TRECVID 2005 by NUS PRIS

Tat-Seng Chua, Shi-Yong Neo, Hai-Kiat Goh, Ming Zhao, Yang Xiao and Gang Wang
School of Computing, National University of Singapore

Sheng Gao, Kai Chen, Qibin Sun and Tian Qi
Institute for Infocomm Research

ABSTRACT

We participated in the high-level feature extraction and search task for TRECVID 2005. For the high-level feature extraction task, we make use of the available collaborative annotation results for training, and develop 2 methods to perform automated concept annotation: (a) a ranked-Maximal Figure-of-Merit (MFoM) method; and (b) a multimodal rankBoost fusion method. We submitted a total of 7 runs based on these two methods. For the search task, we focus on improving our previous retrieval system by utilizing an event entity model derived from relevant external resources. In addition, we also make use of the various high-level feature extraction results contributed by various participating groups to help in the re-ranking step. We submitted a total 6 runs in the automated search category. The evaluation results show that our event-based approach is effective in human/event queries and that the high-level features is useful for general queries.

1. HIGH LEVEL FEATURE EXTRACTION TASK

We explore two methods to perform high-level feature extraction. The first is based on a ranked-Maximal Figure-of-Merit (MFoM) method that has been successfully employed in text categorization. The second employs HMM for high-level features extraction, follow by rankBoost fusion to fuse with other modality features.

1.1 Ranked Maximal Figure-of-merit (MFoM)

Conventional approaches for semantic concept detection is to train a binary classifier (e.g. SVM, Boosting) from the training set by optimizing generalized classification error or maximizing the likelihood. In the high-level extraction task, our concern is ranking the relevant shots as high as possible. Therefore, the basis of this approach lies in learning an optimal ranking function in terms of the mean average precision (MAP) from the development dataset, given any type of multimedia content representation (text from ASR, visual feature, etc). Here we develop an algorithm for training the ranking function with the goal of optimizing the MAP. This algorithm is similar to our work in (Gao et al., 2003 & 2004) and the ROC optimization for classifier (Cortes & Mohri, 2003; Yan, et al, 2003), where the objective function to be optimized is derived using an approximation of the interested metric for evaluation. A good measure for ranking is the Wilcoxon-Mann-Whitney statistic, which is equal to the area under the ROC (AUC), defined over the training set as:

$$U = \frac{\sum_{i=}^{m}\sum_{j=1}^{n} I(x_i, y_j)}{mn} \qquad (1)$$

where, $I(x_i, y_j) = \begin{cases} 1: & x_i > y_j \\ 0: \text{Otherwises} \end{cases}$, with x_i and y_j being the scores from the classifier for the *i-th* out of the *M* positive samples, and the *j-th* out of the *N* negative examples. A classifier is then trained by maximizing ***Eq. (1)*** using a gradient algorithm. A sigmoid function (*see* ***Eq. (2)***) is used to approximate the correct ranking count, $I(x_i, y_j)$, with

$$S(x_i, y_j) = \frac{1}{1 + e^{-b(x_i - y_j)}} \qquad (2)$$

where b is a constant function. After smoothing, ***Eq. (1)*** becomes a differentiable function. The smoothed ***Eq. (1)*** is a function of the parameters of the classifier (embedded by x_i and y_j) and will be the objective for optimisation. Because it is highly non-linear, the gradient descent algorithm is applied to find its solution as in (Gao et al, 2003 & 2004). The ranking optimisation algorithm, named Rank-MFoM, is derived from MFoM learning in (Gao et al, 2003 & 2004). We submitted 4 runs using the Ranked-MFoM algorithm. They are based on: (a) only text feature; (b) only

texture feature; and (c-d) fusions of both features using two different settings (e.g. b , the learning rate, and the iteration cycles) for the Rank-MFoM algorithm. The results are presented below.

Run A (TRECVID Run 6): The text only run. The linear classifier is trained by Rank-MFoM on the shot-level text documents comprising the ASR outputs within 3-window shots. A lexicon with 3,464 terms is extracted from the ASR outputs in the development set after the removal of stop words and both very rare and frequent words. Each text document is represented using the tf-idf feature. In this run, shots without the ASR outputs are ignored.

Run B (TRECVID Run 7): The texture only run. The image is uniformly segmented into 77 grids with a size 32x32. We extract a 12-dimensional texture feature (energy of log Gabor filter) from each grid. Subsequently, we apply Gaussian Mixture Model (GMM) to model the texture distributions for the positive and negative shots. The ranking function is the log-likelihood ratio between the positive and the negative class, defined for an image as,

$$L(X) = \frac{1}{77}\left(\sum_{t=1}^{77} \log P(x_t|C_0) - \sum_{t=1}^{77} \log P(x_t|C_1)\right) \quad (3)$$

$$\text{with } p(x|C_i) = \sum_{j=1}^{M} w_j^i \cdot \mathrm{N}\left(m_j^i, \Sigma_i^i\right), i = 0,1 \quad (4)$$

where M is the number of Gaussian mixture (here we use M=4); $N\left(m_i^j, \Sigma_i^j\right)$ is the *j-th* Gaussian distribution with the mean, m_i^j, and covariance, Σ_i^j, for the positive *i-th* class; and w_i^j the corresponding weight for each Gaussian component. Here C_0 is the positive class and C_1 is the negative class. For each concept, the parameters, w_i^j, m_i^j and Σ_i^j, are estimated using the Rank-MFoM from images available in the development set.

