Multimodal Photo Annotation and Retrieval on a Mobile Phone

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

JieJun Xu*
Vision Research Lab
University of California at
Santa Barbara
CA 93106, USA
jiejun@cs.ucsb.edu

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

ABSTRACT

Mobile phones are becoming multimedia devices. It is common to observe users capturing photos and videos on their mobile phones on a regular basis. As the amount of digital multimedia content expands, it becomes increasingly difficult to find specific images in the device. In this paper, we present a multimodal and mobile image retrieval prototype named MAMI (Multimodal Automatic Mobile Indexing). It allows users to annotate, index and search for digital photos on their phones via speech or image input. Speech annotations can be added at the time of capturing photos or at a later time. Additional metadata such as location, user identification, date and time of capture is stored in the phone automatically. A key advantage of MAMI is that it is implemented as a stand-alone application which runs in real-time on the phone. Therefore, users can search for photos in their personal archives without the need of connectivity to a server. In this paper, we compare multimodal and monomodal approaches for image retrieval and we propose a novel algorithm named the Multimodal Redundancy Reduction (MR2) Algorithm. In addition to describing in detail the proposed approaches, we present our experimental results and compare the retrieval accuracy of monomodal versus multimodal algorithms.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*Signal processing*; I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms

Algorithms, Measurement, Experimentation

*This work was performed when the author was visiting Telefónica Research at Barcelona

MIR'08, October 30–31, 2008, Vancouver, British Columbia, Canada.

Keywords

Multimodal Indexing, Mobile Search and Retrieval, Digital Image Management

1. INTRODUCTION

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 also becomes increasingly difficult for a user to find specific images in the device. In order to tackle this problem, we have proposed a multimodal image annotation, indexing and retrieval prototype named MAMI [1, 2]. 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 or image input without the need of connectivity to a server.

Desired content is typically searched for by providing a monomodal input query (text, audio or image). Therefore, retrieval algorithms do not usually take full advantage of the multimodal information stored with the images. In this paper, we study the impact on image search and retrieval of multimodal fusion techniques. Moreover, we present a novel multimodal redundancy reduction algorithm where a monomodal input query makes use of multimodal information in order to return more relevant results to the user. In addition to describing in detail the proposed approaches, we present our experimental results and compare the retrieval accuracy of monomodal versus multimodal algorithms.

In the area of image multimodal tagging and retrieval, there are several research projects that propose the mobile phone as an interface for tagging and/or retrieval of the user's multimedia data [3, 4, 5, 6, 15]. We shall highlight three pieces of work that are particularly related to ours. Wang *et al.* [6] propose a multimodal (spatial, social and topical) mobile phone-based system that allows users to browse through their digital library as well as their social networks and surroundings. Their mobile application acts as a browser with a connection to a server. In addition, they correlate pictures according to multiple modalities, but the desired images are still searched for according to a single modality. Xie *et al.* [15] provide a nice description of how a mobile phone with a server-based architecture can perform both image and audio search queries. They present several

application examples for both image and audio separately. However, their work does not exploit multiple modalities to enhance the results and proposes a server-based architecture instead of performing such tasks locally on the phone, as is the case of the system presented in this paper.

The most relevant project to ours is probably Fan's *et al.* [3] system, where mobile web search is enhanced via multimodal input. Their system sends the image taken by the camera phone to find images on the web that are similar to it, *e.g.* from a product the user wants to get information for, etc.. Users can refine their search by providing input text queries. Multimodality is used in the web search refinement step, in order to increase the relevance of the results. They do not use speech input.

The system presented in this paper (MAMI) exploits the availability of multiple modalities (audio and visual) to increase retrieval accuracy of the user's personal image database. In addition, MAMI's processing is carried out entirely on the mobile phone, without the need of a remote server.

This paper is structured as follows: First, Section 2 presents the multimodal processing algorithms and Section 3 describes in detail the proposed multimodal retrieval algorithms. Our experiments are presented in Section 4. Finally, Section 5 summarizes our findings and outlines our future work.

