An Improved Method for Image Retrieval Using Speech Annotation

Jiayi Chen[1], Tele Tan[2], Philippe Mulhem[3], Mohan Kankanhalli[4]
[1,2]Laboratories for Information Technology
21 Heng Mui Keng Terrace, Singapore 119613
[3]IPAL-CNRS
21 Heng Mui Keng Terrace, Singapore 119613
[4]School of Computing, National University of Singapore
10 Lower Kent Ridge Crescent, Singapore 119260
(0065) 6874 6777, (0065) 6874 2584, (0065) 6874 8212, (0065) 6874 6738
{jiayi, teletan, mulhem}@lit.org.sg, mohan@comp.nus.edu.sg

ABSTRACT

In this paper, we present a system for the image indexing and retrieval using speech annotations based on a pre-defined structured syntax. In addition to the introduction of N-best lists for index generation, a query expansion technique is explored to enhance the query terms and to improve retrieval effectiveness. By adding the most probable substitutions for the query terms, more relevant images are distinguished from the data collection. This approach is particularly helpful to deal with those less frequently used words, including out-of-vocabulary (OOV) words, which are very common for names of people and places. Experiments on a collection of 1,200 photos show that the retrieval effectiveness is increased considerably for segment of individual domain on ***People***, ***Location*** and ***Event***. With this method, the average value of precision versus recall over a combination of segments has improved significantly, from 50% to 72.4%.

Keywords

Speech recognition, speech signal processing, image retrieval, multimedia information retrieval.

1. INTRODUCTION

With the increasing popularity of digital cameras, more and more consumers are able to possess a large collection of still images in their computers. Consequently, the demand for a system to handle all these images effectively and efficiently has been emerging. Content-based image retrieval (CBIR) is one of the conventional methods that address the task of image indexing through low level features like color, texture and shapes and retrieving desired images from their content attributes [2, 9, 15].

Most image indexing and retrieval systems have been developed to deal with specific user requirements. For example, the QBIC [2] and Photobook [9] systems were developed to help user exploit the large image content from the web, while VisualSEEK [15] allows the user arrange queries according to color regions. In the medical image retrieval system developed by Kak et al [12], domain knowledge about the structure of the anomalies is used to design appropriate detector. Likewise, the Object Probe [20] calls for a tuned feature set to classify home photographs into a few classes of objects. Recently, people have also begun to investigate the semantics information in image features [3].

Nevertheless, compared to the low-level content-based properties, most home users of digital cameras prefer a more direct and intuitive description of the photographs, such as the people/objects in the photo, location where the photo was taken, event related to the photo-taking, date/time of the photo or other pertinent information, to help them recall old photos. Those high level characteristics are not influenced by the orientation or posture of a person, are not sensitive to his/her hairstyle or glasses, neither are they changed by the photo's nature of close-up or long shot, which are universally great concerns for image processing or face recognition technology. However, most importantly, these advantages satisfy the users' demand for a robust and consistent mode to describe photos.

Traditionally users have to type such high level content manually and repeatedly to annotate their photos. However, with the availability of a built-in microphone in most digital cameras, nowadays users can very naturally talk about their photos on the spot and record these annotations into computer readable audio files. Advance in automatic speech recognition (ASR) technologies have already made it possible to automatically transcribe speech files with reasonable accuracy. Therefore transcriptions of those annotations by ASR software provide an alternative and reliable approach to complement existing methods of image indexing and retrieval and replace the tedious work of manual typing as well.

