TREC 2003 Video Retrieval and Story Segmentation Task at NUS PRIS

Tat-Seng Chua, Yunlong Zhao, Lekha Chaisorn, Chun-Keat Koh, Hui Yang, Huaxin Xu
School of Computing, National University of Singapore

Qi Tian
Institute for Infocomm Research, Singapore

ABSTRACT

This paper describes the details of our systems for story segmentation task and search task of the TREC-2003 Video Track. In story segmentation task, we propose a two-level multi-modal framework. First we analyze the video at the shot level using a variety of low and high-level features, and classify the shots into pre-defined categories using a Decision Tree. Next we perform HMM analysis in order to identify news story boundaries. The two-level framework has been found to be effective in overcoming the data sparseness problem in machine learning. In the search task, we perform news video retrieval by integrating multiple intra-video features and external knowledge sources. The retrieval is performed in three stages. Stage 1 uses mainly question-ansering style text retrieval technology. It analyses the text query issued by the users and extracts relevant video stories based on ASR, and external resources like WordNet and related news articles on the web. The second stage acts as a concept filter, which eliminates the irrelevant video shots in the stories retrieved by text query system. The third stage re-ranks the retrieved shots using the image and video retrieval techniques with relevance feedback. Our system emphasizes the automated retrieve process. The experiments demonstrate the effectiveness of the story segmentation system and video retrieval system.

1. Introduction

Our team from the National University of Singapore (NUS PRIS) participated in the story segmentation task and search task of the TREC-2003 Video Track. This paper describes our research on these two tasks. Section 2 presents the design of our multi-modal, two-level story segmentation and classification framework, and the experimental results. Section 3 describes the news video retrieval system that we have developed by integrating multiple video content features and external knowledge sources. Finally we conclude the paper in Section 4.

2. Video Story Segmentation Task

2.1 Overview of the System Components

The key design consideration of our story segmentation framework is in devising a 2-level scheme to analyse the video contents using multi-modal features [2]. The use of 2-level scheme helps to alleviate the data sparseness problem in statistical learning. The two levels are: shot classification level, and story segmentation level. The basic unit of analysis is the shots, and we model each shot using a combination of high-level object-based features (face, video text, and shot type), temporal features (background scene change, speaker change, motion, audio type, and shot duration), and low-level visual features (color histogram). At the shot level, we employ the Decision Tree to classify the shots into one of the predefined genre types. At the story level, we perform HMM analysis to detect story boundaries using the shot genre information, as well as time-dependent features based on speaker change, scene change and cue-phrases. The overall story segmentation scheme is shown in Figure 1.

In addition, we also classify the detected news stories into the class of "news" and "miscellaneous"..We adopt a heuristic rule-based technique to classify the detected stories.

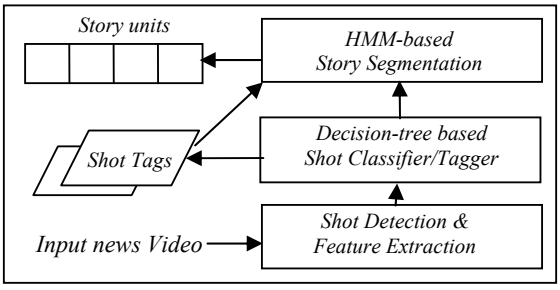


Figure 1: **Overall system components**

2.2 The News Video Model and Shot Categories

This step aims to devise an appropriate model for news video, and to determine the complete set of categories to cover all shot types. The categories must be meaningful so that the category tag assigned to each shot is reflective of its content and facilitates the subsequent stage of segmenting and classifying news stories. To achieve this, we use the class taxonomy of TV Any-Time model as the guide. In addition, we studied the structures of typical news video and the set of categories employed in related works. We arrived at the following set of 12 shot categories*: Intro/Highlight, Anchor, 2Anchor, Meeting/Gathering, Speech/Interview, Live-reporting, Sports, Text-scene, Special, Finance, Weather, and Commercial* as proposed in our previous paper [2]. In addition to these categories, we introduced additional categories to capture the specific shots used frequently in TRECVID videos, i.e. CNN and ABC news. The five additional categories introduced are *"LEDS*": to represent lead-in/out shots; "TOP": to model top story logo shots; "SPORT": to capture sport logo shots; "PLAY": to represent play of the day logo shots; and "health": to model health logo shots. Thus, the total number of shot categories is 17 which cover all essential types of shots in this collection. Some categories are quite specific such as the Anchor or Speech categories. Others are more general like the Sports or Live-reporting categories.

For completeness, we also subdivided the sports story into sub-stories depending on different types of sports. This is also a requirement of TRECVID for story segmentation task.

2.2.1 The Selection of Features

In order to support the tasks of shot classification and subsequent story segmentation, we selected the following set of features that are essential to differentiate one class from the others. The other consideration in selecting these features is that they can be automated using existing tools. The features are:

a. Color Histogram: It models the visual composition of the shot, and is particularly useful to resolve several scenarios in shot classification. This feature is used in the detection of "Weather", "Finance", "Anchor", "2anchor", "TOP", "SPORT", "LEDS", "PLAY", and "health" shots.

b. Scene change: This feature indicates whether there is a change of scene between the previous and current shots. It is derived by computing the difference in color histograms of key frames between the current and previous shots.

