The Value of Stories for Speech-Based Video Search

Bouke Huurnink
ISLA, University of Amsterdam
Kruislaan 403
Amsterdam, The Netherlands
bhuurnin@science.uva.nl

Maarten de Rijke
ISLA, University of Amsterdam
Kruislaan 403
Amsterdam, The Netherlands
mdr@science.uva.nl

ABSTRACT

Anecdotal evidence suggests that story-level information is important for the speech component of video retrieval. In this paper we perform a systematic examination of the combination of shot-level and story-level speech, using a document expansion approach. We isolate speech from other retrieval features, and evaluate on the 2003–2006 TRECVID test sets with a set of 94 natural language queries. Our main finding is that that the use of story information significantly improves retrieval performance compared to shot-based search, increasing overall mean average precision by over 65%.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3 .3 Information Search and Retrieval; H.3.4 Systems and Software; H.4 [**Information Systems Applications**]: H.4.2 Types of Systems; H.4.m Miscellaneo us

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Story information, video news retrieval, temporal mismatch

1. INTRODUCTION

The traditional unit of multimedia retrieval is the shot, a cohesive visual unit. The shot lends itself well to visual analysis, but is less useful when employing speech for retrieval. There is often a temporal mismatch between the moment in which an object appears on screen and the moment that an object is mentioned in the dialogue [18]. For example, a shot might show a news anchor introducing an about George Bush, and be followed by a shot of George Bush giving the State of the Union address. When the temporal mismatch is not taken into account, a text-only search will return the shot of the anchor in the studio, while the user wants to see a shot of George Bush.

To some extent this can be corrected by incorporating visual analysis in search. However, there are still some classes of objects (for example, named people), that remain very difficult to find using only visual analysis [16]. Here, the speech component of video, in the form of automatic speech recognition (ASR) transcripts, plays a vital role. In order to achieve good performance for searches containing these types of objects textual retrieval must be used. In this paper we examine the use of surrounding information to improve speech-based retrieval of shots.

Recent advances in video analysis have lead to an increased ability to detect the boundaries between broadcast stories. Anecdotal evidence indicates that the use of story boundaries is extremely valuable for multimedia retrieval, but—as far as we are aware—there has been no systematic examination into the potential of story windows for improving multimedia retrieval effectiveness. The goal of this paper, therefore, is to provide such a systematic examination. To this end we propose a document expansion based combination model as a natural extension of the shot-story paradigm, and use this model to evaluate the weighted combination of shot transcripts with speech from the surrounding story window. Our main finding is that the weighted combination of shot text with the surrounding story text can result in significant gains in overall retrieval performance.

The paper is organised as follows. Section 2 gives background to the use of stories in multimedia retrieval. Section 3 outlines our model for document expansion to include story information. Section 4 describes the experimental setup and design, and Section 5 describes the results and analysis. Finally, conclusions are presented in Section 6.

2. SHOTS, STORIES AND THE TEMPORAL MISMATCH

As the fundamental units to be examined in this paper, the shot and the story unit warrant further investigation. Within video analysis a *shot* is commonly defined to an uninterrupted camera recording sequence, and this is the definition adopted for the remainder of the paper. The shot is easily detectable and visually coherent. It provides relatively consistent colour, motion, spatial, and texture features, making it useful for multimedia analysis and retrieval based on visual features. However, it can can be problematic when performing retrieval using speech, due to the temporal mismatch. When an object is mentioned in the speech

CIVR'07, July 9–11, 2007, Amsterdam, The Netherlands.

channel of a shot, it may appear in the visual channel of a different shot.

The strong temporal relatedness of sequenced shots [19], is an aspect of video structure that has proved useful for speech-based retrieval of shots. Given the mention of an object in the speech channel, there is an increased probability of that object appearing in the surrounding shots, as illustrated for the case of named people in [18]. The international TRECVID [14] benchmark offers insight into how state-of-the-art multimedia systems have utilised shot proximity for the speech component of retrieval. dealing with this fundamental temporal mismatch. A number of 'window-based' methods have been used, including text from surrounding windows of, for example, n words [15], seconds [6], or shots [5]. However, these window-based methods by necessity make a somewhat arbitrary selection of window size. Put differently, such selection methods do not take into account the natural storytelling units contained in a video sequence, and may include text from totally unrelated news items.

