Chinese University of Hong Kong at TRECVID 2006: Shot Boundary Detection and Video Search

Steven C. H. Hoi[1], Lawson L. S. Wong[2], Albert Lyu[3]

[1]Dept. of Computer Science & Engineering,
Chinese University of Hong Kong
chhoi@cse.cuhk.edu.hk

[2]Dept. of Computer Science
Stanford University
lsw@stanford.edu

[3] Year 12
Chinese International School
albertworld@gmail.com

Abstract

In this paper, we describe our methodologies and empirical evaluations for the shot boundary detection and automatic video search tasks at TRECVID 2006. For the shot boundary detection task, we consider a simple and efficient solution. Our approach first applies adaptive thresholding on color histogram differences between frames to select candidates for shot boundaries, then runs several tests to dismiss less likely ones. These tests proved to be too finely tuned to the TRECVID 2005 shot boundary detection task data, producing mediocre results on the 2006 data set. For the video search task, we propose a novel multimodal and multilevel ranking scheme for video ranking. Different from traditional ranking schemes, most of them are based on supervised approaches, our approach suggests a semi-supervised ranking method which can exploit both labeled and unlabeled data effectively for the ranking task. At the meantime, our multilevel approach makes the semi-supervised ranking method efficient for large-scale search task in practice. We will evaluate the empirical performance of our approaches and give comments of our solution.

1. Introduction

In TRECVID 2006, our group (Chinese University of Hong Kong) participated in two tasks, i.e. shot boundary detection and automatic video search. For the shot boundary detection task, a simple and efficient method is considered in this year's solution. The main idea is based on adaptive thresholding on color histogram differences between frames. Our approach is simply based on color histogram features, varying the combination of tests and parameters used in the post-detection stage to remove likely false positives. Though these methods produced satisfactory results on the TRECVID 2005 data, there was a significant drop in the performance on this year's data.

For the automatic video search task, we proposed a novel multimodal and multilevel ranking solution in order to rank video shots effectively and efficiently in the video search task. The main idea is to model the video data structure and search task by graphical models. Based on the graphical models, we can fuse the information from multiple modalities effectively over the graphs in the search task. We then address the common challenge in video retrieval task, i.e., small sample learning problem. To attack this challenge, we formulate the ranking task by semi-supervised learning in order to exploit both labeled and unlabeled data. Since semi-supervised ranking method may not computational prohibited for large-scale problems, we then propose a multilevel ranking solution, which makes the semi-supervised ranking method efficient for large-scale retrieval task.

The rest of this paper is organized as follows. Section 2 presents our methodology and empirical evaluation in the shot boundary detection task. Section 3 gives our approach for automatic video search task and also the empirical evaluations. Section 4 concludes our work.

2. Shot Boundary Detection

This section describes the shot boundary detection tools that we applied. Difference measures, adaptive thresholding candidate selection, and false positive detection methods are discussed.

2.1 Difference Measures

To produce the color histograms, we tried using both the RGB and HSV color spaces. This choice however did not appear to be significant from testing results. The gray-level histogram was eventually used as it is relatively fast and produced most of the better results.

We experimented with Euclidean distance, color moment, and Earth Mover's Distance (EMD) measures to calculate color differences between frames. The former two performed rather poorly as they were prone to being under-sensitive to true positives but over-sensitive to false-positives. As a result, they achieved average recall and precision rates of ~0.5 when tested on the 2005 data. The EMD method, however, was able to produce better results, as it was sensitive to most transition-like changes. Though it also produced more noise than the other two measures, this was not problematic when adaptive thresholding was applied.

2.2 Adaptive Thresholding

False positive cases which involve camera/object movement or flashes often give unsmooth, noisy EMD data. To distinguish such cases from the 'cleaner' true transitions, we used adaptive thresholding, taking into account the mean and standard deviation of EMD values in the neighborhood of peaks [1]. For every 11-frame window, the peak EMD value is taken to be a shot boundary candidate, and the following threshold is applied:

$$T_{adaptive} = T_{mean} * \max(\mu_l, \mu_r) + T_{sd} * \max(\sigma_l, \sigma_r) , \qquad (1)$$

where T_{mean} and T_{sd} are threshold multiplying factors. Through experimentation, it is optimal for T_{mean} and T_{sd} to be greater than 2 and 5 respectively; any larger values only adjust the recall and precision rates slightly with no overall improvement. In noisy cases, both the mean and standard deviation tend to be rather high, and it is unlikely that a random peak in such cases can exceed this threshold. However, true cut EMD values tend to be much higher than this threshold, and so adaptive thresholding works very well for short transitions, reaching recall and precision rates of about 0.9 on the TRECVID2005 data.