Run C (TRECVID Run 5): The fusion run 1. Since the output of the texture-based classifier is at the key-frame level, we simply choose the maximal likelihood ratio score among all key-frames for a shot as its shot-level output. We then combine the output of the text-based classifier to form a 2-dimensional vector. For shots without the ASR outputs, we treat it as the missing feature rather than removing them from the training set for fusion. The 2-dimension vector is fused by training a single mixture GMM using the Rank-MFoM algorithm.

Run D (TRECVID Run 3): The fusion run 2. This run is different from Run C as the shot-level score is taken to be the average of all key-frames for the shot.

To illustrates the effectiveness of Rank-MFoM as compared to standard SVM using text feature, we perform a preliminary run on a subset of videos in the development set. Table 1 tabulates the MAP scores for the 10 concepts for the top 2000 shots. The results indicate that Rank-MFoM out-performs SVM on auto concept annotation task.

Table 1: Preliminary results for Rank-MFoM vs. SVM on a subset of TRECVID 2005 Development Set

	38	39	40	41	42	43	44	45	46	47	All
Rank-MFoM	0.0091	0.0158	0.0759	0.0291	0.0076	0.0083	0.0175	0.0004	0.0997	0.0399	0.0303
SVM	0.0060	0.0102	0.0045	0.0003	0.0037	0.0091	0.0055	0.0000	0.0405	0.0133	0.0093

Table 2: Performance of Runs in terms of MAP from TRECVID evaluation

	38	39	40	41	42	43	44	45	46	47	All
Run A	0.0449	0.0169	0.0907	0.0498	0.0703	0.0231	0.0287	0.0054	0.0735	0.0638	0.0467
Run B	0.0727	0.0198	0.0527	0.0155	0.0990	0.1143	0.0693	0.0033	0.0769	0.0412	0.0565
Run C	0.0999	0.0426	0.1540	0.0809	0.1326	0.0989	0.0962	0.0102	0.1906	0.1189	0.1025
Run D	0.0872	0.0323	0.1575	0.0795	0.1332	0.0887	0.0939	0.0072	0.2249	0.1070	0.1011

Table 2 lists the results of the 4 runs on the TRECVID 2005 test set. The results clearly show that the fusion runs (Run C and Run D) perform significantly better than those runs employing only individual feature (Run A and Run B). The overall MAP of the texture-feature-only run is also slightly better than that of the text-feature-only run. We observe from the results that fusion has a positive effect on the overall performance, except for concept 43, i.e *Waterscape/waterfront*. The results show that the Rank-MFoM is effective in fusing different modality features for

the detection problem. For these submissions, we only focus on using the extracted text and texture features. In our future work, more features will be introduced and fused. In addition, we will look at the association between different modalities and features to further improve the detection performance.

1.2 Multimodal RankBoost Fusion

This approach aims to fuse a combination of both low and high level features using the rankBoost algorithm. The list of features to be fused include: text features from ASR, audio genre, face information, shot genre, image matching, and visual concepts. The detection of the first five concepts follows standard techniques as describe din Section 2.1. We will elaborate on the extraction of visual concepts using HMM here.

To detect visual concepts within each keyframe image, we employ HMM to model the association between concepts and their positive images. This is done by segmenting each key-frame into fixed 4x4 blocks and followed by clustering them. The visual features used to represent each image block include color histogram, edge histogram, and the adaptive matching pursuit feature for texture. More specifically, they are: Luv color histogram, adaptive Matching Pursuit texture features and edge histogram (Shi et al, 2004). The dimension of the resulting visual feature vector is 90. Our method treats each block in a key-frame separately and a data point in low-level visual space is just a 90-dim block vector. We then perform k-means clustering (with k=500) to the block vectors of the training images. Each training image is tokenized by dividing the image into a set of regular-sized blocks, each represented by a feature vector, and quantized into a visual cluster. As shown in ***Figure 1***, the content representation of each image is modeled as having been stochastically generated by a HMM, where each state will generate some blocks with similar features. In addition, the spatial transactions and co-occurrences between fixed-size blocks within an image can be utilized to help in detecting the feature concepts.

