Multimodal and Mobile Personal Image Retrieval: A User Study

Xavier Anguera, Nuria Oliver & Mauro Cherubini
Telefónica Research
Via Augusta 177, 08021 Barcelona, Spain
{xanguera,nuriao,mauro}@tid.es

ABSTRACT

Mobile phones have become multimedia devices. Therefore it is not uncommon to observe users capturing photos and videos on their mobile phones. As the amount of digital multimedia content expands, it becomes increasingly difficult to find specific images in the device. In this paper, we present our experience with MAMI, 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 without the need of connectivity to a server. In this paper, we focus on our findings from a user study aimed at comparing the efficacy of the search and the ease-of-use and desirability of the MAMI prototype when compared to the standard image browser available on mobile phones today.

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. Mobile device users typically browse their personal multimedia libraries on standard mobile devices by scrolling through image thumbnails or by manually organizing them on folders and browsing through the folders. 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 personal multimedia data management problem is the addition of annotations or metadata to the content [1, 2, 3, 8], 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 [5], cell-ID to group closeby images [12, 13, 7] or content-based image retrieval and user verification to achieve high-level metadata [6]. Hazen *et al.* [2] describe a speech-based annotation and retrieval system for digital photographs. Their system is 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.

Finally, in the area of mobile image search and retrieval we shall highlight the work of Gurrin *et al.* [11], where pictures are annotated with GPS location and time, and clustered for faster search. They show in a user study that this method outperforms standard browsing, and present a search interface on the device. Their work differs from ours in that they do not use speech annotations on the mobile device and their user study was done in a desktop application and therefore did not take into account any factors derived from using the mobile phone to accomplish the task.

We have developed a mobile phone 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. The focus of this paper is an empirical evaluation of the MAMI prototype when compared to a standard image browser in the context of an image search task.

2. SYSTEM DESCRIPTION

MAMI is a multimodal mobile phone prototype for capturing, indexing, searching and retrieving personal photographs. Figure 2 illlustrates the two interfaces available in MAMI. The capture and indexing interface (depicted on the left of the Figure) allows the user to take pictures and input speech annotations associated with the picture. In the indexing step, each photo is stored in a local database together with a collection of metadata associated with it, including: time, date and location at the time of capture, user identity, speech annotation with associated audio features and image-based features.

The search and retrieval interface (depicted on the right of the Figure) allows the user to input the search query via speech for the image (s)he wants to retrieve. In order to achieve a light-weight local search, we use pattern matching techniques via the Dynamic Time Warping (DTW) algorithm [4] [10] over all indexed annotations. Once the search is finished, MAMI displays the 4 images (4-best) with the closest speech annotations to the input utterance, as can be seen on the Figure (showing the results for an example query "beach"). For a detailed description of MAMI's architecture and speech processing algorithms we refer the reader to [9] where a more complete explanation of the system is given.

3. USER EVALUATION

Prior research in the area of personal digital archive management [3] identifies the following three different types of searches applicable to personal collections: (1) searching for all photos associated with a certain event; (2) searching for a specific photo (known); and (3) searching for all photos sharing some attribute (*e.g.* containing a certain person or location). In the user study presented in this paper, we restricted our evaluation to case (2) above, *i.e.* to users searching for a specific known photo.

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 1 illustrates a few exemplary images used in our experiments. In a previous user study [9] with 23 participants, they user a phone application to add predefined speech annotations to the images, resulting in a total of 235 speech annotations (5 annotation repetitions per image) per user. Users recorded these annotations in a variety of background conditions (noises and reverberation) , all inside a corporate building. In addition, we created a *large* database with 235 pictures. Note that the large database included all 47 images in the small database. All the images in the large database also belonged to one of the 6 above-mentioned categories.

The goal of the user study presented in this paper was to better understand the advantages and disadvantages of the

Figure 1: Exemplary images of each of the 6 categories used in our experiments.

MAMI prototype, when compared to the standard image browser available on mobile phones today. Next, we shall summarize materials and methods applied and discuss the results and conclusions of the user study.