2. MULTIMODAL PROCESSING ON A CAMERA PHONE

In this paper, we focus on the impact that having multiple modalities has in annotating, searching for and retrieving personal pictures that are stored on the user's mobile phone. We have developed a mobile phone-based prototype named MAMI (Multimodal Automatic Mobile Indexing) [1, 2] that allows users to easily perform these tasks in their mobile phones. When the user takes a picture with MAMI, (s)he can add an acoustic tag at the time of capture. This audio tag is associated with the image and it is indexed in a local database. Upon indexing, the system computes and stores acoustic and image feature vectors (descriptors) and adds additional metadata information such as location, date and time of capture and user ID.

At a later time, when the user desires to search for and retrieve a specific image, (s)he can query the system via a speech query or a sample image. In both cases, the query is processed by MAMI to compute the query's descriptors. These descriptors are compared to all other descriptors in the local database and MAMI retrieves the 4 pictures (4-best) that best match –*i.e.* whose descriptors are the closest to– the user's query.

In addition, the user can click on any of the retrieved images and query the MAMI prototype to show images similar to the chosen one, according to a variety of dimensions: location, time, visual and acoustic tag similarity.

Note that all the processing in the MAMI prototype is done locally on the mobile phone. This constitutes one of the advantages of this prototype over other systems, as it does not depend on network availability or server access constraints. With the MAMI prototype, all indexing can be done at the time of capture and therefore no information is forgotten by the user. In addition, MAMI's audio processing is speech independent and therefore highly suitable to handle proper names that typically do not exist in standard dictionaries. Local processing, however, poses several challenges, particularly in terms of the mobile phone processing capability. The MAMI prototype overcomes these limitations by using optimized and effective algorithms, both for the image and audio processing modalities.

We direct the reader to our previous work [1, 2] for a detailed description of MAMI's interface and audio processing. The focus of this paper is the addition of image processing and multimodal image retrieval. In particular, we propose the *Multimodal Redundancy Reduction Algorithm* (MR2) to select the most representative pictures of the contexts that the user might be looking for. The MR2 algorithm is designed to avoid retrieving unnecessary repetitions of pictures from the same context. As shown in Section 4, this algorithm boosts the system's accuracy in returning the right picture(s), *i.e.* the picture(s) that the user was interested in. A more detailed explanation of the MR2 algorithm is given in Section 3. In the next subsections, we present an overview of the acoustic and image feature extraction algorithms used in the MAMI prototype, as well as the distance metrics proposed in the audio and image spaces.

2.1 Acoustic Processing

Audio input is used in the MAMI prototype either for tagging or searching. In both cases, the audio recording is carried out via a push-to-talk method. The audio tags are stored in disk as .wav files. A typical audio tag contains a variable amount of silence at both ends, together with some click and background noises, depending on how hard the user clicks the stop-start buttons. These silence and noise segments are filtered out using a simple energy-based speech/non-speech detector with a variable threshold to accommodate different background conditions. Then, 10 Mel Frequency Cepstral Coefficients (MFCC) [9] are extracted every $20ms$, including CMN (Cepstral Mean Normalization) and excluding the C0 component. This choice of acoustic features was designed to optimize the discriminative power of the audio descriptor, while keeping the feature extraction as fast as possible. The obtained feature vectors are stored in memory for later use.

In order to compare two audio tags, the Dynamic Time Warping (DTW) algorithm is applied to their acoustic feature vectors, using the Euclidean distance between individual frames. The choice of DTW was driven by its efficiency and accuracy in pattern matching repetitions of the same utterance by the same speaker. In our experience [1], this feature representation and distance metric has been robust to typical background noises, such as urban outdoor noises.

When the user looks for a specific image via an audio tag, the MAMI prototype performs a DTW comparison between the input query's feature vector and all stored audio feature vectors. In order to speedup this computation, we constrain the amount of time warping to twice the tag's length. The final result of this comparison is the distance, normalized by the overall number of frames.