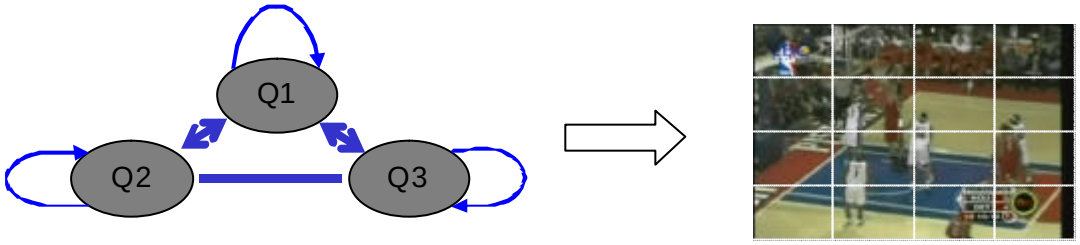


Figure 1: HMM for Concept *i*.

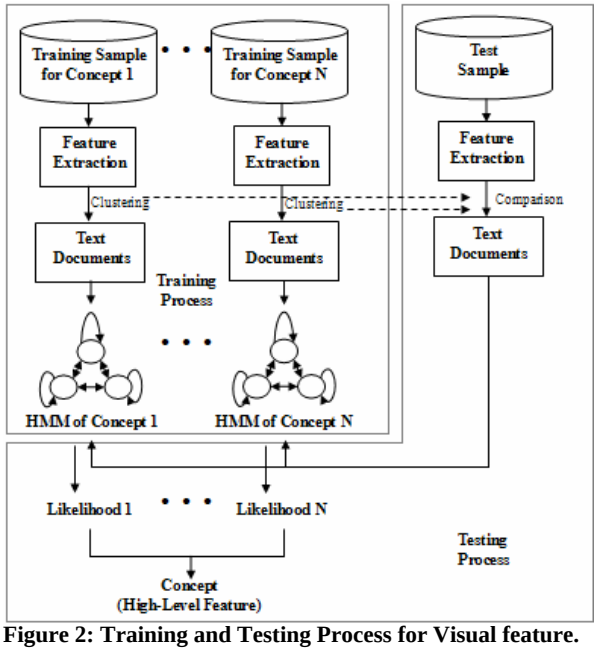


Figure 2: Training and Testing Process for Visual feature.

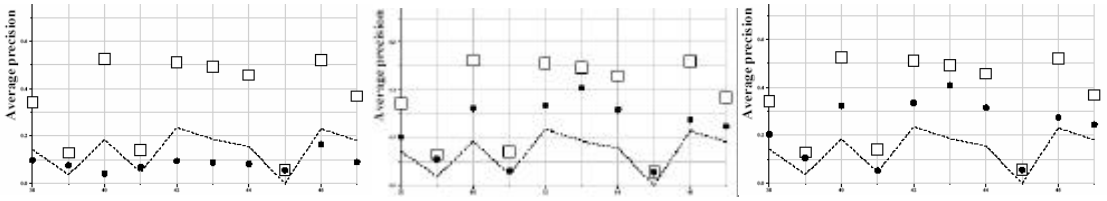
Given a test image, its concept representation is derived by comparing its blocks' feature vectors to the cluster centralities of k-means. Consider the overall image content representation as an instance of a HMM, the likelihood of this instance being generated by each concept HMM is computed. To annotate this image, the concept with HMM yielding the highest likelihoods is selected. After this process, we build several HMMs, one for each co-occurrence concepts pair, to obtain its confidence probabilities for improving automatic image annotation results. The training and testing processes are illustrated in ***Figure 2***.

Given the set of extracted features in the training images, we perform RankBoost fusion to combine different multimodal features for auto concept annotation. Further details of RankBoost fusion can be found in (Zhao et al, 2006). We submitted 3 runs base don this technique. They are:

Run E (TRECVID Run 4): The text-only run. This run is the baseline run and consists of only textual features extracted from the ASR of the positive shots and negative shots. SVM-lite is used for ranking the shots.

Run F (TRECVID Run 2): The first multimodal run. This multimodal run uses textual features, image matching and HMM visual concept detection for fused using RankBoost.

Run G (TRECVID Run 1): The second multimodal run. This multimodal run uses all multimodal features from Run 2 as well as audio genre, shot genre and video OCR.



Run E Run F Run G

Figure 3. Performance of Runs E, R and G in terms of MAP from TRECVID evaluation.

As expected, Run E does not perform well and most of its performance is below the median. With the addition of other multimodal features, we observe a significant increase in performance in Run F and Run G. The overall performance of our best run, Run G, is able to achieve a MAP of 0.22. For future work, we will be looking into using other fusion techniques, such as hierarchical HMM, that can leverage on the spatial relations between different regions/blocks within images for more effective fusion.

