TRECVID 2004 Search and Feature Extraction Task by NUS PRIS

Tat-Seng Chua, Shi-Yong Neo, Ke-Ya Li, Gang Wang, Rui Shi, Ming Zhao and Huaxin Xu
School of Computing, National University of Singapore

Qi Tian, Sheng Gao and Tin Lay Nwe
Institute for Infocomm Research

ABSTRACT

This paper describes the details of our systems for feature extraction and search tasks of TRECVID-2004. For feature extraction, we emphasize the use of visual auto-concept annotation technique, with the fusion of text and specialized detectors, to induce concepts in videos. For the search task, our emphasis is two-fold. First we employ query-specific models, and second, we employ multi-modality features, including text, annotated visual concepts, OCR output, shot classes and specialized detectors to perform the search. Our search pipeline is similar to that employed in text-based definition question-answering approaches. We first perform query analysis to categorize the query into the categories of: {PERSON, SPORTS, FINANCE, WEATHER, DISASTER and GENERAL}. From these categories, we induce a number of constraints on the search process, including: (a) the type of multi-modality features to use or emphasize; (b) the key concept terms in text query to use; and (c) the video classes, such as sports or anchor person etc to use or exclude in the search results. The results on 60 hours of test video from TRECVID 2004 evaluation demonstrate that our approaches are effective.

1. Introduction

This year we participate in both feature extraction and manual search tasks. For feature extraction, our emphasis has been on the application of auto-concept annotation technique, with the fusion of text and specialized detectors, to induce concepts in videos. We adopt the by-now rather standard approach to perform auto concept annotation. Our approach first segments the images into fixed 4x4 blocks, followed by the clustering of blocks, before learning the associations between concepts and block clusters using a probabilistic SVM. The visual features used for image block includes color histogram, edge histogram, and the adaptive matching pursuit feature for texture. The training data is derived from TRECVID 2003 development set with pre-assigned concepts. Although our earlier research has demonstrated that the use of segmented regions and constraint-based clustering based on language model have been effective, we were unable to apply these techniques due to efficiency consideration. In order to enhance the effectiveness of the essentially visual-based auto-concept annotation approach, we employ two additional techniques. First, we use the feature name to induce additional context terms by mining related text terms in ASR and closed caption texts in the training data set. We then employ the expanded set of text terms to locate useful groups of shots at the speaker change level. Second, we employ specialized visual detector, such as the face detector, to provide further evidence for face-related classes. Our technique is generic for all feature classes except for person-related classes, where the face detector is used.

For the search task, our emphasis is two-fold. First we employ query-specific models, and second, we employ multi-modality features, including text, annotated concepts from earlier feature extraction task, OCR output, shot classes and specialized detectors, such as face detector, face recognizer and speaker detector etc., to perform the search. Our search pipeline is similar to that employed in text-based definition question-answering approaches. We first perform query analysis to categorize the query into the categories of: {PERSON, SPORTS, FINANCE, WEATHER, DISASTER and GENERAL}. From these categories, we induce a number of constraints on the search process, including: (a) the type of multi-modality features to use or emphasize; (b) the key concept terms in text query to use; and (c) the video classes, such as sports or anchor person etc to use or exclude in the search results. For example, for concept related to Person, we emphasize the use of named entities in text, the occurrence distribution model between faces and named entities, face detector, face recognizer, and speaker detector, and exclude shots of classes anchor-person, commercial, sports, weather and finance. Given the key terms in query, we induce the context of query in two ways: (a) by using ASR text surrounding the sample videos supplied with the text query; and (b) by extracting terms with high mutual information values from relevant articles retrieved from Google News, constrained by language resource called WordNet. The expanded query is used to retrieve relevant video segments at the speaker change level. We then employ the query model to incorporate appropriate high-level audio and visual features to re-rank the results. The model for fusion is dependent on the type of query as described above. Optionally, we perform pseudo relevance feedback by using the top ten ranked results as relevant to perform one round of relevance feedback. We submitted 6 runs by varying the use of different combinations of multi-modality features, including

the use of just text, and pseudo relevance feedback. Our results indicate that the use of query specific model on the full set of multi-modality features, together with one round of pseudo relevance feedback, produces the best results.

2. Feature Extraction Task

The following sections describe the details of our approach for feature extraction task. Our proposed framework combines text and visual modalities together with face and timing information.

2.1 Visual Feature Extraction Technique

For visual modality, we begin with the annotated development corpus of TRECVID 2003. First, we select low-level visual features that can be automatically extracted and are essential to differentiate one class from the others. The visual features used are: Luv color histogram (69), adaptive Matching Pursuit texture features (16) [17] and edge histogram (6). Thus, the dimension of a visual feature vector is 91. These features are extracted from fixed-size 4x4 blocks within key-frames. We then develop two methods to detect feature concepts, and these methods are named as "Image Level Method" and "Block Level Method". The training pipeline for visual feature extraction is shown in *Figure 1*. The definition of classes C_1, C_2,…, C_{11} are depicted in *Table 1*. The pipeline of testing step is the same as that of training step.

