USING SPEECH ANNOTATION FOR HOME DIGITAL IMAGE INDEXING AND RETRIEVAL

JIAYI CHEN¹, TELE TAN², PHILIPPE MULHEM³

¹,²Laboratories for Information Technology, 21 Heng Mui Keng Terrace, Singapore 119613
³IPAL-CNRS, 21 Heng Mui Keng Terrace, Singapore 119613

Abstract: - We present in this paper a system for image indexing and retrieval through speech annotation made directly on a digital camera. The annotations are based on a pre-defined structured syntax and transcribed by an automatic speech recognition engine. N-best lists are incorporated in the index generation and query expansion process to introduce probabilistic model and compensate recognition errors. We also apply a practical noise cancellation algorithm to enhance the speech signal collected directly through the built-in microphone on a digital camera. Experiments show that by utilizing automatic query expansion and noise cancellation together, the overall system performance improve significantly to be comparable to the lab environment collected data.

Key-Words: - Speech Recognition, Speech Signal Processing, Spoken Document Retrieval, Image Retrieval

1 Introduction

With the increasing popularity of digital cameras, consumers are able to possess a larger collection of still images than ever before. Consequently, the demand for a system to effectively handle all these images of family style has been emerging.

Low-level attributes like texture and color have been widely used in current content based image processing systems. For example, the QBIC [3] and Photobook [11] systems are developed to help user exploit the large image content from the web, while VisualSEEK [16] allows the user arrange queries according to color regions. In the medical image retrieval system developed by Kak et al [13], domain knowledge about the structure of the anomalies is used to design appropriate detector. Likewise, the Object Probe [21] calls for a tuned feature set to classify home photographs into a few classes of objects. People have also begun to investigate the semantics information in image features [4].

Despite the popularity of content-based features, their abstractness hinders the easy comprehension by a common user. Instead, most people would prefer a more direct and intuitive description of their photographs to help them recall old photos, such as the people/objects in the photo, location where the photo was taken, event and date/time related to the photo-taking and/or other pertinent information. 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. Moreover, these advantages satisfy the users' demand for a robust and consistent mode to describe photos.

Traditionally high level image annotations are made by tedious manual typing. But as more digital cameras being equipped with a built-in microphone, and due to advances in automatic speech recognition (ASR) technology, it is now feasible to make speech annotations on the spot and transcribe the recorded audio files into textual format by an ASR system later. There have been numerous works on automatic speech annotation of digital spatial media. Stent and Loui [19] use free-style annotation to index and organize a collection of photos with an event segmentation scheme to group photographs into unique events. Lienhart [7, 8] introduces an on-the-fly annotation language to automatic acquisition of video abstracts and sets up two microphones to differentiate the environmental audio. Nack [10] describes a simulation of a handheld device for the annotation of simple semantic information. Show&Tell [17, 18] uses speech annotations to
index and retrieve both personal and medical images, while FotoFile [6] extends this to more general multimedia objects.

In contrast with those existing applications, we stress the importance of integrating together speech recognition and information retrieval technologies to ease the user’s job of managing home albums. Since image retrieval is virtually embodied in the form of spoken document retrieval, ASR accuracy plays a crucial role in the overall system performance. We deal with the anticipated recognition errors by enhancing query terms acoustically and semantically and apply a noise cancellation algorithm to boost the speech signal. The performance can be verified by experiments on a collection of 500 photos with annotations made through a digital camera.

2 System Overview

2.1 System Architecture

The system architecture is shown in Figure 1. Images and their speech annotations are separated by a multimedia splitter and stored as different files. While the speech files are sent directly to the database with corresponding images, they are also processed to reduce the noise in the signal transcribed by an ASR engine into textual form with a preset number of N-best lists and segmented into different parts according to the pre-defined syntactic structure. The resulting index terms for each part are kept in the multimodal database as well.

When retrieving image with keyword-based queries, it is very likely that a term present in both query and spoken annotation is misrecognized in the transcription session as another term for indexing. Such mismatch between the index and query terms may cause some truly relevant images to fail to be found. To decrease the unfavorable performance dependence on the recognition accuracy, query expansion techniques that intend to generate more terms semantically or acoustically related to the original query terms are involved. By using the original and expanded query terms together, more relevant photos are found at higher ranks.