For long gradual transitions (GTs), however, peak values tend to be rather low, and neighboring values are also rather high, which make them appear like false positive noisy cases. Very low threshold factors had to be used to select candidates, greatly reducing the power of adaptive thresholding. We therefore had to use another collection of techniques to remove less likely long gradual candidates from the results pool.

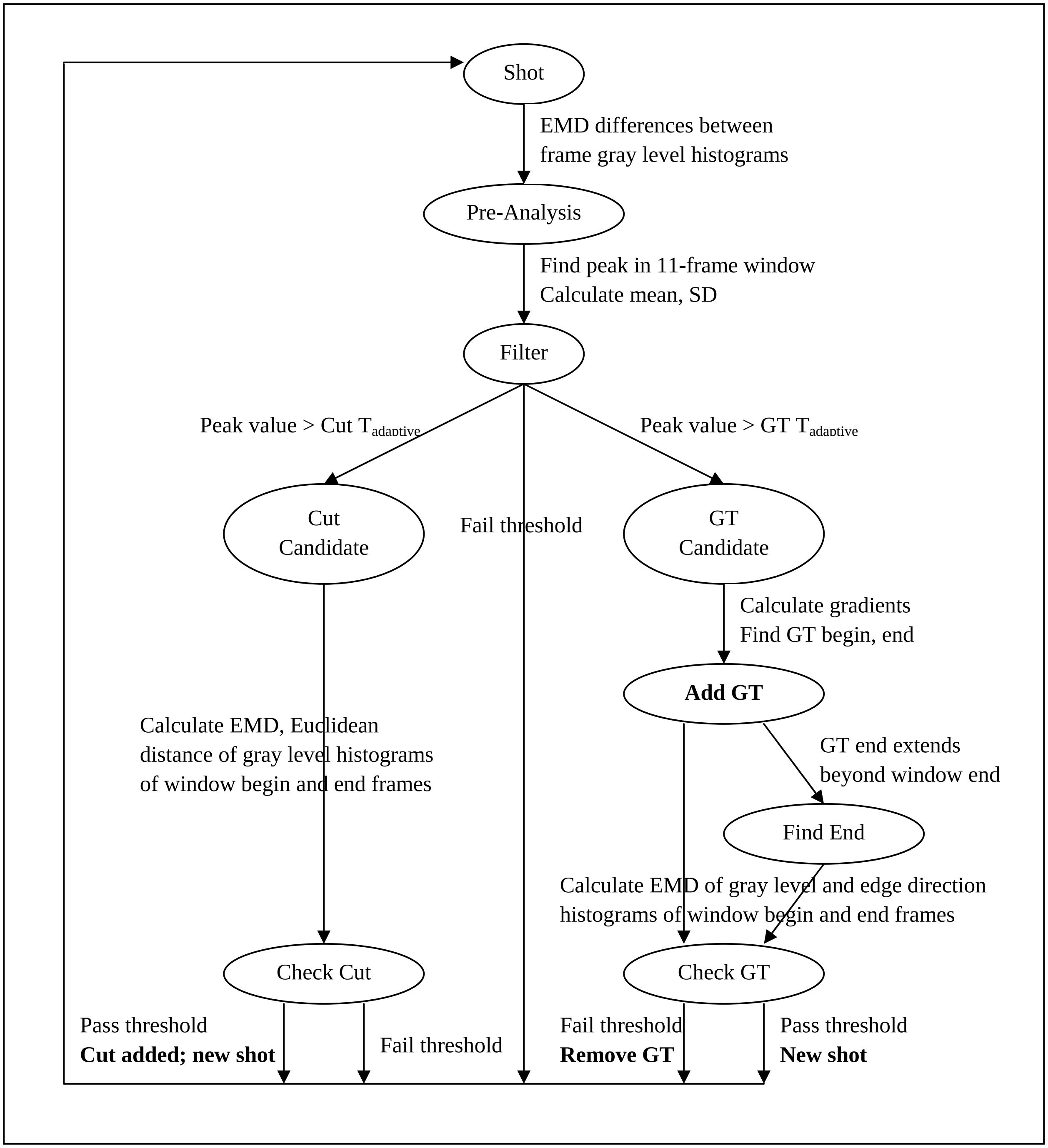
Shot
EMD differences between frame gray level histograms
Pre-Analysis
Find peak in 11-frame window
Calculate mean, SD
Filter
Peak value > Cut T_adaptive
Peak value > GT T_adaptive
Cut Candidate
Fail threshold
GT Candidate
Calculate gradients
Find GT begin, end
Add GT
Calculate EMD, Euclidean distance of gray level histograms of window begin and end frames
GT end extends beyond window end
Find End
Calculate EMD of gray level and edge direction histograms of window begin and end frames
Check Cut
Check GT
Pass threshold
Cut added; new shot
Fail threshold
Fail threshold
Remove GT
Pass threshold
New shot