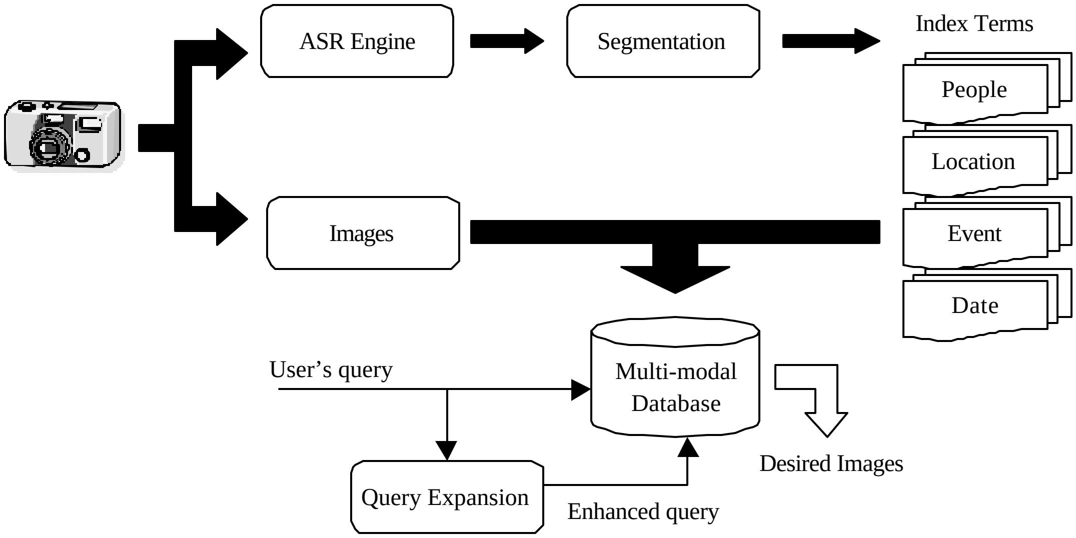


Figure 1. System Architecture

There have been numerous works on automatic speech annotation of digital spatial media like images and video. The work by Stent and Loui [18] uses free-style annotation to automatically index and organize a collection of photographs. They propose an event segmentation scheme that is applied to group photographs into unique events. Lienhart [5, 6] gave a brief introduction of an on-the-fly annotation language for the automatic acquisition of video abstracts and sets up two microphones to differentiate the environmental audio with descriptive annotations. Nack [8] describes a simulation of a handheld device for the annotation of simple semantic information. Show&Tell [16, 17] uses speech annotations to index and retrieve both personal and medical images, while FotoFile [4] extends this to more general multimedia objects.

In contrast with existing techniques that mainly addressed solutions from the standpoints of input speech styles and applications, our study intends to integrate speech recognition and information retrieval technologies together to ease the user's job of managing his home made photo collections. Meanwhile, a small subset of possible recognition results produced from the speech-to-text transcription, also known as the N-best list, is employed in our system instead of only one hypothesis to establish a probabilistic model.

Prior to this work, we have proposed the use of structural speech syntax to annotate photographs into four index fields; e.g. event, location, people and time/date [1]. A novel application called ***SmartAlbum*** has been developed to make use of mixed modality signals [19]. Based on the previous achievements, in this paper we come up with a solution to overcome the recognition errors and boost the retrieval effectiveness further.

ASR system accuracy plays a crucial role in the overall system performance. When speech recognition is of reasonable quality, the retrieval effectiveness from the speech-based index can be quite comparable with that obtained from the manually generated index. However, it is very common to encounter recognition errors even with the best ASR systems, especially when there are severe background noise, adverse recording conditions and disfluent or non-native speech. As a result, the transcription of annotation may not contain all words that were actually spoken, or may contain those that were not spoken. When index terms produced by these poor ASR transcriptions are used for a query, the retrieval effectiveness will be definitely degraded.

To deal with this problem, we introduce a query expansion technique based on a deeper exploration of the recognition process. By enhancing the original query terms with its most probable substitution errors deduced from the N-best lists, the system performance acquires substantial improvement.

The rest of this paper is organized as follows. Section 2 gives an overview of the system and the syntax of speech annotation. Section 3 explains in details the utilization of N-best lists for the task of indexing and retrieval. The query expansion technique is introduced in Section 4. Section 5 depicts our experiment settings and results, while Section 6 presents a further discussion and analysis. Finally, Section 7 concludes this paper.

2. SYSTEM OVERVIEW

2.1 System Architecture

The system architecture is shown in Figure 1. Upon downloading the images along with their speech annotations from the digital camera to a regular PC, a multimedia separator splits the two modes (image and speech) and stores them as different files. The image files can be transferred for further processing using content-based techniques or directly sent to the database, while the speech files are transcribed by the ASR engine to its textual form. A certain number of N-best lists are also kept as the by-product of ASR process. The transcription of a whole annotation sentence, along with its alternative lists, is then segmented into different parts according to a pre-defined syntactic structure and time stamps provided by the recognition process. Index terms are generated for each part and stored in the database with the corresponding images. The speech structure and segmentation process will be detailed in the following subsections.

With this indexing scheme, a user is able to look for photos of a specific person, at a specific place, taken at a special event or during a certain period of time. A typical query can be like "Give me photos of Tom on vacation in Beijing". However, the image retrieval process is highly dependent on the speech recognition accuracy. If a term of considerable significance to image descriptions is missed in the indexes due to misrecognition, those truly relevant images will fail to respond to the query. Therefore, a query expansion technique that utilizes word- and phrase- based alternatives is introduced. Those enhanced query terms, which are also obtained from the recognition results, reveal the general behavioral patterns of ASR engine and compensate for the influence of recognition errors on the retrieval effectiveness. More relevant photos will be returned at higher ranks when using the original and expanded query terms together.

2.2 Structured Speech Syntax

Structured speech has been used extensively in many speech-activated devices like the cell phone and handheld devices. The high recognition accuracy of these implementations is assured by restricting the dictionary of commands and words. Similarly, since most home users of digital cameras are mainly concerned with people in the photo, when, where or why it was taken, a short description that doesn't require a large vocabulary set will suffice. Such descriptions can contain ordered and related information, especially when a structured syntax is introduced for this purpose. Hence we proposed four index fields, i.e., ***People***, ***Event***, ***Location*** and ***Taken_on*** (Instead of the familiar label ***Date/Time***, the 3-syllable word ***Taken_on*** is used to better differentiate itself with all other words.) as constituents of the structure. The four field words act as leading tags. When making annotation, they are spoken out followed by a detailed description or a list of elements for this field. For example, the annotation of a photo taken in Beijing during the vacation can be made as follows.

