively compare the feature combinations, we computed the average percentage of correctly classified words for all speakers. We chose this metric as it optimizes the system with respect to word classification error, which is the desired performance metric for our application.

Table 1 shows the performance obtained for a range of acoustic features, including the use of 10 or 20ms frames, Cepstral Mean Normalization (CMN), variance normalization and the inclusion of the 0th MFC coefficient ($C0$), linked to the energy of the utterance. As shown on the Table, the optimum set of parameters does not include $C0$ nor variance normalization, and uses CMN. When comparing 10ms with 20ms, 10ms gets a slight gain in performance, at the expense of higher computational cost.

Table 1: Performance in word recognition for a range of acoustic features.

#MFCC	Frame	C0	CMN	Vnorm.	% correct
13	10ms	0	0	0	93.52%
13	10ms	0	0	1	92.45%
13	10ms	0	1	0	95.16%
13	10ms	1	0	0	89.74%
13	20ms	0	1	0	94.77%
10	20ms	0	1	0	94.77%

Figure 4 (a) illustrates the impact of the number of MFC coefficients in the system's performance. All MFCC were obtained from 20 Mel-scale filters. After reaching 10 MFCC, performance saturates and the only improvement comes in with additional computational cost. Therefore, in MAMI's implementation we have chosen the acoustic features with a frame period of 20 ms and 10 MFCC per frame.



(a) (b)

Figure 4: (a) MFCC selection based on % of word correctness. (b) Word confusion matrix for one of the best database contributors..

3.3 Distance Metric

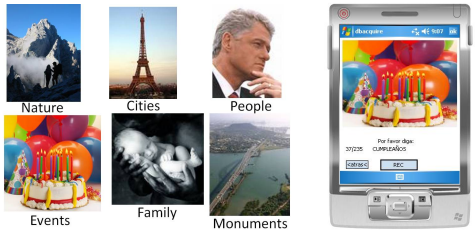
In order to compare two sets of acoustic feature vectors, the Dynamic Time Warping (DTW) metric was selected due to its simplicity, versatility and computational lightness compared to any other speech recognition technique. DTW was extensively used in speech recognition before the upcoming of Hidden Markov Model (HMM)-based stochastic techniques [5]. While HMM modeling can achieve higher performances in a recognition task, it requires lots of training data, high computational cost and is limited to recognizing a pre-defined vocabulary set.

In the MAMI prototype, DTW is implemented allowing any speech utterance to be warped up to double its size. For each pair of acoustic feature vectors, a final distance is obtained by computing the optimum mapping between both words via dynamic programming. In order to speed up the processing, highly unlikely frame pairings are excluded from the computation by applying a global constraint –which consists of a combination of a Sakoe-Chuba (band) mask with an Itakura (romboid) mask [7]– to the distance matrix between both words' frames.

4. PRELIMINARY EXPERIMENTS

To validate the MAMI prototype, we created a small digital image library of 47 images, belonging to one of 6 different

categories: nature, cities, people, events, family and monuments. Figure 5 (a) illustrates some exemplary images used in our experiments.



(a) (b)

Figure 5: (a) Exemplary images of each of the 6 categories used in our experiments. (b) Screenshot of the data collection application.

In order to create a database of speech annotations, we recruited 23 volunteers (14 males, 9 females, with ages ranging from 24 to 42 years old) from a large corporation. All volunteers owned a mobile phone and carried out a variety of positions within the company, including intern students, software engineers, project managers, researchers, administrators and HR specialists. While the audio annotations were done in Spanish, 4 participants did not speak Spanish well and/or had Spanish as their mother language. We developed a data collection application as shown in Figure 5 (b). The application displayed the 47 images of the database in random order. A text label with the annotation that the participants were asked to provide was shown under each image. After seeing each image with its corresponding text annotation, participants pressed the "RECORD" toggle button to record their speech annotation. Once they finished, they pressed "DONE" and the application proceeded to show the next image to annotate. This process was carried out 5 times for each participant. Therefore, at the end of the data collection study, we had a total of 235 speech annotations for each of the 23 participants.