***c Audio*:** This feature is very important especially for Sport and Intro/Highlight shots. For Sport shots, its audio track includes both commentary and background noise, and for Intro/Highlight shots, all the narrative is accompanied by background music

e. Motion activity: We classify the motion into *low* (like in an Speech/Interview shot where only the head region has some movements), *medium* (such as those shots with people walking), *high* (like in sports), or *no* motion (for still frame or Text-scene shots).

f. Shot duration: This feature was employed in both shot classification and news story classification. It helps to resolve the ambiguities between "news" and "misc'" stories.

***g. Face*:** We extract in each shot the number of faces detected as well as their sizes. Shots with one or two faces detected are further differentiated into Anchor, 2Anchor (shots with 2 anchor persons), or other shots. The size of the face is used to estimate the shot types.

***h. Shot type*:** We divide the shot type into *closed-up, medium-distance* or *long-distance* shot based on the size of the face detected in the frame.

i. Videotext: A text-scene shot typically contains multiple lines of centralized text such as the results of a soccer game. Hence, for each shot, we simply extract the number of lines of text appear in the key frame and determine whether the text is centralized

j. Cue-phrase: We include typical cue-phrases that appear at the beginning of the news stories. Thus for each shot, we want to know whether such cue-phrases are present or not.

2.3 The Classification of Shots

News is a rather structured media with regular structures. It consists of a wide variety of shot types arranged in a well-defined sequence designed to convey the information clearly to a wide range of audiences. Certain shot types like commercials, studio anchor person, finance and weather shots etc, have well-defined and rather fixed temporal-visual characteristics. They can best be detected using specific detectors. For the rest of the categories, a learning based approach using Decision Tree is used for their classification. The following sub-sections describe the varing detectors that we used and the decision tree learning process.

2.3.1 Commercial detection

Commercial blocks and individual commercials are usually preceded and ended with a sequence of black frames and audio silence. Also, the ASR recognition rate during the commercials is usually low, as there is more background music/noise. Hence, commercials tend not to have any recognized ASR outputs. The process of commercials detection is therefore accomplished in the following two steps; a) black frames detection using color histogram; and

b) commercials block detection using clustering technique based on a combination of black frames, silence and low ASR confidence level.

2.3.2 Identifying Anchor and 2Anchor shots

For most news video, we observe that anchor persons always appear in three different positions, i.e. left, center, or right position. Thus, in order to eliminate those shots with face detected but are unlikely to be *Anchor* shots, we use the number of faces detected, their sizes and positions to identify the *Anchor* and *2Anchor* shots.

For shots where the detected face satisfies our thresholds for position and size, we extract their LUV color histogram and perform clustering using the single-link clustering algorithm. Since the number of clusters needed to obtain optimum result varies from video to video, we process the key frames for each video starting with 2 clusters and increasing the number of clusters by one, until the largest cluster contains less than or equal to 24 shots (average number of anchor shots for one video in the development set). The cluster with the largest number of shots will be the *Anchor/2Anchor* shots. Finally, we separate the *Anchor* from *2Anchor* shots by detecting the number of faces.

2.3.3 Visual-based shot detection

Visual-based shots are the shots that have distinct visual characteristics depending on their programme categories and broadcast stations. They are regularly aired in certain time slots within the broadcast news. Examples of these visual-based shot categories are: "Finance", "Weather", LEDS, "health" logo, "SPORT" logo, and "TOP" (Top stories) logo. We use the 176-Luv-color-histogram as the feature, and employ image matching and video sequencing techniques developed in our lab to perform the detection.

2.3.4 Rule-based Shots Detector using Decision Tree

The remaining shots are classified using Decision Tree. The feature vector used for each shot is of the form:

$$S_i = (a, m, d, f, s, t, c) \qquad (1)$$

where a is the class of audio, $a \in$ {t=speech, m=music, s=silence, n =noise, tn = speech + noise, tm= speech + music, mn=music+noise}; m *is* the motion activity level, $m \in$ {l=low, m=medium, h=high}; d *is* the shot duration, $d \in$ {s=short, m=medium, l=long}; f *is* the number of faces, $f \geq 0$; s *is* the shot type, $s \in$ {c= closed-up, m=medium, l=long, u=unknown}; t *is* the number of lines of text in the scene, $t \geq 0$; *and* c *is* set to "true" if the videotexts present are centralized, c $\in$ {t=true, f=false}.

2.4 Story Segmentation and Classification

As part of the requirements from TRECVID, we have to perform story segmentation based on different set of features: (i) using only video and audio features; (ii) using only ASR; and (iii) based on the combination of video, audio and ASR features.