> ***People*** David Tom
> ***Location*** Beijing
> ***Taken_on*** 18th April 1995
> ***Event*** China vacation

Although it sounds somewhat different from the normal mode of speech, it's the most effective and straightforward way to keep the annotation clean, concise and with little information loss. The advantages of syntactic annotation have been echoed by others [5, 6], whereby they described the video content of a camcorder through a cluster-based annotation structure. However, they did not provide implementation details about their techniques.

2.3 Speech Segmentation

Speech recognition errors are almost inevitable due to various environmental conditions and speaker's utterance. It has already been proven that a minor recognition error of a single word could possibly lead to disastrous degradation in later retrieval. Even when we are fortunate enough to get the accurate transcription, confusion between the content of each field would still make the retrieval work harder. Suppose there is an image with the following speech annotation:

> ***Event*** Melbourne Visit
> ***Location*** Angela's Home
> ***People*** Rebecca
> ***Taken_on*** 17th February 1996 at 11.43 a.m.

When we are going to find all images of Angela, the above image without Angela will be retrieved as well. This is certainly not the result we'd like to see.

Therefore, separating the whole transcription into individual segments will be greatly helpful to localize recognition results within one specific field and prevent those errors from affecting each other. According to the above mentioned syntax definition, an entire annotation sentence can be divided into 4 segments, with one segment corresponding to one field. Upon segmentation it is also feasible to perform different post-processing and indexing schemes on each segment. For example, word-stemming algorithms can be applied to ***Event*** and ***Location*** field, but not to ***People***. The indexing scheme of ***Taken_on*** field can be different from others. The segmentation also makes it possible to re-recognize each segment with more specific grammars.

Successful segmentation is ensured by the accurate localization of the leading tag of each field that indicates the beginning of next field and/or the ending of the previous one. These words are always present in each annotation and trained with particular emphasis to ensure their correct recognition. Two possible solutions, SAPI[1]-based and KWS (Keyword Spotting) -based, are proposed and compared on a subset (400 photos) of our database collection as a pre-requisite step.

The keyword spotting approach is based on the comparison of audio signal features with the reference pattern templates. Those tags are the keywords to be spotted. Considering the limited vocabulary in the system, classic dynamic time warping algorithms are adequate to provide an efficient way for pattern distance minimization problem with embedded time-normalization and alignment [10]. The feature vector comprises 12 Mel Frequency Cepstral Coefficients and their first derivatives. Mahalanobis distance (covariance weighted distance) is employed as the distance metric to measure the similarity between two patterns.

[1] Microsoft Speech Application Programming Interfaces, also known as Microsoft Speech SDK.

On the other hand, SAPI-based approach allows us take advantage of the time stamps provided by the low-level interfaces of the ASR engine (SAPI) for each word recognized. Time stamps are the indicators marking the time when the engine begins or finishes processing the utterance. After identification of time stamps of each tag, other words can be conveniently classified. This method directly deals with recognition results and outperforms the former substantially, as shown in Table 1. Therefore, the more straightforward SAPI-based mechanism is adopted for our subsequent implementation.

Table 2. Accuracy of tag detection using KWS- and SAPI-based techniques

	People	**Location**	**Taken_on**	**Event**
KWS-based	90.7%	68%	86%	71%
SAPI-based	100%	99%	100%	100%

3. INDEXING AND RETRIEVAL WITH N-BEST LISTS

As its name implies, N-best list represents a small subset of possible recognition results as the engine's next N-best guesses for a particular utterance. When a correct interpretation cannot be found in the top hypothesis, it can be frequently obtained from the N-best alternatives as long as the ASR engine has been reasonably trained. This feature is available with most commercial engines that compatible with MS-SAPI 4.0 including IBM ViaVoice and Dragon NaturallySpeaking. Figure 2 shows the typical recognition errors and their N-best lists. The typical value of N in our implantation is 6. A detailed discussion and experimental results have been reported in [1].

Based on different properties of each field, text processing like stopword removal and word stemming are applied to the transcriptions along with their N-best lists. Word-based index terms are then generated from the processed documents. Although in most IR systems words are scrambled and the content of documents is lost, we preserve the order of words within each segment through adding their time stamps to the database to facilitate query expansion and phrase-based retrieval.