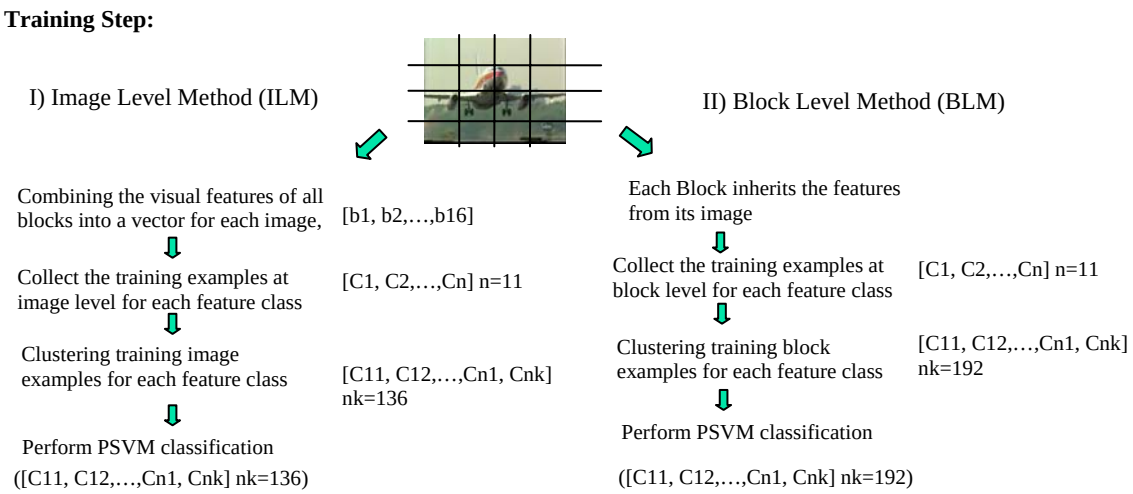

Figure 1: Pipeline of Visual Feature Extraction

Table 1: Feature Concepts Classes

Boat/Ship (C_1)	Madeleine Albright (C_2)	Bill Clinton (C_3)	Train (C_4)	Beach (C_5)	Basket scored (C_6)
Airplane takeoff (C_7)	People walking/running (C_8)	Physical violence (C_9)	Road (C_{10})	Unknown (C_{11})	

Image Level Method (ILM) constructs a feature vector at the image level for each key-frame by arranging the visual features of the blocks in a pre-defined order. Thus, a data point in visual feature space represents an entire key-frame. The advantage of ILM is that spatial associations between fixed-size blocks within a key-frame can be utilized to help detecting the feature concepts.

Block Level Method (BLM) treats each block in a key-frame separately and a data point in low-level visual space of BLM is just a 91-dim block vector. Each block inherits the feature concept from its key-frame. The advantage of BLM is that it should be more effective in detecting object-based feature concepts as compared to IML due to its localized nature. For both methods, the clustering results in respective feature space are used as inputs to a probabilistic SVM (PSVM) [18]. The number of clusters we derived is 136 for ILM and 192 for BLM.

2.2 Text-based Feature Extraction Technique

The text modality aims to tackle two problems in our system. They are: a) what text unit is appropriate to infer visual concept by utilizing text information (ASR and CC); and b) given a feature, how to select informative keywords for such a feature class. Borrowing the ideas from text categorization, we construct a pipeline of term extraction techniques to solve the above problems. The pipeline is shown in *Figure 2*.

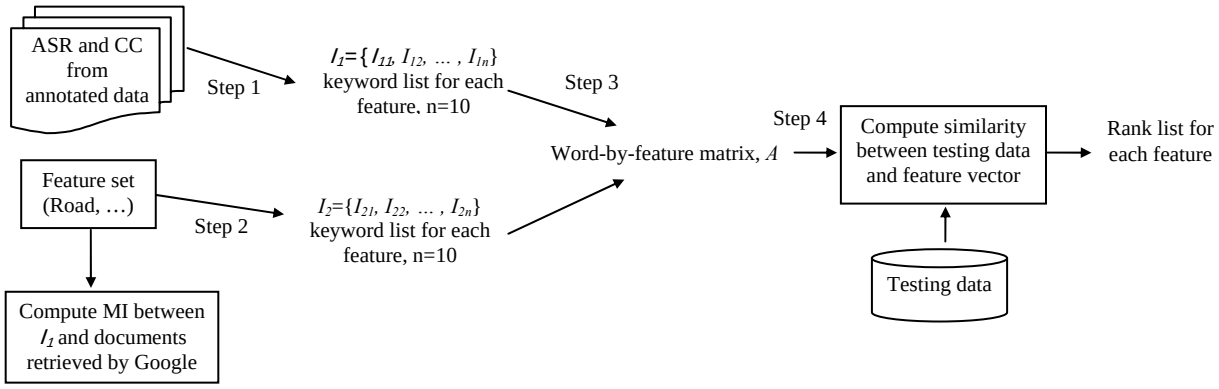

Figure 2: Pipeline of Text-based Feature Extraction

For problem (a), the appropriate text unit could be a shot, a speaker change interval or a story. Based on our empirical study of different text unit size, we choose speaker change level as the basic text unit for detecting feature concepts. The other reason for using speaker change as analysis unit is that such boundaries are readily available from ASR output. In our system, each feature concept corresponds to a category.

Due to insufficient training examples and the diversity of keywords for news, we use two methods, DF (Document Frequency) and MI (Mutual Information) [21], to select informative keywords for each feature concept. The training data comes from ASR and CC of the development corpus of TRECVID 2003, where the text unit at the speaker change level has been annotated by feature concept vectors, $F = \{ f_1, f_2, \ldots, f_{10}\}$, which stands for the textual feature vector of the 10 visual concepts.