This small database was used for acoustic parameter selection as previously explained. Among all contributors to the database, six of them obtained 100% word correctness using the selected features. On the other hand, the worse contributor had a 74.4% word correctness, which could be explained by the high reverberation of the room where these recordings were made.

Figure 4 (b) shows the confusion matrix for all words recorded by one of the participants with best performance. Rows and columns in the matrix correspond to each of the annotations in the database. Each cell shows the log average of DTW distances for all instances of one word with the other. The darker the shading, the closer both words are in terms of DTW distance. Confusion matrices for other contributors did not diverge from this one in terms of relations between words. Note how some words in the diagonal tend to differ in terms of DTW from other iterations of the same word. In addition, there are a few word pairs that are considered similar by the system, returning thus a low DTW. This is the case of *madre(26)-madrid(27), madre-padre(33), pablo(32)-padre, fuente(20)-puente(40) and playa(37)-plaza(38)*. We plan to leverage image features to help disambiguate these acoustically similar words.

Finally, we carried out a small user study where we asked 7 of the users who had contributed to the database to search 12 random words from their prerecorded picture database (totaling 47 pictures), 2 for each of the presented categories, in a different acoustic environment from the one they recorded in. From a total of 73 queries, only 5 did not return the requested picture in the 4-best results (resulting in a **93.5**% word correctness). In addition, **70**% of all queries returned the requested picture as the **top** result. Users were very pleased with the use of speech as input modality, with how quick the search was performed and the fact that it was a stand-alone application (no connection to a server required). While these results are preliminary, they are certainly very encouraging.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have presented MAMI (Multimodal Automatic Mobile indexing), a mobile-phone based prototype for assisting users in annonating and searching their digital pictures. Experiments performed with a recorded database and a small user study indicate performances finding the desired pictures greater than 90%. Some areas for future work include scaling the system for fast searching in very large pictures databases, adding image-based features and other contextual information to improve the search results and allowing multiple-word annotations and search queries.

6. REFERENCES

[1] M. Davis. *Readings in Human-Computer Interaction: Toward the Year 2000*, chapter Media Streams: An Iconic Visual Language for Video Representation, pages pp. 854–866. Morgan Kaufmann, 1995.

[2] T. Hazen, B. Sherry, and M. Adler. Speech-based annotation and retrieval of digital photographs. In *Proceed. of INTERSPEECH 2007*, 2007.

[3] B. Manaris, V. MacGyvers, and M. Lagoudakis. Universal access to mobile computing devices through speech input. In *Proceed. 12th Intl. Florida AI Research Symposium (FLAIRS-99)*, pages pp. 286–292, 1999.

[4] P. Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*, pages pp. 378–388, 1976.

[5] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proced. of IEEE*, 77(2):pp. 257–286, 1989.

[6] K. Roden and K. Wood. How do people manage their digital photographs? In A. Press, editor, *Proceed. of CHI 2003*, pages pp. 409–416, 2003.

[7] S. Salvador and P. Chan. Fastdtw: Toward accurate dynamic time warping in linear time and space. In *KDD Workshop on Mining Temporal and Sequential Data*, 2004.

[8] K. Toyama, R. Logan, A. Roseway, and P. Anandan. Geographic location tags on digital images. In A. Press, editor, *Proc. of Intl. Conf. on Multimedia*, 2003.

[9] L. Wenyin, S. Dumais, Y. Sun, and H. Zhang. Semi-automatic image annotation. In *Proc. of Interact 2001*, 2001.

[10] T. Y. S. V. H. N. Wilhelm, A. and M. Davis. Photo annotation on a camera phone. In A. Press, editor, *Proceed. of CHI 2004*, 2004.

[11] P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In A. Press, editor, *Proceed. of CHI 2003*, 2003.