According to [13], all recognition hypotheses above certain threshold are approximately equally capable of estimating the content of the spoken utterance. Suppose there are n_a speech annotations for the equal number of images. Each annotation utterance is separated into n_s segments (n_s = 4 in this study). For term t in segment s of annotation a, its term presence $\hat{i}_{a,s,t}$ and term count $\hat{c}_{a,s,t}$ can be estimated as

$$\hat{i}_{a,s,t} = \frac{1}{N}\sum_{n=1\ldots N} i^{(n)}_{a,s,t} \tag{3.1}$$

Original Speech Annotation:

People David Tom ***Location*** Beijing
Taken_on 18th April 1995 ***Event*** China vacation

Transcription (misrecognized words are underlined):

People dated town ***Location*** 18
Taken_on 18th April 1995 ***Event*** China vacation

N-best results for misrecognized words:

1-best	2-best	3-best	…	N-best
dated	slated	**David**	-	
town	**Tom**	from	-	
18	taking	**Beijing**	-	

Figure 2: A simple representation of N-best lists

$$\hat{c}_{a,s,t} = \frac{1}{N}\sum_{n=1\ldots N} c^{(n)}_{a,s,t} \tag{3.2}$$

where a varies from 1 to n_a, s from 1 to n_s, t from 1 to n_t, N is the total number of alternatives exploited for each segment, n represents one of the N hypotheses. $c^{(n)}_{a,s,t}$ and Boolean-valued $i^{(n)}_{a,s,t}$ are the term count and term presence within each hypothesis n respectively.

Given this equal probability model, the expected document length over all N hypotheses is

$$\hat{l}_{a,s} = \frac{1}{N}\sum_{n=1\ldots N} l^{(n)}_{a,s} \tag{3.3}$$

with the document length for each hypothesis n, denoted as $l^{(n)}_{a,s}$, being obtained according to [13].

$$l^{(n)}_{a,s} = \left(\sum_t \left(c^{(n)}_{a,s,t}\right)^a\right)^{1/a} \tag{3.4}$$

where a is the constant exponent to compress its dynamic range and set to 3 in the experiments. Term weight of t can be measured as the inverse document frequency (IDF)

$$idf_{s,t} = \log_2\left(\frac{1}{n_a}\sum_a \hat{i}_{a,s,t}\right) \tag{3.5}$$

Meanwhile, terms in the query q are also classified according to the syntactic structure to enable relevance evaluation between the query and annotations on the segmental basis. A modified TF-IDF (Term Frequency – Inverse Document Frequency) formula is defined as

$$r\hat{e}l(q,a)=\prod_{s}\frac{1}{\hat{l}_{a,s}}\left(\sum_{t}b_{q,s,t}\hat{c}_{a,s,t}idf_{s,t}\right) \qquad (3.6)$$

Here $b_{q,s,t}$ stands for the count of term t in the s-th segment of query q.

As shown in equation (3.6), relevance values on individual segment are multiplied to give the overall evaluation instead of being summed up. By this means, only photos satisfying all queries in every segment are pulled out, while those that only meet the requirements well in some fields but have no relevance in others are not returned. This manner is in accordance with what a user would normally expect as the search results.

4. IMAGE RETRIEVAL WITH QUERY EXPANSION

Query expansion technique complements traditional information retrieval methods. Its main objective is to add words that are effective in distinguishing relevant documents from the data collection. Many approaches have been proposed to enhance the entire concept of the query through a thesaurus and other devices [7, 11, 14]. In our system, a simple concept hierarchy has been established to include some *a priori* knowledge of common sense. For example, in order to find all photos taken during USA holiday as an event query, the corresponding location query can be automatically expanded to include more specific phrases like “Disney world”, “Grand Canyon” and “Las Vegas”. Similarly, finding photos of New York City also means that images related to “the Statue of Liberty”, “Battery Park” and “Ellis Island” are desired.

However, the introduction of query concept alone is still not sufficient to tackle the problem of recognition errors. Because of the error-prone nature of recognition process, there would be no surprises if a word in the utterance is absent in the transcription, or a word in the transcription is not actually uttered. With our existing method, all images relevant to a query can be found if their annotations are indexed under these query terms either through correct recognition or in the form of N-best lists. But such ideal situation could be easily corrupted by frequent misrecognitions, which create wrong index terms for the annotation files. In this case, even the expansion of query concept is helpless to improve system performance for the lack of semantic clues in the corrupted transcriptions

For instance, the spelling and pronunciation style of word “Beijing” is quite different from the general recognition patterns inside a standard ASR system for English. Across all annotations that contain the utterance of this word, its corresponding textual form doesn't necessarily always appear in either the top-1 hypothesis or the N-best transcriptions as expected in Figure 2. In addition to the loanwords, there are also other proper nouns like names of people and places in the image annotations that are less frequently used in continuous speech recognition. Moreover, the annotations may probably involve words that are even not included in the engine's built-in vocabulary, which are also known as out-of-vocabulary (OOV) words. Although OOV words can be added to the vocabulary and then go through repeated training to enhance its speech model, we find that the engine is still reluctant to select these words as possible recognition choices. All of these words share the same difficulties in generating right index terms and thus need to be compensated by another query expansion technique for low recognition accuracy.

While N-best lists help us in looking for uttered words in more places and creating more accurate indexing terms, they can also reflect the word-word associations between different ranks. Three types of recognition mistakes, namely substitutions, deletions and insertions, have been widely accepted to evaluate the quality of ASR software. Among them substitution errors are the most valuable to indicate the engine's inherent recognition patterns. Under the concept of N-best lists, all the candidate interpretations in the list for the same utterance can be considered as a potential substitution error for the original word.