2. FULLY AUTOMATIC SEARCH TASK

In this year's search task, our group submitted a total of 6 runs. The 6 runs comprise a required baseline text run and 5 other runs with various features and techniques. We will briefly describe the features extracted during pre-processing stage and the overall retrieval framework. From past experiences, it is clear that the bottleneck of an ideal retrieval lies in the amount of usable semantic obtained. To complement the small amount of usable information available, we introduce an event-based approach to support the retrieval. We illustrate the approach with an example. Given a query from last year's topic: "*Find shots that contain buildings covered in flood water*." It will be beneficial if we can have an idea which are the locations that experienced flood (and their occurrence times), and re-formulate the query to search for videos that contain these locations irrespective of whether the term "flood" is mentioned or not. As there is an abundance of information on the Web, such event-absed information can be mined especially for news. Furthermore, our preliminary results showed that our event-oriented mdoel is also highly effective on Person-X detection (Zhao et al, 2006). In particular, for this year's multi-lingual news video corpus, certain non-English names like (*Mahmoud Abbas, Allawi Iyad*, etc) cannot be easily recognized in non-English speeches or translated to English text. This gives rise to high error rate for the recognition and translation of such names, which will greatly affect the number of retrievable relevant shots especially when the person's name plays an important part. With the use of event information, we can make use of location and time to recover these missing shots. Intuitively, we can predict the presence of these people in the news stories by considering their close proximity in terms of location and time. Locations are seldom misrecognized or wrongly translated even for spoken documents since they are not as vulnerable to errors as person's names.

2.1 Content Pre-processing

The search process uses a set of multi-modal features as described below.

2.1.1 Automatic Speech Recognition and Machine Translated Text

We make use of the Chinese and Arabic machine-translated (MT) text provided by TRECVID as well as the English ASR from Microsoft (given by CMU). The ASR in Chinese and Arabic are not used as we would like to focus on English text processing only. Some of the difficulties we encountered in using the Chinese and Arabic MT texts are that they are prone to errors and the translated texts are often not meaningful. The basic unit for retrieval of MT texts is set to be a phrase, which is the output of MT. To make the retrieval process consistent for the English news ASR, we also divide the speech segments of English ASR according to approximate phrase length. This is done by either considering a long pause or a change of speaker.

2.1.2 Video OCR

The video OCR output is obtained from CMU VOCR systems. As OCR output contains numerous insertion, deletion and mutation errors, we integrate minimum edit distance (MED) matching to maximize the precision and recall of name matching in OCR (Chua et al, 2004).

2.1.3 Annotated High Level Features from High-level Feature Extraction Task

With the availability of high-level feature detection results from other groups, we make use this information to rank shots more effectively. We apply RankBoost to combine result from the various groups to obtain a new rank-list for each high-level feature. As different groups use different techniques for detection, it is inappropriate to merge them without considering their common basis. RankBoost fusion can avoid this problem by only utilizing the relative scores from rank lists of each system. Thereafter, we obtain the new shot rank list by selecting the top 500 shots.

2.1.4 Face Detection and Recognition

Our face detection is based on a cascade of boosted classifiers using an extended set of Haar-like features (Lienhart et al, 2002), which is an extension of Viola's fast face detector (Viola et al, 2001). By analyzing the statistics of name entities appearance in the training set, we identify a list of top 15 most frequently appearing persons. Thereafter, we build a face recognition model for each of the selected person. The face recognition algorithm we sued is based on 2DHMM (Eickeler et al, 2000; Chua et al, 2004).

2.1.5 Audio Genre

Audio information is one of the important features for news video. In addition to obtaining ASR from it, we can alsoutilize it to help improve retrieval. The primary benefit originates from the different identifiable audio genres. Audio genres encode some form of semantic information which is very useful for detecting actual shot genre. Music for example may provide clues to advertisements or news headline reporting. In our search task, we identified 7 audio classes to be used. They are: cheering, explosion, silence, music, female speech, male speech, and noise. For our audio genre detection, we make use of zero crossing rates, mel-frequency cepstrum coefficient (MFCC), centroid and roll off point energy (Jiang et al, 2000). In our experiments, we observed that some audio genres only occur in specific scenes. For instance, a scene with explosion and gunshots has a low probability of containing music as well. The only exceptions are advertisement scenes which often contain music and other genres. Thus, segmenting the audio signal into periods of different genres, especially periods of speech and non-speech, is essential for improving retrieval accuracy.

2.1.6 Shot Genre

The detection of shot genre by itself is a useful tool in providing partial semantic for news video. We follow the implementations and techniques as described in (Chaisorn et al, 2002). The defined genres in our system are: sports, finance, weather, commercial, studio-anchor-person, general-face and general-non-face.