Figure 1. Shot boundary detection algorithm

2.2 False Positive Detection

A distinguishing characteristic between cuts, long GTs, and false positives is the smoothness of their EMD values across time. Cuts appear similar to delta functions, with almost zero values around a very high peak. Long GTs, especially those that change regularly (dissolves, wipes) have smooth gradients. Finally, false positives, mainly comprising of random motion or object appearances, have fluctuating values, often creating large jumps. The gradient around a peak can therefore be used to distinguish the transition type, dismissing it in the case of false positive behavior. Since cuts have already been treated by adaptive thresholding, it is possible to use the gradient solely to determine whether a candidate is more likely to be a gradual transition or false positive based on the gradient on both sides of the peak value. In fact, few GTs are lost this way, while a significant number of false positives are removed. However, this still does not exclude a majority of more structured false positives such as rapid camera rotation.

Another basic check that proved useful is to calculate the difference between two frames before and after the candidate transition. Assuming that the boundaries of the transition are determined correctly, the frames at each boundary should have a relatively large difference value. Since there are only very few candidates compared to frames in the video, more complex tests can be applied without consuming too much overall time. Both the EMD and Euclidean distance measures for color histogram difference are used for cut candidates; as for GT candidates, the EMD and edge histogram are used. As these methods involve additional features from the videos, a second pass is required. Although these false positive detection methods also eliminated some true positives, in particular the more color-invariant candidates, on average only one true positive is lost for every three false negatives eliminated.

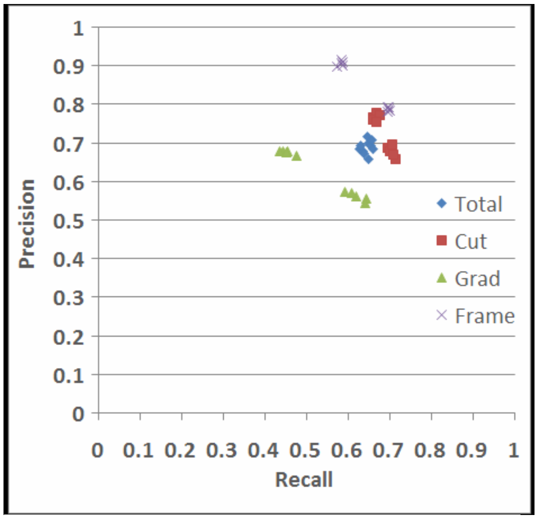

Figure 2. TRECVID 2006 Results

2.3 Overall Algorithm

Our shot boundary detection algorithm is outlined in Figure 1 below. As described previously, candidates are first selected using adaptive thresholding. In the case of a GT candidate, the gradients of the EMD values around its peak are calculated to find the beginning and end of shot. Systematic checks are finally applied to the candidates to filter out likely false positives.

2.4 Experimental Results

We submitted 10 runs, each applying the above algorithm with slightly different parameters for adaptive thresholding and GT gradient methods. The evaluation result is listed in Table 1. The odd and even-numbered groups differ significantly; however, within each group the results vary very little. The odd/even distinction is caused by the GT gradient threshold; even-numbered runs used a lower threshold than odd-numbered ones, hence allowing more GT candidates and creating a significantly higher GT/frame recall, while also pulling down the GT/frame precision. In contrast, changing the adaptive thresholding parameters, which creates differences within the groups, had little effect as seen by the insignificant differences within the groups. This was expected as all values were above the minimal values of $T_{mean} = 2$, $T_{sd} = 5$ (see Section 2.2).