2.2 Image Processing

Edge-derived features have traditionally been an important and computationally light-weight approach to characterize image content. MAMI's image processing module uses the Edge Histogram Descriptor (EHD) of the MPEG-7 Visual Standard for measuring similarity between images. The EHD is designed to capture the spatial distribution of edges in an image by computing a histogram that represents the

frequency and the directionality of the brightness changes in the image [13].

To extract the EHD, a given image is first sub-divided into 4×4 sub-images. Each sub-image is further divided into non-overlapping image blocks, which are the basic units for edge extraction. The number of image blocks is typically fixed (*e.g.* 1100). Therefore, the block size depends on the resolution of the image. Each of the image blocks is then classified into one of the five types of edges, namely: vertical, horizontal, 45-degree diagonal, 135-degree diagonal or non-directional. This classification is performed by applying edge detectors with the corresponding directions on the image block and selecting the one with the strongest (highest in value) response. If the response is above a certain preset threshold (*e.g.* 11), the block is classified as an edge of the respective orientation. Otherwise, it is classified as a non-edge block [8, 13]. All the blocks in a sub-image are classified and a 5-bin histogram is constructed to characterize the distribution of edges in the sub-image. The end of this process yields an edge histogram with a total of 80 (16×5) bins, since there are 16 (4×4) sub-images. This 80 dimensional histogram constitutes the image feature vector that is stored with the rest of the image metadata for later use.

In order to compare two images, the Euclidean distance is applied to their image feature vectors, such that the smaller the distance, the more similar the images are. When the user looks for a specific image via an image sample, the MAMI prototype computes the Euclidean distance between the input image's visual feature vector and all stored image feature vectors.

3. MULTIMODAL INFORMATION FUSION

The availability of multimodal metadata allows for the formulation of image retrieval algorithms that combine this information intelligently. In this paper, we focus on the combination of image and audio features in the context of an image search task.

The first subsection presents the three fusion techniques that were tested in our experiments. The second subsection proposes a new algorithm named *Multimodal Redundancy Reduction Algorithm* (MR2) that takes advantage of multiple modalities to increase the accuracy of the image search results.

3.1 Multimodal fusion techniques

There has been extensive previous work in the area of multimodal information fusion for multimedia data search [7, 14]. Typically, multimodal fusion algorithms carry out the fusion at the feature or at the decision levels. When fusing at the feature level, a large dimensional feature vector is created with the features from all modalities, followed by PCA for dimensionality reduction and/or ICA for identifying statistically independent sources. Finally, a classifier is applied.

Fusion techniques at the decision level, however, process the data from each modality independently and apply a fusion algorithm when making the classification decision. Some examples include product combination [11], weighted-sum [12], voting and min-max aggregation [7].

In this paper, we have considered fusion techniques at the decision level as they allow the use of different algorithms for characterizing and comparing the feature vectors in each of the modalities. In particular, we have opted for the weighted-sum (linear combination) approach, due to its simplicity and reasonable level of tolerance to noise in the input data. We shall describe next three alternative implementations of the weighted-sum that use different weights in the linear combination.

The first fusion alternative (FUSION1) explored is shown in equation 1

$$D_f(i,j) = W\frac{D_{dtw}(i,j)}{max(D_{dtw}(i,\cdot))} + (1-W)\frac{D_{ehd}(i,j)}{max(D_{ehd}(i,\cdot))} \quad (1)$$

where $D_f(i,j)$ is the fused distance between images i and j, $D_{dtw}(i,j)$ is the DTW distance between the acoustic tags of images i and j, $D_{ehd}(i,j)$ is the visual distance between images i and j, W is the weight and $max(D_{dtw}(i,\cdot))$, $max(D_{ehd}(i,\cdot))$ are the maximum values of the pairwise distances between image i and all other pictures in the acoustic and image spaces.

The value of weight W is determined empirically from real data. Figure 1 shows the average accuracy of this algorithm in our multimodal image dataset[1] when varying the weight W from 0 (only using image information) to 1 (only using acoustic information). The optimum value is found at $W = 0.7$.