First, in Step 1 of *Figure 2*, we employ DF (Document Frequency) to select informative keywords for a feature concept, $I_1 = \{I_{11}, I_{12}, \ldots, I_{1n}\}$ where $I_{1i} \in I_1$, and $(1 \le i \le 10)$, represents the keyword set for feature $f_i \in F$. The keywords whose DF is less than a predefined threshold are removed from $I_{1i} \in I_1$. However, DF has the drawback that when the number of training examples is very small or the feature concept class contains many diversified keywords, it is not very effective in measuring the importance of a keyword to a category. Thus to overcome this problem, we use Mutual Information (MI) [15] between keywords from ASR and CC, and documents retrieved from Google search engine, to select top k $(k=5)$ for each feature concept category. Here, $I_2 = \{I_{21}, I_{22}, \ldots, I_{2n}\}$ where $I_{2i} \in I_2$, and $(1 \le i \le 10)$ represents the keyword set for feature $f_i \in F$ shown in step 2.

Step 3 constructs a *word-by-feature matrix*, $A = [a_{ij}]$. The element of A, a_{ij}, is a weight between keyword and its feature defined by *Equation (1)*:

$$a_{ij} = \begin{cases} 0, & w_i \notin I_{1j} \cup I_{2j} \\ 1, & w_i \in I_{1j} / I_{2j} \quad w_i \in I_{2j} / I_{1j} \\ 2, & w_i \in I_{1j} \cap I_{2j} \end{cases} \qquad (1)$$

The basic idea behind such a definition is that keywords supported by both DF and MI measures are more informative, and therefore should be assigned higher weights. In Step 4, given a test data $T = \{ T_1, T_2, \ldots, T_s\}$ at the speaker change level from the test corpus, we compute the similarity for T with respect to each feature $f_j \in F$ as:

$$Sim(T, f_j) = \sum_{i=1}^{m} a_{ij} \times \arg\max_{t_k \in T}(MI(w_i, t_k)) \qquad (2)$$

Here we use the documents retrieved from Google search engine to compute $MI(w_i, t_k)$ the mutual information between the keyword w_i and $t_k \in T$, $m = |I_1 \cup I_2|$ We use *Equation (2)* to rank the text-based results at the speaker change level.

2.3 Overall Framework and Results

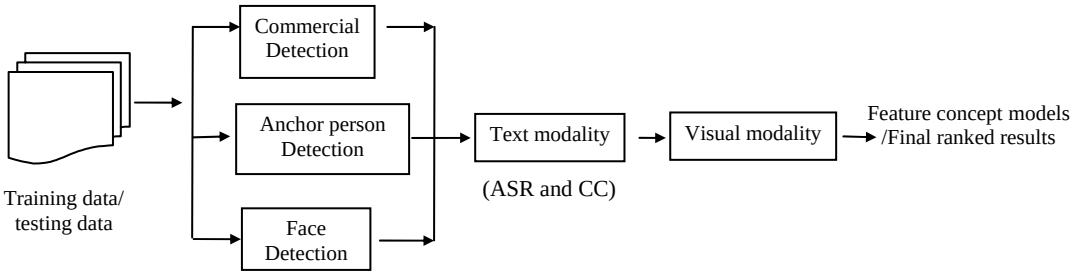

Figure 3: Framework for Feature Extraction

Our overall framework is shown as *Figure 3*. We use the methods in [3] to perform commercial, anchor person and face detection at the shot level. We then use the text analysis as outlined in *Section 2.2*. to select a ranked list of text units at the speaker change level. As each speaker change utterance contains an average of 4-5 shots, we need additional features to re-rank the text retrieval output at shot level. Here we use the similarities of visual modalities to compute the final ranked list of shots for each feature category $f_j, f_j \in F$.

Table 2 summarizes the use of different combinations of modalities to detect different feature concepts. The reasons that most of the features can be detected well by IML is because the spatial associations between blocks are important to detect the concepts. Thus, in the future, we plan to develop a mathematical model to describe such a spatial associations to improve the performance of BLM. *Figure 4* summarizes our final evaluation results for TRECVID 2004, which shows that most of our results are above the median level.

Table 2: Respective Features with their methods

Features	Methods
Bill Clinton	Text + Face detector/recognizer + Time Info
Boat, Madeleine Albright, Train, Airplane takeoff, People walking/running, Road, Physical violence	Text + ILM
Basketball scored	Text + BLM
Beach	IML

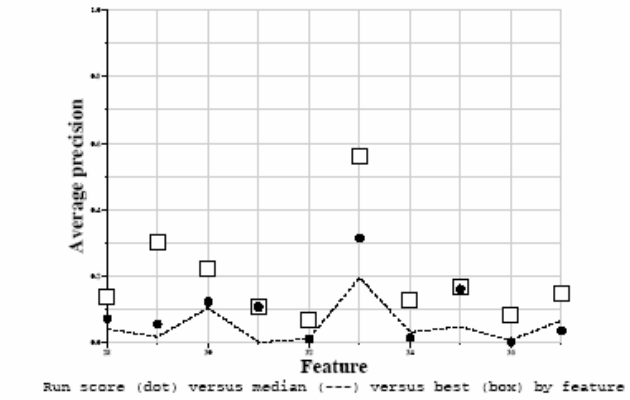

Figure 4: Breakdown of Results

3. Automatic/Manual Search Task

For the search task of TRECVID 2004, we focus on the design and implementation of an automatic news video retrieval system. We emphasize on the multi-class question analysis of text query to determine the appropriate retrieval methods. Subsequently, we perform expansion and utilize external resources like WordNet and Web to provide supplemental knowledge that may not be available in the video contents. *Figure 5* shows the overall framework of the retrieval system. In the following sub-sections, we will introduce the structure of our news video retrieval system.

