TRECVID 2006 by NUS-I^{2}R

Tat-Seng Chua, Shi-Yong Neo, Yantao Zheng, Hai-Kiat Goh, Sheng Tang*, Yang Xiao and Ming Zhao
School of Computing, National University of Singapore

Sheng Gao, Xinglei Zhu, Lekha Chaisorn, Qibin Sun
Institute for Infocomm Research, Singapore

ABSTRACT

NUS and I^{2}R joint participated in the high-level feature extraction and automated search task for TRECVID 2006. In both task, we only make use of the standard TRECVID available annotation results. For HLF task, we develop 2 methods to perform automated concept annotation: (a) fully machine learning approach using SVM, LDF and GMM; and (b) Bi-gram model for Pattern Discovery and Matching. As for the automated search task, our emphases this year are: 1) integration of HLF: query-analysis to match query to possible related HLF and fuse results from various participating groups in the HLF task; and 2) integration of event structures present implicitly in news video in a time-dependent event-based retrieval. The proposed generic framework involves various multimodal (including HLFs) features as well as the implicit temporal and event structures to support precise news video retrieval.

1. HIGH LEVEL FEATURE EXTRACTION TASK

1.1 Visual Feature Extraction

The visual features used in our systems are listed in the below:

- Global color correlogram (*GCC*) in HSV space: 324-dimension.
- Co-occurrence texture extracted from global gray-level co-occurrence matrix (*GLCM*): 64-dimension.
- 3-D global color histogram in HSV (*HSV*): 162-dimension.
- 3-D global color histogram in RGB (*RGB*): 125-dimension.
- 3-D global color histogram in LAB (*LAB*): 125-dimension.
- Gabor filter (2-scale, 12 orientations) based texture feature (*Gabor*): 48-dimension extracted from one of 5*5 patches uniformly segmented of the image.

The textual features used are described in Section 2.

1.2 Machine Learning approach by SVM, LDF and GMM

For each type of extracted feature, we make use of SVM classifier (*SVM*), linear discriminative function (*LDF*) classifier or Gaussian mixture model (GMM) for training. The details are summarized in Table 1. Thus, there are 9 visual classifiers trained for each concept. The SVM is trained using the SVM^{light} tool (Joachims, 2002) and the LDF and GMM classifiers are trained using the AUC maximized learning algorithm (Gao & Sun, 2006). Based on our experiments on the development set of TRECVID 2005, we find AUC maximized learning algorithm works better than the best-tuned SVM system.

Table 1: Description of classifiers (+: the classifier is available for the feature)

	SVM	LDF	GMM
GCC	+	+	
GLCM	+	+	
HSV	+	+	
RGB	+		
LAB	+		
Gabor			+

For each image, we use the 451 (9*39) features from the feature-space and then train a SVM classifier for each concept. PCA or LSA is also applied to reduce the selected feature space dimensions. We follow the approach in (Gao et al, 2007) which uses a knowledge-based method to retain only informative components. This is done by extracting the pair-wise concepts association among the 39 concepts based on the development set. The strength of the association, *Str*, for the target concept *A* and the concept *B* is defined as,

* Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

$$Str = \frac{\#(A,B)}{\#(A)}$$

where #(A, B) is the number of shots relevant with both *A* and *B*, and #(A) is the number of shot relevant with *A* which is the conditional probability of *B* on *A*. Higher value of *Str* means concept *B* has a stronger relation with *A*. This relation is however asymmetric to many concepts. For example, in the TRECVID 2005 corpus, *outdoor* is closely related to concept *airplane* (an airplane shot is usually outdoor). But on the other hand, *airplane* relative relation to *outdoor* is ranked at 25th for which the most related concept is the *person* concept. We therefore make use of this property (for Run4) to extract the concepts with the high association and keep the corresponding components of the model-based feature while ignoring other components. Besides using highly associated concepts, there are many concepts that have never co-occurred with the target concept. They also convey much information about the target concept. We exploit the corresponding concept detectors as a filter to prune the most impossible ranking shots for a specific target concept (Run5). A final run combining results from Run5 and the textual SVM classifier is fused for Run6 using a weighted linear method.