During video production shots are generally chained together to form *stories*, each of which communicates a cohesive message. Story units are commonly used to create structure in many types of produced media, including motion pictures, sitcoms and documentaries [21]. Broadcast news is also composed of story units, in the form of news items. Where shots are visually coherent, stories are semantically coherent. Shots within a news story may suffer from the temporal mismatch, but shots containing an object in the visual channel should have a good chance of having that object described in the speech channel of another shot in the same story. Recent advances in broadcast news analysis have resulted in reasonably accurate automatic story boundary detection algorithms [3]. However, shots rather than stories remain the usual unit of video retrieval.

Recently, efforts have been made to use the story unit to improve video retrieval. Three of the four state-of-the-art systems that achieved the best performance in the 2006 TRECVID automatic search task incorporated surrounding story text in some way [1,2,4]. All three of these systems incorporated a search on story-level speech with other textual and multimodal search features. Due to the large number of features used to search, as well as the small number of topics used for evaluation, it is somewhat unclear to what extent the addition of story boundaries to text retrieval improves the results. Yang and Hauptmann [20] have attempted to correct for the temporal mismatch by using story boundaries as cutoff points for speech. They learn co-occurrence relationships between visual and textual features, and use these to assign descriptive words from the speech channel to the shots containing the object that is describe in the visual channel. They find some success in assigning words to the correct shots, but do not provide an evaluation on the effect of this on retrieval performance.

3. COMBINATION FRAMEWORK

From the previous section we can see that there is some evidence that story units can be used to improve video retrieval. In this section we describe a formal model for exploiting the story unit in textual news video retrieval, recasting the problem within the context of document expansion. We choose to focus on speech-only retrieval, in order to avoid the clouding influences that may be caused by the many features interacting in a fully multimodal video search engine.

Specifically, when performing retrieval on a set of multimedia shots, we aim to take into account the speech occurring in the story surrounding each shot. One approach would be to simply associate each shot with the words surrounding it in the story, as was done in, e.g., [1, 2]. It is reasonable to assume, however, that the ASR surrounding a shot should be given a different weighting to the shot itself. Document expansion is a technique originating from spoken document retrieval that allows for this in a natural way [13]. In this approach, a document is expanded and re-weighted with related text at indexing time. Traditionally, this approach is used to augment the original document with text from multiple related documents that have been obtained by some form of feedback. Instead of using feedback we hypothesise that the text of all shot transcripts within the surrounding story window is relevant for the current shot, and use and use these shots as the related documents, giving each shot an equal weighting.

The technique outlined in [13] was specific to the vector space model approach and a certain implementation of relevance feedback. Tao et al. [17] propose a more general model for document expansion, outlined below, on which we build. To perform document expansion, we use a corpus E to determine additional information about every document d in a corpus C. At indexing time we use word counts from d and from documents in E to create a 'pseudo-document,' d'. The word counts in d' are adjusted from those in d according to:

$$c(w|d') = \alpha \cdot c(w,d) + (1-\alpha) \cdot \sum_{e \in E} (\gamma_d(e) \cdot c(w,e)), \quad (1)$$

where α is a constant, e is a document in E, γ is our confidence that e provides information that is useful for d, and $c(w,d)$ is the number of occurrences of w in d.

Placing this model in the context of shot and story combination, we define two collections of equal size:

- C the collection of shots, associated with the speech that occurs within the shots.
- E is the collection of stories, where each shot is associated with the speech that occurs in the story surrounding the shot.

There is then a direct one-to-one mapping from C to E, where every shot s has one corresponding story document S in E. This eliminates the need for γ, which is now either 1 (for the related story document) or 0 (for all other story documents), and gives the following simplified version of Eq. 1:

$$c(w|d') = \alpha \cdot c(w,s) + (1-\alpha) \cdot c(w,S). \quad (2)$$

This, then, is the framework in which we design and perform our combination experiments.