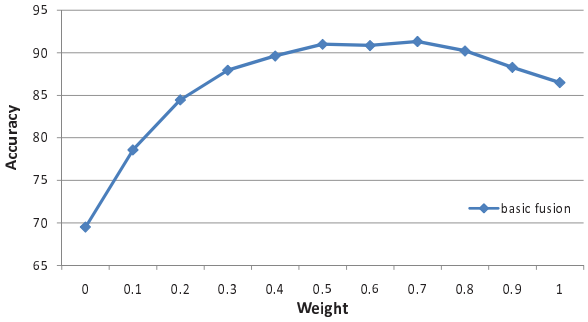

Figure 1: Selection of the weight W in the first fusion algorithm.

The second fusion alternative (FUSION2) does not use the weight W, but normalizes each modality by its mean μ and standard deviation σ, as shown in Equation 2:

$$D_f(i,j) = \frac{D_{dtw}(i,j) - \mu_{D_{dtw}(i,\cdot)}}{\sigma_{D_{dtw}(i,\cdot)}} + \frac{D_{ehd}(i,j) - \mu_{D_{ehd}(i,\cdot)}}{\sigma_{D_{ehd}(i,\cdot)}} \quad (2)$$

Equation 2 makes sure that the probability density function (pdf) –assumed to be a normal distribution– of each modality's output values overlaps around 0 and is scaled to have unit variance.

Finally, the third fusion alternative (FUSION3) adds an extra weight to each modality (W_{dtw} and W_{ehd}) to Equation 2. These weights are designed to compensate for the likely non-Gaussianity of the pdf's both in the image and audio metric spaces. Given $D_{dtw}(i,.)$ and $D_{ehd}(i,.)$, we first compute the total dynamic range of each of the distances. Note that outliers (if any) have been previously filtered out. Next,

[1]See section 4 for a detailed description of the dataset.

we compute the number of times that the distances in each modality fall below the 50% of such dynamic range [10]. The weights W_{dtw} and W_{ehd} are computed as the ratio of these values between the two modalities.

3.2 Multimodal Redundancy Reduction Algorithm

The multimodal annotation capability of the MAMI prototype augments the user personal images with additional metadata: audio and image feature vectors, date and time, location and user identification. However, a typical picture search task will only include a monomodal input, either by means of an acoustic query (*i.e.*the user provides the audio tag associated with the desired picture) or an image query (*i.e.* the user provides an exemplary picture that is similar to the desired image).

In this Section, we present a *Multimodal Redundancy Reduction* Algorithm (MR2) that takes advantage of the multimodal metadata in search mode to improve the accuracy of the search.

The MR2 algorithm attempts to maximize the probability that a desired picture will appear in the N-best pictures which are retrieved by the MAMI prototype and consequently shown to the user. To do so, the MR2 algorithm exploits the information contained in the modality *not used* in the input query (*e.g.* image for audio queries and viceversa) to avoid showing *redundant* images to the user, *i.e.* images that were taken in the same context and display similar content.

Note that redundant images are very common in personal image databases, because users typically take more than one picture in the same scene (and context) to ensure that the desired content has been captured. Also note that once the user has identified the desired image, (s)he should be able to select it and retrieve *related* images from a variety of perspectives, including visual, acoustic, temporal and geographical.

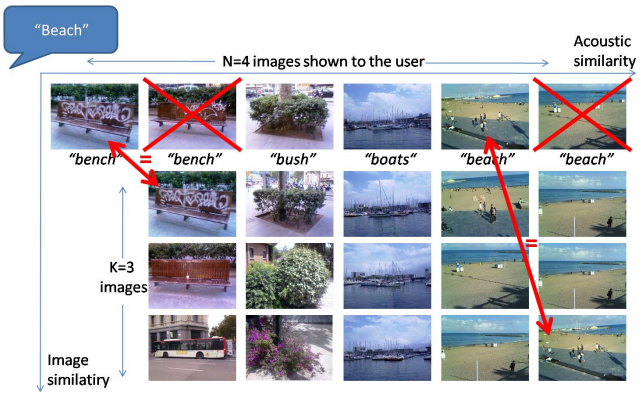

Figure 2: Multimodal Disambiguation algorithm.