2.1.7 Story Boundary

Each news story can be considered a basic unit for summarization, browsing or further analysis. The story-level boundaries are detected by (Hsu et al, 2005). The basic algorithm employs a visual cue cluster construction (VC3)

process. This process is dependent on prosody features extracted from speech (Hsu et al, 2005) and theoretically based on the information bottleneck principle (Hsu and Chang, 2005). It has been shown in TRECVID 2004 that this technique is quite effective in the story segmentation task. In order to handle the different production styles of each news language, story segmentation is performed in a language-dependent manner. There are separate segmentation programs trained for each language - English, Chinese, and Arabic.

2.1.8 Spatial-temporal Information from News Video Stories

Every news story is a description of an event. Typically an event will have location, time of occurrence, people involved and a detailed description of what has happened. Using the story boundaries, we are able to extract sets of semantically meaningful terms that describe each given story or event. This is done through the analysis of ASR and MT. More specifically, Named Entities (NE) such as location, time and persons' names are extracted. For each story, we assume that people whose names are mentioned in the story should be related to the event. As such, we allow each event entity to contain more than one key person. As for location, we need to identify the main event location based on all mentions of locations in the story. In the case of only one mentioned location, it is easily solved. However if several locations are mentioned, we will group the locations according to their actual spatial relations in the real world. For instance, if "Iraq" and "Baghdad" are both mentioned, "Baghdad" will be chosen as the main event location since Baghdad is the capital of Iraq. In general, a more specific place is preferred over a general location in the same area. For cases where multiple cities or countries are involved, we will select the location which appears first. This is based on the intuition that news story tend to report the location of event happenings at the start of news. The time information we use is the date of the news video when no time information are mentioned in the story. Cue terms such as "two days ago", "tomorrow" are also used through a rule-based process to extract the actual event dates. In situations where there may be more than one date, we allow a multiple-dated event with a maximum allowance of 3 days. We tested the effectiveness of the event extractor on our composed training data and found that these rules are able to work with 85% confidence. The spatial temporal information for an event is found to be effective for supporting search task.

2.2 Retrieval Framework

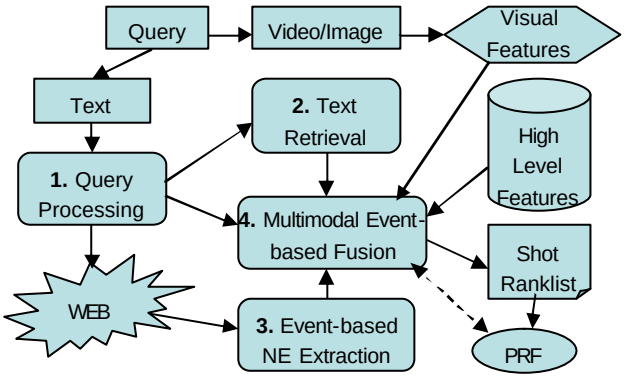


Figure 4: Overall Retrieval Framework

The event-based approach is similar to the traditional systems that perform query expansion to induce the context of the query. However, these traditional systems are tuned to extract keywords that have high correlation to the query. Even though these methods are effective, as the number of keywords used is limited, it is not invariant across time. Also, the additional keywords induced may only be related to news articles in a certain period of time. For example, suppose there are a few earthquakes events in the video corpus, it is unlikely that we will be able to extarct a linear set of keywords that can cover all the news articles for different earth quake events. The use of event-based retrieval can be seen as a structured form of keyword based query expansion. It relies on named-entities and their correlations in time and space, rather than just keywords with high mutual information to the query. It is thus able to cover different aspect of queries relating to multiple events more effectively. It also tends to be more invariant to time.

Figure 4 illustrates the complete retrieval process. The framework consists of four main stages: (a) query processing, (b) text retrieval, (c) event-based NE extraction, and (d) multimodal event-based fusion.

2.2.1 Query Processing and ASR Retrieval

Besides extracting keywords for retrieval, we also induce information such as the query-class and explicit constraints. Seven query classes are used: {Person, Sports, Finance, Weather, Politics, Disaster and General}. The General-type query class is designed as a container class to capture all queries not classified into the other 6 categories. As most such queries are unlikely to be related to events, we will not be carrying out event-entity re-ranking for this query class. Examples of queries under the General class are: "Find shots of cars with road", " Find shots of people entering building" . The relevant shots to these queries can often appear in any footage and they do not correspond to meaningful events.

In addition to event-based modeling, traditional query expansion is also carried out to induce query context from sample shots provided with the query and from parallel texts. The expanded terms are words which have high mutual information with the original query terms. WordNet (Fellbaum 1998) is used to reselect those expanded terms which lie in the same synset as the original terms. After query processing, the next step is ASR retrieval. More information on query processing and retrieval can be found in our previous work **(**Chua et al, 2004**)**.

