Term Selection and Query Operations for Video Retrieval

Bouke Huurnink and Maarten de Rijke

ISLA, University of Amsterdam,
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
{bhuurnin,mdr}@science.uva.nl

Abstract. We investigate the influence of term selection and query operations on the text retrieval component of video search. Our main finding is that the greatest gain is to be found in the combination of character n-grams, stemmed text, and proximity terms.

1 Introduction

Widespread availability of digital video technology has led to increasing amounts of digital video data. Evidently, we need retrieval systems and algorithms to help us search through it. Automated Speech Recognition (ASR) transcripts are often the only text information source available for video search; the challenge is that this text is subject to speech recognition errors and that the spoken language used is substantially different from that in written query text. Additionally, written queries may request information contained in the visual modality.

Past experiments have shown that ASR transcriptions of video are a valuable source of information [8]. Despite this fact, most research in video retrieval centers around multimodal analysis. Typically, only standard text retrieval methods are used for search of speech extracted from video. As part of an agenda aimed at optimizing the textual components of video retrieval, we report on experiments with different types of text representation for video retrieval. Specifically, we consider character n-grams (i.e., sub-word units), stemming, and proximity terms (i.e., multi-word units) and determine their impact on text retrieval effectiveness for video search. Our main finding is that video retrieval performance can be significantly improved through combination of term selection strategies.

2 Experimental Setup

For evaluation purposes we use the TREC Video Retrieval Evaluation (TRECVID) [8] datasets from 2003–2006. The combined dataset yields over 300 hours of news broadcast video which has been automatically segmented into over 190,000 shots—the basic units of retrieval. It is accompanied by 95 topics, each consisting of a short natural language description of the visual content that is desired from relevant shots. For each topic a ground truth has been manually created. Other

components of the TRECVID dataset utilised for our retrieval experiments include ASR transcripts, machine translation text, and news story boundary annotations, all of which have been generated automatically.

Each shot is associated with the (English language) transcription of words spoken during that shot, as well as the transcription of the news story in which it occurs. By incorporating the surrounding news story we compensate for the temporal mismatch between the occurrence of an entity in speech and its occurrence in the associated video. We focus on the effects of textual term selection for video retrieval. Therefore we choose to use an existing, well researched, vector space model as our retrieval mechanism, rather than tweaking parameters for this specific task. Indexing and retrieval is done using Lucene [4]. Evaluation of the ranked shots is done using Mean Average Precision (MAP) [8].

3 Experiments

We ran three sets of experiments, one on character n-gram tokenization, one on stemming, and one on the use of proximity terms. For each set of experiments we determine the winning algorithm and perform a comparison to the text baseline. Our baseline representation is an index containing words as they occur in the ASR with stopwords removed. To see to what extent the representation methods just listed have complementary effects, we ran several combination experiments.

Character N-Gram Tokenization. Character n-gram tokenization has been shown to boost retrieval in certain situations [5]. E.g., the use of character 4-grams has proved useful for retrieval from English newspapers [2]. We investigate the effects of n-gramming at different sizes of n (and in different combinations). We follow the tokenization strategy used in [5], creating overlapping, cross-word character n-grams, using values for n from 3 to 7.

Stemming. For our stemming experiments we used the Porter stemming algorithm [7] to normalise morphological variants of words.

Proximity Terms. Our third set of experiments follows [6], who found the use of proximity to be beneficial in web queries, prioritising results in which query words occur close together. Here we experiment with word n-grams, varying the magnitude of n and the proximity required between query words. We investigate the effects of retrieving consecutive sequences of 1 to 7 query terms, and combinations thereof. We also explore the effects of proximity terms—where terms are required to occur in a window of n words—and vary the window size.

Run Combinations. Here, we evaluate all linear combinations of runs produced using the optimal term selection strategies determined in the previous three experiments [3].

4 Results and Analysis

Table 1 provides an overview of the best performing settings for each of the methods identified in the previous section, and for their combinations. Due to

Table 1. Best performing settings per method and for combinations; Δ indicates the percentage change compared to the baseline. Significant changes are indicated with * and ** (two-tailed Wilcoxon Matched-pair Signed-Ranks Test, improvements at the 0.05 and 0.01 level, respectively).

