University of Sheffield at TRECVID 2006 High-level Feature Extraction

Siripinyo Chantamunee Yoshihiko Gotoh
Department of Computer Science, University of Sheffield, UK
{s.chantamunee, y.gotoh}@dcs.shef.ac.uk

Abstract

We present our approach to TRECVID 2006, high-level feature extraction task. We submitted one run with type 'A', annotating all required 39 features. The approach was based on textual information extracted from speech recogniser and machine translation outputs. They were aligned with shots and associated with high-level feature references. A list of significant words was created for each feature, and it was in turn utilised for identification of a feature during the evaluation. In this notebook paper, we describe the approach and the results we obtained. We also describe the problems we encountered during the system development, some of which were critical to the system performance.

1 Introduction

We participated in TRECVID 2006 as a part of the University of Glasgow team[1]. The workshop annually promotes challenging tasks on content-based information retrieval [1]. The following four tasks were run this year: shot boundary determination, high-level feature extraction, search, and rush exploitation. This paper describes the overview of our work on the high-level feature extraction task.

Our approach was aiming at extraction of relevant features based on outputs from automatic speech recognition (ASR) and/or machine translation (MT) systems, underpinned by the mature progress made in the area of text and speech processing. Our assumption was that textual data often carried very important information that described the corresponding video shots. However, ASR and MT systems were still far from human's level, and we were interested to see if ASR errors and translation errors could cause any reduction in performance for the high-level feature extraction task.

The approach was benefitted by the ASR and MT dataset provided. We also utilised shot boundary reference in order to segment video into shot-based units. Shots were then aligned with text from ASR and MT systems. Finally, the feature reference was used to build a list of significant words for that feature. We submit one run with type 'A', that annotated all of 39 features using the approach outlined above. It was evaluated by NIST and the results were returned using the inferred average precision [3].

2 Approach

Outputs from ASR and MT systems were rich information sources. It was hoped that, by associated them with feature annotation and shot boundary reference, we would be able to identify many of, if not all, video shots relating to the given features without relying on other modalities. Figure 1 illustrates the architecture of the system. It consisted of several stages — broadly, data pre-processing stage and feature extraction stage (for training the system with 2005 data); the latter was paired with testing stage (with 2006 data). Finally the evaluation was performed by NIST.

2.1 Data Pre-Processing

Data pre-processing was concerned with extraction of textual attributes. The textual descriptors were provided however, they required some pre-processing to put them together, partially due to differences in formatting. ASR and MT data were aligned with shot units by employing speaker time and the shot boundary reference (referred to

[1]We would like to thank the University of Glasgow for kindly letting us be in their team.