4. EXPERIMENTAL SETUP AND DESIGN

The combination framework described in the previous section gives us a basis for performing a systematic examination of the potential of story windows for improving multimedia retrieval effectiveness. In this section we specify our research questions and describe the experiments we set up to address these questions.

4.1 Research Questions

The questions we aim to answer concerning the usage of story information for multimedia retrieval are the following:

Q1 Can stories help improve news video retrieval effectiveness? If so, what weighting should we give their associated speech?

Q2 In which cases are story clues particularly detrimental, and in which cases are they particularly useful?

Q3 How stable is the optimal setting of α in Eq. 2 across different retrieval parameters?

Q4 Does it make a difference in terms of the benefit of stories for news video retrieval whether manual or automatic story boundary detection is used?

4.2 Setup

4.2.1 Data

For evaluation purposes we use the TRECVID datasets from the 2003–2006 benchmarking evaluations. These test collections yield over 300 hours of English, Arabic, and Chinese news broadcast video have been created. Associated with the test collections are automatically generated boundary annotations for over 190,000 shots—the basic unit of retrieval. Also included are ASR transcripts for the entire datasets, and in the cases of Arabic and Chinese videos, machine translations of those transcripts to English. In our experiments, we only consider the (machine translated) English-language transcripts.

A total of 97 official TRECVID topics have been created for the 2003–2006 test sets. Each topic consists of one or two natural language sentences describing the visual content that is desired from relevant shots, as well as multimodal examples. It is also accompanied by a ground truth of relevant shots from the associated test set. After an examination of the ground truth we eliminated three topics, leaving us with a total of 94 topics for evaluation.[1]

We use a combination of manually and automatically generated story boundaries in our experiments. For the 2003 and 2004 test sets we utilise the manually annotated story boundary ground truth created for the TRECVID story boundary detection task [14]. For the 2005 and 2006 editions we utilise automatically generated story boundary annotations donated by Columbia University, which were shown to be effective in the 2004 story boundary detection task [8].

4.2.2 Experimental design

For each shot in our total collection we define two textual units. The first unit is the ASR that occurs during the shot. The second unit is the ASR that occurs in the story that surrounds the shot; observe that for any given shot, the two types of units have an empty intersection. At indexing time all text is stemmed using the Porter [12] stemmer, and commonly occurring stop words—including the TRECVID-specific stopwords 'find' and 'shots' that occur in every topic—are removed.

[1]Topic 0146 was eliminated as there were no relevant shots in the test set. Topics 0118 and 0119 were eliminated as the (C-SPAN) videos containing the relevant shots were not accompanied by ASR transcripts in that year, rendering it impossible to find the relevant shots through a search on the transcripts alone.

For retrieval it is possible to use any method that incorporates term statistics, thanks to the generic nature of our document expansion framework. For the purposes of this paper we restrict ourselves to a single retrieval method—language modeling. We choose language modeling as it is a theoretically transparent retrieval approach and has been shown to be competitive in terms of retrieval effectiveness [7, 11, 22]. Here, a query is viewed as having been generated from an underlying document specific language model, where some words are more probable to occur than others. At retrieval time each document is scored according to the estimated likelihood that the words in the query were generated by a random sample of the language model underlying the document. The underlying word probabilities are generally estimated from the document itself (using maximum likelihood estimation) and combined with background collection statistics to overcome zero probability issues. This combination process is known as *smoothing*, and in our experiments we assess two common smoothing techniques—Jelinek-Mercer and Dirichlet [22].

Using the textual units described above we created three types of runs:

baseline Parameter optimised set of searches on shot ASR only. (Considering Eq. 2, this is equivalent to expanding documents using $\alpha = 1$.)

story_only Set of searches on surrounding story ASR only, using the optimised baseline parameter settings. (This is equivalent to expanding documents with $\alpha = 0$.)

combined Multiple sets of searches performed with incremental values of α ($0 < \alpha < 1$), using the optimised baseline parameter settings.

To measure the quality of our runs we used the standard mean average precision (MAP) metric, based on ground truth provided by the TRECVID organizers.

