Multi-modal Information Retrieval from Broadcast Video using OCR and Speech Recognition

Alexander G. Hauptmann
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213 USA
+1 412 268 1448
alex@cs.cmu.edu

Rong Jin
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213 USA
+1 412 268 1448
rong@cs.cmu.edu

Tobun Dorbin Ng
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213 USA
+1 412 268 4499
tng@cs.cmu.edu

ABSTRACT
We examine multi-modal information retrieval from broadcast video where text can be read on the screen through OCR *and* speech recognition can be performed on the audio track. OCR and speech recognition are compared on the 2001 TREC Video Retrieval evaluation corpus. Results show that OCR is more important that speech recognition for video retrieval. OCR retrieval can further improve through dictionary-based post-processing. We demonstrate how to utilize imperfect multi-modal metadata results to benefit multi-modal information retrieval.

Keywords
Multi-modal Video Information Retrieval, Speech Recognition, Optical Character Recognition OCR

INTRODUCTION: Information retrieval from video: Speech Recognition and OCR

Video is a rich source of information, with aspects of content available both visually and acoustically. While image information can be exploited though content-based image retrieval [7], this paper focuses on information retrieval (IR) from textual representations of video documents. Previous experiments on information retrieval from speech recognition transcripts [6] found that as long as speech recognition has a word error rate better than 35%, then IR is only 3-10% worse than from perfect text transcriptions. Similarly, experiments have shown that IR on documents recognized through optical character recognition (OCR) with a character error rate of 5% and 20% only degrades by X% compared to perfect text retrieval [2]. This paper examines multi-modal IR from video documents where visible text is recognized with OCR *and* speech recognition is performed on the audio track. We also compare post-processing steps to improve the IR effectiveness of the highly errorful OCR data.

The TREC Video Retrieval Corpus

In 2001, the Text REtrieval Conference (TREC) performed the first video information retrieval evaluation [5] using a corpus of 11 hours of MPEG-1 encoded broadcast video. We report experiments with a fully automatic system on the 34 known item queries: where the evaluation could be done automatically and the top 100 system results were scored. The unit of retrieval was an automatically determined 'shot', i.e. a time range between two shot changes such as editing cuts and fades. A result shot was 'relevant' if it had reasonable overlap to a relevant target shot, at least 33% of the length of the returned item overlapped and less than 33% of the returned item was outside the target shot.

Query:
Find pictures of Harry Hertz, Director of the National Quality Program, NIST

OCR:
H,arry Hertz a Director
aro 7 wa-,i,,ty Program
,Harry Hertz a Director
Speech: We're looking for people that have a broad range of expertise that have business knowledge that have knowledge on quality management on quality improvement and in particular

Figure 1: A sample query and a relevant result found through OCR

Speech Recognition

The Informedia Digital Video Library audio processing component uses the Sphinx speech recognition system, a state-of-the-art large vocabulary, speaker-independent speech recognizer [1]. We used a 64000-word language model derived from a large corpus of broadcast news transcripts. Previous experiments had shown the word error rate on this type of mixed documentary-style data with frequent overlap of music and speech to be just over 30%.

Video OCR

Our video optical character recognition system [4] only sampled every 3rd image from the video. Initially, text regions are detected by searching for clustered sharp edges using horizontal differential filtering. Filtering is performed to reduce background noise. The potential text region is then extracted as a black and white image and submitted to a commercial optical character recognition package for the final stage of recognizing the text. On this video corpus, the word accuracy for detected text was estimated to be 27%.

Correcting OCR errors

We explored two different methods for correcting errors in the OCR transcriptions, both applied only to unmatched query words. The first method generates a new set of n-

gram strings to match the unedited the OCR transcriptions. These n-gram strings include strings with an edit distance of 1 character (1 deletion, insertion or substitution) and all possible n-gram substrings with at least 3 characters.

The second method used the dictionary spelling correction provided in MS Word. Through a program interface to MS Word 2000, an OCR-recognized string was expanded into its possible "corrected" spellings. We only expanded OCR words that MS Word had flagged as incorrectly spelled. This conservative expansion dramatically reduced spurious word candidates and avoided false matches.