3.1 Material

3.1.1 Hardware and Software

All participants used the same hardware and software. The hardware consisted of an HTC Touch^{TM} Smartphone running Windows Mobile 6.0. The phone had the digital photo database previously described, including the speech annotations for each participant. Both the MAMI prototype and the standard image and video browser were available on the phone.

Figures 2 and 3 depict the user intefaces for MAMI and the standard browser respectively. Note that in the user study, MAMI was only used in *search and retrieval* mode. Therefore, participants only used MAMI's search interface, depicted on the right of Figure 2. As seen on the Figure, this interface allows users to provide via speech input the tag or annotation associated with the image that they are interested in. In the case of the Figure, the user said *beach* and, after pushing the *search* button, the system returned the 4 images whose annotations were the closest to the input (*e.g. beach*).

Finally, Figure 3 depicts the standard browser interface where 12 thumbnails are visible at the same time. In this case, users navigate through the repository via a scrollbar on the right.

3.1.2 Participants

A total of 17 participants (5 female and 12 male) completed the user study. All participants were familiar with both hardware and software, as they had participated a few months ago in a previous user study that used the same hardware and software [9]. They had originally been recruited by email advertisement in a large telecommunications company.

(a) (b)

Figure 2: MAMI's interfaces: (a) indexing and (b) search.

Figure 3: Standard image browser interface.

Before starting the study, participants filled out a questionnaire about their demographic information and digital photo taking habits, both with digital cameras and mobile phones.

Their ages ranged from 26 to 43, with an average of 30.5 years. Their occupations included engineers, graduate students, managers and finance experts. All participants owned a mobile phone, 15 participants owned a digital camera and 13 owned a camera phone. When asked about their picture taking habits, the majority of participants (60%) reported taking pictures with a digital camera 6 to 10 times a month, with a significant difference between the average number of pictures taken during a regular day (7) and a vacation (600). With respect to the camera phone, the majority of participants (85%) reported using their phone to take pictures 5 or less times a month. Similarly as with the digital camera, there was a significant difference between the average number of pictures taken during regular days (2.7) and a vacation (25.1).

The usage model for digital cameras and camera phones seems to be such that the portable device is mostly used to *capture* the pictures, but not to store, browse or search for them. In our study, 11 users (73%) *exclusively* used their PCs to store, search and browse their digital pictures, and only 1 participant reported *exclusively* using the device to carry out such task. The reasons for storing the pictures on the PC included saving space, and the ability to classify and upload them to a web server, to share them via email, to print them and to creating a backup. The average user satisfaction in browsing and searching for pictures with the current technology was 2.8 (SD= 1.0) on a 5-point Likert scale, where 1 corresponded to *not satisfied at all* and 5 to *very satisfied*.

3.2 Methodology

This study was dedicated to compare the MAMI prototype with the standard photo browser available on current Windows Mobile 6.0 devices. In particular, we wanted to: (1) compare MAMI's efficacy (objective and subjective) in helping users find a specific photo; (2) evaluate how easy it was to use MAMI's interface when compared to the standard image browser; and (3) determine the impact that the number of photos in the digital library has on a search task.

3.2.1 Task

Participants were asked to carry out a picture search task four times on the mobile phone, under four different conditions: (1) MAMI with a small database; (2) MAMI with a large database; (3) standard browser with a small database; (4) standard browser with a large database.

Each search task consisted of finding 15 randomly-selected pictures from the pool of 47 unique pictures that composed the small image database. They were given to the participant in written right before staring every task. The list of pictures changed from condition to condition. The order of execution of the conditions was counter-balanced across participants to avoid any bias.

As participants started the study, they filled out a pre-study questionnaire and were instructed in the use of both MAMI (in search mode) and the standard image browser. Next, they were shown all the pictures in the small database, together with their associated speech annotation tags in written. This was done to allow the users to remember the tags associated with each image that they had previously recorded as two months had passed in average between their recording and this test. Finally, they logged into the MAMI system, such that we could: (1) log all their interactions for further analysis; and (2) load their speech annotations for each of the images in the small database.