For parameter optimization we proceeded as follows. We optimised the baseline run with a parameter search over the Jelinek-Mercer and Dirichlet smoothing methods. The optimal baseline overall setting (in terms of MAP score) was obtained using the Jelinek-Mercer smoothing method with λ set to 0.8. For the *story_only* and *combined* runs, we did not perform further optimizations, but simply took the optimal settings from the baseline: Jelinek-Mercer smoothing with $\lambda = 0.8$. For the combination run performance was calculated at α increments of 0.05.

For significance testing we used the (two-tailed) Wilcoxon matched-pair signed-ranks test and looked for difference at the 0.01 level.

5. RESULTS AND ANALYSIS

In this section we provide answers to the research questions raised in the previous section, and provide a topic-level analysis of our results.

Q1 Finding the optimal combination

To answer our first research question (Can stories help improve news video retrieval effectiveness?), we used the *combined* run to compare the MAP across varying weightings. This is shown in Figure 1, where the *baseline* run corresponds to $\alpha = 1$, and the *story_only* run to $\alpha = 0$. Performance increases as α increases from 0, but drops sharply

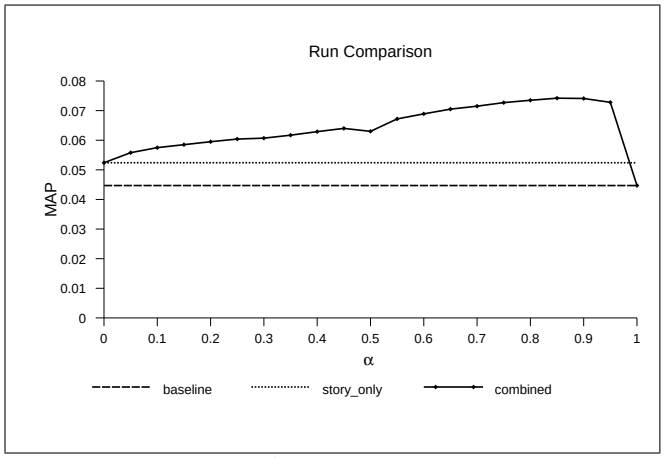

Figure 1: Overall MAP scores for varying values of α.

Table 1: Overview of the relative performance of the baseline shot only search, story only context search, and the optimal combined shot-story search with $\alpha = 0.85$. When counting the number of topics where the MAP score increased or decreased, we only take changes with a magnitude > 0.01 into account.

Measure	Baseline	Story Only	Combined
MAP	0.0447	0.0524	0.0742
Δ	–	17.1%	65.9%
# increased topics	–	25	34
# decreased topics	–	23	10
# unchanged	–	46	50

when α approaches 1. The optimal weighting is reached at $\alpha = 0.9$.

Table 1 gives a summary of the *baseline* run, the *story_only* run, and the optimal set of searches from the *combined* run. We see that the runs using the stories improve over the baseline (significantly so in the case of the combined run) and that the combined run improves over the story run (again, significantly so). So we can conclude that stories can indeed help improve retrieval effectiveness. It is worth stressing that search utilising only the story text—which does improve over the baseline, although not significantly so—does *not* contain the shot text.

Q2 Topic case studies

To answer our second research question (In which cases are story clues particularly detrimental, and in which cases are they particularly useful?), we consider the topic-level changes in MAP when comparing the optimal *combined* run to the *baseline* run. A topic-level overview is given in Figure 2, showing that more topics benefit from story optimisation than not. Furthermore, the magnitude of gains in MAP is greater than the magnitude of the losses. In Table 1 we see that, after discounting topics with very small changes in MAP, optimisation causes 34 topics to increase in MAP and 10 topics to decrease. Not all topics were affected to a great degree by changing the shot-story combination. We hypothesise that this is caused by a lack of sensitivity to any kind of textual retrieval, and that these topics may especially benefit from the visual search.

To help us understand why story context can be better for retrieval than a search on the actual text within a shot, we provide an analysis of two topics that benefit from shot-story combination search (on the extreme left-hand side in Figure 2), and one topic (taken from the extreme right-hand side in Figure 2) where performance degenerates.

Topic 0116, 'find shots of the Sphinx,' profited the most out of all topics and increased in MAP from 0.1667 to 0.6745 upon addition of story speech. We examined the retrieval result data and found that in the test set there were only three shots containing the word 'Sphinx.' These were the only results returned by the shot-based search. However, these three shots were part of two lengthy news items about the Sphinx, each containing many shots. When story level text was included, a total of 28 shots were returned, 12 of which contained the Sphinx. In this way recall was increased and MAP was dramatically improved. Topic 0153, 'find shots of Tony Blair,' also benefited from the shot-story combination approach, increasing in MAP from 0.0787 to 0.3839. This was caused by a single long news item about Tony Blair. In this story half of the total of 20 shots showed Tony Blair. Only 4 of the shots contained the words 'Tony' or 'Blair,' and of those 4 shots, only one contained an actual image of Tony Blair. When searching shot text only, the single video shot containing both a visual and a textual representation of Tony Blair. When using the shot—story combined text, all 20 shots were returned as the top 20 results, resulting in a very high MAP score. In this case searching on story speech improved recall by returning shots that did not contain matching keywords, though they did contain the subject of the topic. Precision was also improved, by including all mentions of the keywords in the overall retrieval score, despite them being spread over many fragments.

Finally, we analyse the topic that was harmed the most by shot-story combination, topic 0106, 'find shots of the Tomb of the Unknown Soldier at Arlington National Cemetery.' The results for this topic decreased in MAP from 0.3260 to 0.1845 upon introduction of story speech. In this case, retrieval performance was harmed by a long story that was not directly about the actual topic. When searching on shot ASR only, relevant shots from a variety of stories about the Tomb of the Unknown Soldier were returned (11 relevant in the top 20). However, when story text was included, the top results were dominated by a story about *preparations for a ceremony* planned around the Tomb of the Unknown Soldier, and so the returned shots were dominated by these preparations, rather than the Tomb itself. The results only contained a total of 4 relevant results in the top 20. In this case it appears that inclusion of an indirectly related story caused the drop in MAP.

Q3 Weighting stability

We turn to our third research question (How stable is the optimal setting of α across different retrieval parameters?).

We duplicated the *combined* run under varying smoothing settings, testing both Dirichlet and Jelinek-Mercer smoothing and varying the parameter values. Figure 3 shows the results of shot-story combination under 12 different smoothing settings. While different smoothing methods results in different MAP scores, the general shape of the combination curve is similar across the different retrieval settings. This curve is characterised by increasing mean average precision as the shot speech is given increasing influence, followed by a sharp drop when story speech is given no weighting at all

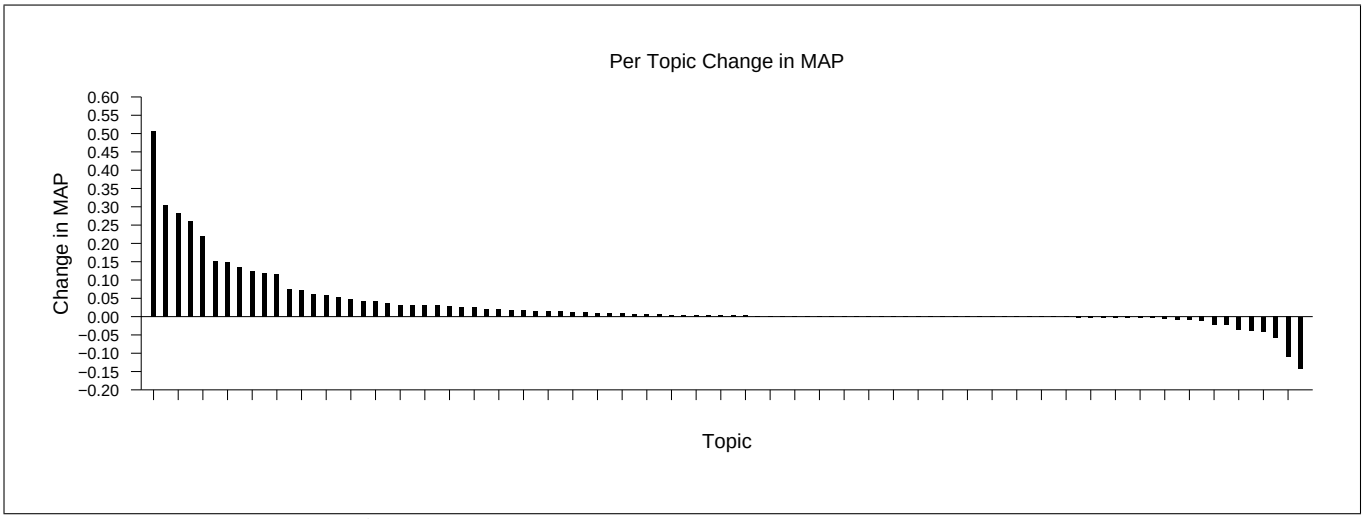

Figure 2: Per topic change in MAP score when comparing the optimal combined search to the baseline, sorted by improvement over the baseline.

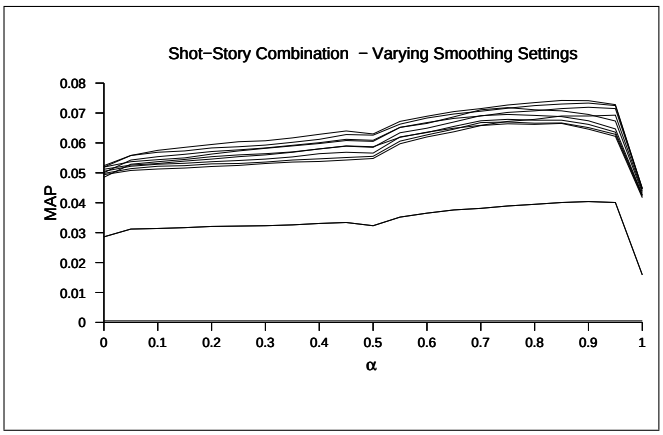

Figure 3: Overall MAP scores for different Jelinek-Mercer and Dirichlet smoothing parameters, varying α for shot and story combination

($\alpha = 1$). The optimal value of α remains constant at close to 0.85 for nearly all retrieval parameter settings, indicating that for the language modeling approach the optimal value for α is fairly stable.

Q4 Manual vs Automatic Boundaries

For our final research question (Does it make a difference ... whether manual or automatic story boundary detection is used?) we examine differences in performance when using automatic versus manual story boundary detection. From our total set of 94 topics, 46 were assessed using manual boundary annotations, and 48 topics were assessed using automatic annotations. MAP scores increased from 0.0622 to 0.0847 (+36.1%) when using manual annotations, and from 0.028 to 0.0618 (+120.5%) when using automatic annotations. While these results are not strictly comparable as they are obtained from different test sets, they seem to indicate that the combination approach works no matter whether the boundaries have been generated automatically or manually. This may imply that the combination approach is robust to story boundary recognition errors.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have examined the effect of combining speech from shots with the speech occurring in their surrounding story windows. Our main finding, evaluated on a set of 94 topics taken from the TREC Video Retrieval Evaluation setting, is that an overall increase in performance of over 65% may be achieved through the weighted combination of shot and story speech. Further analysis shows that increased recall is an important factor in the increase of retrieval performance. Another finding is that the optimal combination weighting of shot and story speech is fairly stable over a number of different language modeling smoothing methods and parameters. Furthermore, it does not seem to make a difference whether manually or automatically generated story boundaries are used. Altogether we conclude that story speech can form a valuable contribution to the text-based component of multimedia news retrieval.

Our experiments here were restricted to the linear combination of shot and story speech. In future work we plan to perform a comparison between this approach and window-based approaches such as those described in Section 2. Also, temporally biased language models have been shown to improve retrieval effectiveness in the setting of standard textual news documents [10], but their role in news videos remains relatively unexplored. We plan to investigate whether the use of temporal shot weighting can provide gains in retrieval performance. Another important avenue of research is the integration of these speech-based results with multimodal features.

7. ACKNOWLEDGMENTS

We would like to sincerely thank Tao Tao, Xuanhui Wang, and Ork de Rooij for their valuable discussions and coding help. 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.

8. REFERENCES

[1] M. Campbell, A. Haubold, S. Ebadollahi, M. R. Naphade, A. P. Natsev, J. R. Smith, J. Tesic, and L. Xie. IBM research TRECVID-2006 video retrieval system. In *TREC Video Retrieval Evaluation Proceedings*, 2006.

[2] S.-F. Chang, W. Hsu, W. Jiang, L. Kennedy, X. Dong, A. Yanagawa, and E. Zavesky. Columbia University TRECVID-2006 video search and high-level feature extraction (draft). In *TREC Video Retrieval Evaluation Proceedings*, 2006.

[3] T.-S. Chua, S.-F. Chang, L. Chaisorn, and W. Hsu. Story boundary detection in large broadcast news video archives: techniques, experience and trends. In *MULTIMEDIA*, pages 656–659, New York, NY, USA, 2004. ACM Press.

[4] T.-S. Chua, S.-Y. Neo, Y. Zheng, H.-K. Goh, Y. Xiao, S. Tang, and M. Zhao. TRECVID 2006 by NUS-I2R. In *TREC Video Retrieval Evaluation Proceedings*, 2006.

[5] A. Hauptmann, M. Christel, R. Concescu, J. Gao, Q. Jin, W.-H. Lin, J.-Y. Pan, S. M. Stevens, R. Yan, J. Yang, and Y. Zhang. CMU Informedia's TRECVID 2005 Skirmishes. In *TREC Video Retrieval Evaluation Proceedings*, 2005.

[6] A. Hauptmann, R. Yan, Y. Qi, R. Jin, M. Christel, M. Derthick, M.-Y. Chen, R. Baron, W.-H. Lin, and T. D. Ng. Video classification and retrieval with the Informedia digital video library system. In *Text REtrieval Conference*, 2002.

[7] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, 2001.

[8] W. H. Hsu and S.-F. Chang. Visual cue cluster construction via information bottleneck principle and kernel density estimation. In *CIVR*, pages 82–91, 2005.

[9] B. Huurnink and M. de Rijke. Term selection and query operations for video retrieval. In *Proceedings 29th European Conference on Information Retrieval (ECIR 2007)*, LNCS, 2007.

[10] X. Li and W. B. Croft. Time-based language models. In *CIKM*, pages 469–475, 2003.

[11] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281, New York, NY, USA, 1998. ACM Press.

[12] M. F. Porter. An algorithm for suffix stripping. In *Readings in information retrieval*, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[13] A. Singhal and F. Pereira. Document expansion for speech retrieval. In *SIGIR*, pages 34–41, New York, NY, USA, 1999. ACM Press.

[14] A. Smeaton, P. Over, and W. Kraaij. TRECVID: Evaluating the effectiveness of information retrieval tasks on digital video. In *ACM Multimedia*, New York, USA, 2004.

[15] J. R. Smith, S. Srinivasan, A. Amir, S. Basu, G. Iyengar, C.-Y. Lin, M. R. Naphade, D. B. Ponceleon, and B. Tseng. Integrating features, models, and semantics for TREC video retrieval. In *Text REtrieval Conference*, 2001.

[16] C. G. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM International Conference on Multimedia*, pages 421–430, Santa Barbara, USA, October 2006.

[17] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In R. C. Moore, J. A. Bilmes, J. Chu-Carroll, and M. Sanderson, editors, *HLT-NAACL*. The Association for Computational Linguistics, 2006.

[18] J. Yang, M. Chen, and A. Hauptmann. Finding person X: Correlating names with visual appearances. In *CIVR*, pages 270–278, 2004.

[19] J. Yang and A. G. Hauptmann. Exploring temporal consistency for video analysis and retrieval. In *MIR*, pages 33–42, 2006.

[20] J. Yang and A. G. Hauptmann. Exploring temporal consistency for video analysis and retrieval. In *MIR*, pages 33–42, New York, NY, USA, 2006. ACM Press.

[21] M. Yeung, B. Liu, and B.-L. Yeo. Extracting story units from long programs for video browsing and navigation. In *ICMCS*, page 0296, Washington, DC, USA, 1996. IEEE Computer Society.

[22] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.