2.4.1 Cue Phrase Detection

In order to make use of ASR features in tasks 2-3, we need to extract two types of cue-phrases, those appear at the beginning of news stories, and those appear in MISC story classes. To extract the list of cue-phrases, we first compile a list of unique n-grams from the ASR transcript in all the story segments. For each n-gram t_i, we calculate, p_b, the probability that the n-gram indicates the start of news stories and p_{misc}, the probability that indicates it is part of a misc-type. The list of p_b and p_{misc} are ranked, and we select the top n-grams with $p(t_i) \geq 0.80$ as the cue-phrases. Examples of begin-cue-phrases in the news corpuses are "checking the hour's"; "good evening i'm"; and examples of *misc* cue-phrases are "Weather forecast is next", "when we come back", "on the score board", etc.

Next, we detect and classify commercial blocks by using the following set of information: (a) *typical timing of commercial within news video*; (b) *long silence duration*; (c) *low Averaged ASR confidence;* and (d) *preceding cue-phrases*. This works well as commercial blocks tend to contain many irrelevant transcribed words and irrelevant information.

Third, we perform post processing to re-align the separation results. From the ASR of the development set, we found that 96% of the story boundaries are located at the silence intervals of ≥ 0.2 seconds. We thus incorporate this knowledge by re-aligning the results from MRA to the closest silence or speaker change using the distance measure:

$$D(y, x) = \frac{\alpha_s}{|y - x|} SilenceDur(y) + \alpha_c SpkrChange(y) \qquad (2)$$

where y: potential boundary; x: detected boundary from MRA;

α_S, α_c: arbitrary weights; SpkrChange(y): 1 if speaker change at y, 0 otherwise.

For the classification of results, segments in the video are classified as *misc* if it is detected as a commercial block or contains *misc* cue-phrases. The remaining segments are labeled as *news*.

2.4.2 The Segmentation Using Video-Audio Based Features (Test i), and Combination of Features (Test iii)

After the shots have been classified into one of the pre-defined categories, we employ the HMM technique to detect story boundaries. We use the shot sequencing information, and examine both the tagged category and appropriate features of the shots to perform the analysis. We represent each shot by: (a) its tagged category; (b) scene/location

change (1= change, 0 = unchanged), and (c) cue-phrase at the beginning of story (1=present of cue-phrase, 0= no cue-phrase).

$$S = [t, l, c] \quad (3)$$

where '*t*' is the tag-ID of a shot; '*l*' is the scene/location change feature, and '*c*' is the cue-phrase feature at the beginning of story. Note that for Test (i) that uses only video and audio features, the cue phrase feature is not used. From Equation (3), it can be seen that for Test (iii) that uses the full set of features, each output symbol is represented by 1 of 17 possible categories, 1 of 2 possible scene/location changed feature, and 1 of 2 cue-phrase feature. This gives a total of 17x2x2 = 64 distinct vectors for modeling using the HMM framework. For more details on our HMM framework, refer to our paper [2].

2.4.3 The Segmentation Using Only ASR Based Features (Test ii)

We divided the task under text segmentation using the ASR result given by TRECVID into four main tasks. They are multi-Resolution Analysis (MRA), cue-phrase detection, commercial block detection, and news classification. Figure 3 depicts the system processes.

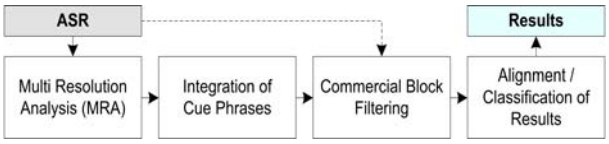


Figure 2: Processes in ASR-based segmentation

For text-based segmentation of the video, we first make use of the multi-resolution analysis and wavelet transformation technique as described in [6] to identify topic chnge boundaries. We adopt the term-based and domain independent approach, which relies only on word variations across segments of text to detect topic change. More details of this method can be found in [6].

2.5 Testing and Results

2.5.1 Training and Test Data

The training and test data are CNN and ABC news of the year 1998. Altogether, there are about 120 hours (240 videos, each with about half an hour in duration) of videos, 112 videos are used as the development set, while the remaining is used as the test set.

2.5.2 Shot Level Classification

We report our results on shot classification based on a subset of TREC videos. In particular, we test on 20 videos, 10 each from CNN and ABC. Our initial result shows that we can obtain an accuracy of about 85%.The accuracy is lower than that of our previous paper because the test set is much larger and from 2 different broadcast stations. Moreover, there are more categories and more techniques have to be incorporated. Our analysis shows that most of the errors are from the detection of those temporal-visual based shot types, for example "LEDS", "TOP", etc. These types of shots typically appear in very short durations, thus our algorithm which is designed to handle longer videos failed to detect them effectively.

2.5.3 News Story Segmentation and Classification

2.5.3.1 News Story Segmentation

We set up five runs to test the use of different combination of features for news story segmentation.

Run 1: Recall-priority run using video-audio feature without ASR, i.e. we use tag-ID and scene/location change features only.

Run 2: Recall-priority run using video-audio feature plus ASR, i.e. we use tag-ID, scene change and cue phrase features.

Run 3: Precision-priority run using the same feature set as Run 1.

Run 4: Precision-priority run using the same feature set as Run 2.

Run 5: Use ASR-based features only.