Table 2: Evaluation Results in inferred MAP

	Run4	Run5	Run6
MAP	0.055	0.037	0.040

1.3 Bi-gram Model for Pattern Discovery and Matching

The multimedia pattern discovery problem can be decomposed into two sub-problems: choosing suitable representations of the raw data, and learning a model from these representations. We first transform the high dimensional continuous video stream into tractable token sequence. As video data is high-dimensional, multi-modal and continuous stream, we first need to decompose the video stream into tractable token sequence. The video key-frames in training video dataset are clustered by the K-means algorithm (*K* is predefined number of clusters) using feature of global HSV auto-correlograms. After the clustering, video key-frames that have similar low-level features are labeled with the same token. In this way, the original input video stream is transformed into token sequences. We also map the test video key-frames into the clusters of training set. Similarly, the test video stream could also be transformed into token sequences. Then, we apply soft matching models based on bi-grams (Cui et al, 2006) to model pattern matching as a probabilistic process that generates token sequences to reveal the video patterns. N-gram language modeling is one important approach which models local sequential dependencies between adjacent tokens. We apply linear interpolation (Manning et al, 1999) of unigrams and bigrams to represent probability of bigrams in our task. The reason is two-fold: (1) to smooth probability distribution in order to generate more accurate statistics for unseen data; and (2) to incorporate conditional probability of individual tokens appearing in specific slots. Considering the “local context token” around <TARGET>, it is modeled as a window centered on <TARGET> according to a predefined size *W=3*: *< token−W, . . . , token−2, token−1, TARGET, token1, token2, . . . tokenW >*. In particular, we model the sequence of pattern tokens as:

$$\Pr(t_1 ... t_L) = \Pr(t_1 \mid \mu) \prod_{i=2}^{L} \left(\lambda \Pr(t_i \mid t_{i-1} | \mu) + (1-\lambda) \Pr(t_i | \mu) \right) = \Pr(t_1, S_1) \prod_{i=2}^{L} \left(\lambda \Pr(t_i \mid t_{i-1}) + (1-\lambda) \Pr(t_i, S_i) \right)$$

where μ stands for the bigram model and *Pr*(t_i, S_i) stands for the probability of token t_i appearing in slot S_i. λ is the mixture weight combining the unigram and bigram probabilities. Note that it uses the conditional probability of a unigram being in a slot to represent unigram probability. This is because the position of a token is important in modeling. Incorporating individual slots’ probabilities enables the bigram model to allow partial matching, which is a characteristic of soft pattern matching. In other words, even if some slots cannot be matched, the bigram model can still yield a high match score by combining the matched slots’ unigram probabilities. As test instances are often different in length, it normalizes the log-likelihood of Equation (1) by the length *l* of the test instance:

$$\mathrm{P}_{norm}(t_1 ... t_L) = \frac{1}{l} \left(\log \Pr(t_1, S_1) + \sum_{i=2}^{l} \log \left(\lambda \Pr(t_i \mid t_{i-1}) + (1-\lambda) \Pr(t_i, S_i) \right) \right)$$

where *l* denotes the number of tokens in the test instance, which could be smaller than the model length *L*. Next, unigram and bigram probabilities are estimated by their maximum likelihood (ML) estimates:

$$P_{ML}(t_i, S_i) = \frac{|t_i(S_i)|}{\sum_k |t_k(S_k)|}; \qquad P_{ML}(t_i \mid t_{i-1}) = \frac{|t_i(S_i) t_{i-1}(S_{i-1})|}{|t_{i-1}(S_{i-1})|}$$