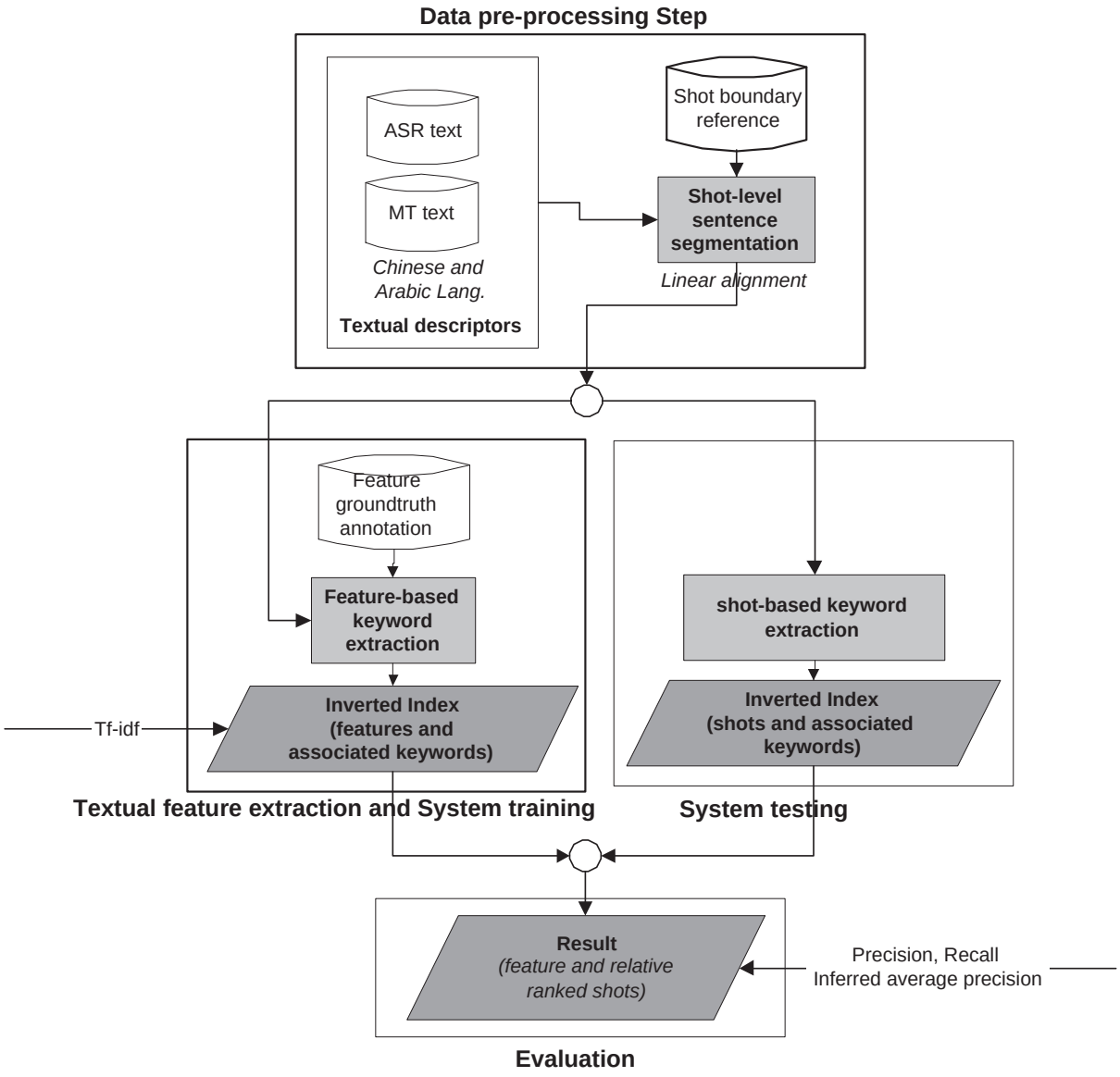

Figure 1: Architecture for high-level feature extraction system.

as 'shot-level sentence segmentation'). It was followed by identification of the most significant words occurred in shots that were labelled with high-level features ('feature-based keyword extraction').

ASR and MT outputs. The ASR transcripts and translations from Chinese and Arabic sources were provided. Time stamp was used to align words to each of individual shots. Stop words were removed and stemming was preformed. We encountered several problems. Firstly, the MT texts did not always correspond to the most relevant video scenes. In some cases, a portion of translations or ASR transcripts was lost from the data provided. Not surprisingly, there were shots without any textual descriptors. In the current implementation, these shots could not be processed. We are considering the use of textual information from adjacent shots in order to alleviate the problem. Information from adjacent shots may also be useful for refining the list of the most significant words.

Common shot boundary reference. The shot boundary reference was released by the TRECVID organiser. The news story is considered as a concatenation of individual video portions. The frames within one motion-camera normally describe the same story. A story may be produced by including all frames from one continuous unit of video. Therefore, shot-level segmentation can provide a reasonable structure for the contents of video.

Feature annotation. Using the feature annotation, we should be able to identify shots that describe the features. The annotation for the TRECVID 2005 data was provided by the MedialMill team [2]. 101 features were annotated for 169 hours of Arabic, Chinese and the US broadcast news, out of which 39 features were involved in this year's

task [1]. A number of shots is extracted for each feature and associated with ASR and MT texts using time stamp information. We realised that there existed shots that did not match the annotated feature. This had caused very serious effect on the performance of the system.

2.2 Textual Feature Extraction

For each word, the *tf-idf* score was calculated. The procedure produced a ranked list of the most significant words for individual high-level features. We found 6 297 significant words for 39 features (161 words per feature on average)[2]. Note that we examined the use of subsets (say, 70% or 85%) instead of using the complete set of significant words for the testing. It was found that there were not significant difference in terms of precision and recall. In practice, a subset might have been be sufficient because it could save space and the processing time.

3 Experiment

3.1 Experimental Design

We derived a list of the most significant words from TRECVID 2005 data, using the annotation of high-level features, produced by the MedialMill team [2], as the reference. ASR transcripts and MT texts were aligned with corresponding shots and the standard textual feature extraction techniques were applied. For evaluation the TRECVID 2006 dataset was utilised. It comprised of 158.6 hours of video in three languages including English, Chinese, and Arabic.

We completed a single run for all of 39 high-level features, using the text based system described earlier. First, occurrences of significant words were examined in shot units. When the extracted words were significant enough, shots were associated with one of high-level features. The final result was a list of ranked shots classified by individual features. The run was an 'A' type, and referred to as 'A_Glasgow.Sheffield01_1'.

3.2 Results and Discussion

Our submission was evaluated by NIST using the inferred average precision. Figure 2 shows the results that compare our scores with minimum, median and maximum scores. On average, our submission resulted in precision for 2000 shots at 0.0119 and for 100 shots at 0.0480. The 475 shots were identified correctly out of 9074 groundtruths. As the result, the inferred average precision was calculated as 0.005.

Problem caused by the erroneous annotation of high-level features. As noted earlier, we noticed that, for TRECVID 2005 data, there existed a number of shots that did not match the annotated high-level features. This has caused a serious effect on our system. We are still investigating the extent of this problem.

Problem caused by news contents. The system was developed from TRECVID 2005 data, and then applied to 2006 data. Because the system relied on occurrences of particular sets of words, changes in news contents from 2005 to 2006 certainly has some effect on the performance.