Table 1. Our Shot Boundary Detection Results at TRECVID 2006

Run	**Total**			**Cuts**			**Graduals**			**Frame Accuracy**		
	Rcl	Prc	F_1	Rcl	Prc	F_1	Rcl	Prc	F_1	Rcl	Prc	F_1
tv6cuhk1	0.648	0.700	0.673	0.659	0.767	0.709	0.620	0.560	0.588	0.697	0.787	0.739
tv6cuhk2	0.627	0.684	0.654	0.695	0.686	0.690	0.444	0.677	0.536	0.586	0.909	0.713
tv6cuhk3	0.647	0.715	0.679	0.667	0.778	0.718	0.591	0.572	0.581	0.695	0.793	0.741
tv6cuhk4	0.649	0.659	0.654	0.713	0.657	0.684	0.476	0.668	0.556	0.586	0.901	0.710
tv6cuhk5	0.655	0.693	0.673	0.660	0.760	0.706	0.643	0.557	0.597	0.699	0.784	0.739
tv6cuhk6	0.634	0.678	0.655	0.700	0.677	0.688	0.455	0.679	0.545	0.585	0.907	0.711
tv6cuhk7	0.657	0.709	0.682	0.675	0.772	0.720	0.608	0.569	0.588	0.696	0.792	0.741
tv6cuhk8	0.631	0.694	0.661	0.705	0.697	0.701	0.434	0.677	0.529	0.584	0.915	0.713
tv6cuhk9	0.660	0.685	0.672	0.668	0.755	0.709	0.640	0.544	0.588	0.694	0.781	0.735
tv6cuhk10	0.638	0.672	0.655	0.708	0.671	0.689	0.451	0.676	0.541	0.573	0.898	0.700

Overall, the results dropped by about 0.1 for both recall and precision compared to the tests run on the 2005 training data. This reflects that some major sources of error existed in the 2006 data that were not anticipated from training results. Cuts suffered the most, dropping from a 0.9 recall and precision level mainly due to a one-frame shift on many two-frame short GTs, and for each such error created both an insertion and deletion. Almost a half of the cut errors arose from this, and a post-evaluation run which only corrected this boosted both the recall and precision rates well over the 0.8 mark. Another major source of errors is the confusion between cuts and GTs, either by concatenating several close cuts and short GTs together, or by splitting a long GT into several cuts. These occur when there is rapid motion between frequent short transitions and when there is unsmooth changing in a long GT respectively. Like the cut errors, each of these problems induce multiple insertions and deletions, so although they do not occur as regularly, about one quarter of all errors arise due to such transition type confusion.

Apart from algorithmic errors listed above, the other major source of error arises from the choice of difference measure used to determine whether there is a shot transition between two frames. For our experiments, we used the EMD between frame color histograms almost as the sole feature, which is quite limited despite being relatively efficient. While color change is a good indication of scene changes, it is particularly prone to creating false positives in shots with rapid motion (including people walking by directly in front of the camera) or animated logos (which may occur many times in a television news clip); on the other hand, only detecting color changes also misses out a significant number of scene transitions with little color change, such as where the background changes very slightly. An expansion of techniques is necessary to detect such cases, either by including more specific tools such as logo detectors, or by changing the approach entirely such as using SVM classification. We would like to attempt these other techniques in the future, while also developing other novel approaches.

3. Automatic Video Search

3.1 Overview of Video Search Tasks

In the video search tasks, we focus our attention on the automatic search task without human interactions or manual information. This is the most challenging and fundamental task in video search tasks. Specifically, there are several well-known difficulties in video search tasks.

First of all, the video search task is a multimodal search problem, i.e., a task of combining information from multiple resources for ranking video shots given some query topic. In

TRECVID, query of each search task usually contains both text and visual examples. This multimodal ranking problem is a long-term challenging issue in video search tasks.

Another problem in video search tasks is the small sample learning problem. Typically, only a few visual examples are provided in each query topic. This poses a challenging issue of learning ranking function with limited number of examples in the search tasks. Many ranking approaches based on machine learning algorithms may suffer significantly from the insufficient training examples.