For the first four runs, we employed HMM framework as described in Section 2.4.2 to locate story boundaries. We divided the development set in training and test sets, and we performed initial experiments by varying the number of states from 4 to 15 to evaluate the results. Our initial results indicate that the number of states equals to 11 gives the best result for Runs 1 and 3, and the number of states equals to 13 gives the best result for Runs 2 and 4. As for Run 5, we perform story segmentation using the ASR based feature only. The experimental results evaluated by TRECVID are presented in Table 1.

Table 1 shows that we could achieve the best recall of 73.7% and best precision of 77.8%. This is one of the best performed system for TREC video evaluation. This performance is lower than the results achieved in our previous paper because of several reasons. First, the test data used here is much larger and more varied than the one we used previously. Second, according to TRECVID guidelines, each submitted boundary must lie within the tolerance of 5

seconds (in both directions) of the reference boundary. This is much stricter than the guideline we used in our previous test.

Table 1: Results of story segmentation based on TRECVID test corpuses

Exp	Type	Total BD	SubB D	Foundi nSub	Found inTruth	Re (%)	Pr (%)
1	1	2929	2919	2156	2105	71.87	73.86
2	2	2929	2825	2199	2158	73.68	77.84
3	1	2929	2812	2132	2084	71.15	75.82
4	2	2929	2731	2166	2127	72.62	79.31
5	3	2929	2433	1402	1383	47.22	57.62

T – type (1=Video+Audio, 2=Video+Audio+ASR, 3=ASR)
Total BD – total number of reference boundaries in ground truth data
SubBD – the submitted boundaries
FoundinSub – Correct boundaries found in our submitted result
FoundinTruth – Correct reference boundaries found in our submitted results
Re – recall (FoundInTruth/TotalBD)
Pr – precision (foundInSub/SubBD)

Table 2: The result of news story classification

Run	T	News Recall (%)	News Precision (%)
1	1	93.60	93.61
2	2	92.36	96.02
3	1	91.78	95.14
4	2	91.57	96.26
5	3	92.21	77.20

2.5.3.2 News Classification

We introduce heuristic rules to classify the detected stories into the classes of "news" or "misc". For the first shot of each detected story, we identify its category. This category was obtained during the shot tagging process as discussed in Section 2.3. The category gives us the clues on whether the detected story is likely to be "news". For example, if the first shot is an *Anchor* shot, then it is likely that this story is "news". However, this is not always true. For instance, the story that begins with *Anchor* shot in which the anchor person is introducing the upcoming news after the commercials. This story is considered as "misc". In this case, we need the shot category information of the current and successive stories. Furthermore, story duration is also important to differentiate the ambiguity between "news' and "misc". Therefore, in order to perform the classification effectively, we also need the shot category information of the successive stories as well as the current story duration. For runs that incorporate ASR (Runs 2, 4 & 5), we use the miscellaneous cue phrases to realigning the story boundaries.

The algorithms/rules for story classification are given below:

a) The Common rues for both ABC and CNN news

```
rule 1. if (Curr = COMMERCIAL),     then the story is  "misc"
rule 2. if (Curr = LEDS),     then the story is  "misc";
rule 3: if (Curr = Intro/Highlight), then the story is  "misc";
rule 4. if (Curr = ANCHOR) and (Next = LEDS) and
   story duration <=TOLERANCE,  then the story is  "misc";
rule 5. if (Curr = ANCHOR) and (Next = COMMERCIAL)
            then the story is  "misc";
rule 6. if (Curr = ANCHOR)
            if story_dur <=TOLERANCE), then the story is
            "misc" ,     else the story is  "news";
rule 7. if (Curr =  2ANCHOR) and (story duration <=
            TOLERANCE), then the story is  "misc";
rule 8. if (Curr = OTHERS),  then the story is  "news";
```

b) The specific rules for CNN news

```
rule 1: if (Curr = ANCHOR) and ((Next = WEATHER)
            or  (Next = HEALTH) or  (Next = 2ANCHOR) or
            (Next = Intro/Highlight)), then the story is  "misc";
rule 2: if  (Curr = SPORT), then the story is  "news";
rule 3: if (Curr = WEATHER), then the story is  "news";
rule 4 if (Curr = HEALTH) and (Next  = HEALTH)
            then the story is  "news";
rule 5: if (Curr = TEXT-SCENE) and (Prev = sport)
            then the story is  "misc";
```

Note: In both algorithms a) and b), **Curr -** first shot of the current story**, Next -** first shot of the next story.

2.5.4 News Story Segmentation based on the ASR

For the classification results, we could achieve the accuracy of 93.6% and 93.6 % for recall and precision respectively.

In our previous paper, we could achieve the accuracy for the story segmentation about 90%. From Table 1, the accuracy from experiment A (using video and audio based features) is lower than that of our previous paper because of several reasons. First, according to TRECVID guidelines, each submitted boundary must lie within the tolerance of 5 seconds (in both directions) of the reference boundary. That is, each submitted boundary is allowed up to 5 seconds late or early than the reference boundary. Second, by using only visual-based cue is not sufficient to locate and classify certain detected stories into "misc". For example, the score summarizing scene which normally appears at the end of each sport reporting, this portion is considered "misc". In general, our algorithm detects the whole chunk of sport including these scenes summarizing the scores as one detected story. Third, there are some miscellaneous words that although appear in news story but this portion of news is considered as "misc". For example, "I am <person name> CNN Headlines news" which appears in *Anchor* shots, this duration of the above phrase is classified as "misc". In order to tackle this problem, only text segmentation and classification can do the

job. Fourth, the test data set in these corpuses are much larger than our test data in the previous paper. There are other guidelines that if we use only visual cues (video and audio), will not be sufficient to perform the story segmentation and classification adequately. Thus, in experiment B (based on the result from experiment A, plus the use of text feature), we could improve our system performance in both recall and precision as can be seen in table 1.