A closer study on the automatic transcriptions shows that when a word is misrecognized, its substitution errors tend to involve only a small number of words. Moreover, words in this set are likely to appear together within the same list. Due to the differences in environmental conditions and user’s articulation, the total number of occurrences of a popular recognition error may even exceed that of its truly uttered counterpart, across all annotation files.

For example, suppose a term *t* is spoken originally in 3 annotation files a_1, a_2, a_3 and only the best three hypotheses are kept, i.e. N = 3. The recognition results for each utterance of t in various contexts may look like

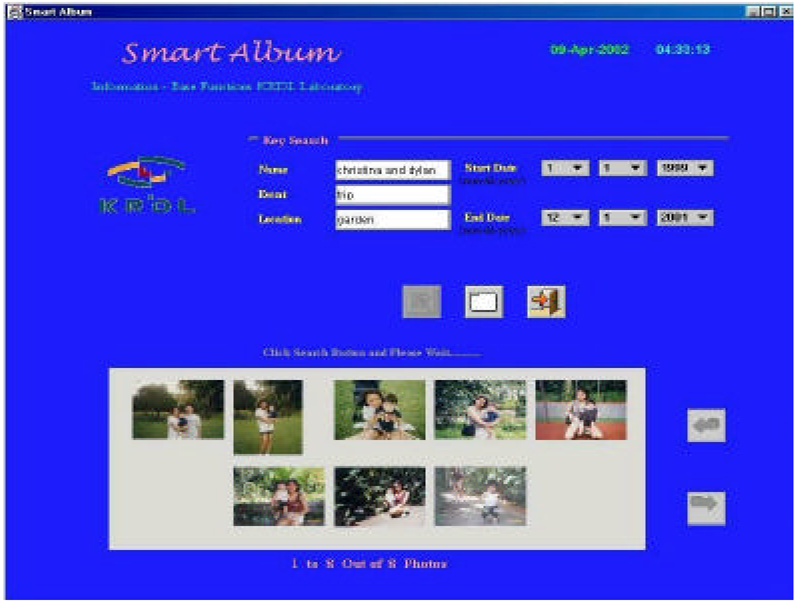
a_1:	**t**	t_1	t_2
a_2:	t_1	**t**	t_3
a_3:	t_1	t_2	t_4

The original term *t* appears at different ranks in a_1 and a_2, while it is absent in a_3. When *t* is input as the keyword for retrieval, only a_1 and a_2 will be found because there is no relevant index term in a_3.

However, among all the 4 possible alternatives for *t*, the most frequent cooccurrence of t_1 and *t* within the same list suggests that the former has a relatively high chance to be the substitution error of the latter. It is thus reasonable and wise to add t_1 as a supplementary query term to enhance the presence of *t.*

More generally, the conditional probability of an uttered term *t* being recognized as another term t_i, denoted as $P(t_i \mid t)$, can be computed when both of them are in the list.

$$P(t_i \mid t)=\frac{P(t_i,t)}{P(t)} \qquad (4.1)$$



Smart Album

09-Apr-2002 04:33:13

Information - Base Function KRDL Laboratory

Key Search

Name: christina and dylan

Event: trip

Location: garden

Start Date: 1 1 1999

End Date: 12 1 2001

Click Search Button and Please Wait.........

1 to 8 Out of 8 Photos

Figure 3. Query interface

Although the total list of alternatives for a single t could be quite long, many of them are merely picked up by chance and not qualified enough to represent the original term on a statistical basis. A practical solution is to keep only a limited number of alternatives with the substitution probabilities higher than a pre-set threshold. Let's define the number of alternatives for t as $M_{t.}$ After augmenting the query with additional M_t terms, the new relevance value between the query and the annotations can be calculated as:

$$r\hat{e}l_{new}(q,a) = \prod_s \frac{1}{\hat{l}_{a,s}} \left(\sum_t b_{q,s,t} c_{new} idf_{new} \right) \quad (4.2)$$

where the new values of term count and IDF are

$$c_{new} = \hat{c}_{a,s,t} + \sum_i \hat{c}_{a,s,t_i} P(t_i \mid t) \quad (4.3)$$

$$idf_{new} = idf_{s,t} + \sum_i idf_{s,t_i} P(t_i \mid t) \quad (4.4)$$

where i varies from 1 to M_t.

With the above equation, the presence of a query term in its relevant annotations can be predicted automatically even if it is not recognized in the N-best transcriptions. Besides, when a term t appears in the transcriptions, the more elements of T are there in its N-best list within one annotation, the more confident we can be of its actual utterance, and the higher the rank of the corresponding image will be in the final retrieval.