Once participants felt comfortable with the software and the images, they started the experiment. They were asked to find the 15 pictures that appeared on the annotations target list in the order shown, while the experimenter measured the task completion time. As mentioned above, participants carried out the search task 4 times, each time with a different (*prototype*, *database*) combination.

After completing the task in each condition, participants were asked to fill out a user satisfaction questionnaire (post-use questionnaire). The experiment lasted about 45 minutes per participant.

In the case of the MAMI prototype participants had 3 trials to find a picture. On average, one picture per subject

was not properly retrieved via the speech interface, mostly due to background noise conditions at the time of capture. In those cases, we penalized *a posteriory* the total retrieval time, by adding the average time that the user needed to find a picture with the standard browser.

In addition, the experiment was set-up as a competition: participants were told that the goal of the experiment was to correctly find all the requested pictures in the *shortest* amount of time. The fastest participant to retrieve the requested pictures in the small and large image database would be awarded a prize.

3.2.2 Measures and Treatments

Three measures were quantitatively evaluated in the study:

1. *Efficacy or speed in the search:* This measure was computed as the total amount of time that users needed to find a predefined set of pictures (task completion time). We also obtained a *subjective* measure of efficacy via a post-use questionnaire.

2. *Ease of use:* This measure was obtained from a subjective evaluation via a post-use questionnaire.

3. *Desirability:* This measure was obtained from a subjective evaluation via a post-use questionnaire.

Therefore, we used a single-measure design with the within-participant independent variables *prototype* (MAMI and Standard Browser) and *database size* (small with 47 pictures and large with 235 pictures). Each participant performed one search task and the four dependent variables were: (a) *task completion time*, measured with a chronometer; and (b) perceived efficacy; (b) *ease of use*; and (c) *desirability*, all measured via the post-use questionnaire.

3.3 Statistical Analysis

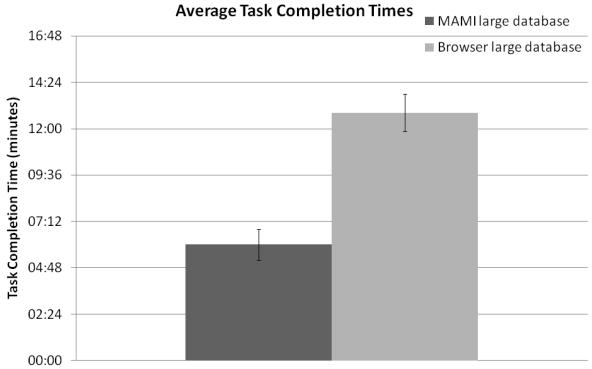
After preliminary analysis of the dataset, we did not find consistent variances in the data, nor the values were normally distributed. Therefore we had to opt for the non-parametric Wilcoxon Signed Rank test [14]. Unfortunately, this type of analysis does not handle a 2×2 factorial design. Therefore, we decided to only look for effects between each prototype used and the four measures (task completion time, perceived efficacy, ease-of-use and desirability). We left the systematic study of an interaction effect between the size of the database and the interface used to future work.

We shall start describing our findings in the large database, as we found stronger effects than in the case of the small database.

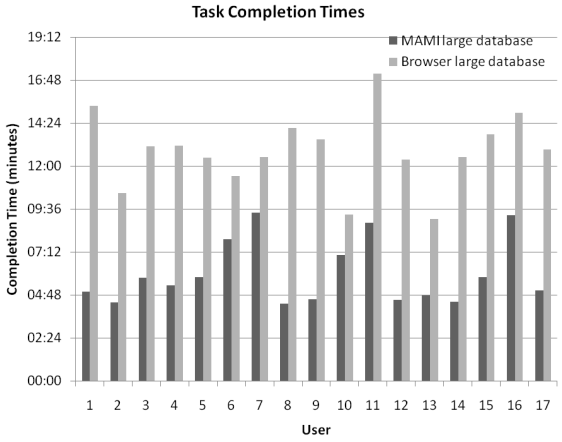
3.3.1 Large Database

1. Efficacy of Search. The test revealed a *significant* effect of the availability of the MAMI interface on the task completion time. Users retrieving the images with the MAMI interface completed the task *faster* than users using the classical interface (0.95 CL, $Z = -3.621$, $p < .001$). The median [quartiles] completion time for subjects without and with the MAMI interface respectively were $775[714, 835]$ and $319[273, 448]$ seconds.