3. Search Task

For the search task of TRECVID 2003, we focus on the design and implementation of an automatic news video retrieval system. We combine the low-level feature descriptors for key frames and video sequences, mid-level concept descriptors like shot categories, and semantics expressed by text. This is to make a good use of all the available multimedia contents in news videos. Meanwhile, we also utilize external resources like WordNet and Web to provide supplemental knowledge which may not be available in video contents. In the following sub-sections, we first introduce the structure of the news video retrieval system. We then discuss the system in details, including the text retrieval process, the derivation of mid-level concepts based on shot classification, as well as the use of image matching and video matching with relevance feedback. Finally we present the experimental results.

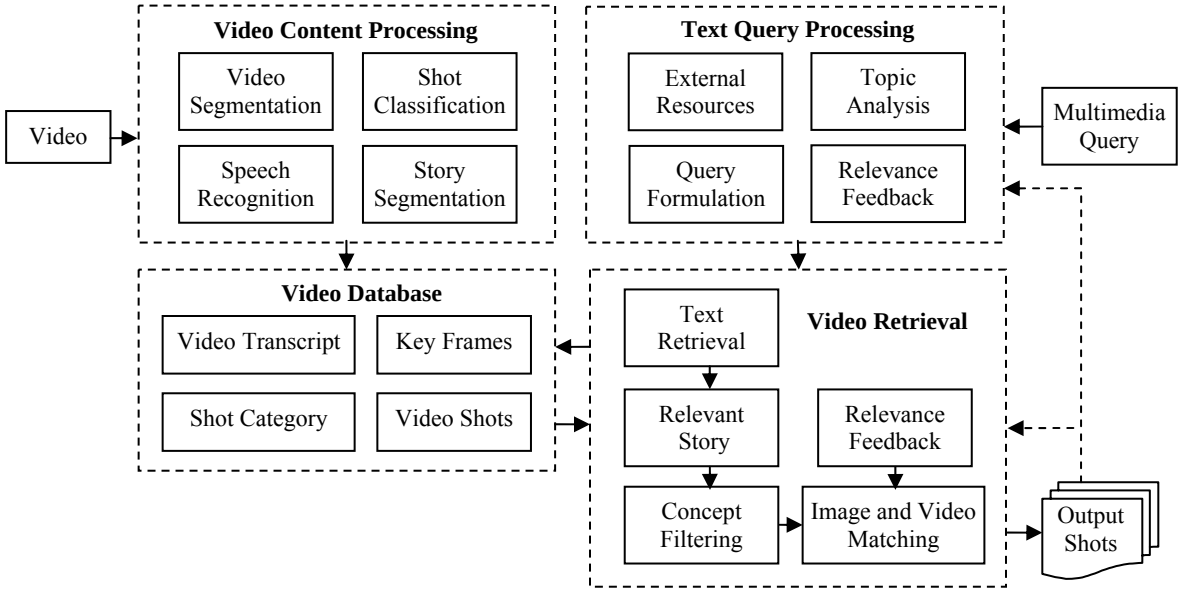


Figure 3. Architecture of the video retrieval system

3.1 Outline of the Approach

Our news video retrieval system aims to provide precise video clips to answer multimedia queries that could be the combination of free text, image and video examples. The outline of the whole system is shown in Figure 3.

The video retrieval consists of 3 stages.

a) The first stage is video story retrieval based on vido transcripts segmented at story level. We rely on question and answer (QA) techniques to extract a pool of relevant stories [1]. Here, we treat the whole segment of ASR in a story as a document, rather than go to sentence level. This is due to the consideration of the probable misalignment between key terms and the visual contents. For example, the visual content in a shot may not match to the key terms in the speech. And the sentence within a shot might be too short to contain suitable words for text retrieval.

At this stage, one important issue is that some visual contents are never mentioned in the ASR. This makes it difficult to infer the meaning of video only with text. To tackle the problem, we employed 4 methods to expand the original query to cover the relevant stories. Various external resources such as WordNet and Web are incorporated in the process. We then vote on the results returned by each method and the story picked up simultaneously by more methods will have higher confidence level.

b) The second stage filters irrelevant shots retrieved in stage 1 based on the basic concepts derived from shot categories and query targets. Here, each story may consist of multiple shots. Among these shots, some may not be good answers to the query. For example, if we want to find shots of Yasser Arafat, those shots with anchor persons are not good choices. Thus we filter the returned story and remove the shots that are irrelevant to the query. This is done by comparing shot-categories and mid-level concepts with the retrieval target type, original query, motion type and various constraints derived by query analysis.

