News Video Retrieval using Multi-modal Query-dependent Model and Parallel Text Corpus

ABSTRACT

This paper describes a fully automated news video retrieval system that is capable of retrieving relevant shots using a multi-media query. The emphasis we adopted is three-fold. First, we explore the use multi-modal features such as speaker identification, video OCR, face recognition and Name-entities in ASR text, along with pseudo relevance feedback, for video retrieval. Second, we employ query modeling similar to that used in text Question-Answering (QA) to classify the text query into different query-classes and use heuristics and pseudo relevance feedback to assign feature-weights to each query-class. Third, we perform query expansion on a parallel set of news articles online to enhance the recall of the retrieval process. The initial results from TRECVID 2004 evaluation indicate that the use of multi-modal query-class dependent model, coupled with external knowledge source, is effective in news video retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance Feedback, Retrieval Models, Search Process

General Terms

Design, Experimentation

Keywords

Video Retrieval, Query Class, Relevance Feedback

1. INTRODUCTION

The efficient retrieval of video is essential to many applications today as multimedia data is increasing at an exponential rate. Numerous techniques have been proposed [3, 9, 26] to handle the indexing, storage and retrieval processes. However, automated video retrieval still poses many research problems because of several reasons. First, the low-level representations of audio and visual information make it hard to provide robust and useable semantics. Second, although we can obtain part of video semantics from ASR (automatic speech recognition) output, they only provide partial semantics that are the main focus of the video. Third, the usage of visual-concepts detector is useful in detecting visual objects such as faces. However, they are too specific and often miss many interesting details. Thus it is clear that none of the features is able to provide sufficient semantics to model the context of news video retrieval comprehensively. The challenge is to develop an appropriate way to fuse all the partial information from various modalities to provide a more robust picture for retrieval.

In the search task in TRECVID 2004, participants are required to submit a ranked list of shots given a multimedia query. The query consists of a short text description and may be accompanied by short video clips and/or images [19]. There are a total of 24 multimedia queries, which express the need for video (not just information) concerning people, things, events, location etc, and their combination. The topics reflect many of the queries that real users pose: request for video with specific people or types of people, specific objects or instances of object types, specific activities or locations, and instances of activity or location. For example one of the queries is: *Find shots of Bill Clinton speaking with at least part of a US flag visible behind him.* The video clip and sample image accompanying the text query is shown in *Figure 1*.

Figure 1. Images of Bill Clinton with US Flag

The test data for TRECVID 2004 evaluation [19] consists of more than 60 hours of news video. There are 3 categories for system evaluation: interactive system, manual system and fully automated system. Our work is based on the criteria of the fully automated system where no user intervention is allowed. Given a query, we first employ query modeling techniques to classify the text component of the multimedia query into multiple query classes. This type of query classification or typing is frequently used in text QA systems in TREC-QA [20]. The QA systems will base on the class of query to apply appropriate retrieval schemes and select possible answer targets. We make use of the same reasoning to classify the news video into different shot categories and select answer shots for each query-class only from possible shot categories. At the same time, each query-class will have its own linear-mixture function to fuse the multi-modal features. We can either manually assign these feature weights or train them from a tagged training set. To complement the short textual query, we utilize Web and news articles from AQUAINT corpus, which falls in the same period of 1998 as the TRECVID news video, to perform query expansion. The purpose of query expansion is to provide an additional list of keywords which will be useful to retrieval. Normally, this additional list of keywords contains words which have high mutual information with the original query terms [12]. They are useful in increasing the recall of ASR text retrieval. We then re-rank the high-recall retrieval results using the multi-modal features. The features that we use include speaker identification, video OCR, name-entities, specific visual-concepts and face recognition. Finally, we perform pseudo relevance feedback (PRF) by treating the top *n* returned shots as relevance. This PRF helps to indicate which feature contributes more significantly. On top of that, we make use of this information to fine tune the feature weights for each query accordingly.

To demonstrate the effectiveness of the retrieval system, we have conducted a series of experiments. These experiments demonstrate that system with multi-modal query-dependent models obtains better results than those with query-independent models. The introduction of PRF and parallel text corpus for query expansion also helps to enhance the overall performance of the system. The system is able to obtain better results than the top performing systems in TRECVID 2004.

2. RELATED WORKS

A huge of amount of work has been done in the area of video retrieval. The earliest systems focus only on textual features. The use of multi-modal features was introduced in recent years. In the TRECVID 2004 search task, *Amir et al* [1] employed speech-based retrieval with re-ranking based on visual features. The visual features include color histogram, color correlogram, color moments, color wavelets, co-occurrence texture, wavelet texture, Tamura texture, edge histogram and shape moment invariant. *Quenot et al* [15] developed a search system that uses a user-controlled combination of three feature mechanism for the manual task. The mechanisms are: keywords, similarity to example images and semantic categories. The semantic categories are automatically labeled accordingly to 15 pre-defined categories, which are chosen out of the 114 video concepts from the collaborative annotation effort of TRECVID 2003. The system uses textual similarity, visual similarity and semantic to rank the shots accordingly to the query. *Westerveld et al* [21] presented the idea of ranking video shots using a generative model inspired by the language modeling approach in information retrieval [17]. The visual model ranks images by their probability of generating samples in query examples. The model is smoothed using background probabilities, calculated by maginalisation over the collection.