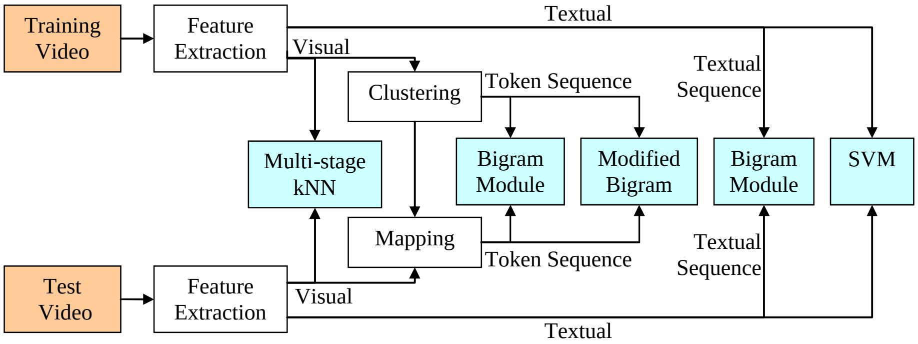


Figure 1: Overall Framework for High-Level Feature Extraction

As Figure 1 showed, our HLF task applies both sequence level and key-frame level methods for visual and textual features. Sequence level methods include bigram module, modified bigram module, while key-frame level methods include multi-stage kNN and SVM. For visual feature, we employ a two-stage kNN classification (Xiao et al, 2007) to find the best *k* key frames for each training key-frame using different visual features and different number of candidates at different stages. We also use bigram module for visual feature to calculate the probability of how the test samples match the revealed video patterns. It aims at soft pattern matching, which models two basic characteristics: local individual slot match degree and sequential dependencies between adjacent tokens. The modified bigram module considers the distance of local context to the <TARGET> for different weight in bigram module. As for textual features, we use standard SVM and bigram module. From the result of validation set which is extracted from the training set, we find that the sequence level methods are more useful when the <TARGET> frame has positive context, or consecutive positive frames occurred. Because some useful context itself makes the high degree of pattern matching through similarity between individual units, or sometimes some useful co-occurrence of context makes the high degree of pattern matching through sequence fidelity. So, for many cases, the whole sequence has high confidence for certain concept although the center unit itself is not so confident, while it is not easy to be identified by key-frame level method. We have designed 3 runs to test the effectiveness.

Run1: Run2 combined with Run3.
Run2: High corresponding test subset for detected typical cluster ID for certain HLF in training set from the clustering result combined with Bayesian fusion of weight and confidence of test samples based on the best models for each HLF derived from validation set.
Run3: High corresponding test subset for detected typical cluster ID for certain HLF in training set from the clustering result combined with the sequential arrangement of the rank list of the best models for each HLF derived from validation set.

Table 3: Evaluation Results in inferred MAP

	Run1	Run2	Run3
MAP	0.059	0.045	0.051

Sequence level and key-frame level methods are accomplished in different concepts. When using the sequence level methods, some useful context itself reaches the high degree of pattern matching through similarity between individual units, and sometimes some useful co-occurrence of context reaches the high degree of pattern matching through sequence fidelity. Such as “Weather”, “Sports” etc. When using the key-frame level methods, it performs better than the sequence level methods when the target unit is high similar to certain training samples but they appeared in the sequence with no useful context, and also don’t have high similarity between individual units with the training sequence. Such as “Charts”.

2. FULLY AUTOMATIC SEARCH TASK

This is our third year participating in the TRECVID automatic search task. Our approach for this year’s automated search task therefore focus on the two above mentioned points. We enhance our query analysis by including 2 extra key query contents namely the query-HLF and query-event, in addition to previously used query-contents like query classification, query typing and query expansion. The query-HLF measures the importance of a HLF with respect to a query by performing descriptive lexical matching and time-dependent mutual information ranking using external articles. This importance will help in the re-ranking of shots together with the confidence of its HLF detection. The query-event links the query to possible event groups generated through an event-clustering algorithm. This linkage reveals the implicit event structures and provides another facet of partial semantic during retrieval. News videos are generally multimedia depictions of real-life happenings which are events in nature. Semantically similar stories in an event-cluster thus form the basis for finding relevant shots as answer shots are generally similar in nature. We submitted a total of 6 runs. Run6 (MAP0.039) is the required baseline run using only text. Run5 (MAP0.042) uses query expansion and relevant news articles of the same period for retrieval. Run4 (MAP0.043) is a HLF-only run using HLF detection results without considering the detection confidence as well as query-inference. Run3 (MAP0.051) enhances Run4 with both detection confidence as well as query-inference. Run2 (MAP0.067) uses both HLF and Text for re-ranking; and Run1 (0.075) uses both HLF and Text based on our suggested event-based model. 2 of our best performing runs are ranked 5th and 6th in the 75 submitted automated search task runs. A breakdown of our system performance shows that it is able to obtain best average precision for 6 of the queries, with also 10 other queries well above the median performance. An analysis of our system performance shows that our retrieval system is able to outperform other systems in Name entity or event-related queries, as well as produce similar comparative results in other classes of queries.