Moreover, a video search task can be computationally intensive since the dataset is usually of large-scale. This is challenging for a real-world video search application, particularly if complicated machine learning algorithms are employed in the video ranking solution. It is important to develop an efficient solution for large-scale problems. In addition to these difficulties, other common issues, such as the short query problem, noisy text data, and semantic gaps also pose a lot of challenges for an automatic video search task.

To tackle these challenges in a unified solution, we proposed a multimodal and multilevel ranking framework for video search tasks. The main idea is to solve the multimodal ranking problem by graph based ranking methodology, which is able to fuse information from multiple resources smoothly in a probabilistic scheme. Further, to tackle the small sample learning problem, we suggested a semi-supervised ranking method using semi-supervised learning techniques for learning ranking functions on both labeled and unlabeled data. To make our solution efficient for large-scale problems, we design the ranking scheme using a multilevel ranking solution. We will explain details of our solution as follows.

3.2 Multimodal and Multilevel Framework for Video Search

The multimodal and multilevel ranking framework is shown in Fig. 3. Basically, our framework comprises four different ranking stages:

(1) Text-based Ranking Stage

The text based retrieval approaches are usually more effective than visual based approaches from past experiences in TRECVID. Therefore, we consider the text-based ranking solution in the first ranking stage. For a text based ranking task, there are two challenging issues. One is the noisy texts, which are usually obtained from automatic speech recognition or OCR techniques. The other is the short query problem. To address these issues, pseudo relevance feedback (PRF) techniques are suggested in our approach to alleviate these challenges.

(2) Nearest Neighbour Reranking Stage

This is the most efficient way of combing visual information for multimodal video ranking tasks. For textual modality, we employ the normalized ranking scores from the text based ranking stage for computing the ranking scores. For visual modality, in which data are often given in vector space, we measure Euclidean distances between normalized training data examples and query data examples for similarity measures. Other more general Mahalanobis metrics may also be considered for better performance.

(3) Supervised Large Margin Reranking Stage

The third ranking stage is based on the supervised large margin reranking method, in which large margin learning methods are used for learning the ranking functions. In our current stage, we employ support vector machines (SVM) [2], the most well-known large-margin learning method with state-of-the-art performance, to learn the ranking function with visual examples in this stage. SVM usually is a binary classification method, which requires training data from both positive and negative classes. However, for the search task, only positive visual examples are provided initially. To address this problem, we consider the negative example from the list of most irrelevant examples ranked from the previous stage by NN ranking, which is also known as negative pseudo-relevance feedback studied in the multimedia community [5].

(4) Semi-Supervised Reranking Stage

The last ranking approach is the semi-supervised reranking method, which learns the ranking functions in exploiting both labeled and unlabeled data examples. The semi-supervised ranking approach is to attack the small sample learning problem, which is challenging for many learning algorithms. The semi-supervised ranking method, exploiting the unlabeled data information, can be more effective than large margin methods, although it may involve more computational cost. Since we consider only a small portion of training examples in the semi-supervised reranking stage, the semi-supervised ranking approach on a small-scale problem can still be accomplished efficiently based on our multilevel ranking solution.

In summary, we propose the novel multilevel ranking framework to learn multimodal ranking functions efficiently based on four different ranking stages using different learning strategies. In the first stage, the text-based ranking method to obtain a set of top M ranked video stories, which are associated with a set of N_1 video shots. In the second stage, the NN ranking method reranks the N_1 shots and outputs top N_2 most relevant video shots. In the third stage, the SVM ranking method reranks the N_2 shots and outputs top N_3 most relevant video shots. In the last stage, the SSR ranking method reranks top N_4 shots of SVM output results. Finally, the multilevel ranking framework returns top k shots for performance evaluation. It is clear that $N_1 > N_2 > N_3 > N_4$.