2.2.2 Event-based NE Extraction

We use the text query from the query processor to retrieve related news articles. This is done by using parallel news articles collected in the same period, as well as Google news search. Morphological analysis is applied on the retrieved documents to derive the Part-of-Speech (POS) information. These POS-tagged documents are passed to the NE extractor module to obtain various NE types such as: Person Name, Location and Time. The technique follows that used in (Yang et al, 2003) and the accuracy of detecting these NEs is in the range of 90%. Intuitively, each news article signifies a single news story and therefore the group of NEs extracted from each document can be denoted as the set of NEs in an " event group" with respect to time. By identifying the Person's Name, Location and Time entities from each video news story, we are able to predict whether a specific "event group" extracted online is related to the targeted news video. Let *E'* denote the set of NEs extracted from the particular "event group" from online news, *P'* denote the set of NE's extract from a news video story, the relationship function *Rel* is given by

$$\mathrm{Re}\,l(P', E') = \sum_{i=Loc,Person,Time} b_i * Y(P'_i \cap E'_i) \tag{5}$$

where β_i is the weight given for different NE type, *Y* is the output number of intersections. Similarly, we can obtain the probability of *NE' or Event'* (given by query) relevance to a news video story in terms of location-time relation.

$$P(NE', Event' \mid P', E') = a_m \,\mathrm{Re}\, l(P', E') \tag{6}$$

where a_m are weights given to different query types.

2.2.3 Multimodal Fusion

The task of combining various modalities appropriately is essential to obtaining good results. Different queries may have very different characteristics and hence require very different feature combinations and various methods of fusion have been tested in (Neo et al, 2005). We use a combination of heuristic weights, and the visual information obtained from the sample shots given to form an initial set of fusion parameters for the queries. Subsequently, we perform a round of pseudo relevant feedback (PRF). This is done by using the top 20 return shots from each query.

2.3 Results and Discussions

We perform a number of runs base don the guideline of TRECVID auto retrieval tasks. The runs are:

Run 5. (The required text-only run). This run is the required baseline text run. We make use of morphological analysis on the text query to obtain the part-of-speech (POS) information. Subsequently, we extract the Name Entities (NE) and nouns phrases from the query to form the keywords for retrieval. The number of keywords in this case is restricted to 4. Using these keywords, we perform a basic retrieval on the ASR and MT using standard tf-idf

function to obtain a ranked list of "phrases". Using the time information of these retrieved phrases, we return the shots that lie in the respective boundaries.

Run 4. (Including other text). Run 4 is also a text-only run. The difference between Run 4 and Run 5 is the use of additional expanded words and context. We derive additional words by using text from given sample videos as well as terms with high mutual information from the parallel news. These additional keywords are then added to the original keyword from Run 5. The shots are returned in a similar manner as in Run 5.

Run 3. (Run 4 with high level features). We introduce the use of high level features. The weights of the shots are boosted in the following manners: (a) if the shot contains the high level feature that is found in the text query; and (b) if the shot contains the high level feature that is found in the given sample video.

Run 2. (Multimodal Run with Pseudo Relevance Feedback (PRF)). This run makes use of the various multimodal features extracted from the video to re-rank shots obtained in Run 3. Using the query-class information derive from the text query, weights are assigned to various multimodal features, similar to previous work in (Chua et al, 2004). The score of each shot is set to:

$$Score(Shot_i) = \sum_{Model_Q} \beta n_i^Q F_i \qquad (7)$$

where Q is the query-class, n_i^Q is the feature parameter weight defined for Q-class, F_i is the confidence score of feature i, and β is the normalizing factor. In this case, a high score will signify that the shot is relevant to the query. Subsequently, we introduce a round of pseudo relevance feedback to adjust the parameter weights n_i^Q based on the top 15 returned shots.

Run 1. (Multimodal Event-based Run with PRF). This run makes use of all multimodal features in Run 2 as well as the fusion with an additional event entity feature (Neo et al, 2006).

Run 6. (Visual only). This run uses only visual features. The purpose of this run is to test the underlying retrieval result if all textual features are discarded.

Table 3 shows the MAP results of our six runs. The best MAP performance of 0.126 was reached when we use all multimodal features with a fusion with Event-entity features (Run 1). *Figure 5* illustrates the performance of our runs with respect to other participating groups in the same category.

Table 3: Evaluation Results in MAP

	Run1	Run2	Run3	Run4	Run5	Run6
MAP	0.126	0.109	0.104	0.076	0.061	0.039

From ***Table 3*** and ***Figure 5***, we can see that a purely text-based run do not perform well (Run 5). Even though it actually performs better than a fully visual feature run (Run 6), Run 5 only achieves a MAP of 0.061. Observation shows that some texts translated from other languages may include meaningless words or phrases that are useless during retrieval. In Run 4, we notice that the usage of a parallel corpus to derive additional context and words can increase the performance by about 25%. The main basis for this improvement is because keywords extracted from the given query are usually few and does not describe the context of the query well. For example query 167, " Find shots of airplane taking off" , does not offer any indication of possible contexts that the answers may appear in. By using the parallel corpus and sample video texts, we can mine additional words like " airport" , " sky", or "British Airways" . These words provide a possible context where the shots are likely to occur in, thereby helping to retrieve more correct shots.