c) The third stage uses image and video matching techniques as additional evidence to re-rank the relevant shots. After the above two stages, we obtain a list of candidate shots with the relevant answers scattered over it. It is important if we can move the more relevant shots to the top-k positions of the rank list. Example-based image and video matching are employed to fulfill this task. To obtain the initial query examples, many image retrieval systems provided functions for users to browse or randomly arrange the examples. But this will be limited in effectiveness and flexibility. The performance tends to be degraded if the examples are not from the search data set, especially when the visual features of examples are far from those in the search data set. This is understandable as the same concept can be expressed with various visual contents. Our solution to the problems lies in two aspects. First, we learn from the online feedback from users to achieve more appropriate image and video examples. Normally, we can improve the performance of image and video matching with good examples. Second, we group the query examples to several clusters according to their visual features. This query-by-groups approach [4] is more appropriate by grouping query examples into multiple clusters according to the similarity of visual features.
The following sub-sections discuss the pre-processing of video contents, and describe each of the above 3 stages in greater details.

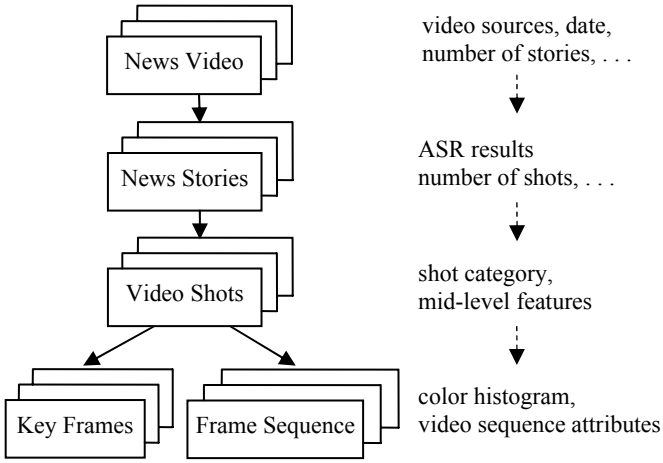


Figure 4. The structure and features of the news video collection

3.2 Processing of Video Content

The basic unit of video analysis and retrieval in this work is a shot. We have the common shot boundary reference from TRECVID. For each shot, at least one key frame is selected as its representation. Techniques introduced in Section 2 perform the story segmentation and shot classification. After analyzing the whole news video collections, we store the relevant information in SQL databases and respective direct files. Figure 4 illustrates the structure of the video collection and the accompanied features. The databases of the whole collection are organized in a layered structure. They store the attributes of the videos, stories, shots and frames, etc.
We maintain the timing information for each word in the ASR results. We also maintain the timing information of the ASR transcript at the story, shot, and speaker levels. This provides us the flexibility and effectiveness of applying text retrieval techniques.

3.3 Stage 1: Video Transcript Retrieval

The task of text retrieval is to select the video stories to answer the query, and analyze the query to obtain the retrieval target types. The following sub-sections details the topic analysis, query formulation and ASR segment retrieval.

3.3.1 Topic Analysis

Given the topics like "Find shots showing flames", we perform part-of-speech tagging, phrase tagging, Named Entity tagging on the topic sentences. The results are used to generate the retrieval target type, original query, motion type and various constraints. In our system, we have 5 basic retrieval target types (PERSON, SPORT, BUILDING, OBJECT, GENERAL), and 2 basic motion types (moving, still).

3.3.2 Query Formulation

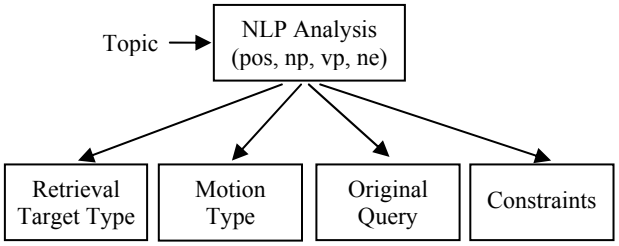
Because the original query (key words) from topic is usually very short and contains little context information, it's very hard to just make use of that to retrieve the relevant video stories. In our system, we perform query formulation

in various ways and combine the strengths of all the formulation methods to get the best matches. They are detailed as follows.

a. Query Formulation by WordNet

WordNet is a natural language resource that has been incorporated into many text mining systems. It contains about 130,000 English words linked to more than 100,000 lexical senses that are interconnected by semantic relations such as *synonym, antonym*, *hypernym* (generalization)*, hoponyms (specification) and holonym* (part-of), etc. In WordNet, English nouns, adjectives, verbs and adverbs are organized into synonym sets (*synsets*)*,* each representing an underlying lexical concept.

We make use of synonyms, hoponyms and hypernyms to enhance the query context by introducing more lexical related terms to the original query.