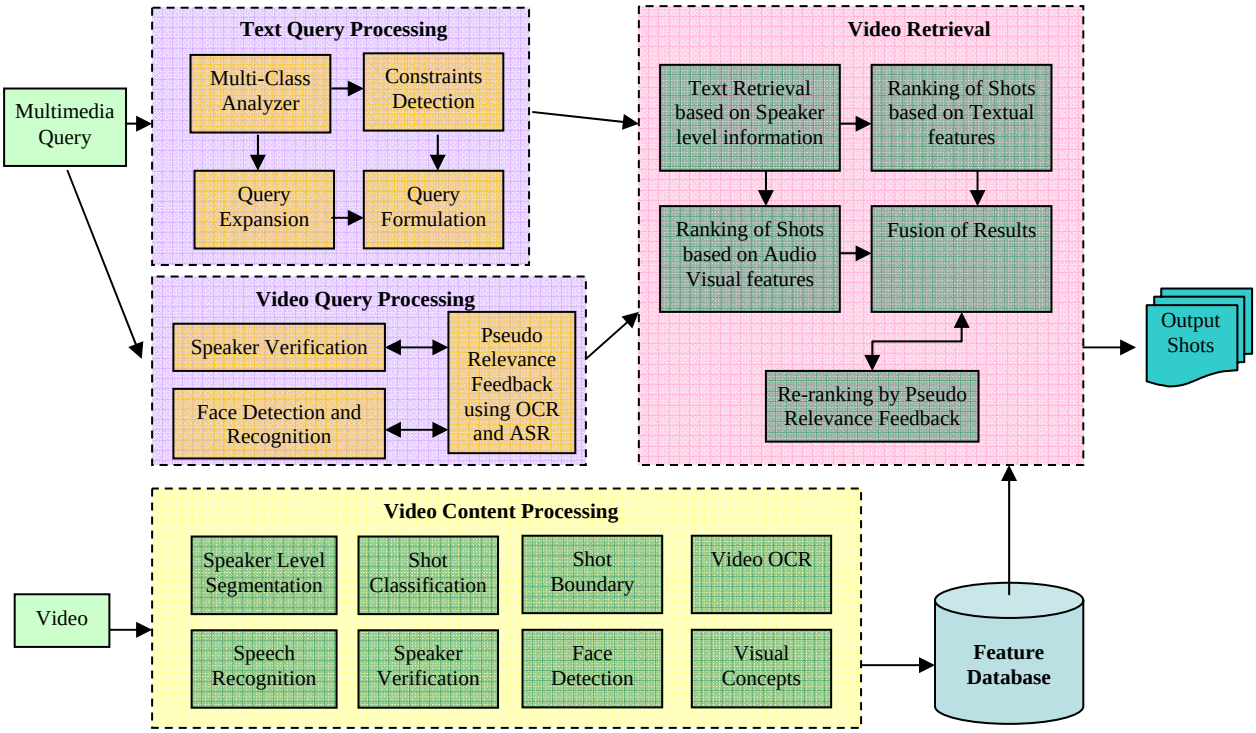

Figure 5: Overview of Video Retrieval System

3.1 Multi-class Query Analysis

Combining search result from multiple modalities is an important task in video retrieval system. This is especially true when we perform manual search task 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. Though manually assigning weights to different features is easy, the use of uniformed query-independent weighting scheme does not address the problem that different query-classes may have very different feature-weights from the rest. 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 we classify queries into several pre-defined classes based on our knowledge of news video. After which we try to optimize the weighting scheme for each of these classes. Besides using individual query-class weighting for the combination of different modality features, the query-class also provide us with the information to reduce the search target size by filtering out those shots belonging to categories that are not related to the query-class.

3.1.1 Defining Query Class

In a query classification scheme proposed by [18], 4 different query classes were proposed within the domain of general news videos. The four classes are Named person, Named object, General object and Scene. For each of the query-class, they performed machine learning to determine the various weights to be given to each feature. For our system, we use a query classification scheme which is closely associated to the news category of the shots. The rationale is simple: if the system is able to detect which news categories that the answer will come from, the chances of retrieving the correct shots are greatly increased. This is because sports queries will naturally require a sport news answer. However, it is not possible to use the short text query to perform query classification to high accuracy. Therefore, we performed our own news video classification into 6 query-classes based on the intuition that they can be classified by using simple rules. The six query classes are:

PERSON: queries looking for a person, together with other constraints. 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 with other constraints. 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"