In general, it is evident that the use of multi-modal features did not bring a substantial increase as compared to text-based system. This is due to the fact that different queries may require very different evidences for support. Motivated by this, query-dependent modeling for retrieval was investigated. *Yan et al* [22] proposed 4 different query classes within the domain of general news videos. The four classes are Named-person, Named-object, General-object and Scene. Given a query, they performed query analysis to categorize the query and employed appropriate query-dependent model to fuse the multi-modal features. For each of the query-class, they performed machine learning to determine the various weights in the linear discriminant model to combine the features. *Yang et al* [24] also highlighted a hierarchical classification approach to categorize free text queries. They tried to predict the answer type from the query and selected answers only from possible video genres. Query classification is also widely used in Text-QA systems [18].

Various techniques commonly used in text retrieval are also tested to complement video retrieval. The usage of parallel text or comparable corpus for supplementing ASR is highlighted by *Yang et al* [24] where they used related news articles to correct video news transcripts. They first mapped a news article to a news video story. For those articles with sufficiently high matching confidence, they extracted a list of Name-Entities (NE) and used them to correct the output of the ASR results.

3. QUERY MODELING

The task of combining various modalities appropriately is essential to obtaining good results. This is especially true when we perform automated search where there is no user feedback in the process. Various learning algorithms have been proposed to optimize the weightings of different modalities from training data [22]. Though assigning same set of query-independent feature-weights to all queries is easy, the use of this scheme does not address the problem that different queries may have very different characteristics and hence require very different feature combination. The ideal case would be to find an optimal weighting scheme for each individual query. However this is clearly not practical as it is not possible to anticipate all possible user queries. Thus it is more reasonable to classify queries into several pre-defined classes like in [22, 24]. We apply our understanding of news video structure and create several suitable query-classes. Thereafter we base on training queries to optimize the feature weights for each of these classes. Besides applying individual query-class feature weights for the combination of different modal features, the query-class also provides us with the information to reduce the search target size by filtering out those shots belonging to categories that are not relevant to the query-class.

3.1 Query Classification Scheme

We adopt a query classification scheme which is closely associated to the news category of the shots. Six non-intercepting classes are proposed as follows: {PERSON, SPORTS, FINANCE, WEATHER, DISASTER, GENERAL}. The GENERAL-class is created to accommodate the queries that do not belong to any of the first five classes. From our experiences with video retrieval, the first five query-classes cover about 50% of the commonly asked video news queries. The other reason for this classification scheme is to create a mapping of query-class to video shot-class. For example: the answer shots for sports queries are normally found in sport news; and similarly for financial news and weather news. These five classes are also chosen because they can be easily classified by using simple heuristic rules based on textual information alone. This is important as it is not possible to perform complex query classification for short text queries.

The six query classes used are:

PERSON: queries looking for a person. For example: *"Find shots of Boris Yeltsin" and "Find Bill Clinton speaking with at least part of US flag visible behind him."*

SPORTS: queries looking for sports news scenes. For example: *"Find more shots of a tennis player contacting the ball with his or her tennis racket."*

FINANCE: queries looking for financial related shots such as stocks, business Merger & Acquisitions etc.

WEATHER: queries looking for weather related shots.

DISASTER: queries looking for disaster related shots. For example: *"Find shots of one or more building with flood waters around it/them"*

GENERAL: queries that do not belong to any of the above categories. For example: *"Find one or more people and one or more dogs walking together"*

Table 1 shows the queries from TRECVID 2004 and their corresponding query-classes.

Table 1. Query Class Mapping

Query from TRECVID	Query-Class
Find shots of a street scene with multiple pedestrians in motion and multiple vehicles in motion somewhere in the shot.	GENERAL
Find shots of one or more buildings with flood waters around it/them.	DISASTER
Find shots of US Congressman Henry Hyde's face, whole or part, from any angle.	PERSON
Find shots of a hockey rink with at least one of the nets fully visible from some point of view.	SPORTS
Find shots of Sam Donaldson's face - whole or part, from any angle, but including both eyes. No other people visible with him	PERSON

3.2 Learning the Feature-weights

Intuitively, each query-class will exhibit different characteristics and require different evidence to induce the answers. For example, speech recognition is important for PERSON class but may not be the case for SPORTS class. For FINANCE and WEATHER, image retrieval plays a more significant role because their key frames tend to be similar which makes image matching techniques more effective. Thus, it is more reasonable to fuse the multi-modal features using a query-dependent model. By simple observation, we can manually determine if a certain feature will be important to that class. However, it is difficult to obtain an optimal weighting manually if there are too many modalities to consider. Therefore we employ a heuristic approach to find the optimal feature weights of each class.