Individual			Combinations		
Method	MAP	Δ	Methods	MAP	Δ
Baseline	0.0609	–	Char. n-grams	0.0596	−2.1
Char. n-grams	0.0574	−5.7	Char. n-grams + stem.	0.0647	+6.3
Stemming	0.0647	+6.2	Char. n-grams + Prox. terms	0.0616	+1.2
Prox. terms	0.0627*	+3.0	Prox. terms + stem.	0.0658	+8.2
			Char. n-grams + stem. + Prox. terms	0.0691**	+13.5

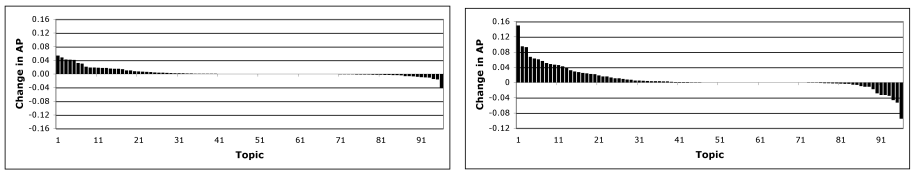

Fig. 1. (Left): Per-topic change in average precision compared to the baseline, using stemming. (Right): Per-topic change in average precision compared to the baseline, using the grand combination strategy. Topics sorted in descending order by change over the baseline.

space limitations we cannot provide detailed tables per method; instead, we briefly discuss the findings per group of experiments.

Character N-Gram Tokenization. The best performing method here is 5-grams (which differs from the best settings reported in [2,5], who found 4-grams to be optimal). It performs below the baseline, but not significantly. The best performing combination of character n-gram tokenizations combines 4, 6, and 7-grams and performs somewhat better, but still below the baseline.

Stemming. Retrieval on stemmed ASR text outperformed all single n-gram techniques, with a MAP of 0.0647; the difference with the baseline was not significant, though: as Figure 1(Left) suggests, all gains were offset by losses of practically the same size.

Proximity Terms. Proximity terms (allowing up to 10 non-query term to occur between query terms) outperformed the baseline, and did so significantly, as they did in [6]. The best results were achieved by using 2 word sequences for the proximity query, and combining these with the baseline run (i.e., "1 word sequences").

Combinations. Turning to combinations of different methods, we observe that the following best combinations all outperformed the baseline (but not significantly): character n-grams plus stemming, character n-grams plus proximity terms, and proximity terms plus stemming. The grand combination that combines runs created using the best settings for each of character n-gramming, stemming, and proximity terms results in a run that significantly outperforms

the baseline (at the 0.01 level). Figure 1(Right) shows the change in topic average precision values using the final grand strategy. Space does not allow an analysis of the effects on individual topics, but it is evident that the majority of topics show an increase in performance.

5 Conclusion

We examined whether term selection alone can be used to significantly improve video retrieval performance. We see that the textual component of video retrieval is similar to other forms of retrieval in that the use of proximity term pairs significantly improves retrieval effectiveness. Stemming also improves performance, while character n-gramming proves not to be directly useful for video retrieval. However, a combination of character n-grams, stemmed terms, and proximity terms results in the best performance.

In future work we plan to explore further avenues for improving the effectiveness of textual search for video retrieval. These include query and document expansion techniques [9] and the linking of textual topics to visual detectors [1].

6 Acknowledgments

Both authors were supported by the Netherlands Organization for Scientific Research (NWO) MUNCH project under project number 640.002.501. Maarten de Rijke was also supported by NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.065.120, 612-13-001, 612.000.106, 612.066.-302, 612.069.006, 640.001.501, and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

References

1. C.G.M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber and M. Worring Adding semantics to detectors for video retrieval. *Journal paper* (submitted).
2. V. Hollink, J. Kamps, C. Monz, and M. de Rijke. Monolingual document retrieval for European languages. *Inf. Retr.*, 7(1-2):33–52, 2004.
3. J. Kamps and M. de Rijke. The effectiveness of combining information retrieval strategies for European languages. In *Proc. 19th Annual ACM Symp. Applied Computing*, pages 1073–1077, 2004.
4. Lucene. The Lucene search engine, 2006. `http://lucene.apache.org/`.
5. P. McNamee and J. Mayfield. Character n-gram tokenization for European language text retrieval. *Inf. Retr.*, 7(1-2):73–97, 2004.
6. G. Mishne and M. de Rijke. Boosting web retrieval through query operations. In *Proc. ECIR 2005*, LNCS 3408, pages 502–516. 2005.
7. M. F. Porter. An algorithm for suffix stripping. In *Readings in Information Retrieval*, pages 313–316. 1997.
8. A.F. Smeaton, P. Over, and W. Kraaij. TRECVID: Evaluating the effectiveness of information retrieval tasks on digital video. In *ACM Multimedia*, 2004.
9. E. Voorhees and J. Garofolo. Retrieving noisy text. In E. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*. 2005.