3.1.2 Query Analysis and Classification

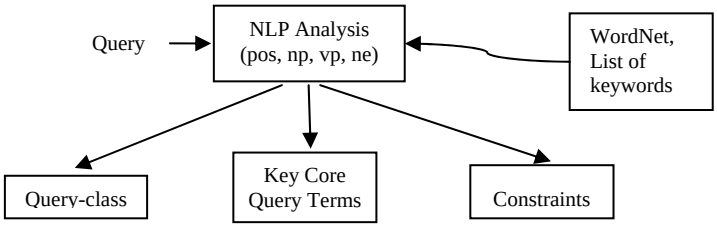

Figure 6: Overview of Multi-class query analysis

As illustrated in *Figure 6,* the 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 on Part-of-Speech(POS), verb-phrase and noun-phrase as is done in [20]. We then extract the main core terms of the query using our core term extractor (the strongest noun or noun phrase), followed by the other possible key terms. After which we employ the Name Entity extractor to identify names of persons, organizations and possible objects in the main query. Given these information, we develop a rule-based query classifier to identify the query-class which is essentially a text categorization problem. For PERSON class, we use mainly the Named Entity result from NE extractor. If a person's name appears in the query as the core term, it will belong to that class. For the other 4 classes other than GENERAL, we extract a list of keywords designed for each class from a set of training samples. Because the query text is usually short, we use simple keyword matching techniques to classify the queries into the various classes. Those queries that do not belong to any of the 4 classes will be classified as GENERAL class. Some of the classifications of the queries for this year's search task are shown in *Table 3*.

Table 3: Query Analysis

Topic	Query-class	Constraints	Core terms	Class
0125	Find shots of a street scene with multiple pedestrians in motion and multiple vehicles in motion somewhere in the shot.	in motion somewhere	street	GENERAL
0126	Find shots of one or more buildings with flood waters around it/them.	with flood waters around it/them	Buildings, flood	DISASTER
0128	Find shots of US Congressman Henry Hyde's face, whole or part, from any angle.	whole or part, from any angle	Henry Hyde	PERSON
0130	Find shots of a hockey rink with at least one of the nets fully visible from some point of view.	one of the nets fully visible	hockey	SPORTS
0135	Find shots of Sam Donaldson's face - whole or part, from any angle, but including both eyes. No other people visible with him	whole or part, from any angle, but including both eyes. No other people visible with him	Sam Donaldson	PERSON

3.1.3 Using the Query Class

From the query-class, we are able to predict the category of news video shots that are likely to contain the answers. For the tagging of news categories, we use the techniques described in [2]. We classify the shots into the categories of general, sports, finance, weather and anchor-person. Once the query-class is determined and the news categories are tagged, it is then possible to perform filtering of non-relevant shots. *Table 4* shows the various query-classes and their corresponding news categories where relevant shots may be found.

Table 4: Corresponding target types for each class

Query-class	Target News Categories
PERSON	General_non-anchor
SPORTS	Sports
FINANCE	Financial
WEATHER	Weather
DISASTER	General_non-anchor
GENERAL	General_non-anchor

Intuitively, each query-class will exhibit different characteristics and will require different evidence to induce the answers. For example, speaker identification 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. Therefore within each of the query-classes, we use labeled training corpus to train the weights of various features. This is similar to the methods describe in [19]. The shots in the training set are first manually labeled accordingly to the 6 query-classes. We extract the features of each shot using our detectors and train the model $\boldsymbol{M_{class}}$ for each class using probabilistic modeling. *Table 5* shows the importance of each modality feature for different query-class $\boldsymbol{M_{class}}$. Subsequently, each $\boldsymbol{M_{class}}$ is used as the base model for our ranking.

Table 5: $\boldsymbol{M_{class}}$ for Fusing Multi-modality Features

Class	**Weight of NE in Expanded terms**	**Weight of OCR**	**Weight of Speaker Identification**	**Weight of Face Recognizer**	**Weight of Visual Concepts(total of 10 visual concepts used)**					
					People	basketball	hockey	water-body	fire	etc
PERSON	High	High	High	High	High	Low	Low	Low	Low	.
SPORTS	High	Low	Low	Low	Low	High	High	Low	Low	.
FINANCE	Low	High	Low	High	Low	Low	Low	Low	Low	.
WEATHER	Low	High	Low	High	Low	Low	Low	Low	Low	.
DISASTER	Low	Low	Low	Low	Low	Low	Low	High	High	.
GENERAL	Low	Low	Low	Low	High	Low	Low	Low	Low	.

3.2 Text Retreival

Because 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 perform query expansion in various ways and combine the strengths of all the techniques to obtain the best matches.