We first collect a set of training queries and their answer video clips from TRECVID 2003 search task. We classify these queries according to the six classes. For those classes that have less than 10 training queries, we manually create more queries for them. We then generate answer-shots for these queries. For those queries used in previous search task, we simply use the ground-truth provided by TRECVID. For those queries that are generated by us, we manually mark 10 to 20 correct shots for each query. This process is assisted by our previous video retrieval system.

Given the set of queries and corresponding answer shots, we can now employ the respective modal detectors to detect each modal feature and capture their respective feature confidence scores. With these confidence scores, we employ multiple linear regression model to find respective feature-parameters for each class.

$$Y = \sum_{\mathrm{mod}\,el_Q} \beta_i^Q F_i \qquad (1)$$

In each class, we treat each of the positive instances as having a real value of α (we set it to 1) and find the optimal parameter set β^q which will best converge all the positive instances to a single value. The reason that we involve only the correct shots is because it is unrealistic to consider the negative instances as there are too many of them. In addition, they may be too diverse. The objective of feature weights is to provide a suitable mixing function for re-ranking the shots returned by the text retrieval engine. *Table 2* shows the importance (high, medium, low or nil) of various features with their respective class. The visual concept in the Table refers to a specific visual object [26].

Table 2. Weights for fusing Multi-modal Features

Class	Person	Sports	Financial	Weather	Disaster	General
NE	High	High	Med	Med	Low	Low
OCR	High	Med	High	High	Med	Low
Speaker	High	Low	Low	Low	Low	Med
Face	High	Med	High	High	Med	Low
VC People	Med	Med	Nil	Nil	Low	High
VC Hockey	Nil	High	Nil	Nil	Nil	Low
VC Water	Nil	Nil	Nil	Med	High	Low
VC Fire	Nil	Nil	Nil	Nil	High	Low
Etc	..	..	..	..	..	..

*VC → Visual Concept

3.3 Query Analysis and Classification

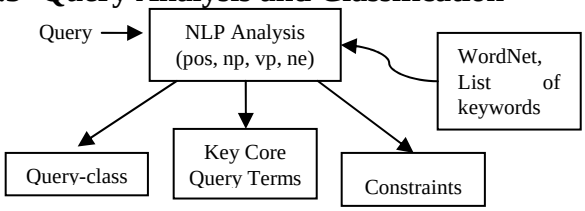

Figure 2. Overview of Multi-class query

As illustrated in *Figure 2,* the text query is first analyzed to derive 3 types of information, the query-class, key query terms and the constraints. We first perform morphological analysis on the given text query to extract information like Part-of-Speech (POS), verb-phrase and noun-phrase similar to [23]. This analysis is crucial as it helps us to understand the query and differentiate between the topic and the constraints. Next, we extract the core-terms of the query. The core-terms are taken as the topic of the query, which are normally the strongest nouns or noun phrases. We also employ Name-entity extractor at the same time to identify entities like names of persons, organizations and possible objects in the main query. These NEs are used together with the core-terms.

Given these keywords, we develop a rule-based query classifier to identify the query-class which is essentially a text categorization problem. For PERSON class, we rely on the existence of Person-type NE, which will signify that the query is person oriented. For the other four classes other than GENERAL, we extract a list of keywords for each class from the set of training samples. These lists of keywords are used to classify the query topic into their various classes. For those queries that do not belong to any of the first 5 classes, we classify them as the GENERAL class.

4. SHOT CLASSIFICATION

One key problem to tackle in the preprocessing stage of news video analysis is shot classification which is used to limit the scope of retrieval. We selected seven suitable shot-categories for the system: sports, finance, weather, commercial, studio-anchor-person, general-face and general-non-face. The shots that do not fall into the first six categories are classified as general-non-face. The rationale for this selection is that we want to exploit the

inherent structure of news video and to follow as closely as possible to the pre-defined query-classes. The categories chosen are also the same as the major categories used in news video story segmentation task [2]. In this work, we employ the techniques described in [3] to tag the news categories automatically..

Once we obtain the query-class from the query classifier, we will be able to carry out the filtering of the non-relevant shots. *Table 3* shows the various query-classes and their corresponding target news categories.

Table 3. Corresponding target types for each class

Query-class	Target News Categories
PERSON	studio-anchor-person, general-face
SPORTS	sports
FINANCE	financial
WEATHER	weather
DISASTER	general-non-face
GENERAL	general-non-face

We make use of text, visual and timing features to detect different categories of shots. Certain shot categories like finance and weather, have well-defined and rather fixed temporal-visual characteristics; they can be detected using specific detection techniques. For sport news, we use a combination of motion, speech and visual to perform the classification [3]. The following subsections describe briefly the algorithms used to detect commercial and anchor shots.

4.1 Commercial Detection

We make use of black frames, cuts, silence, cut rate and quality of ASR outputs. The algorithm first detects black frames with sufficiently long audio silence. It then "looks" ahead for the next block of black frames. If the time interval from this block of black frames to the next falls into within a certain time interval, this implies that the commercial may fall within this region. The algorithm then analyzes the cut rate and examines the ASR output of shots residing within these two blocks of black frames. If it detects sufficiently high cut rate and low confidence in the ASR output, it will classify these shots into commercial category.