Figure 2 illustrates the MR2 algorithm via an example of an audio query searching for pictures containing the acoustic tag: *"beach"*. The top row in the Figure, shows the 6-best results, according to the acoustic distances between audio feature vectors, as described in Section 2.1. Underneath each of the 6-best images, we show their associated audio tag for reference. From these results, the MAMI prototype would show the user the best $N = 4$ pictures as depicted in Figures 3(a) and (b).

In this example, the tags *bench*, *bush* and *boats* appear in the top 4 results, due to ambiguities in the audio processing. In the case of monomodal query processing, the MAMI prototype would retrieve and present to the user the $N = 4$-top images, *i.e.* several pictures from the same context and none from the desired topic, as seen in Fig. 3(a). The MR2 algorithm, however, considers the *image* similarity between the retrieved pictures, marking as redundant –*i.e.* belonging to the same context– the 2nd and 6th best images. This is done by computing for each picture in the N-best list (except for the first one) its K-best ($K = 3$ in this example) pictures according to the non-queried modality. Images in the N-best list are discarded if any of the pictures in their corresponding K-best list has been selected before. The MR2 algorithm will mark them as redundant. For example, the K-best list for the 2nd-best picture in the top row of the Figure –labelled with the tag *bench*– contains the 1st-best picture. Therefore, the 2nd-best image is considered redundant, as the system has already included a similar image in the list of images that will be shown to the user. As a result of running the MR2 algorithm in this example, the user would be presented with the set of 4-best pictures shown in Fig. 3(b). Note how, in the case of the MR2 algorithm, each image belongs to a different context and there is at least 1 image from the desired context.

Figure 3: Final 4-best images displayed to the input query *beach*: (a) Monomodal processing and (b) Multimodal Redundancy Reduction algorithm.

Finally, the MR2 algorithm is described in Alg. 1. As it will be shown in Section 4, this algorithm increases image retrieval accuracy when compared to monomodal methods. In addition, the computational complexity associated with the MR2 algorithm is insignificant, as all pair-wise multimodal distances can be performed off-line.

4. EXPERIMENTAL RESULTS

In this section, we describe the multimodal database and the experiments performed to test the impact of using multiple modalities when searching for pictures. We also evaluate the effectiveness of the MR2 algorithm proposed in this paper when compared to a monomodal query approach.

Algorithm 1 MR2: Multimodal Redundancy Reduction Algorithm

Input: a given query Q (either acoustic or image)
while $n \leq N$ **do**
  Retrieve the next closest image X_n to Q
  if $n = 1$ **then**
    Set X_1 as the first output.
  else
    Compute the closest K images ($Y_{m=1...K}$) to X_n given the modality *not used* in query Q
    if ($Y_{m=1...K}$) contains any images in $X_{1...n-1}$ **then**
      discard X_n
    else
      include X_n to the output and $n = n + 1$
    end if
  end if
end while
Output: list of N most relevant pictures shown on screen

4.1 Multimodal Database

In order to carry out the experiments, we collected a multimodal database. Six different participants (one female, five male, with ages ranging from 25 to 40) were asked to use the MAMI prototype for a few days. They were asked to take *at least* 2 exemplary pictures –with their corresponding audio annotations– of a set of categories or contexts, *e.g.* monument, street, building,car, beach, etc. The number of pictures taken by the users ranged between 52 and 212 pictures, with an average of 91. Figure 4 depicts a few exemplary pictures from the multimodal database, together with their associated audio tags.

Figure 4: Examples of pictures taken by the participants with their annotations.