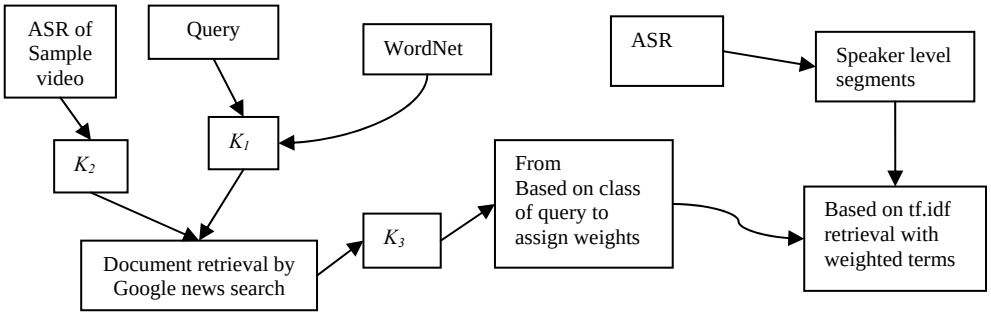

Figure 7: Text Retrieval with Query Expansion

The unit of retrieval is a single speaker change, which is provided by LIMSI ASR system [9]. As illustrated in *Figure 7*, we first extract non-trivial keywords from the query to form a bag of initial key words. To ensure recall, these keywords are also expanded by its glossary and Synset from WordNet to form K_1. For each query, there are some video samples provided together with the query. The ASR surrounding the video samples is extracted to form K_2. Next, we perform query expansion based on K_1 and K_2 by using the documents retrieved from Google search engine to extract terms that co-occur frequently with K_1 and K_2. We union the results of query expansion together with K_1 and K_2 to form the final weighted term set K_3. We then re-weight K_3 terms based on knowledge of query-class. For example, higher weights are given to Name Entities of type Human_Person if the query class is of type SPORTS. We use K_3 to perform standard retrieval in the test corpus at the speaker change level. A set of ranked speaker change level segments are retrieved for further re-ranking based on other modality features.

3.3 Classification of News Shots

Corresponding to our query-class definition, we pre-classify the shots of news video into 6 categories: sports, financial, weather, commercial, anchor-person and general (non-anchor). Text, visual and timing features are used to detect different categories of shots. Certain shot types like commercials, finance and weather, have well-defined and rather fixed temporal-visual characteristics and can be detected using specific detection techniques. For sport news, we use a combination of motion, speech and visual to perform the classification [2]. As for the rest of the news that

does not fall into the above four categories, we classify them as general news, which comprises anchor and non anchor shots. The following sub-sections describe the details of various detection techniques that we have used.

3.3.1 Commercial Detection

We use black frames, silence, cut rate and low confidence in ASR outputs (TRECVID2003 data) as the features. The algorithm first detects black frames with sufficiently long audio silence. Second, the algorithm searches for the next block of black frames. Third, it determines the cut rate and the confidence level of the ASR output of shots residing within these two blocks of black frames. If it detects sufficiently high cut rate and low confidence values in the ASR output, the algorithm will classify these shots into commercial category.

3.3.2 Weather/Finance Detection

As the weather and finance shots have distinct visual characteristics and appear in similar temporal location, we use 176 LUV color histogram to model the contents of representative key frames of each category, together with domain knowledge (estimate the search space by finding where is the block of Weather or Finance). The algorithm compares the test key frames with representative key frames stored in the database using image matching technique. We then assigned the majority-category of the top *n* retrieved representative key frames to the test key frame.

3.3.3 Anchor shot Detection

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* 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* shots.

3.4 Extraction of Other Modality Features

We employ a number of modality features to help in determining the answers. This includes face recognition, video OCR, speaker identification and visual concept detection. The visual concept detection is discussed in *Section 2*.

3.4.1 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 a heuristic for face recognition. Appearance time distribution indicates the probability of person-X's appearance in different time of a video clip, while appearance shot distribution indicates the probability of person-X's appearance in different shots beside 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 detect a face, we apply a 2DHMM to classify them [5]. For each of the detected faces in the test set, we use Viterbi Algorithm to calculate the possibility of each model. Then, we consider the top three models, which are denoted as M_1, M_2, M_3. where And P_{M1}, P_{M2}, P_{M3} are their probabilities. We use θ_{personX} to indicate if a model is that of personX.

$$\theta_{\text{personX}}(M) = \begin{cases} 1 & \text{if } M \text{ is personX's Model} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The probability of face recognition is:

$$P_{face} = \frac{1}{1111}\left(\theta_{\text{personX}}(M_1)\cdot 10^3 + \theta_{\text{personX}}(M_2)\cdot 10^2 + \theta_{\text{personX}}(M_3)\cdot 10^1 + (P_{M1} - P_{M2})\cdot(2\theta_{\text{personX}}(M_1) - 1)\right) \quad (4)$$

The final probability of person's appearance is $P_{\text{personX}} = P_{time} \cdot P_{shot} \cdot P_{face}$.

3.4.2 Video OCR

The OCR results are donated from CMU [16]. Even though there were a number of insertion, deletion and mutation errors, video OCR prove to be a good feature as it is able to give precise information on the appearance of certain human person. Therefore we have integrated 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 overall effectiveness of OCR and MED.

3.4.3 Speaker Identification

The speaker identification follow the techniques describe in [11]. In addition, we introduce a pseudo relevance feedback loop-back using face detector and OCR. We first extract the MFCC features of the speech segment and trained a model for each speaker using HTK [12]. In the first training instance, we use of the speech segments in the sample video to derive M_1. Next, we use the ASR from LIMSI to retrieve possible speech segments that could be made by speaker X using means of text retrieval and expansion as mentioned above. Then, the possible segments are run through M_1. The results of the identification are further justified by using the Video OCR and face detector. We then make use of the highly possible segments as our new training instances to obtain M_2. Finally we use M_2 to identify all other speech segments within the test set.