2.2 Structured Speech Syntax & Speech Segmentation

Structured and semi-structured data like XML have been found important to various web and database applications, and are widely used in handheld devices like cell phones and camcorders [7, 8]. Similarly, to annotate photos with a digital camera, a structured syntax will suffice to cover most necessary information in a concise and effective manner with little information loss. It consists of four index fields, 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.). Each of these words acts as a leading tag that is spoken out before a detailed description of this field. Typically, the sample annotation of a photo taken at a friend’s home during a trip may look as follows.

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

However, speech recognition is susceptible to various uncertainties due to different environmental conditions and speaker’s utterance. A minor recognition error might lead to severe degradation in retrieval. Even when we are lucky enough to get accurate transcription, confusion between the content of each field would still make the retrieval work harder. For example, image related to the previously mentioned sample annotation will be
Figure 2. A simple representation of N-best lists returned to respond to a query finding all images of Angela.

Thus we introduce a segmentation step to separate the whole transcription into individual segments by identifying the leading tag of each field. Relevant content can now be analyzed on a segmental basis and recognition results are localized within one field. We have experimented with two solutions to this step, namely SAPI1- and KWS (Keyword Spotting) –based methods. Details can be found in [2, 23] and will not be repeated here. Table 1 shows the performance of both methods. It is clear that SAPI-based mechanism is superior in accuracy and adopted in following implementation.

Table 1. Accuracy of tag detection

<table>
<thead>
<tr>
<th>People</th>
<th>Location</th>
<th>Taken_on</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAPI-based</td>
<td>100%</td>
<td>99%</td>
<td>100%</td>
</tr>
<tr>
<td>KWS-based</td>
<td>90.7%</td>
<td>68%</td>
<td>86%</td>
</tr>
</tbody>
</table>

3 Image 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. With a reliable ASR engine, when a correct interpretation cannot be found in the top hypothesis, it can be frequently obtained in the N-best alternatives. This feature is available with most commercial engines compatible with MS-SAPI 4.0 like IBM ViaVoice and Dragon NaturallySpeaking. Figure 2 shows the typical recognition errors and their N-best lists.

Once the N-best lists are generated, all subsequent processing can be borrowed from standard document-based routines, such as stopword removal, word stemming and index term generation. According to [14], we assume the equal capability of recognition hypotheses above certain threshold to estimate the utterance content, thus introducing a probabilistic model to measure relevance based on TF-IDF (Term Frequency – Inverse Document Frequency). A detailed discussion and experimental results have been reported in [1, 2].

4 Image Retrieval with Automatic Query Expansion

Automatic query expansion techniques address the issue of word mismatching by adding words that are effective in distinguishing relevant documents from the data collection without user intervention. Quite a few methods have been proposed to enhance queries through thesauri and relevance feedback techniques [5, 9, 12, 15, 22]. Nevertheless, most existing methods make use of words with similar meaning and deal with large full-text collections, which are greatly different from the case of home album annotations. Instead, it is more sensible to discover word relationships and exploit the statistical cooccurrence of similar-sounding data from the parallel recognition hypotheses.

4.1 Query Expansion with Thesaurus

A thesaurus is a set of items (phrases or words) plus a set of relations between these items. Within our domain of application, a simple thesaurus is built to include some a priori knowledge of common sense mostly related to family activities. Semantically, we can use aliases and synonym pairs like “vacation” and “holiday” to account for their interchangeability in annotations. Conceptually, general geographical and chronological relationships can be discovered and organized through hierarchy, as the example shown in Figure 3. The retrieved image can then be expected to be more capable of describing the query.

4.2 Query Expansion with N-best Lists

With our existing method, images relevant to a query can be found only if their index terms or the counterparts in the thesaurus can be matched against the query terms. However, speech recognition is an error-prone process and may frequently create wrong index terms. The corruption of transcriptions can be
particularly serious when the annotation involves a
lot of loanwords, names of people and places that are
seldom used elsewhere, or even words outside the
engine’s inherent vocabulary. In such case we need
to consider query expansion techniques other than
the thesaurus level.