4.2 Multimodal fusion experiments

The multimodal experiments were performed in a per speaker basis, given that the MAMI prototype is intended to be a personal system. For each picture in each of the participant's databases, we searched for the best (or N-best depending on the test performed) set of pictures (different from the one taken as query) with smallest distance to the input image, either in the image or acoustic space. A metric of accuracy is reported as the % of times that the correct picture –*i.e.* the picture comes from the same context as the desired one– was retrieved.

We performed a first set of experiments where we searched for the closest picture given a multimodal query (audio and visual features) as input. The distance was measured in each modality independently and then using the three previously described fusion algorithms.

User	image	audio	FUSION1	FUSION2	FUSION3
user1	81.1%	90.6%	**95.8%**	**95.8%**	92.5%
user2	84.9%	**98.8** %	**98.8%**	**98.8%**	**98.8%**
user3	51.9%	61.5%	**73.1%**	**73.1%**	69.2%
user4	85.5%	87.0%	92.8%	**94.2%**	89.9%
user5	46.9%	**96.9%**	**96.9%**	92.2%	**96.9%**
user6	66.7%	84.1%	**90.5%**	85.7%	84.1%
average	69.5%	86.5%	**91.3%**	90.0%	88.6%

Table 1: Retrieval accuracies for all users and retrieval algorithms for 1-best output.

Table 1 presents the retrieval accuracy results for the individual modalities and for the three fusion alternatives proposed. Results are given at the individual participant and average level and correspond to considering only the first returned result (1-best).

As seen on the Table and in the case of the individual modalities, audio always outperforms image processing. Among the fusion techniques, FUSION1 obtains the best average score, but has the burden of requiring recomputation of the linear combination weight W when there are changes in the database. FUSION2 and FUSION3 do not require any tuning. FUSION2 achieves slightly better performance than FUSION3 on average, but for some speakers its score falls below the best of the individual modalities. FUSION3 always equals or outperforms the best individual modality and therefore it is considered to be the most robust approach over all options.

In the current MAMI prototype implementation, the user receives back the 4-best images matching the query. The user will typically consider that the system behaved correctly if the image he/she was expecting is in the displayed set. Intuitively, it seems that the system's performance could be increased the more images it would show to the user. Figure 5 shows the average accuracy for each of the retrieval algorithms shown above as a factor of the number of images (N) returned as possible matches. As expected, the system's accuracy increases with N. In a standard mobile phone screen, $N = 9$ would probably be the maximum number of images that could be shown and still be easily recognized by the user.

As can be seen in the Figure, the system always behaves worse on average in the monomodal case than with any of the proposed fusion techniques. It is interesting to note that image-based retrieval increases performance much faster than any other approach, reaching over 93% accuracy at $N = 9$[2]. Among the three different fusion techniques, FUSION1 obtains the best results until $N = 5$, when FUSION2 –*i.e.* automatic weight selection via variance/mean normalization– becomes the best choice. It seems that all fusion techniques reach a stable accuracy point for $N > 5$.

4.3 Multimodal Redundandy Reduction

As previously explained, the MR2 algorithm has two design variables: (1) N, the number of pictures that will be

[2]For appropriate viewing we have trimmed the image accuracy values for $N < 3$

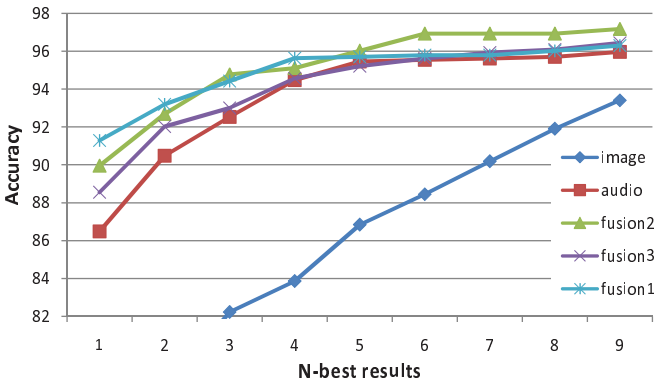

Figure 5: Accuracy for all evaluated techniques as a function of N-best output

shown to the user; and (2) K, the size of the list of *similar* pictures according to the complementary modality to that being used in the query.