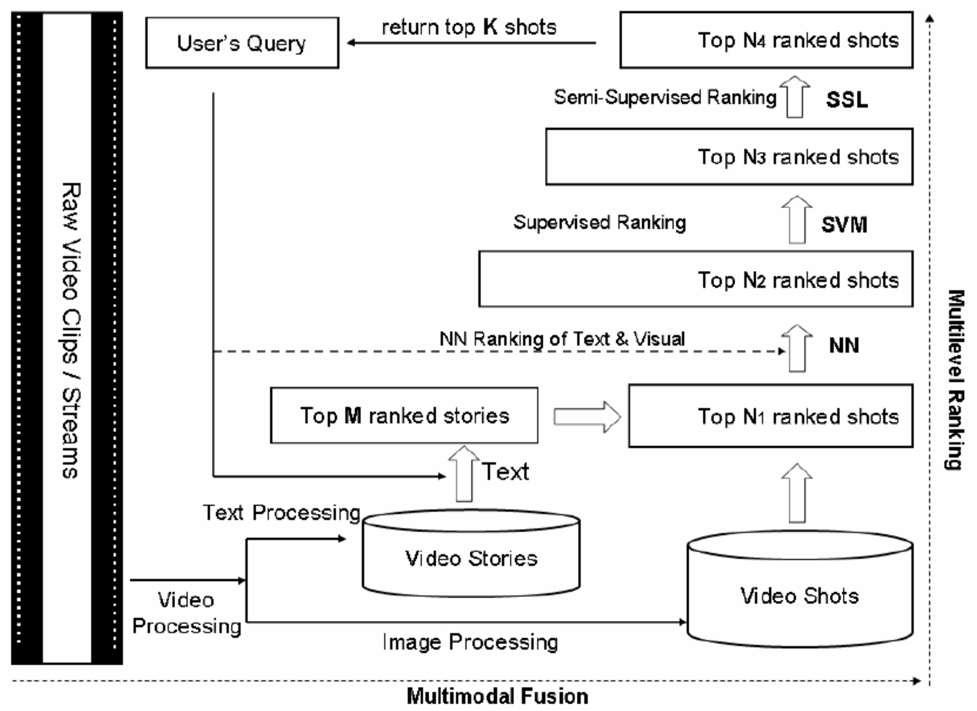

Figure 3. A Multimodal and Multilevel Ranking Framework

3.3 Story Segmentation and Textual Processing

Texts were extracted from the ASR and MT procedures. It is a difficult issue for relating the text information to specific video shots. We are aware the fact that shots in a same video story are usually more likely to be relevant. Therefore, we consider to segment the video text by relating the text information to video shots in a story level, i.e., video shots in a same story share the same text information. To this purpose, we consider the automatic video story segmentation method [6]to detect the boundaries of video stories.

Once the text stories are obtained, all text stories and query texts are parsed by a text parser with a standard list of stop words. The Okapi BM-25 formula is used as the retrieval model together with pseudo-relevance feedback (PRF) for text search. In our implementation, the Lemur toolkit was adopted for textual processing and indexing in our experiment [7].

3.4 Visual Feature Representation

Three kinds of visual features are considered in our approach: color, shape and texture. For color, we employed Grid Color Moment (GCM). Each image is partitioned into 3*3 grids, in which three types of color moments are extracted for each grid. Therefore, an 81-dimensional color moment is adopted for color features.

For shape, we used edge direction histogram. A Canny edge detector is used to get the edge image and then the edge direction histogram is computed. Each histogram is quantized into 36 bins of 10 degrees each. An additional bin is used to count the number of pixels without edge information. Hence, a 37-dimensional edge direction histogram is used for shape features.

For texture, we studied Gabor feature. Each image is scaled to 64*64. Gabor wavelet transformation is applied on the scaled image with 5 scale levels and 8 orientations, which results in 40 subimages. For each subimage, three moments are computed: mean, variance, and skewness. Thus, a 120-dimensional feature vector is adopted for texture features.

In total, a 238-dimensional feature vector is employed to represent each key frame of video shots.

3.4 Experimental Results

In the empirical evaluations, we compare our multimodal and multilevel ranking (MMML) solution with other two popular approaches, Nearest Neighbour (NN) search and Support Vector Machines (SVM). In our official submission list, three submissions with different settings of MMML solutions were submitted. These approaches adopted some different scale sizes in different stages. Specifically, the **MMML**^{1} approach is based on N_1=10000, N_2=4000 and N_3=200. The **MMML**^{2} approach is based on N_1=10000, N_2=6000 and N_3=200. The **MMML**^{3} approach is based on N_1=20000, N_2=10000 and N_3=200. All combination coefficients of multiple modalities were simply fixed to average constants.

Table 2 shows the official results evaluated by TRECVID. First of all, we see that the MAP result of the text baseline is quite low compared with previous results reported in TRECVID. Our similar approach applied in TRECVID 2005 achieved much better results. We explain this reason is that due to the quality of text transcripts and video story segmentation algorithms. Noisy text data is likely to influence the quality of retrieval performance particularly for ill-defined video data. This may show that TRECVID 2006 data is noisier than the previous years.