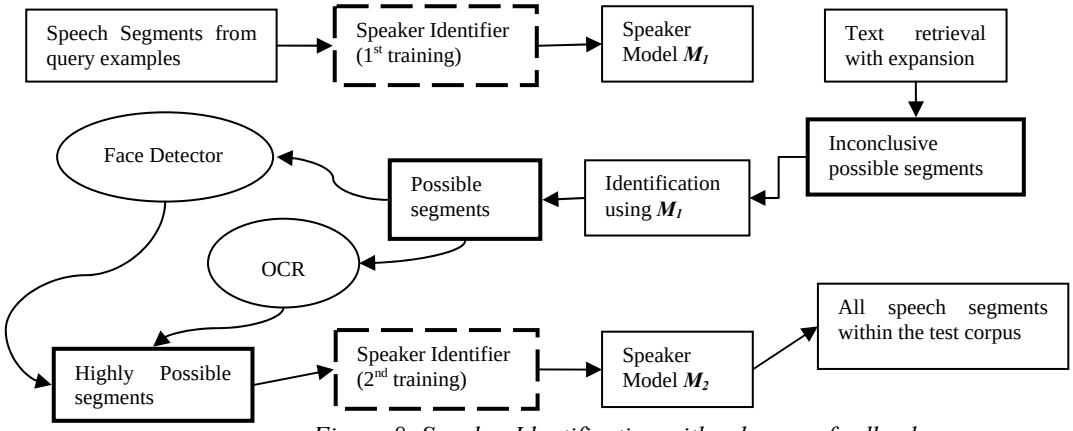

Figure 8: Speaker Identification with relevance feedback

3.5 Fusion and Pseudo Relevance Feedback

For each query-class described in Section 3.1, we train a model M_{class}. Each model defines a different set of coefficients α_i that model the importance of different modality features for that query-class. After obtaining the results from different modality feature detectors, we combine the scores for shot *S* using *Equation (5)*.

$$Final\_Score(S) = \sum_{allmoalities} \alpha_i^M * Score_i \quad \text{where} \quad \sum_{allmoalities} \alpha_i^M = 1 \qquad (5)$$

For pseudo relevance feedback, we make use of the top *n* returned shots as our positive instances. Here we set *n* to 10 empirically. We use the textual information in these *n* shots to obtain a list of additional keywords K_4. We then perform a similarity- based retrieval using both K_3 and K_4, and re-rank the respective shots. Currently, we do not perform pseudo relevance feedback on non-text features.

3.6 Evaluations

We submitted 6 manual runs to TRECVID 2004 for evaluation. They are:

Run1: Use textual output from ASR and close caption and text descriptions of the topics.

Run2: Use textual output from ASR and text descriptions of the topics with expansion from external resources.

Run3: Employ textual output from ASR and text descriptions of the topics with expansion and OCR, visual concepts with shot classification and speaker identification.

Run4: Utilize textual output from ASR and text descriptions of the topics with expansion and OCR, visual concepts, shot classification, speaker identification and face recognition.

Run5: Utilize textual output from ASR and text descriptions of the topics with expansion and OCR, visual concepts, shot classification, speaker identification and face recognition with more emphasis on video OCR

Run6: Combine textual output from ASR and text descriptions of the topics with expansion and OCR, visual concepts, shot classification, speaker identification and face recognition with one round of pseudo-relevance feedback

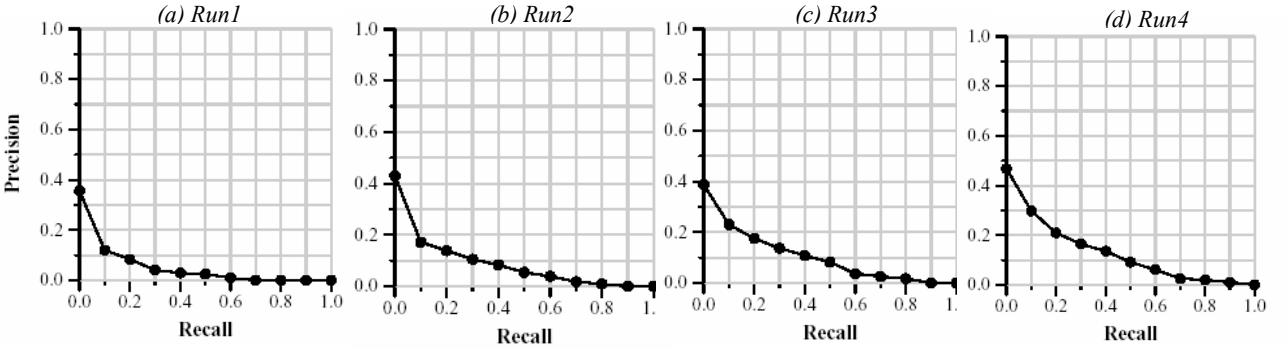

Figure 9: The First 4 Runs

Run1 is a baseline text search run where we only involve the query and ASR. This run gives us the baseline retrieval capabilities of using only text query. Run2 demonstrates the usefulness of query expansion to extract context of short queries. In Run3, we combine some of the modality features such as video OCR, speaker identification and visual concepts. The improvement is significant as we can see that at 0.1 recall point, the precision doubles as compared to Run1. Adding visual features and shot classification to complement text retrieval helps in certain queries which is not possible to obtain answers by using text alone. This is particularly true for PERSON class queries, where the target shot classes are general news (non-anchor) and video OCR and speaker identification have been demonstrated to be very useful. Thus in Run4, we utilized all modality features including face recognition to obtain even better performance.