In addition, the MAMI interface also had a *significant* effect in the *perceived* efficacy of the search (0.95 CL, $Z = -3.695$, $p < .001$). In other words, not only participants were faster, but also felt that the MAMI interface allowed them to find the images *faster* than the traditional interface. The median [quartiles] values for the perceived efficacy without and with the MAMI interface were $1[1, 2]$ and $4[3.5, 4]$, respectively, where we used a 5-point Likert scale[1].



(a)



(b)

Figure 4: (a) Task completion times for all users with the large database. (b) Average task completion times with the large database.

2. Ease-of-use and Desirability. Ease-of-use and desirability were determined by the post-use questionnaire, via a 5-point Likert scale. The Wilcox test revealed a *significant* effect of the availability of the MAMI interface on the ease-of-use. Users found the MAMI interface to be easier to use than the standard interface (0.95 CL, $Z = -3.641$, $p < .001$). The median [quartiles] values for subjects with-

[1]In the following, all Likert-scale values correspond to: 1 being *not at all* and 5 being *absolutely*.

out and with the MAMI interface respectively were $1[1,2]$ and $4[4,4]$.

Finally, we also found a *significant* effect of the availability of the MAMI interface on desirability. Users expressed stronger desire to use the MAMI interface than the standard browser (0.95 CL, $Z = -3.562$, $p < .001$). The median [quartiles] values for subjects without and with the MAMI interface respectively were $1[1,2]$ and $4[3,4.5]$.

Figure 5 summarizes the results of the subjective measures.

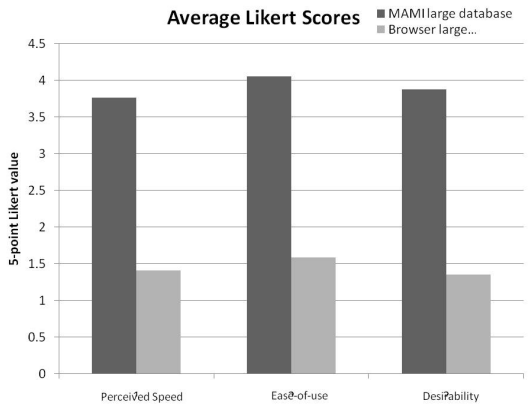


Figure 5: Average subjective measures for MAMI and standard browser in the large database.

3.3.2 *Small Database*

Our results with the small database reflected weaker interactions between the prototype used and each of the measures.

1. Efficacy of Search. The test revealed a *significant* effect of the availability of the MAMI interface on the task completion time. Users retrieving the images with the MAMI interface completed the task *slower* than users using the classical interface (0.95 CL, $p < .001$). The median [quartiles] completion time for subjects without and with the MAMI interface respectively were $167[131,199]$ and $264[223,293]$ seconds.

However, there was no significant effect of the prototype used on the perceived efficacy or speed of search. The median [quartiles] values for the perceived efficacy without and with the MAMI interface were $3[3,4]$ and $3[4,4]$, respectively.

2. Ease-of-use and Desirability. Ease-of-use and desirability were determined by the post-use questionnaire, via a 5-point Likert scale. We did not obtain significant differences in the perceived ease-of-use and desirability measures without or with the MAMI interface. In the case of ease-of-use, the median [quartiles] values for subjects without and with the MAMI interface respectively were $3[4,4]$ and $4[4,4]$. Finally, in the case of desirability, the median [quartiles] values for subjects without and with the MAMI interface respectively were $2[2,3]$ and $3[4,4]$.