Another important reason is the worse performance of story segmentation algorithm. We adopted the segmentation results provided by Columbia University, which employed some clustering algorithms on the video story segmentation task. Since the TRECVID 2006 data is not well-structure, it may importantly influence the segmentation performance.

Table 2. Experimental Results of Automatic Video Search at TRECVID 2006

Methods	**MAP**	**TOP5**	**TOP10**	**TOP15**	**TOP20**	**TOP30**
Text baseline	0.0284	0.1333	0.1708	0.1444	0.1292	0.1153
Text + NN	0.0377	0.2000	0.2000	0.1972	0.1750	0.1514
Text + SVM	0.0380	0.1917	0.1875	0.1750	0.1625	0.1486
MMML^1	0.0387	0.1833	0.2042	0.1833	0.1771	0.1611
MMML^2	0.0390	0.1917	0.2083	0.1861	0.1833	0.1667
MMML^3	0.0406	0.2083	0.2042	0.1889	0.1708	0.1583

For a comparison of several solutions in our approaches, we can see that the NN approach of combining visual information is significant to improve the text baseline method. This shows that the visual features are effective. By examining the SVM performance, we found it slightly improve the performance of NN search. This result was not significant and a bit surprising from our expectation (much better results are achieved in our test on TRECVID 2005). The main reason may be the difference of kernel parameter setting. By examining the MMML solutions with NN search and SVM approach, we found that our MMML solutions achieved better results in all cases. When the number of examples in the first stage is larger, e.g., **MMML**^{3} , the improvement is quite significant. This shows that our MMML is effective for improving the performance of traditional approaches by combining the unlabeled data in the learning tasks. In fact, we also tested other situations with better parameter validation schemes, better results were achieved in our empirical tests after the official evaluations.

Figure 4 and Figure 5 show the results of comparisons of our approaches to others' results. From the results, we can see that our solution achieved similar performance compared with median results. In the 178 query topic, the performance is worse than the median result. But in 195and 196 cases, our approaches were significantly better than the median results.

3.5 Discussions

We proposed a novel multimodal and multilevel ranking solution for video search tasks. Our solution is better than traditional NN search and SVM approaches, which shows that our method is effective. We also found the dataset of TRECVID 2006 is quite different from the TRECVID 2005. Our text baseline approach did not achieve good results compared with our empirical results in TRECVID 2005. We explain the reason may be the ill-defined text data and problematic video story segmentation algorithms. In future work, we can improve these problems so as to improve the overall performance of our solution. We believe once better text retrieval performance is achieved, our MMML solution is able to achieve much better results compared with other existing results.