4.2 Face/Anchor Shot Detection

For most news video, we observe that the anchor persons always appear in three different positions, i.e. left, center, or right positions. Thus, in order to differentiate those shots with face to anchor person shots, we use the number of faces detected, their sizes and positions. 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. This helps in deciding if the shot is studio or non-studio, as studio accounts for majority of such shots. Since the number of clusters needed to obtain optimum result varies from video to video, we process the key frames for each video starting from 2 clusters and increasing the number of clusters by one, until the largest cluster contains less than or equal to 24 shots (which is the average number of anchor shots for one video in the development set). The cluster with the largest number of shots will be the studio-anchor shots and the rest will be general-face shots.

5. TEXT RETRIEVAL

Text retrieval is a critical module in the retrieval pipeline as it will determine the initial answer candidate set. Therefore, we emphasize on optimizing the recall to make sure we obtain most of the answer-shots within the initial retrieved set. To achieve this, we need to tackle two problems. The first is the expansion of query in order to incorporate the context and linguistic variations of initial query. The second is to determine the unit of retrieval – either at shot-level or story-level. Intuitively, story-level is better as it contains a sufficiently large body of coherent text. Since story-level units are not available, we instead use the speaker-change intervals as pseudo-stories. Speaker-change information is available reasonably accurately as part of the ASR output [8].

5.1 Using Parallel Text Corpus for Query Expansion

As the original query is short and contains little contextual information, it is hard to just make use of this query to retrieve most relevant video stories. In our system, we try to induce more keywords by obtaining a list of keywords that has high mutual information with the original query terms. This is done by gathering news articles using the online news search engine such as Google. However, as the online search engines only index recent news articles, it is not possible to obtain good contextual terms for queries. To overcome this problem, we also perform the query expansion using a parallel set of text articles which is in the same year as the video data. In this way, the text articles used will be relevant to the actual video news. Therefore, they will provide useful keywords which will eventually be helpful to retrieving the relevant video shots from ASR. As each of the video news is dated, it is reasonable to assume that we can locate parallel news from external news archives automatically.

5.2 Speaker Change Unit

In this research, we use the speaker-change tags extracted from ASR results from LIMSI ASR [8] to identify boundaries of "pseudo" story units. . The main reason for using speaker-change as unit of retrieval is because speaker-change information is readily available from the ASR output. Also, current speech recognition systems cannot insert punctuations accurately into the sentences, and hence there is no effective way to segment the continuous string of words from ASR into reasonable units for retrieval. Speaker change is also preferred over silence intervals as we want to prevent over segmentation and the consequent lost in recall. We observe that using speaker change instead shot change provides more contexts and thus render the retrieval more effective.

5.3 Text Retrieval Process

As illustrated in *Figure 3*, we first extract non-trivial keywords from the query to form a bag of initial key words K_1. K_2 is extracted from the ASR surrounding the video samples, which are provided as part of the multimedia query. In order to extract the context of this query, we use K_1 and K_2 to retrieve relevant news documents either from online search engine or parallel text articles. We extract from the documents terms that co-occur frequently with K_1 and K_2. As the set of expanded terms is big, we only select terms that exist together in the same synset as the original query terms. This synset information is obtained from WordNet [6]. Besides co-occurrences, we also apply morphological analysis on these documents to extract the name-

entities. The name-entities, as well as the selected expanded terms will be merged with K_1 and K_2 to form K_3. Finally, K_3 is used to retrieve passages from text corpus at speaker-change level based on ASR output. A set of ranked speaker-change level segments, along with the shots enclosed within these segments, are retrieved for further re-ranking based on other multi-modal features.

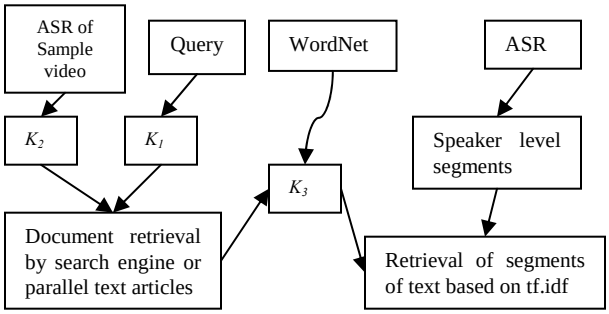

Figure 3. Text Retrieval with Query

6. MULTI-MODAL FEATURES

From experimentation, we know that the average number of shots in a speaker-change segment is about 5 and within these shots, only 1 or 2 shots may be relevant. In order to improve retrieval performance, we need to employ other sources of information or techniques to eliminate irrelevant shots and perform overall re-ranking of retrieved shots. Here, we use a number of multi-modal features to help in further analysis. These features include face recognition, video OCR, speaker identification and visual concept detection. We briefly describe the algorithms used to extract these features.