In addition to using the common features for video search tasks as described in (Chua et al., 2004), we evaluated our system on the utilization of high level features. The full set of high level features is obtained from the publicly available corpus after the completion of high level feature extraction task. By employing these high level features, we are able to track the occurrences of different concepts in the given shots. This information is especially useful if the given query is directly related to one or more of the available high level features. An example would be query 168: " Find shots of a road with one or more cars". Since the annotated set of high level features contain the concept " car" , we can use this knowledge to boost the weights of shots that contain the term or concept " car". This

significantly improves the results of any such queries. Other examples of important high level features include " people walking/running", " explosion" , " maps" , "building" , and " sports". *Figure 6* illustrates the results of applying high level features (Run 3) as opposed to using only text (Run 4).

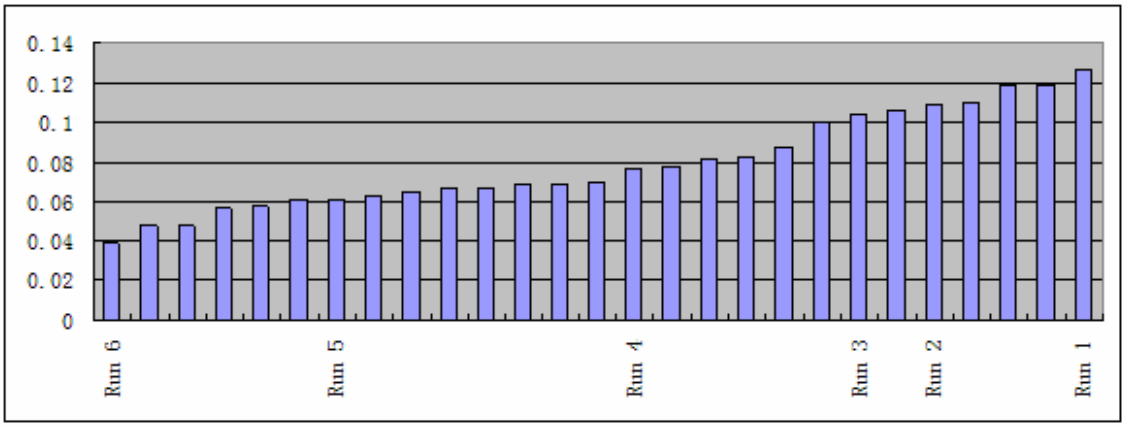


Figure 5: MAP performance of our system as compared to other systems.

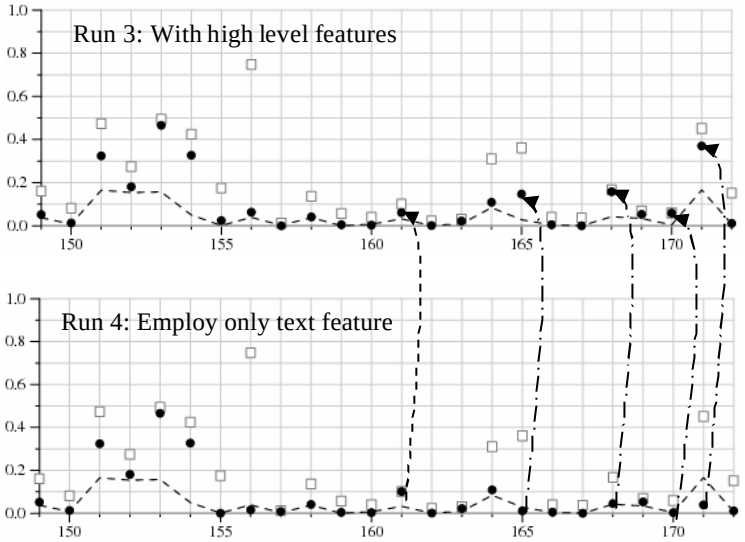


Figure 6: Comparison between Run 3 and Run 4 to illustrate the Effects of High Level Features. (Note: the arrows depict significant increase or decrease in performance.)

In can be seen in *Figure 6* that the majority of the queries that improve through the use of high level features are general queries. In particular, queries 165, 168, 170 and 171 have shown significant improvement. The only exception comes from query 161 " Find shots of people with banners or signs" that registers a degradation in performance from Run 4 to Run 3. We conjecture that this is due to the introduction of " People Marching/Walking" concept in the process of ranking. This introduces a lot of noise to the results as there are many shots containing people walking/marching but with no banners. This indicates that the high level concepts need to be used appropriately. In addition, better inference should be carried out to infer the most relevant concept to be used for re-ranking.