In order to test the MR2 algorithm, we used the same multimodal database presented above.

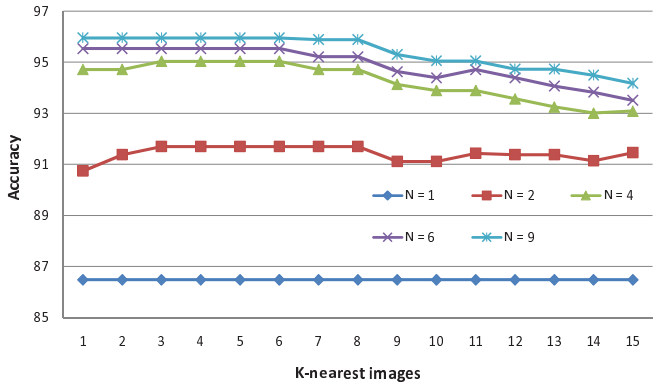

Figure 6: Audio query MR2 Algorithm as a function of N and number of redundant elements considered(K)

Figure 6 shows the accuracy as a function of K for several values of N (*i.e.* $1, 2, 4, 6$ and 9 images presented to the user) and in the case of an acoustic input query. By design, the algorithm never alters the first picture shown. Therefore $N = 1$ always gives the same result, which is equivalent to using only acoustic information in the system. For $N > 1$, accuracy increases as N increases. For $N = 4, 6, 9$, accuracy remains almost constant for $K < 7$ and starts decreasing afterwards. This is explained by the fact that the larger the number of images used for disambiguation, the higher the probability to include images that are not related to the input image. Therefore, the larger the K, the larger the number of images that would be excluded from the output list without being redundant. The optimum value for K depends on the size of the database and on the average number of redundant images that it contains. The database used in our experiments had 2 to 5 redundant images for each image, depending on the user and the context. Performance does not seem to suffer in the range $K\epsilon[1, 7)$.

If we consider the inverse case, *i.e.* an image query with audio used to eliminate redundancy, we obtain a similar plot and similar optimum values for K. Therefore, we set the MAMI prototype to work with $N = 4$ pictures shown to the user, and $K = 4$ pictures used in the redundancy elimination algorithm. Results for these settings in the two possible input query modalities are shown in table 2.

4.4 Overall Results

Next, we shall summarize the results of our experiments. Table 2 depicts the accuracy of the MAMI prototype when showing the 4-best pictures for each of the proposed techniques: monomodal, MR2 algorithm with $K = 4$ and the best result obtained with the fusion algorithms.

Query	MonoModal	MR2	Fusion(best)
audio	94.48 %	95.04%	95.63%
image	83.87 %	84.92%	

Table 2: Results for all Presented Techniques

The MR2 algorithm outperforms the monomodal approaches. Its point of departure is the monomodal query result, and reduces the redundancy of the results by using the complementary modality. On a per-user analysis, we observe that the impact of the MR2 algorithm is more significant as the monomodal approach is more prone to errors. In the case of audio queries, the worst performing user in our study obtained an audio monomodal accuracy of 88.4%, whereas his accuracy improved to 90.3% when using the MR2 algorithm. In the case of image queries, the worst performing user obtained accuracies of 75.0% and 78.1% for the monomodal and MR2 cases respectively. Finally, note that the multimodal fusion techniques obtain the best results when compared to any of the other approaches. However, they require a multimodal input query, consisting of an image and an audio tag.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a mobile, multimodal image annotation, indexing and retrieval prototype named MAMI. It allows users to annotate and search for digital photos on their camera phones via speech or image input. The focus of this paper has been the multimodal nature of MAMI. In particular, we have proposed and evaluated two multimodal approaches to image retrieval: First, we have compared the accuracy of three multimodal fusion algorithms. Secondly, we have described a novel algorithm for multimodal redundancy reduction (MR2 algorithm). In order to validate and compare the various approaches, we have created a multimodal image database with 546 pictures for 6 users.