6.1 Face Recognition

Our approach to detect Person-X uses three sources of information: appearance time distribution, appearance shot distribution and face detection. First, we filter out shots of anchor person and commercials. Then, we make use of the appearance time and shot distribution as heuristics for face recognition. Appearance time distribution indicates the probability of person-X's appearance at different time of a video clip, while appearance shot distribution gives the probability of person-X's appearance in different shots beside the shots where person-X's name appears. We make use of the labeled training corpus to obtain various models for each person. For each shot in the testing data where we detected a face, we apply a 2DHMM to classify them [4].

6.2 Video OCR

The OCR results are tagged by CMU [9]. Even though there were a number of insertion, deletion and mutation errors, video OCR proves to be a good feature as it is able to give precise information on the appearance of certain human persons. Therefore we integrate a minimum edit distance (MED) matching to maximize the precision and recall of name-matching in OCR. We use 10 videos (5 CNN, 5 ABC for the development videos) and a general set of name entities terms to test the overall effectiveness of OCR and MED.

6.3 Speaker Identification

The speaker identification is similar to the techniques described in [9, 11]. In addition, we introduce a pseudo relevance feedback loop-back using face detector and OCR to ensure that the match is correct. First, we extract the MFCC features of the speech segment and train a model for each speaker using HTK [10]. In the first training instance, we use the speech segments in the sample video to derive $\boldsymbol{M_1}$. Next, we use the ASR from LIMSI to retrieve possible speech segments that could be made by speaker X by performing text retrieval with query expansion as described above. The possible segments are then tested using model $\boldsymbol{M_1}$. The results of the identification are further justified by using video OCR and face detector modules. Only shots that satisfy all the criteria are used as new training instances. Finally we train the speaker identification module using the new instances and use it to identify all other speech segments within the test set.

6.4 Visual Concept Detection

For visual concepts detectors, we first need to determine which visual concepts to use. This is because the detection rates of some concepts can be as low as 5%, and will be too unreliable for use in fusion. For this, we have chosen 10 visual concepts which have a detection rate of over *10%*. Such visual concepts include fire, water-body, hockey, basketball, greenery, people etc. We employ the techniques described in [26] to detect these visual concepts.

7. OVERALL SYSTEM ARCHITECTURE

Our system consist of 3 main parts: preprocessing of video data, retrieval of relevant video shots and pseudo relevance feedback.

7.1 Preprocessing of Video Data

For the preprocessing of the video data, we perform the following seven processes: speaker level segmentation based on ASR, shot boundary detection, video OCR, shot classification, speaker identification, face detection and recognition and visual concepts detection. The features of each shot are stored in a database for retrieval.

7.2 Retrieval

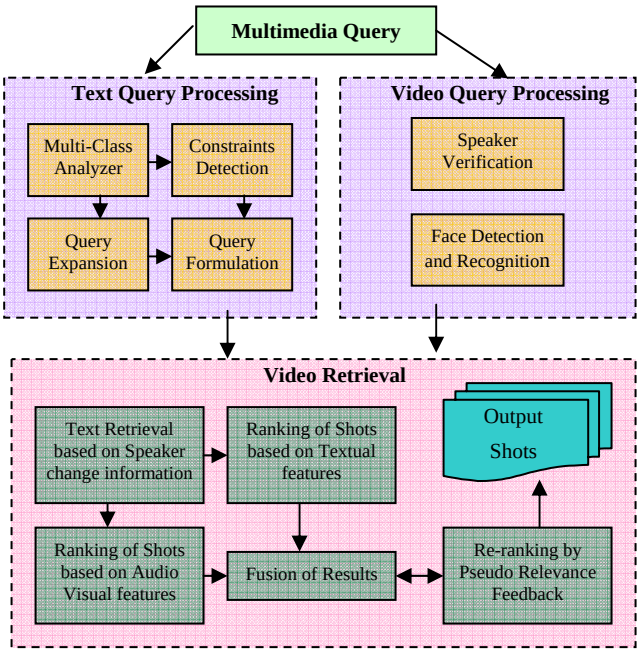

Figure 4. Retrieval Framework

Figure 4 illustrate the complete retrieval process. The retrieval system comprises 2 steps, the query processing step and the retrieval step. The text query is analyzed accordingly to the steps described in S*ection 3*. For video query, we extract the audio stream for speaker identification and image for face recognition. The retrieval step involves the retrieval of ASR text at the speaker change level. The individual shots are then rank by the multi-modal features given by *Eqn. (2)*.

$$Score(Shot_i) = \sum_{Model_Q} \beta n_i^Q F_i \quad (2)$$

$$\text{where } \sum_{Model_Q} \beta n_i^Q = 1$$

We obtained the βn^Q by normalizing β^Q obtained from *Eqn. (1)*. F_i is the confidence score of feature *i.* In this case, a high score will signify that the shot is relevant to the query.

7.3 Pseudo Relevance Feedback

PRF has been shown to bring significant improvement to text retrieval systems [20]. We assume that the top *n* retrieved shots to be relevant and use them to perform a round of pseudo relevance feedback based on the text in these shots. Here we set *n* to 10 empirically. In addition, we also use the relevance information to fine-tune the multi-modal feature-weights for use in *Eqn.* (1-2).