Firgure 5. Overview of topic analysis

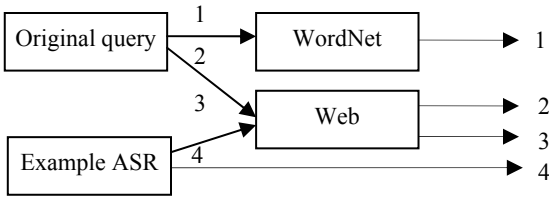


Figure 6. Illustration of query formulation

b. Query Formulation by WWW

With the huge amount of information available in the World Wide Web, it is possible that we find related information for a particular query. We downloaded CNN news during April 1998 to June 1998 available in the World Wide Web and use them as another external resource to generate more context-related query. Terms that appear frequently and close to the original query in the WWW news are expanded to form new query.

c. Query Formulation by Example ASR

ASR results from the example videos are important evidences to recognize the targeted ASR segment. We treat example ASR as a long query. We perform Stopword removal and NE recognition, only meaningful named entities are passed to the retrieval system as a query.

d. Query Formulation by Example ASR and WWW

In order to maximize the precision of our system, we combine the knowledge coming from example ASR and WWW. Terms appearing frequent and close to example ASR query are expanded to form the new query.

A single run is produced by given the retrieved shots via each of the above methods a vote. A story retrieved by more methods will have a higher weight since it is more relevant to the search topic in various ways.

3.3.3 ASR Segment Retrieval

Based on story segmentation, we segment the ASR for the test video set into stories, and indexed by MG. We then make use of the various methods described earlier to retrieve the relevant stories. In the interactive run, we also introduce negative and positive relevant feedback to modify the constraints and query.

3.4 Stage 2: Video Shots Filtering with Shot Categories and Mid-Level Concepts

Stage 1 returns a list of candidate video stories based on video transcripts. We need to refine the results for at least two reasons. First, the text retrieval usually makes a decision based on the occurrences of keywords. Although it is possible to rank the retrieval results, the discrete nature of words makes it difficult to use the nearest neighborhood as a criterion to measure the similarity. The relevant answers may be scattered in the rank list. Ideally, we would like to move these relevant answers to the top-k positions in the rank list. Second, there are possible misalignments between the speech and visual content. If we only rely on the occurrence of relevant words to decide if the corresponding shot is relevant, the results will be unreliable.

Taking the above into consideration, we need to remove the irrelevant shots or re-rank the relevant ones from those retrieved by the text retrieval techniques. This is done by matching the category and mid-level concepts within the shots against the expected retrieval target type, motion type and various constraints derived by query analysis. We measure the dissimilarity between the mid-level features of a shot and the query target by computing the Euclidean distance, $d = \sum_{i=0}^{K-1} \omega_i (B_q[i] - B_t[i])^2$, where B_q and B_t are *K*-dimensional query target and mid-level feature vectors, respectively. If *d* is less than a threshold, then the shots will be kept as candidate shots for further processing. Otherwise, it will be eliminated as irrelevant shots.

Further improvements can be achieved by inferring the exact meaning of the query by relevance feedback in text retrieval.

3.5 Stage 3: Image and Video Matching with Relevance Feedback

Like searching on the WWW, we believe that it is unrealistic and infeasible to ask users to browse all the retrieved results. More frequently, users will only browse the top-k items in the returned list of shots. In this case, it would be necessary perform re-ranking so that the relevant shots are moved to the top-k in the rank list and push the irrelevant shots to the back. This can be done by using example based image and video matching techniques in the interactive search process. Initially, the user needs to check the first 100 shots and mark the relevant shots. The system then generates the image and video queries based on the content of relevant shots. The visual queries will be used to re-rank the ranked list.

3.5.1 Image matching

We compute the similarity between the query image and the key frames of the shot. We employ the image retrieval technique introduced in [5] to perform this task. The color distribution of an image is represented as separate coherent and non-coherent color histograms. In general, coherent regions tend to correspond to part of objects within an image while non-coherent pixels tend to come from image background. Among various color space, we choose the Luv color space and decompose it with 176 cubes. In this way, we have 176 distinct Luv colors. In order to take into account the similarity between the different Luv colors, a perceptually similar color matrix is defined.

Normally, it is difficult to achieve a successful search only with one-time image matching with color histograms. An effective approach to improve this is to conduct the relevance feedback so as to reformulate query based on evaluation of the previously retrieved images. For more details of the image matching technique, please refer to [5].

Generally, the same idea can be illustrated by multiple query images with apparently different visual appearance. For example, a white cat and a black cat share the common concept. But they are different in terms of color distribution. The traditional two class example-based image retrieval tends to be ineffective to handle this situation. It would mixture the examples with different color distributions into one refined query. We follow the idea of Query-by-Groups from Nakazato et al. [4] and maintain multiple positive groups of examples and one negative group of example in the relevance feedback process.

3.5.2 Video matching

Besides image matching, we also use video sequence matching to compare the similarity between a candidate shot and a query shot. In addition, we locate the relevant segment in the candidate shot labeled as relevance by user in the interactive video search process. This segment can be used as new example for video matching. More details about the video matching technique can be found in [3].

3.5.3 Combination of image matching and video matching

We make the final decision based on the combination of the results from the image matching and video matching. The overall similarity value is defined as $\mathbf{S} = w_I\mathbf{S}_I + w_V\mathbf{S}_V$, where $w_I + w_V = 1$. The shots are ranked in terms of their similarity values from high to low.