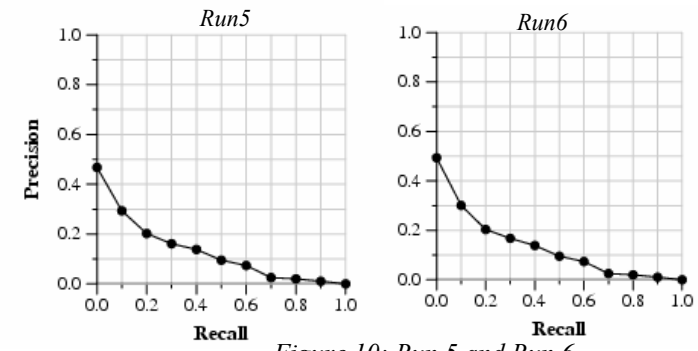

Figure 10: Run 5 and Run 6

In Run5, we give OCR and speaker identification modules higher weights during the fusion stage. And in Run 6, we perform a round of pseudo relevance feedback using the top 10 returned results.

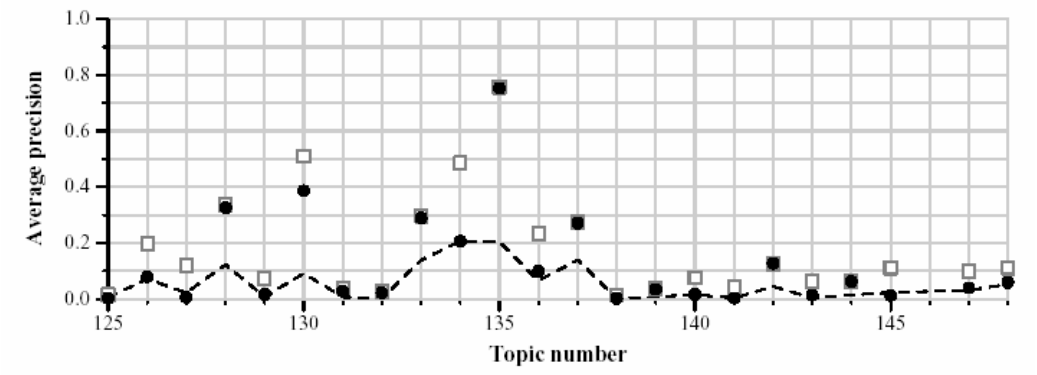

Figure 11: Run 6 results breakdown

Figure 11 shows the result of our best run, which has achieved a mean average precision (MAP) of 0.124. We achieve best or close to the best results for 10 of the queries. From the evaluations, our systems are generally able to perform well for queries related to human and sports.

4. Conclusion

This paper discusses the details of our participation in TRECVID 2004 evaluation. We described the framework and techniques we employed for feature extraction and automated/manual search task. For feature extraction, our system focused on the application of visual auto-concept annotation technique, with the fusion of text and specialized detectors, to induce concepts in videos. In the Search task, we focused on the use of query classes to associate different retrieval models for different query classes. We employed multi-modality features including text, OCR, image, visual and audio features to support the retrieval. The evaluation result shows that our system has good overall performance, and we performed especially well for human and sports class queries.

5. Acknowledgments

The authors would like to thanks Institute for Infocomm Research (I2R) for the support of the research project "Intelligent Media and Information Processing" (R-252-000-157-593), under which this project is carried out.

References

[1] J. Boreczky, and L. Wilcox, "A Hidden Markov Model Framework for Video Segmentation Using Audio and Image Features." In the proceedings of the International Conference on Acoustics, Speech, and Signal Processing (Seattle, WA), Vol. 6, 1998, pp. 3741-3744., May 12, 1998

[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, TRECVID 2003 Workshop, pp. 129-134, November 17-18, 2003

[4] 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.

[5] M. Y. Chen and A. Hauptmann, "Searching for a specific person in broadcast news video," in proceedings of the International Conference on Acoustic, Speech and Signal Processing, May 2004, vol. 3 pp. 1036-1039.

[6] T.-S. Chua, C. Chu and M.S. Kankanhali. Relevance feedback techniques for image retrieval using multiple attributes. Proceedings of IEEE International Conference on Multimedia Computing and Systems (ICMCS'99). Florence, Italy. Jun 1999. 890-894.

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

[8] 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. 1pp.119-139, August 1997.

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

[10] Google Search Engine. http://www.google.com

[11] A Hauptmann, R. Jin., and T. D. Ng. Video Retrieval using Speech and Image Information. In Proceedings of Electronic Imaging Conference (EI'03), Storage and Retrieval for Multimedia Databases, Santa Clara, CA, January 20-24, 2003

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

[13]C. Kenneth., and P. Hanks. Word Association Norms, Mutual Information, and Lexicography. In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics. 1989

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

[15]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.

[16]T. Sato, T. Kanade, E. Hughes, M. Smith, and S. Satoh "Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Caption", ACM Multimedia Systems Special Issue on Video Libraries, February, 1998.

[17]R. Shi, H. Feng, T.-S. Chua, C.-H. Lee. An Adaptive Image Content Representation and Segmentation Approach to Automatic Image Annotation, CIVR 2004, pp. 545-554.

[18]Support Vector Machine: http://bach.ece.jhu.edu/svm/ginisvm.

[19]R. Yan, J. Yang, and A. G. Hauptmann. Learning Query-Class Dependent Weights for Automatic Video Retrieval

[20]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.
[21]Y.M.Yang, and J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization, Proceeding of the Fourteenth International Conference on Machine Learning, pp. 412-420, 1997.