7.3.1 Text PRF

When using the ASR text, although we can determine which sentences may be important by locating keywords, we cannot determine if they are actually answer shots. The rationale of applying PRF on text is that we want to give higher weights to terms that co-exist with the selected shots.

We extract the textual information in these *n* shots to obtain a list of additional keywords K_4. We then use K_4 to modify the original text query $q^{(0)}$ to obtain a modified query $q^{(1)}$ using a version of vector-space relevance feedback formula [16] as follow:

$$q^{(1)} = q^{(0)} + \alpha \sum D_i \quad (3)$$

where D_i denotes the set of ASR associated with the pseudo relevant shots. We then use $q^{(1)}$ to perform a new round of similarity-based retrieval to obtain a new ranked list of shots.

7.3.2 Multi-Modal PRF

Ideally, the feature weights for each of the features of a query-class should tell us its level of importance. However, for certain queries, this is not true. One such query is "*Find shots of Boris Yeltsin*". Although this query belongs to PERSON class, it is rather different from the rest of the queries in that class. This is because Boris Yeltsin did not speak in the given training and test data. This contradicts the high weights assigned to the speaker identification feature for PERSON class. Therefore we introduce the idea of using PRF to further fine-tune the feature-weights. This PRF will decrease the weights of certain features deem not important.

We first treat the top *n* returned shots as our positive instances. These shots will be used to determine if we should decrease the weight of certain modal feature. The criterion for decreasing the weight is as follows:

$$\sum_{i<n} (F\_Confidence_A(Shot_i) < Feature_A^{threhold}) < k \quad (4)$$

$Feature_A^{threhold}$ is determined empirically such that if the confidence-score of feature-A exceed $Feature_A^{threhold}$, then it is likely that feature-A will exist. We count the total number of shots in the top *10* returned-shots whose feature-confidence is less $Feature_A^{threhold}$. If the total count of a particular feature is less than *k* (we set it to 2), it means that this feature does not play a significant part in the selection process. We then discard the use of this feature to prevent errors from accumulating and redistribute the weights to the rest of modalities.

8. EVALUATION

We design a series of tests to evaluate the effectiveness of query modeling, the use of external parallel text corpus for query expansion, and pseudo relevance feedback.

8.1 Experiment Setup

We follow the same evaluation methodology as in TRECVID 2004 search task [19]. The performance measure used is the mean average precision (MAP). This performance measure is widely used for system evaluation in information retrieval over large corpuses where recall rate is hard to determine. For the evaluations, we follow strictly to the criteria of fully automated search where no user intervention is allowed. The training data available is 140 hours of news video during early 1998, and the test data is 60 hours from the last 3 months of 1998. We use the same set of queries as in TRECVID 2004 search task for the experiments. For each of the query, a maximum of a 1,000 shots is returned. The result is then evaluated using the ground-truth provided.

In order to evaluate different aspects of the system, we carry out 3 series of tests based on: (a) text; (b) multi-modal features; and (c) pseudo relevance feedback. For comparison, we also compare our results with that of *Amir et al*. [1], one of the top performing systems in TRECVID 2004.

8.2 Text-based Retrieval

We design a series of tests that use only the ASR text for retrieval. These tests are used to evaluate the following three premises. First, we want to know to what extend can text-only features perform and this will provide a baseline performance for video retrieval. Second, we want to highlight the effectiveness of query-dependent retrieval model. Third, we want to contrast the performance of using different types of external information sources- general web versus targeted web (parallel corpus) for query expansion. Since the unit of retrieval is the speaker-change segment, the shots that fall within these segments are returned and ranked according to the respective segments. We carry out 3 runs as follows:

T1) Baseline (basic text retrieval with no query expansion).

T2) T1 with query expansion (either general web or parallel corpus).

T3) T2 with the use of query-dependent retrieval mode.

Table 4. Text Runs in terms of MAP

Runs	(a) web for expansion	(b) parallel corpus for expansion
T1	*0.038	-
T2	0.047	0.058
T3	0.071	0.078

The results of the experiments are present in Table 4. From the Table, we can see that by using only keyword matching techniques with no query expansion (*T1*), we could achieve a MAP of only 0.038. By performing query expansion, we achieve a significant improvement in performance over the baseline. In particular, the results of *T2* run that use parallel text for query expansion are superior over the use of general web. This is to be expected as parallel text comes from the same period as news video and it is able to bring in more relevant context.

Results from *T3* run show that the use of query-dependent retrieval model could further enhance the performance by about 50%. Overall, the best run of the experiment comes from the one that uses query-dependent model supplemented with parallel text for query expansion.

8.3 Multi-Modality Retrieval

This series of tests utilize the rest of multi-modal features in retrieval. The aim is to ascertain our hypothesis that while text feature is useful for retrieving most relevant shots, multi-modal features are essential for re-ranking the results and hence improving the precision. Thus, the multi-modal features will only be used to perform re-ranking of shots from the text runs results. The runs we performed include:

AV1) T3 with the use of video OCR and shot classification.