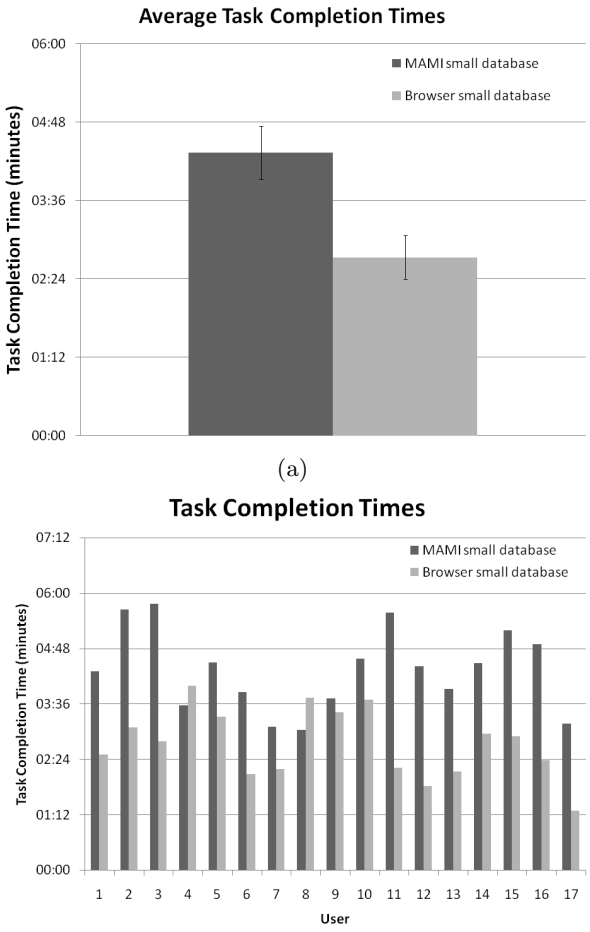


Figure 6: (a) Task completion times for all users with the small database. (b) Average task completion times with the small database.

Figure 7 summarizes the results of the subjective measures.

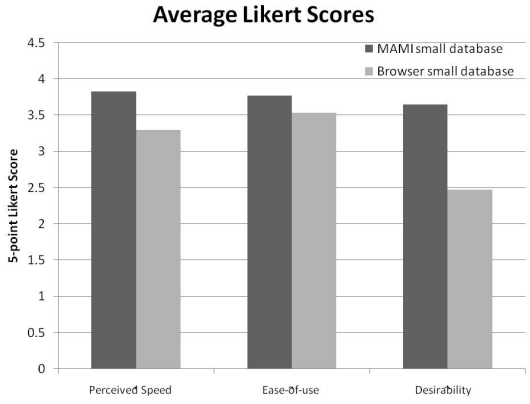


Figure 7: Average subjective measures for MAMI and standard browser in the small database.

3.3.3 Analysis of user logs

As described above, participants had to log in before starting the study. During the study, we logged all searches and system results for the MAMI prototype. Each log entry contained: a timestamp, the annotation corresponding to the desired picture, the pictures that were retrieved and their score. In addition to the three measures previously described (efficacy, ease of use and desirability), we carried out an exhaustive study of the participants' logs.

In particular, we were interested in analyzing MAMI's error rate as measured by the % of time that MAMI failed to retrieve the desired picture in the four-best pictures. We were also interested in understanding the impact that the size of the database had on MAMI's failure rate.

The percentage of pictures not found when performing the search using the MAMI prototype is shown in Table 1. The first row on the Table corresponds to the small database, while the second row corresponds to the large database. For each of the two databases, the first three columns depict the percentage of time that the user made *exactly* one, two or three errors. The last column contains the overall aggregate percentage of making any error using the system. Finally, the number of erroneous pictures appears in parenthesis.

The words that accumulated the most errors for all users were *estatua*, *padre*, *playa* and *acueducto*. In the case of *padre*, it was mostly confused with *madre*. We are still analyzing the potential reasons for failure for the rest of the words.

Database	1 error	2 errors	3 errors	overall
Small	7.8% (1.2)	4.7% (0.7)	8.6% (1.3)	21.2%
Large	4.7% (0.7)	2.7% (0.4)	5.5% (0.8)	12.9%

Table 1: Error rate in finding pictures using the MAMI prototype.