5. EXPERIMENTS

A collection of over 1,200 photos is used in our experiments. All of them are taken from 10 albums of a family spanning a period of more than 9 years. The database represents a typical home collection of photos such as holidays, birthday parties, picnics, park visits, wedding and graduation ceremonies, campus life, home, family type photos and others. Each photo is annotated in both textual and audio forms. Currently the audio data are collected under the lab environment by one female speaker with the length of each file less than 15 seconds. The voice annotations are transcribed by Dragon NaturallySpeaking[2] Preferred version 5, separated from the image parts, converted to the defined formats, and then segmented as mentioned in Section 2.3. The number of alternative lists explored, N, is set to 6 in the experiments.

According to our syntax definition, the content in ***Taken_on*** field is only concerned about numbers and names of months that can always get high accuracy by an engine of quality. Transcriptions of this field are then passed to a date parser, which outputs the date information in a unified format to facilitate date retrieval. For each of the other fields, the index terms are generated as described in Section 3.1.

Figure 3 illustrates how a user can submit keyword-based textual query through the ***SmartAlbum*** interface. The sample in this figure asks for photos of "Christina and Dylan" in the "garden" during a "trip" taken "from 01/01/1999 to 12/01/2001". 8 photos are returned from the system. By listening to the annotation of

[2] http://www.dragonsys.com/

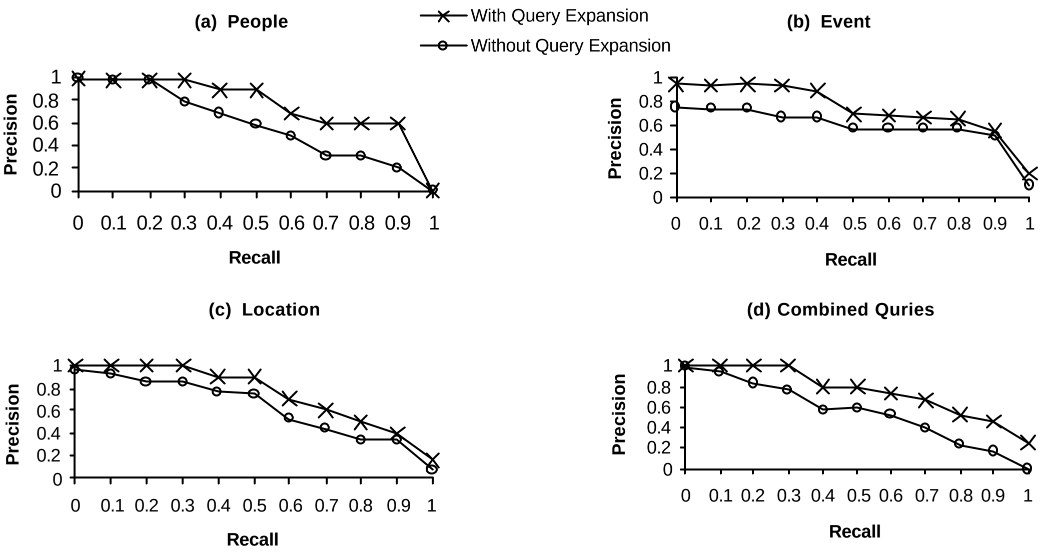


Figure 4. Precision vs. recall curve for 4 types of queries

each photo, we can confirm their relevance to the requirement on each field.

Being aware of the performance deterioration caused by recognition errors, especially when those special names of people and places are misrecognized as non-proper words, query expansion is applied as described in Section 3.2. The threshold of cooccurrence probability is set to 0.7 in order to ensure the reliability of extracted substitutions. In addition, the maximum number of expanded terms is also defined to be equal to the value of N, which is 6 in this case. Figure 4 shows the precision-recall curves for queries performed individually on each of the three fields as well as on a combination of all four fields. 10 queries are generated for each individual segment respectively and there are 15 inputs for a combination of different fields.