Retrieval using:	ARR	Recall
Speech Recognition Transcript only	1.84 %	13.2 %
Video OCR only	5.21 %	6.10 %
Video OCR + Speech Recognition	6.36 %	19.30 %
VOCR w. n-gram post-processing	5.89 %	11.81 %
VOCR w. dictionary post-processing	5.93 %	7.52 %
Speech+VOCR with n-gram post-processing	5.11 %	16.07 %
Speech + VOCR with dictionary post-processing	7.07 %	20.74 %

Figure 2. Results for Video Retrieval using Speech Transcripts and OCR

Information Retrieval method and evaluation metric

Our retrieval used the OKAPI formula [3] in Equation (1)

$$Sim(Q,D) = \sum_{qw \in Q} \left\{ \frac{tf(qw,D) \log\left(\frac{N - df(qw) + 0.5}{df(qw) + 0.5}\right)}{0.5 + 1.5 \frac{|D|}{avg\_dl} + tf(qw,D)} \right\} \quad (1)$$

where $tf(qw,D)$ is the term frequency of word qw in document D, $df(qw)$ is the document frequency for the word qw and avg_dl is the average document length for all the documents in the collection.

Because our collection contains only small numbers of relevant items, we adopted the average reciprocal rank (ARR) as our evaluation metric, as in [2]:

$$ARR = \left\{ \sum_{i=1}^{k} i / r_i \right\} / N_r \quad (2)$$

For a given query, there are a total of N_r relevant items in the collection. If the system retrieves k relevant items, they are ranked as $r_1, r_2, \ldots, r_k$.

ARR rewards relevant items near the top of the retrieval list and de-emphasizes relevant items near the bottom of the list. Since the formula divides by the total number of relevant items for a given query, 'easier' queries with more answer items are not favored over 'difficult' queries where only a few answer items are relevant.

Recall is measured as the number of relevant items found over the total number of relevant items.

Results and Discussion

The results in Figure 2 show that speech transcripts do much worse than OCR in ARR (1.84% vs. 5.21%) but not recall (recall 13.2% vs. 6.10%). Combining OCR and speech transcripts increased ARR and recall to 6.36% and 19.30% respectively. The n-gram post-processing improved the OCR output to 5.89% ARR (11.81% recall). Similarly, dictionary-based post-processing yielded 5.93% ARR and 7.52% recall.

Interestingly enough, combining the n-gram post-processed OCR with the speech transcripts (ARR of 5.11% and recall of 16.07%) did not improve the retrieval effectiveness. But the dictionary-based post-processing method, which on its own had about the same precision and 40% *lower* recall than the n-gram method, provided a more effective combination with the speech transcripts at 7.07% ARR and 20.74% recall. This is about a 10% increase over the previous best combination. N-gram OCR correction initially appeared as good as the dictionary method, but much worse in combination with speech transcripts, possibly due to over-generation of word candidates.

Overall, the queries presented a very challenging task for an automatic system. While the ARR and recall numbers seem small, we should note that for about one third of the queries *nothing* relevant was found by ***any*** of the automatic systems participating in the Video Retrieval Track.

In conclusion, our most surprising finding is the dominating importance of OCR over speech recognition in this video retrieval task. This surprise was perhaps due to queries that were designed for video documents and not merely text transcripts. A possible explanation is that OCR text appears directly inside a relevant image, while relevant words can be spoken in the vicinity near the relevant video clip, but not directly during the target shot.

REFERENCES

1. Singh, R., Seltzer, M.L., Raj, B., and Stern, R.M. "Speech in Noisy Environments: Robust Automatic Segmentation, Feature Extraction and Hypothesis Combination", *ICASSP*, Salt Lake City, UT, May 2001.
2. Kantor, P. and Voorhees E.M, Report on the Confusion Track, "The Fifth Text Retrieval Conference, (TREC-5) 1997.
3. Robertson S.E., et al.. Okapi at TREC-4. In The Fourth Text Retrieval Conference (TREC-4). 1993.
4. Sato, T., Kanade, T., Hughes, E., and Smith, M. Video OCR for Digital News Archive. In *Proc. Workshop on Content-Based Access of Image and Video Databases.* (Los Alamitos, CA, Jan 1998), 52-60.
5. The TREC Video Retrieval Track Home Page, http://www-nlpir.nist.gov/projects/trecvid/
6. Voorhees, E. and Harman, D., The Eighth Text Retrieval Conference(TREC-8), Gaithersburg, MD, 2000 http://trec.nist.gov/pubs/trec8/t8_proceedings.html
7. A. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," IEEE Trans. PAMI, 22(12), pp.1349-1380, Dec. 2000.