AV2) AV1 with the addition of face recognition and speaker identification.

Table 5. Multi-modal Retrieval Runs in terms of MAP

Run	(a) web for expansion	(b) parallel corpus for expansion	Amir [1]
T3 (Baseline)	0.071	0.078	0.057*
AV1	0.086	0.096	--
AV2	0.119	0.123	0.109**

*Based on a single fusion run **Based on the overall best run

From the experimental results in *Table 5*, we see a gradual increase in performance with the use of more multi-modal features. Results in *AV1* reveal that the use of shot classification and video OCR acts as a precision enhancement device in improving the MAP to 0.096 with the use of parallel corpus. Analysis of results shows that SPORTS-class queries attain the largest improvement with the average increase in precision of about 30%. We attribute this to the use of shot categories to constrain the results to only sports category and the use of OCR.

With the utilization of face recognition and speaker identification modules, we see a further significant increase in performance in *AV2* to about 0.123. This is attributed to the vast improvement in performance of the PERSON-class queries. Most of the shots with faces in this case are highly possible answer candidates, and the use of speaker identification also helps in the re-ranking process.

In comparison, our AV2 run is able to obtain 20% better performance than the best run from Amir [1]. Overall, our results and that of *Amir et al*. [1] confirm that the use of multi-modal features is essential in improving the overall performance.

8.4 Pseudo Relevance Feedback Runs

For this round of tests, we focus on using PRF to improve the quality of text and multi-modal retrievals. For this series of tests, we use *AV2* as baseline and perform the following:

PRF1) AV2 with one round of Text-based PRF.

PRF2) AV2 with one round of multi-modal PRF (see Eqn. 4).

PRF3) PRF2 together with one round of Text-based PRF.

Table 6. Pseudo Relevance Feedback Runs in terms of MAP

Run	(a) web for expansion	(b) parallel corpus for expansion
PRF1	0.124	0.126
PRF2	0.125	0.128
PRF3	0.127	0.130

From Table 6, we observe that the use of text-based PRF lead to some improvement in MAP scores. We notice that the DISASTER-class of queries benefit the most from PRF, while the GENERAL-class queries obtain slightly worse results, partly due to poor quality of initial retrieval. The improvement in performance with the use of multi-modal PRF (in *PRF2* run) is more significant as compared with *AV2*. We notice substantial improvement in the performance for PERSON-class, SPORT-class and DISASTER-class queries. This increase is attributed to the better assignments of feature-weights to individual queries.

Finally, in *PRF3* we combine both PRFs. As expected, the performance of the system improves to an MAP value of 0.130. In all cases, we observe that PRF is very useful for queries in which we are able to obtain good relevant top ranked shots. This is because the answer shots generally shared very similar features.

8.5 Analysis of Results

In general, the system is not able to perform shot filtering using the given text constraints. For example: "*Find shots of US Congressman Henry Hyde's face, whole or part, from any angle.*" Even though the system is able to differentiate the topic and the constraints, it is not able to interpret the constraints as it requires thorough understanding of the language. Also, it is quite hard to translate a text constraint into a multimedia one. Often, the system will ignore the constraints and perform the ranking only base on the topic. In addition, this classification scheme is not sufficiently fine-grained and therefore it is not possible to capture many different topics other than the 5 pre-defined genres. Many queries are classified into GENERAL class where in fact they are very different from one another. Hence, finer classification is necessary. However, this requires a deep understanding of domain knowledge and a large amount of training queries.

Analysis of our results shows that our system performs well in PERSON and SPORT classes. This is because name-entities in text, faces, OCR etc are extremely useful in these classes. In contrast, classes such as DISASTER and GENERAL generally perform badly. Besides being too general as mentioned above, there is also a lack of suitable accurate visual concept detectors for these classes. Textual features are practically useless in pinpointing the relevant shots and visual counterparts only propagate more errors. One way to counter this problem is to introduce more multi-modal features that are specific to these classes.

Using the pseudo relevance feedback to alter the weights is useful but it may not work for some specific cases. According to the algorithm, we will lower the weights of certain features which we deem to be unimportant to the ranking based on the frequency of occurrences. However, this may not always be the case. For example, if a person rarely speaks in the video data, we will lower

the weights of the speaker identification module. However, in the rare occasion that this person actually speaks and is detected correctly by the speaker identification module, we want to make use of this information. Thus, we need to develop an adaptive way to modify weights of features depending on confidence.

9. CONCLUSION

This paper discussed the framework and techniques employed in our automated news video retrieval system. Our contribution is three-fold. First, we employed multi-modal features including text, OCR, image, visual and audio with pseudo relevance feedback to support the retrieval. Second, we explored the use of query classes to associate different retrieval models for different query classes. Third, we explored the use of parallel text for query expansion. Our initial results show that we are able to improve the performance of the system with the use of additional features and techniques. Our system performs better than the best results published in TRECVID 2004 search task. Our research points to many promising areas to improve the performance of video retrieval. We can therefore conclude that while text is useful in retrieving most relevant shots (recall), the use of multi-modal features and PRF are essential to improve the precision of the overall system.