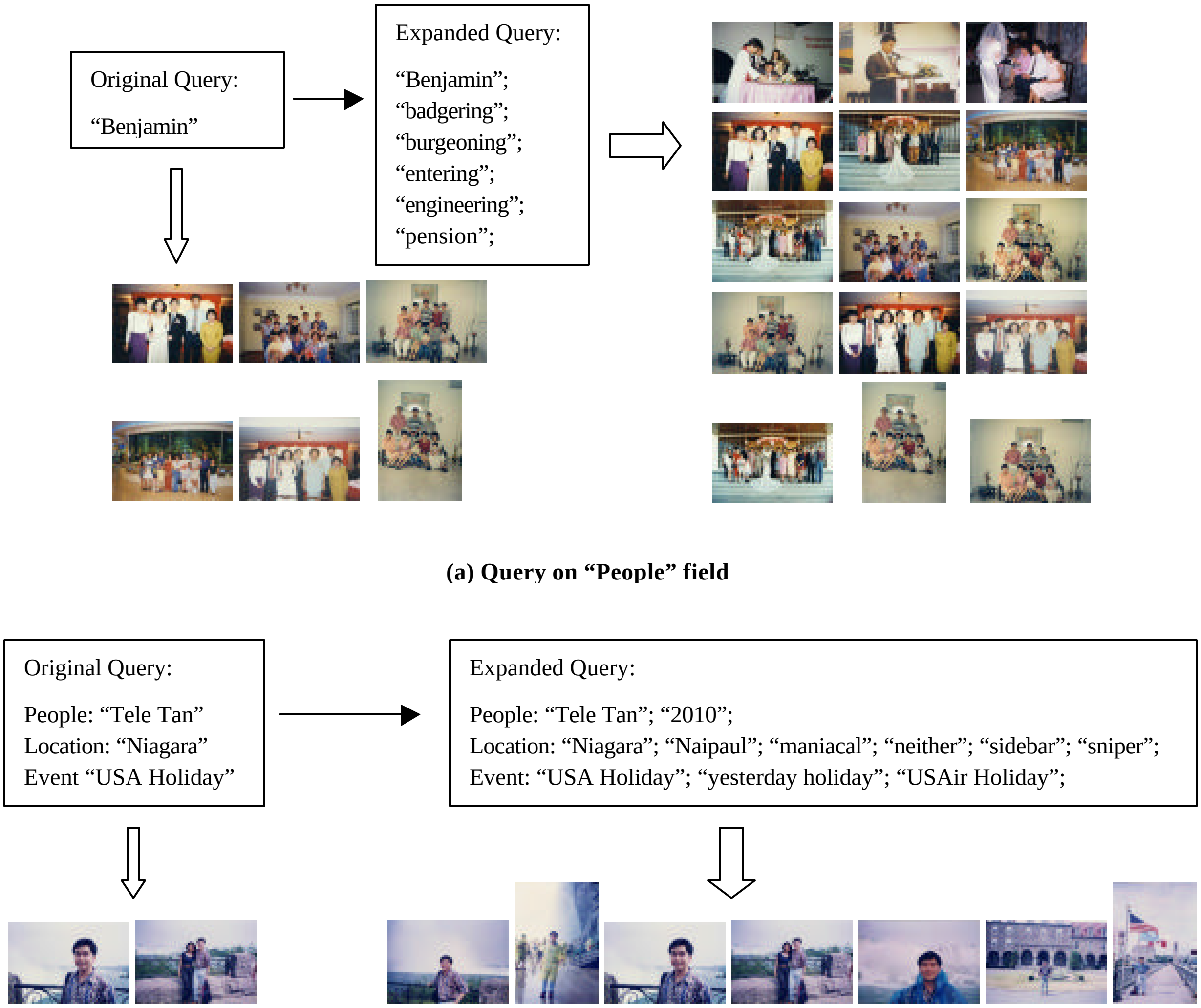
6. DISCUSSIONS

From Figure 4, we observe that the retrieval system with query expansion performs significantly better than without the expansion. More precisely, the average value of precision over recall increases from 52.7% to 71.5% on the field of ***People***, from 57.0% to 71.6% on ***Event***, and from 58.7% to 71.6% for ***Location***. When the queries are launched on the combination of various fields, we observe a more significant gain from 50.0% to 72.4% on average.

The relatively higher figures of improvement for ***People*** in comparison with other domains can be attributed to the frequent appearance of words that cannot be found the engine's inherent vocabulary, particularly when there are names specific to some nation or language. The augmented queries help the system retrieve far more relevant photos when higher value of recall is required. As for the other two fields, the expansion of query shows more advantages at lower recall because more relevant images are ranked higher in the final list. When the queries are placed on different fields simultaneously, the more stringent criteria result in the fewer number of photos retrieved that can meet all the requirements in each queried field. This can explain why the improvement for these combined queries is better than that of individual field.

It is also worthwhile to mention the versatility of timestamps in the system implementation. As described in Section 2.3 and Section 4, these parameters are useful to both determine which segment a term belongs to and identify the potential substitution for an utterance. In addition, they also play an important role in the expansion of phrase-based queries. Word-based index terms are usually weak in dealing with queries for an entire phrase. The situation is made even worse when forming an enhanced phrase query. However, with the exploitation of time stamps, only terms with adjacent beginning and ending marks are concatenated to form an enhance phrase. Thus utterance breaking or overlapping is prevented and the integrity and completeness of the resultant phrase can be ensured.

Two more examples are demonstrated in Figure 5, which include a single-field word-based query and a multi-field query consisting of both word and phrase. Retrieved photos are displayed from left to right and from top to bottom with their ranks of relevance in descending order. As shown in this figure, when the original term has fewer errors during the transcription, the number of its expanded terms will probably be less than the pre-defined value 6, which is the case for the sample multi-field query on ***People*** and ***Event***. On the contrary, a frequently misrecognized word usually needs a larger set of substitution terms to enhance its presence and tends to have their cooccurrence probabilities distribute fairly



(a) Query on "People" field

(b) Query on a combination of fields

Figure 5. Retrieval results with and without query expansion

evenly, like the query terms "Benjamin" and "Niagara" in this example.

In Figure 5(a), while only 6 out of 29 relevant photos with "Benjamin" are found based on the original query, that number is boosted to 15 after applying query expansion. In Figure 5.(b), the number of retrieved photos increases by more than 3 times from 2 to 7.