Another interesting metric extracted from the log files is the impact of presenting the user with an N-best choice of possible matching results and the effect of the value of N. For all successful cases (*i.e.* the desired picture appears in one or more of the four-best results), 77.3% of results returned it as the 1st-best, 89.96% returned it within the 2-best and 95.53% within the 3-best. Note how the 2-best case has already a good coverage of correct cases. As shown in the user study, standard browsing interfaces seem to be appropriate (fast and easy to use) in the case of a small number of pictures to search from. Therefore, presenting the user with the 4-best pictures allows for an increased accuracy with almost no usability penalty. In future work we plan to explore user interfaces with larger N.

4. DISCUSSION, CONCLUSIONS AND FUTURE WORK

As users capture and store an increasing number of images (and videos) in their mobile phones, there is a need for fast and easy-to-use multimedia search interfaces. Traditional desktop methods such as browsing and/or folder classification are not necessarily appropriate in the context of a mobile phone.

In this paper, we have presented our experience with MAMI, a speech-based mobile phone prototype for capturing, annotating, indexing and searching photos. We have described the results of a user study that aimed at comparing MAMI with the standard image browser available on today's phones. In particular, we were interested in measuring the efficacy, ease-of-use and desirability of each of the interfaces in the context of a search task. We were also interested in understanding the effect that the number of pictures in the image database had in each of these three measures.

The results of our study are somewhat intuitive: when the user's database had a large number of images (235 in our case), we found a *very significant positive effect* of the use of the MAMI interface in each of the measures. Users were and felt that they were *faster* finding the desired pictures with MAMI than with the standard browser. In addition, users found MAMI's interface to be easier-to-use and more desirable than the standard browser.

However, when the image database had a small number of pictures (47 in our case), users were faster with the standard interface than with MAMI. Interestingly, users gave higher scores in all of the subjective measures to MAMI than to the standard browser despite being slower with MAMI. This mismatch between perceived speed and actual speed could be due by differences in the way users interact with each interface: In the MAMI prototype, the user simply provides a speech input and most of the retrieval time is caused by MAMI's audio processing and search algorithms. However, when using the standard browser the main delay comes from the user struggling to find the desired pictures. Therefore, the user needs to spend *longer* time *actively* looking for the pictures with the standard interface than with MAMI.

Some areas for future work include scaling the system for fast searching in very large pictures databases and optimizing the processing and search algorithms for smaller delays. We also plan on adding image-based features and other contextual information to improve the search results and allowing multiple-word annotations and search queries. Finally, we intend to carry out a field study with a large number of users during an extended period of time and also a comparison of the voice tagging system with a text-based system to get an insight on the added value of using voice instead of text.

5. 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] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] X. Anguera and N. Oliver. MAMI: Multimodal Annotations on a Camera Phone. In *Proc. of MobileHCI*, 2008.

[10] H. Sakoe and S. Chiba. Dynamic Programming Algorithm optimization for Spoken Word Recognition. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, pp. 43-49, 1978.

[11] C. Gurrin, G.J.F. Jones, H. Lee, N. O'Hare, A.F. Smeaton, N. Murphy Mobile Access to Personal Digital Photograph Archives. In *Proc. of MobileHCI*, 2005.

[12] S. Ahern, S. King, M. Naaman, R. Nair, and J.H. Yang. ZoneTag: Rich, Community-supported Context-Aware Media Capture and Annotation. In *Mobile Spatial Interaction workshop (MSI) at the SIGCHI conference on Human Factors in computing systems (CHI 2007)*, 2007.

[13] A. Hwang, S. Ahern, S. King, M. Naaman, R. Nair, and J. Yang. Zurfer: Mobile Multimedia Access in Spatial, Social and Topical Context. In *proceedings, Fifteenth ACM International Conference on Multimedia, (ACM MM 07)*, 2007.

[14] F. Wilcoxon. Individual Comparisons by Ranking Methods. Biometrics Bulletin, Vol. 1, No. 6 (Dec., 1945), pp. 80-83 Published by: International Biometric Society