For future work, we will refine our query-dependent classes and models to ensure that it is more robust and able to capture more types of typical user queries. We will further investigate the use of PRF to fine-tune the weights and possibly to learn the weights on the fly. We will also explore the use of more sophisticated learning methods to learn the query-dependent models.

10. REFERENCES

[1] A. Amir, J.O. Arillander, M. Berg, S.F. Chang, W. Hsu, G. Iyendar, J. R. Kender, C.Y Lin, M. Naphade, A. Natsev, J. R. Smith, J. Tesic, G. Wu, R. Yan and D. Zhang, *IBM Research TRECVID-2004 Video Retrieval System*. In the Notebook Paper, 82-91, TRECVID 2004.

[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. Chaisorn, C.-K. Koh, Y.-L. Zhao, H.-X. Xu, T.-S. Chua, T. Qi. *Two-Level Multi-Modal Framework for News Story Segmentation of Large Video Corpus*. In TRECVID 2003 Workshop, 129-134, Nov 2003.

[4] M. Y. Chen and A. Hauptmann. *Searching for a specific person in broadcast news video*. Proc. of the Int'l Conf on Acoustic, Speech and Signal Processing, Vol. 3, 1036-1039. May 2004.

[5] T.-S. Chua, C. Chu and M.S. Kankanhali. *Relevance feedback techniques for image retrieval using multiple attributes*. Proc. of IEEE Int'l Conf on Multimedia Computing and Systems (ICMCS'99), Florence, Italy, 890-894, Jun 1999.

[6] C. Fellbaum. WordNet: *An Electronic Lexical Database*. MIT Press. 1998.

[7] Y. Freund and R. E. Schapire, *A Decision-theoretic generalization of online-learning and an application to boosting.* Journal of Computer and System Sciences, Vol. 55, no. 1, 119-139, August 1997.

[8] J.L. Gauvain, L. Lamel, and G. Adda. *The LIMSI Broadcast News Transcription System*. Speech Communication, 37(1-2): 89-108, 2002.

[9] A Hauptmann, R. Jin and T. D. Ng. *Video Retrieval using Speech and Image Information*. Proc. of Electronic Imaging Conference (EI'03), Storage and Retrieval for Multimedia Databases, Santa Clara, CA, Jan 2003.

[10] Hidden Markov Model Toolkit: *http://htk.eng.cam.ac.uk/*

[11] H. Jiang, T. Lin and H.J. Zhang. *Video segmentation with the Support of Audio Segmentation and classification*. ICME'2000-IEEE Int'l Conf on Multimedia and Expo, NY, USA, Jul 2000.

[12] C. Kenneth and P. Hanks. *Word Association Norms, Mutual Information, and Lexicography*. Proc. of the 27th Annual Meeting of the ACL, 1989.

[13] L. Lu, S. Z. Li, and H.J. Zhang. *Content-based audio segmentation using support vector machines*. Proc. ICME 01, Tokyo, Japan, 956-959, 2001.

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

[15] G. M Quenot, D. Mararu, S.Ayache, M. Charhad, L. Besacier, M. Guironnet, D. Pellerin, J. Gensel and L.Carminati. *CLIPS-LIS-LSR-LABRI Experiments at TRECVID 2004*. In the Notebook Paper, 24-39, TRECVID 2004.

[16] G. Salton and C. Buckley. *Improving retrieval performance by relevance feedback*. Journal of the American Society or Information Science, 288-297, 1990.

[17] Special Interest Group on Information Retrieval SIGIR, *http://www.acm.org/sigir/*

[18] E.M Voorhees. *Overview of the TREC 2004 Question Answering Track*. In the Notebook of the Thirteen Text Retrieval Conference (TREC 13), TRECVID 2004.

[19] TRECVID, TREC Video Retrieval Evaluation, *http://www-nlpir.nist.gov/projects/trecvid*

[20] TREC, Text Retrieval Conference, *http://trec.nist.gov*

[21] T. Westerveld and A. P. Vries. *Experiment Results Analysis for Generative Probabilistic Image Retrieval Model*. Proc. of SIGIR 2003, Canada, Jul 2003.

[22] R. Yan, J. Yang, and A. G. Hauptmann. *Learning Query-Class Dependent Weights for Automatic Video Retrieval*. Proc. of ACM MM, New York, Oct 2004.

[23] H. Yang, T.-S. Chua, S. Wang and C.-K. Koh. *Structured use of external knowledge for event-based open-domain question-answering*. Proc. of SIGIR 2003, Canada, Jul 2003.

[24] H. Yang, L.Chaisorn, Y. Zhao, S.Y. Neo, T.S. Chua, *VideoQA: question answering on news video*, Proc. of ACM MM, Berkeley, 632-641, Nov 2003.

[25] Y.M. Yang, and J. O. Pedersen. *A Comparative Study on Feature Selection in Text Categorization*, Proceeding of the Fourteenth Int'l Conf on Machine Learning, 1997.

[26] Reference is removed for purpose of anonymous review.