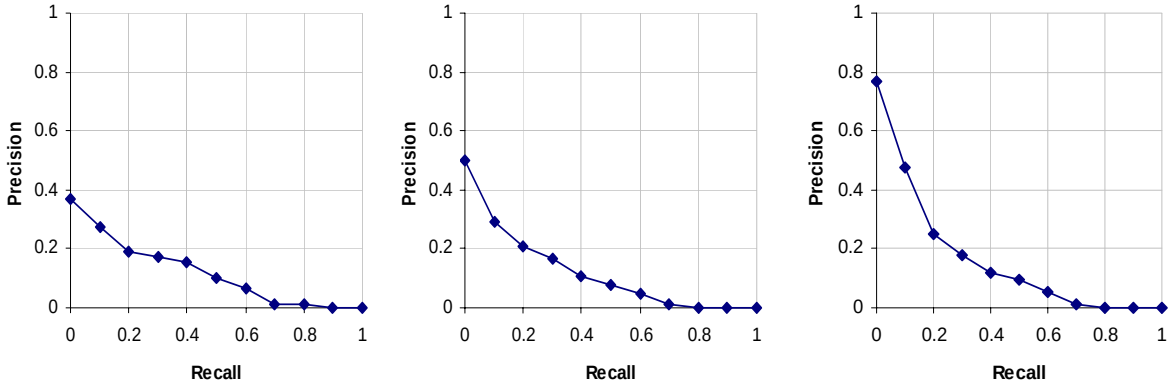
3.6 Experimental Results

We submitted 3 runs to TRECVID 2003 for evaluation. They are a manual run based only on the textual output from ASR and on the text of the topics, a manual run based on textual and visual output of test videos and the topics, and an interactive run based on textual and visual output of test videos and the topics.

Figure 7 illustrates the charts of precision versus recall of each of the three runs. The three charts (a), (b) and (c) correspond to the three runs accordingly. After comparing charts (a) and (b) from two manual runs, we can see that adding visual features to pure text retrieval helps to increase the precision of the retrieval. After comparing chart (c) to charts (a) and (b), we can conclude that the learning from users' feedback in the interactive run does improve the precision of the retrieval. The experimental results demonstrate the effectiveness of the retrieval system and satisfy our initial intentions. However, it should be noted that the recall of the system is capped by the performance of the text retrieval process, since it is the first layer to provide the pool of retrieved stories for further refinement.

4. Conclusion and Future Work

From the evaluation from TRECVID 2003, we could achieve the accuracy of 72.62% and 79.34% for story segmentation for recall and precision respectively. As for news classification, we could achieve the accuracy of 93.6% and 93.6% for recall and precision respectively. The results demonstrate that our two-level multi-modal framework is very efficient.



(a) **(b)** **(c)**

Figure 7. The charts of precision and recall of our three submitted runs.

As for the search task, we proposed a multi-layered structure for news video retrieval. Users can submit formatted multimedia queries to the system and interact with the system. The system integrates the video story segmentation, shot category classification, text retrieval, image and video matching. The external resources like WordNet, Web are employed to generate more context-related query. Intermediate concept vector helps to restrict the shots returned. Image and video matching with relevance feedback helps to move the relevant video shots to the top-k positions in the rank list. The experimental results demonstrate the effectiveness and feasibility of the processes.

Further researches will focus on the following issues. First, we need to improve the accuracy of the ASR results by correcting with available resources, such as closed-caption-based transcripts and on-line news articles. Second, we are looking at higher order statistical techniques such as the hierarchical HMM to perform news story segmentation. We need more comprehensive concept vector than what we have now. Third, we need to expand the capability of the image and video matching techniques. The current approaches based on low-level features suffer from the lack of discrimination power, especially when the examples are not from the search data set.

5. Acknowledgment

The authors would like to acknowledge the support of the National Science and Technology Board and the Ministry of Education of Singapore for the provision of a research grant RP3960681 under which this research is carried out. The author would also like to acknowledge the support I2R for the support of funding and for the helps in many ways. Last, the authors would like to thanks Feng Huamin, Lee Chee Wei, Liu Bin, and Hung Wendong for their helps through out this research.

6. References

[1] H.Yang, T.-S.Chua, S.Wang and C.-K.Koh. Structured use of external knowledge for event-based open-domain question-answering. 26th Int'l ACM SIGIR Conference . 2003.

[2] L.Chaisorn, T.-S Chua and C.-H.Lee. The segmentation of news video into story units. IEEE Int'l Conf.on Multimedia and Expo . 2002.

[3] L.Chen and T.-S Chua. A match and tiling approach to content-based video retrieval. IEEE Int'l Conf.on Multimedia and Expo , 417-420. 2001.

[4] M.Nakazato, C.Dagli and T.S.Huang. Evaluating group-based relevance feedback fro content-based image retrieval. Int'l Conf.on Image Processing . 2003.

[5] T.-S.Chua and C.Chua. Color-based Pseudo-object for image retrieval with relevance feedback. Intl. Conf. on Advanced Multimedia Content Processin. 148-162. 1998.

[6] Y.Li and T.-S.Chua. Multi-resolution analysis on text segmentation. Master degree thesis, School of Computing, National University of Singapore . 2001.