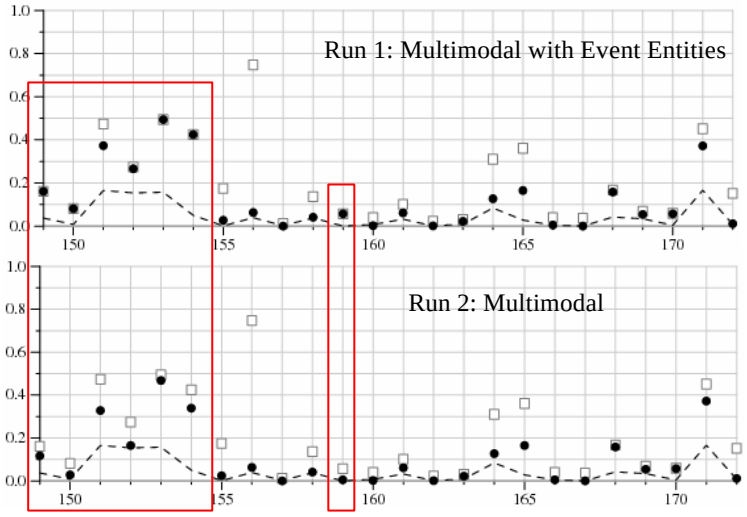


Figure 7: Applying Event-based approach to non-general queries; significant improvement for queries shown in boxes.

For Run 2 and Run 1, we took advantage of the multimodal features available in videos to help improving the search process. The only difference between both runs is that the latter makes use of event-based retrieval during fusion. Though Run 2 demonstrates the effectiveness of multimodal fusion and pseudo relevance feedback, it is only able to achieve a MAP of 0.109. In Run 1, we are able to increase the performance by about 16% to 0.126. From *Figure 7*, we can see that the improvements appear more significantly for the first few questions. Specifically, these queries are about human person, such as "Tony Blair" and " Mahmoud Abbas" . This shows that our event-based tracking process is effective when used for search about humans that tends to be more event-oriented. In general, our approach works for queries which contain either a specific detectable event or event-related information.

References

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

S.F. Chang, R. Manmatha, T.S. Chua. Combining Text and Audio-Visual Features in Video Indexing, In IEEE ICASSP 2005, Philadelphia, PA, March 2005

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.

C. Cortes & M. Mohri. AUC optimization vs. error rate minimization, In NIPS, 2003.

T.S Chua, S.Y. Neo, K.Y Li, G. Wang, R. Shi, M. Zhao and H. Xu, TRECVID 2004 Search and Feature Extraction Task by NUS PRIS, TRECVID, 2004.

Stefan Eickeler, Stefan Muller, and Gerhard Rigoll. Recognition of JPEG Compressed Face Images Based on Statistical Methods, Image and Vision Computing, Vol.18, pp.279-287, 2000.

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

S. Gao, W. Wu, C.H. Lee, & T.S. Chua. A maximal figure-of-merit approach to text categorization, In SIGIR, pp.174-181, 2003.

S. Gao, W. Wu, C.H. Lee, & T.S. Chua. A MFoM learning approach to robust multiclass multi-label text categorization, In ICML, 2004.

W. H. Hsu, L. Kennedy, S. F. Chang, M. Franz, and J. Smith, "Columbia-IBM News Video Story Segmentation In TRECVID 2004," Columbia ADVENT Technical Report 209-2005-3, New York 2005

W. H. Hsu and S. F. Chang, "Visual Cue Cluster Construction via Information Bottleneck Principle and Kernel Density Estimation," The 4th International Conference on Image and Video Retrieval (CIVR), Singapore, July 20-22, 2005

H. Jiang, T. Lin and H.J. Zhang. Video segmentation with the Support of Audio Segmentation and classification, In ICME-IEEE Int'l Conf on Multimedia and Expo, 2000.

R. Lienhart and J. Maydt. An Extended Set of Haar-like Features for Rapid Object Detection. In Proceedings of the IEEE Conference on Image Processing, pages 900-903, 2002.

S.Y. Neo, H.K. Goh, T.S. Chua. Multimodal Event-based Model for Retrieval of Multi-Lingual News Video, to appear in IWAIT, 2006.

S.Y. Neo, T.S. Chua. Query-dependent Retrieval on News Video, *MMIR'05* workshop *in SIGIR'05*, 2005.

C. Petersohn. Fraunhofer HHI at TRECVID 2004: Shot Boundary Detection System, TREC Video Retrieval Evaluation Online Proceedings, TRECVID, 2004.

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.

P. Viola and M. Jones. Robust Real-time Object Detection. In IEEE ICCV Workshop on Statistical and Computational Theories of Vision, 2001.

L. Yan, R. Dodier, M.C. Mozer, & R. Wolniewicz. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic, In ICML, 2003.

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

M. Zhao, S.Y. Neo, H.K. Goh, T.S. Chua. Multi-Faceted Contextual Model for Person Identification in News Video, to appear in MMM, 2006.