Problem caused by alignment. Time stamps were utilised to align ASR and MT text to shot segments. Our assumption was that, within a shot, significant words would occur that described that particular shot. Clearly, this assumption was not quite correct. There were many occasions that some words could be strongly related to the next or the previous shot. For example, there were cases whereby anchors appeared in a studio shot was talking about the contents of a report in the next shot. We are currently experimenting the alignment using the speaker information.

Problem caused by the number of features. We have applied the same approach to all of 39 high-level features. The question is — would it be possible to apply a single scheme to many different kinds of features? Clearly, we might be able to achieve better by focusing on one particular feature at the cost of the rest of features. But that luxury cannot always be expected. For the current submission, we developed a system solely based on textual information. It is likely that the overall performance would be improved by combining multiple approaches in the multiple modalities, and now we are looking at this direction.

4 Conclusions

We presented our first attempt for TRECVID high-level feature extraction task using information derived from ASR and MT data. We submitted one run for 39 features, from which 20 features were scored by NIST. Dur-

[2] Stemming and stopping were applied at the earlier stage.

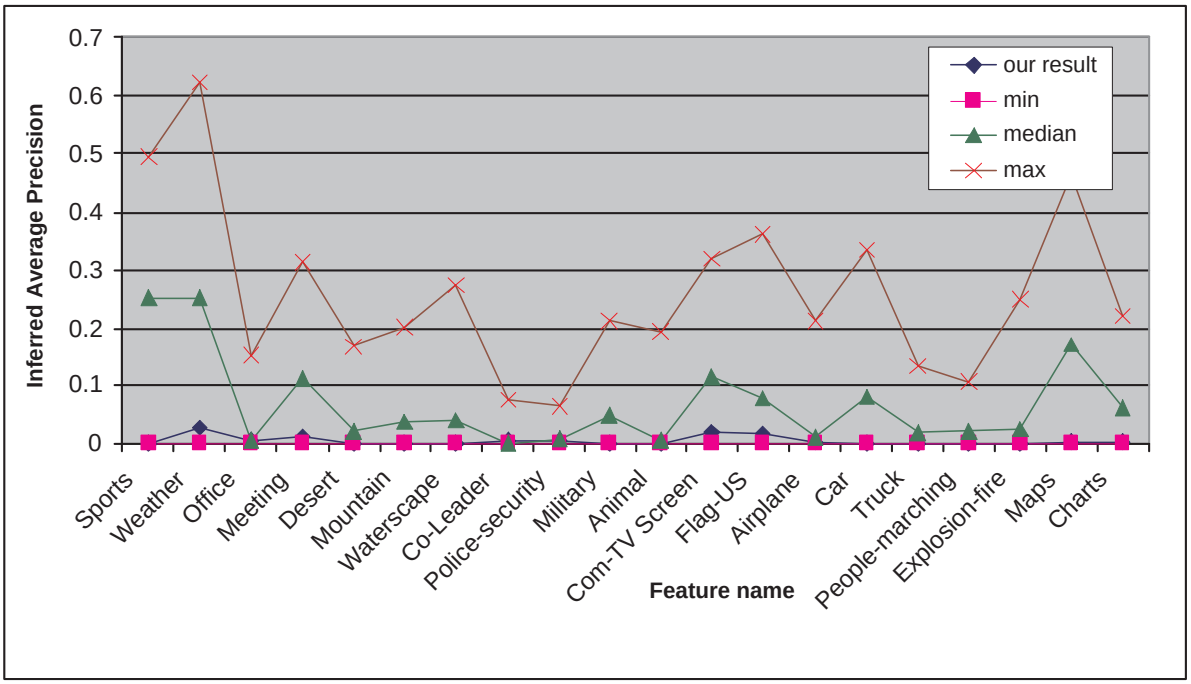

Figure 2: The inferred average precision scores for selected 20 features.

ing the system development, we have encountered several problems, some of which were critical to the system performance. We are currently analysing the results obtained, aiming at further development in the area.

References

[1] NIST. *2006 TREC Video Retrieval Evaluation*, [Online] Available http://www-nlpir.nist.go/projects/tv2006. September, 2006.

[2] Cees G.M. Snoek, M. Worring, Jan C. van Gemert, J. Geusebroek, and Arnold W.M. Smeulders. *The challenge problem for automated detection of 101 semantic concepts in multimedia*. In Proceedings of the ACM International Conference on Multimedia, Santa Barbara, USA, October 2006.

[3] E. Yilmaz, and Javed A. Aslam. *Estimating Avarage Precision with Incomplete and Imperfect Judgments*. In Proceedings of the fifteenth ACM International Conference on Information and Knowledge Management (CIKM). November, 2006.

[4] C. Petersohn. *Fraunhofer HHI at TRECVID 2004: Shot Boundary Detection System*. In TREC Video Retrieval Evaluation Online Proceeding. TRECVID. 2004.