Our experiments have shown that:

1. The multimodal fusion algorithms have higher accuracy than their monomodal counterparts.

2. The multimodal redundancy reduction (MR2) algorithm also has higher accuracy than the monomodal approaches on image retrieval. More importantly, the MR2 algorithm is able to improve accuracy in the

cases where the monomodal retrieval algorithm exhibits poor performance.

3. The MR2 algorithm augments monomodal queries with information from the complementary modality to the one used in the query. All the necessary information can be precomputed off-line. Therefore, the MR2 algorithm's computational cost is equivalent to that of monomodal approaches.

We believe that multimodal approaches to multimedia indexing and retrieval in mobile phones will play a crucial role in the years to come. Therefore, some of our lines of future work include:

1. The development of a client-server architecture in the MAMI prototype, such that users would be able to upload their mobile pictures and search for their digital content in the server with increased capabilities.

2. The implementation of a clustering approach in the image and audio spaces to increase the efficiency and scalability of the search algorithms.

3. The deployment of a user study to compare the proposed multimodal approach to traditional and text-based approaches to multimedia information retrieval.

4. The inclusion of other modalities for more accurate retrieval, *e.g.* location, time and date, social context, etc.

6. REFERENCES

[1] X. Anguera and N. Oliver. MAMI: Multimodal annotations on a mobile phone. In *Proceed. of Intl. Conf. on Mobile HCI (MobileHCI-08)*, 2008.

[2] X. Anguera, N. Oliver, and M. Cherubini. Multimodal and mobile personal image retrieval: A user study. In *Proceed. of SIGIR Workshop on Mobile Information Retrieval (MobIR'08)*, 2008.

[3] X. Fan, X. Xie, Z. Li, M. Li, and W.-Y. Ma. Photo-to-search: Using multimodal queries to search the web from mobile devices. In *MIR '05: Proceedings of the Multimedia Information Retrieval Workshop*, pages 143–150, Singapore, 2005.

[4] C. Gurrin, G. J. F. Jones, H. Lee, N. O'Hare, A. F. Smeaton, and N. Murphy. Mobile access to personal digital photograph archives. In *MobileHCI '05: Proc. of the 7th int. conf. on Human computer interaction with mobile devices & services*, pages 311–314, New York, NY, USA, 2005. ACM.

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

[6] A. Hwang, S. Ahern, S. King, M. Naaman, R. Nair, and J. Yang. Zurfer: mobile multimedia access in spatial, social and topical context. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 557–560, New York, NY, USA, 2007. ACM.

[7] J. Kittler, M. Hatef, and R. Duin. Combining classifiers. *Intl. Pattern Recognition*, pages 897–901, 1996.

[8] B. S. Manjunath, J. R. Ohm, V. V. Vinod, , and A. Yamada. Color and texture descriptors. *IEEE Trans. Circuits and Systems for Video Technology, Special Issue on MPEG-7*, 11(6):703–715, Jun 2001.

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

[10] E. Moxley, T. Mei, X.-S. Hua, W.-Y. Ma, and B. Manjunath. Automatic video annotation through search and mining. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, Jun 2008.

[11] D. Tax, M. van Breukelen, R. Duin, and J. Kittler. Combining multiple classifiers by averaging or by multiplying? *PR*, 33(9):1475–1485, September 2000.

[12] K. Ting and I. Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10:271–289, May 1999.

[13] C. S. Won, D. K. Park, and S.-J. Park. Efficient use of mpeg-7 edge histogram descriptor. *ETRI*, 24(1):23–30, Feb 2002.

[14] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith. Optimal multimodal fusion for multimedia data analysis. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 572–579, New York, NY, USA, 2004. ACM.

[15] X. Xie, L. Lu, M. Jia, H. Li, F. Seide, and W.-Y. Ma. Mobile search with multimodal queries. *Proceedings of the IEEE*, 96(4):589–601, April 2008.