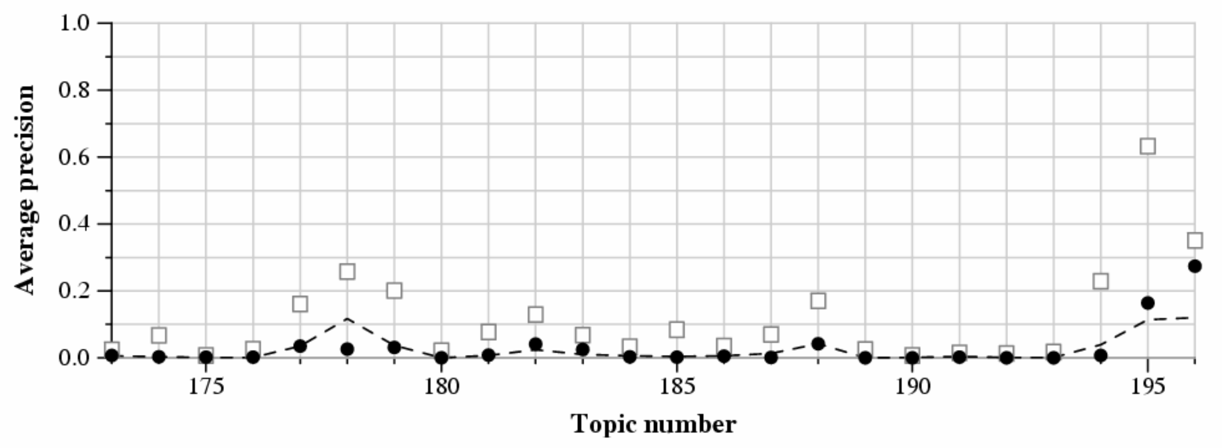

Figure 4. Experimental Result of Text Baseline Method in Official Evaluation

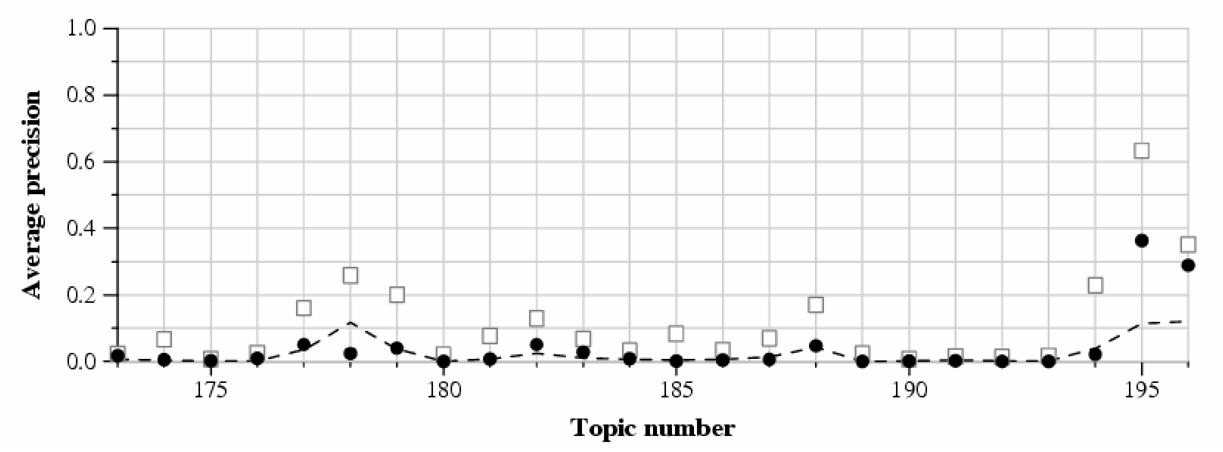

Figure 5. Experimental Result of Multimodal and Multilevel Ranking Method in Official Evaluation

4. Conclusion

We summarized our participation experience at TRECVID 2006 on the shot boundary detection and automatic video search tasks. For the shot boundary detection, we consider a simple and efficient approach based on adaptive threshoding methods. While our approach achieved good results in TRECVID 2005, the approach produce median results in the TRECVID 2006. In future, we can improve the solution by using more effective features and better learning methods. For the video search task, we proposed a novel multimodal and multilevel ranking solution for video search tasks. Different from traditional supervised learning methods in video ranking problems, our semi-supervised ranking method is effective to exploit both labeled and unlabeled data in the ranking task. The empirical results showed that our method is better than traditional nearest neighbor search and support vector machines methods. We also addressed some reasons of explaining the factors impacting the performance of our overall solution and indicate the directions to improve our work in future.

Acknowledgement

We would like to thank Shan Li for help in data collection. Lawson L. S. Wong and Albert Lyu participated in this project when they were interns at the Chinese University of Hong Kong. This project was under supervision by Prof. Michael R. Lyu. The work described in this paper was fully supported by a grant from the Shun Hing Institute of Advanced Engineering (SHIAE), The Chinese University of Hong Kong.

References

[1] Y. Yusoff, W. Christmas, and J. Kittler. Video Shot Cut Detection Using Adaptive Thresholding. In *Proc. British Machine Vision Conference*, p. 362-371, September 2000.

[2] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[3] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of The Twentieth International Conference on Machine Learning (ICML-2003)*, Washington, DC, 2003.

[4] S. C. H. Hoi and M. R. Lyu. A semi-supervised active learning framework for image retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, San Diego, CA, USA, 2005.

[5] R. Yan, A. Hauptmann, and R. Jin. Negative pseudo-relevance feedback in content-based video retrieval. In Proceedings of ACM Multimedia Conference (MM 2003), Berkeley, CA, USA, 2003. ACM Press.

[6] Winston H. Hsu and Shih-Fu Chang, “Visual cue cluster construction via information bottleneck principle and kernel density estimation,” in *Proc. CIVR*, Singapore, 2005.

[7] http://www.lemurproject.org/