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 frequent and
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. More specifically, we compute the conditional
probability of uttered term $t$ being recognized as $t_i$ as

$$P(t_i | t) = \frac{P(t_i, t)}{P(t)}$$  \hspace{1cm} (1)

The total list of alternatives for a single $t$ might
be quite long, but many of them are merely picked
up by chance and not qualified enough to represent
the original term on a statistical basis. Practically,
we define a threshold for the substitution probability
to keep a limited number of alternatives. Meanwhile,
it is also reasonable and wise to add those terms with
higher probabilities as supplementary query terms to
enhance the presence of original $t$. Suppose $t$ can be
augmented by $M_t$ alternatives, the TF-IDF based
relevance value between query $q$ and annotation $a$
for segment $s$ can be calculated following the
notations used in [2]

$$r_{el}(q, a) = \prod_{s} \frac{1}{M_t} \left( \sum_{i} b_{q,s,i} c' i df^r \right)$$ \hspace{1cm} (2)

where

$$c' = \hat{c}_{a,s,i} + \sum_{i} \hat{c}_{a,s,i} P(t_i | t)$$ \hspace{1cm} (3)

$$idf^r = idf_{s,i} + \sum_{i} idf_{s,i} P(t_i | t)$$ \hspace{1cm} (4)

with $i$ varying from 1 to $M_t$. Here $\hat{c}_{a,s,i}$ and $b_{q,s,i}$ are
term count of $t$ in $a$ and $q$, respectively, $idf_{s,i}$ is a
measurement of term weight of $t$.

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 $M_t$ 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 Noise Cancellation on Speech Signal

The built-in microphone on a digital camera is
usually omni-directional, which may collect much
environmental noise and degrade the speech quality
when used directly for audio recording. It is thus
necessary to reduce the noise before the speech
signal is processed by the ASR engine.

There have been quite a few works on the noise
reduction and speech enhancement, among which
spectral subtraction method is one of the simplest
and most widely used. The basic idea is to estimate
the noise level based on the non-speech intervals,
and subtract it from the input signal on the frequency
domain. If we denote the input noisy signal, the
estimated noise and enhanced speech as $x(k), n(k)$
and $s(k)$, and denote their respective spectrum as
$X(\omega), N(\omega)$ and $S(\omega)$, then we will have

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{concept_hierarchy.png}
\caption{An example of concept hierarchy}
\end{figure}
where \( g(\omega) \) is defined as the gain factor on the power domain. In practice, to avoid the value of \( g(\omega) \) going negative when the estimated noise level exceeds the input signal, we often assign a threshold, which is between 0.1 and 0.01.

6 Experiments

We establish our experiments on a collection of 501 photos taken from 5 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, home and campus life, etc [23]. Annotation of each photo is collected using the built-in microphone on a Canon S30 digital camera. All audio data are recorded by one female speaker, with length of each file less than 20 seconds. Upon being downloaded to PC, all files are converted to a certain format and processed by the noise cancellation algorithm described in Section 5. The enhanced speech annotations are then transcribed by Dragon NaturallySpeaking\(^2\) Preferred version 5, segmented to generate index terms as mentioned in Section 2.3. The number of alternative lists explored, \( N \), is also set to 6 in the experiments.

Figure 4 shows the precision versus recall curves for 16 queries that consist of both single- and multi-field criteria, with and without noise cancellation before and after query expansion. The results of the same queries for data collected by a headset in lab environment are also shown for reference.

7 Discussions

According to Figure 4, the retrieval effectiveness on data with noise cancellation is superior to that without cancellation. The average precision before query expansion is 41.26\% for camera-generated data, while that value is increased to 54.14\% on the noise-cancelled data. After query expansion, the same measurement is improved to 54.29\% and 62.07\% for data without and with noise cancellation, respectively.

In comparison with the reference data collected in lab environment, before query expansion the average precision of noise-cancelled data is about 5\% less. However, for recall values lower than 0.3, the precision of the latter is even higher or nearly equal to the former, as shown in Figure 4(a). After query expansion, it is observed from Figure 4(b) that retrieval on noise-cancelled data outperforms the lab data until the recall value is greater than 0.5. Most noticeably, its average precision of 62.07\% is very close to that of 62.32\% obtained from the reference data, which leads to the conclusion that by utilizing low-level noise cancellation and high-level query expansion techniques together, the results for data from the camera microphone are almost as good as those collected by a headset microphone.

8 Conclusion and Future Works

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 and transcribed by an ASR engine. A probabilistic model is explored in the form of N-best list for index

\(^2\) http://www.dragonsys.com/
generation and query expansion. Considering the performance degradation caused by audio recording using the embedded microphone on a camera, we apply a practical method of spectrum subtraction to enhance the speech signal. The overall performance in terms of average precision versus recall can be improved significantly by utilizing query expansion and noise cancellation techniques together.

At present our scheme does not consider the semantic associations between speech annotations and image content features. However, it will be an interesting topic to study the audio and visual relationship for multimedia data-mining and multi-modal information discovery. It may also encourage us to extend this application to other multimedia modalities such as video and music collections.

Reference: