IDIAP RESEARCH REPORT

Effect of Segmentation Method on Video Retrieval Performance

David Grangier [1] Alessandro Vinciarelli [2]

IDIAP–RR 04-83

December 2004

[1] IDIAP, CP 592, 1920 Martigny, Switzerland, grangier@idiap.ch
[2] IDIAP, CP 592, 1920 Martigny, Switzerland, vincia@idiap.ch

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny − Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 04-83

Effect of Segmentation Method on Video Retrieval Performance

David Grangier Alessandro Vinciarelli

December 2004

Abstract. This paper presents experiments that evaluate the effect of different video segmentation methods on text-based video retrieval. Segmentations relying on modalities like speech, video and text or their combination are compared with a baseline sliding window segmentation. The results suggest that even with the sliding window segmentation, acceptable performance can be obtained on a broadcast news retrieval task. Moreover, in the case where manually segmented data are available for training, the approach combining the different modalities can lead to IR results close to those obtained with a manual segmentation.

1 Introduction

The problem of video retrieval is becoming increasingly important with the availability of large multimedia databases in various domains (e.g. broadcast news archives, video conference databases, meeting recordings). An effective approach to this problem relies on the use of texts that can be automatically extracted from the original data [10]: the video data are first transcribed into a continuous stream of words (using Automatic Speech Recognition, ASR or Optical Character Recognition, OCR), the transcription is then segmented into smaller units, called documents and the resulting textual documents are used by the retrieval system to access the original media. This approach has two main advantages: the extracted texts are appropriate for retrieval as they are related to the semantic content of the video and the use of a text retrieval system allows one to benefit from previous works in this well-established research domain [1].

This work focuses on the effect of the segmentation step: different segmentation techniques are compared in the context of video retrieval. These techniques are based on the detection of different audio, video or text cues that possibly indicate a topic change (e.g. segmentations based on speaker change, video shot transition or vocabulary change). Our experiments are performed over TRECVid corpus [9] which is, to our knowledge, the largest annotated video database available to the research community. The video corpus ($\sim$ 110 hours broadcast news recordings) is segmented according to each technique and their retrieval performances are compared over a retrieval task using TREC Spoken Document Retrieval (SDR) queries [4].

In order to compare retrieval results obtained with different segmentations, we use two alternative evaluation methodologies that corresponds to two possible scenari. In the first approach, a fully automatic system is considered: the system outputs a ranked list of video segments that the users can watch to find the information of interest. In the second approach, a semi-automatic system that requires more effort from the users is considered: the system outputs a ranked list of time pointers. In this case, each pointer indicates the presence of a possibly relevant video segment but does not provide the exact location of the segment boundaries. The users should hence browse the video data located around each top-ranked time pointer to identify the relevant segments.

The rest of this paper is organized as follows: section 2 presents the segmentations we evaluated, section 3 introduces the evaluation methodology, section 4 describes the experiments and the results, section 5 draws some conclusions.

2 Segmentations

The goal of a segmentation process is to identify document boundaries in a video recording. Ideally, each document should be short enough to be only about a single topic (i.e. for any query, a document is either entirely relevant or entirely non-relevant) while being long enough to allow the IR system to determine whether it is relevant or not.

The use of a sliding window (i.e. extracting l seconds of video every s seconds) is possibly the most simple method to segment the corpus. However, the video data contain several cues that might be more appropriate to identify document boundaries: audio (e.g. speech/non speech detection, speaker change detection), text (e.g. vocabulary change detection) and image information (e.g. shot transition detection) can indicate a topic change. In the following, different approaches relying on such cues are presented.

The audio signal conveys information that can be of interest to segment the corpus: speech/non speech transitions may indicate a change of topic as the speaker is likely to take a short break before speaking about a new topic, speaker changes might be a useful cue as well. To evaluate such segmentations, LIMSI [5] partitioning system is used: for speech/non speech segmentation, a Gaussian Mixture Model system is trained on labeled training data and each test recording is segmented using Viterbi decoding with a minimum duration constraint. For speaker segmentation, a maximum likelihood segmentation clustering process is performed: this algorithm does not need any training data, nor any prior knowledge about the number of speakers and moreover, it has been shown to lead to good results on broadcast data.

The transition between shots (i.e. sequences filmed by a single camera without interruption) can also be appropriate to determine document boundaries as the producer is likely to use shot transition to emphasize changes of topic. We thus perform a shot segmentation with CLIPS system which relies on the combination of three classifiers (cut detector, photographic flash detector and dissolve detector) [7]. The detected shot boundaries are then post-filtered to include a minimum duration constraint.

Vocabulary changes also supply useful information for segmentation: the set of specific terms which are repeated during the course of a given topic discussion is likely to be replaced by a different set when a topic change occurs. TextTiling algorithm [6] relies on this assumption and detects a topic change at a given point when the number of terms shared between left and right context of the candidate point is low. In this work, TextTiling is used over the ASR transcriptions (LIMSI [5]) of the video data.

In the preceding, different modalities (audio, image and text) are used individually to segment the video data. It might also be useful to combine different information. For that purpose, we used NUS segmenter [2]: a Hidden Markov Model (HMM) system is trained on the manually segmented TRECVid development set. The input features of the HMM consist of the outputs of three different classifiers (a shot category classifier based on audio/video feature, a location/scene change detector based on video features and a cue phrase detector based on ASR transcriptions).

The following section introduces the evaluation methodology we used to compare the effect of each of the above segmentation on video retrieval performances.

3 Evaluation Methodology

Our goal is to evaluate the effect of various segmentation methods on a retrieval task. For that purpose, the standard IR evaluation is not suitable: this methodology has been introduced in the context of digital text retrieval where the segmentation problem does not exist (e.g. a newspaper archive is already segmented in articles, a website is segmented in pages, etc) and it assumes that the same segmentation has been used for annotation (i.e. when the human assessor determine which documents are relevant to a given query) and for evaluation [1]. In our case, the human assessors have manually segmented the corpus to determine the relevant document boundaries while automatic segmentation techniques have been used for evaluation.

We hence introduce a different methodology that allows the comparison of retrieval results from different segmentations. Our approach evaluates a system which outputs a ranking of video segments. Ideally, the top-ranked segments should correspond to the relevant segments that have been identified by human assessors. Along the obtained ranking, we measure precision as the percentage of *retrieved time* that is actually relevant and recall as the percentage of *relevant time* that has been retrieved:

$$P = \frac{T_{rs}}{T_s} \text{ and } R = \frac{T_{rs}}{T_r}$$

where T_s is the amount of time retrieved by the system, T_r is the total amount of time that have been judged as relevant by human assessors and T_{rs} is the amount of time retrieved by the system that is actually relevant.

For a more complete evaluation, we also used a second method (initially introduced in the context of TREC SDR [4]) in which a system that outputs a ranking of time pointers is evaluated. Each pointer indicates the presence of a possibly relevant video segment but does not provide the segment boundaries. A pointer appearing at position n in the ranking is considered relevant if and only if it verifies the two following properties: it refers to a relevant segment (i.e. the pointer is located in a time segment identified as relevant by human assessors) and, in this case, this relevant segment has not been retrieved above in the ranking (i.e. no pointer appearing above position n refers to this segment). According to this method, precision and recall are defined as follows:

$$P = \frac{PT_{rs}}{PT_s} \text{ and } R = \frac{PT_{rs}}{N_r}$$

where PT_s is the number of pointers retrieved by the system, N_r is the number of time segments that have been judged to be relevant by human assessors and PT_{rs} is the number of pointers that refer to a relevant segment that has not been retrieved previously.

These two alternative evaluation methods corresponds to different application environments. In the first case, a fully automatic system is evaluated: the users only submit their query and watch the top-ranked video segment. In the second approach, a semi-automatic system is considered: after submitting their query, the users obtain a ranking of time pointers rather than segments which means that they should, in addition, browse the video data located around each top-ranked time pointer to identify the segment boundaries. The two methods are referred as *EvalTime* and *EvalPointer* in the following.

4 Experiments and Results

This section describes the experiments we perform and the results we obtain. Our goal is to compare video retrieval results when using different segmentation methods (see section 2). We use TRECVid 2003 corpus (American Broadcast News) and TREC SDR queries for that purpose. TRECVid corpus [9] is, to our knowledge, the largest annotated video corpus available ($\sim$ 110 hours). As our queries (from TREC SDR) have not been created for TRECVid data but for a larger audio corpus (TDT2 which includes TRECVid data), we remove the queries that do not have any relevant data in TRECVid corpus. Parameter tuning is performed using TREC8 queries over TRECVid development set while evaluation is performed using TREC9 queries over TRECVid test set (see table 1). The retrieval performances obtained using each segmentation (see section 2) are measured according to both *EvalTime* and *EvalPointer* methodologies (see section 3). Moreover, each method is compared to the baseline sliding window segmentation (win) according to Wilcoxon sign rank test. We also evaluate the I.R. performance obtained when using a manual segmentation (i.e. news story segmentation performed according to TDT guidelines [3]). The IR system used for evaluation is based on OKAPI formula [8] and uses LCA query expansion [11] (using Tipster as parallel corpus). The results obtained are presented in the following.

Both *EvalTime* (table 2) and *EvalPointer* (table 3) evaluations suggest that only NUS and manual segmentation are performing better (according to Wilcoxon sign test with 95% confidence) than the baseline sliding window segmentation. The good results of NUS and Manual might be due to the fact that both of them segment the video into news stories which have also been used to define the relevance judgments. This is confirmed by the fact that their advantage is less important for *EvalPointer* results: *EvalTime* counts every second as relevant or not according to the relevance judgment boundaries while *EvalPointer* only relies on time pointers (central point of the document) rather than exact boundaries which makes it less dependent on the mismatch between the segmentation and the relevance judgment boundaries.

When looking at the other segmentation techniques, shot and speaker segmentation are leading to worse results than win for *EvalTime* but not for *EvalPointer* which can highlight an over-segmentation problem: the segments that are detected as relevant are too short to account for a significant part of the relevant data, whereas in the case of *EvalPointer*, the time pointers can be

	Training	Test
Duration (min)	3390	3210
N. of queries	34	35
T_r (min)	8.1	9.3
N_r	5.7	6.6

Table 1: Corpus size, number of queries, average relevant duration per query, average number of relevant pointers

Segmentation	AvgP (%)	Wilcoxon test
Win	26.5	-
Speech/Non sp.	27.2	same as win
Speaker	19.6	worse than win
TextTiling	28.0	same as win
Shot	11.6	worse than win
NUS	35.4	better than win
Manual	44.8	better than win

Table 2: Average Precision (*EvalTime*)

Segmentation	AvgP (%)	Wilcoxon test
Win	21.6	-
Speech/Non sp.	24.2	same as win
Speaker	22.6	same as win
TextTiling	23.5	same as win
Shot	25.3	same as win
NUS	28.4	better than win
Manual	35.9	better than win

Table 3: Average Precision (*EvalPointer*)

useful to identify the location of relevant segments. This is further confirmed when looking at average document length: shot and speaker segmentation are leading to the shortest documents (respectively $6s$ and $19s$ on average while manual and sliding window segmentation are $66s$ and $100s$ respectively). The TextTiling results are not significantly better than win which was not expected as this technique has been successfully used to segment digital texts into passages about different topics [6], which is appropriate for retrieval. The text data extracted from broadcast news might be too short to detect a significant amount of repeated terms on which the algorithm relies (see section 2). The presence of ASR recognition errors might further emphasize this problem since it can lead to detect fewer repetitions than there actually are in the data.

When the top positions of the ranking are evaluated according to *EvalTime* (table 4), the results show that the use of the IR system allows the user to identify part of the relevant data with little manual effort: for any segmentation technique, the user can identify more than 1.5 min of relevant material while looking at 5 min of video (which is small compared to the corpus size, 53.5 hours). The same kind of conclusions can be drawn when looking at *EvalPointer* results (table 5): for any segmentation technique, there is at least one pointer to a relevant document in the top 5 pointers.

These results seem to outline that a text retrieval system, although developed for manually segmented digital text data, can be useful even when using a simple sliding window segmentation. This suggests that the sliding window segmentation which requires little tuning (only two parameters: window length and shift) could be suitable for several types of data for which no manually segmented data are available to train a more complex automatic segmentation tool, like NUS (e.g. meeting recordings).

5 Conclusions

In this paper, we have presented experiments to evaluate different segmentation methods in the context of video retrieval. In text-based video retrieval systems, segmentation plays a central role: the video corpus is first transcribed into text data (using ASR and/or OCR), this transcription is then segmented into documents and the resulting textual documents are then indexed by a text IR system that gives the possibility to access the video data corresponding to a given text query. The segmentation should

Segmentation	Ptop 5 min	Wilcoxon test
Win	30.6	-
Speech/Non sp.	29.8	same as win
Speaker	26.7	same as win
TextTiling	31.7	same as win
Shot	18.5	worse than win
NUS	39.7	better than win
Manual	43.7	better than win

Table 4: Precision at top 5 min (*EvalTime*)

Segmentation	Ptop 5 pointers	Wilcoxon test
Win	19.4	-
Speech/Non sp.	22.9	same as win
Speaker	23.4	same as win
TextTiling	21.7	same as win
Shot	24.0	same as win
NUS	28.0	better than win
Manual	32.6	better than win

Table 5: Precision at top 5 pointers (*EvalPointer*)

lead to documents that allow the IR system to identify whether a text segment is relevant or not for any given query.

Four segmentations relying on different criteria have been performed (speaker, speech/non-speech, video shot and text based segmentations). A technique combining different modalities has also been evaluated (NUS HMM system). All these approaches have been compared to a baseline sliding window segmentation.

The retrieval performance of each segmentation has been evaluated by performing a retrieval task (TREC SDR queries on TRECVid broadcast news data) and using IR measures that have been modified to take into account the segmentation problem. The results suggest that the multimodal system (NUS) is leading to the best IR performance with results close to those obtained with the manual segmentation. However, the training of this system requires that a part of the corpus has been manually segmented into topics. In absence of such expensive data, the sliding window segmentation seems to be an appropriate solution: although simple (only two parameters, window length and shift, should be tuned), this method leads to results similar to those obtained with segmentations based on audio, video or text alone.

This work has focused on a broadcast news corpus and it would be interesting, as a future work, to perform the same type of experiments on less structured data like meeting recordings.

6 Acknowledgments

This work is supported by the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2).

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] L. Chaisorn, T.-S.Chua, C.-K. Koh, Y. Zhao, H. Xu, and H. Feng. A two-level multi-modal approach for story segmentation of large news video corpus. In *TRECVid Workshop*, 2003.

[3] C. Cieri, D. Graff, M. Liberman, N. Martey, and S. Strassel. The TDT-2 text and speech corpus. In *DARPA Broadcast News Workshop*, 1999.

[4] J.S. Garofolo, G.P. Auzanne, and E.M. Voorhees. The TREC SDR track: A success story. In *Content-Based Multimedia Information Access Conf.*, 2000.

[5] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Comm.*, 37, 2002.

[6] M. Hearst. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Comp. Linguistics*, 23, 1997.

[7] G. Quenot. TREC-10 shot boundary detection task: Clips system. In *TREC*, 2001.

[8] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *TREC*, 1994.

[9] A. Smeaton, W. Kraaij, and P. Over. TRECVid 2003: an introduction. In *TRECVid Workshop*, 2003.

[10] T. Westerveld, A. P. De-Vries, A. Van-Ballegooij, F. De-Jong, and D. Hiemstra. A probabilistic multimedia retrieval model and its evaluation. *J. on Applied Signal Proc.*, 2, 2003.

[11] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *SIGIR*, 1996.