7. CONCLUSION

A system for the image indexing and retrieval with speech annotations has been described in this paper. High level information of images is annotated in the form of speech with a pre-defined structural syntax. Considering the error-prone nature of ASR process, N-best lists are explored for both index generation and query expansion. The query terms are enhanced through the addition of its most probable alternatives to improve retrieval effectiveness. Experiments on a collection of 1,200 photos shows that the approach is especially useful in retrieving images whose annotations contain a number of words unfamiliar to the engine. The overall performance in terms of average precision versus recall is improved from 50.0% to 72.4%, while the gains on individual domain of ***People***, ***Event*** and ***Location*** are 35.7%, 25.7% and 22% respectively.

In addition to query expansion using substitution terms, other techniques involving relevance feedback processing are being investigated in our context of speech annotations. The semantic associations between speech annotations and image features are also interesting. While the speech data used in this experiment is recorded with a normal microphone directly onto a PC, we will next proceed to test out the system on a more realistic backdrop; i.e. speech data capture directly onto the embedded microphone of a digital camera (e.g. Sony DSC-S70 model). It can be expected that a pre-processing step for the input signal is necessary to improve sound quality and produce reliable transcriptions.

8. REFERENCES

[1] J. Chen, T. Tan and P. Mulhem, A Method for Photograph Indexing using Speech Annotation, in Proc. of the Pacific Rim Conference on Multimedia, Oct 2001, Beijing, 867-872.

[2] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, and B. Dom, Query by Image and Video Content: The QBIC System, IEEE Computer, 28, 23-32, Sep 1995.

[3] K. Barnard and D. Forsyth, Learning the Semantics of Words and Pictures, International Conference on Computer Vision, vol 2, 408-415, 2001.

[4] A. Kuchinsky C. Pering, M.L. Creech, D. Freeze, B. Serra, and J. Gvvizdka, FotoFile: A Consumer Multimedia Organization and Retrieval System, in Proc. of the CHI'99 Conference on Human Factors in Computing Systems, Pennsylvania, U.S., 496-503.

[5] R. Lienhart, Dynamic Video Summarization of Home Video, SPIE 3972: Storage and Retrieval for Media Databases 2000, Jan 2000, 378-389

[6] R. Lienhart, A System for Effortless Content Annotation", IEEE CBAIVL 2000, 45-49, June 2000; also MRL Technical Report, Jan. 2000

[7] M. Mitra, A. Singhal, C.Buckley, Improving Automatic Query Expansion, SIGIR '98, 206-214, 1998

[8] F. Nack, W. Putz, Semi-automated Annotation of Audio-Visual Media in News, Dec 2000, http://www.gmd.de/publications/report/0121/Text.pdf

[9] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Content-based Manipulation of Image Databases. In SPIE Storage and Retrieval for Image and Video Databases II, number 2185, Feb 1994.

[10] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, 141-239, Prentice Hall, 1993.

[11] M. Sanderson, B. Croft, Deriving Concept Hierarchies from Text, SIGIR '99, 206-213, 1999.

[12] C. Shyu, A. Kak, C. Brodley, L. Broderick, Testing for human perceptual categories in a physician-in-a-loop CBIR system for medical imagery. IEEE Workshop of Content-Based Access of Image and Video Databases, 18-22, 1999.

[13] M.A. Siegler, Integration of Continuous Speech Recognition and Information Retrieval for Mutually Optimal Performance, Ph.D Thesis, Computer Science Department, Carnegie Mellon University, 1999. http://www.cs.cmu.edu/~msiegler/publish/PhD/thesis.ps.gz

[14] A. Singhal, F. Pereira, Document Expansion for Speech Retrieval, SIGIR '99, 34-41, 1999.

[15] J. R. Smith and S.-F. Chang, VisualSEEk: a Fully Automated Content-Based Image Query System, Proceedings, ACM Multimedia '96 Conference, Boston, MA, November 1996, 87-98.

[16] R.K. Srihari, Z. Zhang, M. Srikanth, B. Han, A. Rao and X. Wu, Multimedia Indexing and Retrieval of Voice-Annotated Consumer Photos, in Proc. of the Multimedia Indexing and Retrieval Workshop, SIGIR '99, University of California, Berkeley, U.S,1-16.

[17] R.K. Srihari, A. Rao, B. Han, S. Munirathnam and X. Wu, A Model for Multimodal Information Retrieval, IEEE International Conference on Multimedia and Expo (II) 2000: 701-704

[18] A. Stent and A. Loui, Using Event Segmentation to Improve Indexing of Consumer Photographs, in Proc. of SIGIR'01, Sep 2001, New Orleans, USA, 59-65.

[19] T. Tan, J. Chen and P. Mulhem, *SmartAlbum* – Towards Unification of Approaches for Image Retrieval, ICPR 2002, to appear.

[20] T. Tan and P. Mulhem, Image Query System Using Object Probes, in Proc. of the International Conference on Image Processing, ICIP 2001, Oct 2001, Thessaloniki, Greece, 701-704.