2.1 Video Features

2.1.1 Automatic Speech Recognition and Machine Translated Text

We make use of the standard set of Chinese and Arabic machine-translated (MT) text provided by TRECVID as well as the English ASR. We index the ASR and MT according to speaker-change segments as well as induced phrase. The speaker change information is available from the ASR output (except for 3 of the videos which does not have ASR). The speaker-change usually provides good coherent text and span over multiple shots (3-4 shots). The phrase is another unit which can preserve the textual coherency when indexing. This phrase information is available through the outputs of the 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. From past experience, phrase level text retrieval can yield better results than shot level text (Neo et al, 2005). 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.

2.1.2 Video OCR

The video OCR output is contributed by CMU Informedia System. OCR outputs are valuable features especially for tracking of known person. 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 High Level Features

With the availability of high-level feature detection results from other groups, we can use this information to rank shots more effectively. We compute the detection confidence of the set of 39 HLFs set based on the detection output from the TRECVID HLF task. Participating systems are required to submit a ranked list of maximum 2000 shots for each HLF of the 39 features with the first shot having the highest confidence. We combine the outputs from 7 of the participating systems to estimate the confidence of a particular HLF using the following formula.

$$Conf(S_c \mid HLF_k) = \alpha \cdot Contains(S_c) + (1-\alpha)\sum_{j} \frac{\max Pos - Pos(S_c)}{\max Pos}$$

where *Contains(S_c)* is an indicator function that checks how many systems j have S_c in their ranked list and the second term produces a normalized score in the range of [0..1] that linearly weights the position (*Pos*) for the shot on the ranked list. From previous experiment, we notice that each HLFs does not have an equal chance in appearing in the video. For example: faces can be seen almost in every shot but a car may only be appearing in 1 out 10 shots and a boat may be lower. It is therefore necessary to account for this phenomenon by pegging the *Conf* to the number of estimated appearance. Based on the supplied ground truth from the development set, we estimate the approximate number of times each HLF will appears in the testing corpus. Subsequently, we impose a limit, given by $2\times A_{HLF}$ (where A_{HLF} = estimated appearance for a particular HLF), to the ranked list of each HLF *Conf*. Any shot which is below $2\times A_{HLF}$ will not be considered.

2.1.4 Near Duplicate Key-frames (NDK)

The task of detecting NDK in video corpus is important as it helps to build up the linkage of relevant news stories across different TV news channel, language, and time. This type of visual information is useful in news video retrieval, topic detection and story threading. In addition, NDK is also use in retrieval for context free video (video without speech) like BBC rushes, home videos and pre-editing video as these video does not contain the Automated Speech Recognition (ASR). However, it is known that the computational cost for such computation especially across large image or video dataset can be intensive. In this work, we use a fast NDK (Zheng et al, 2006) implementation which makes use of a pre-clustering step to group similar keyframes in a video corpus together and then perform NDK within each individual clusters. This clustering step is based on a set of globally invariant image features like the autocorrelogram of the transformation of color intensities. The transformation makes the color features invariant to illumination change by normalizing color intensity with its average intensity and variance. We then apply SIFT-based image matching within the cluster for NDK detection. The time taken for the NDK on the a video corpus containing eighty thousand shots only require a processing time of 90 hours on a standard machine. The result is a set of clusters as well as the SIFT matching results within each clusters which provides a basis of linkage between shots in a visual context manner.

2.1.5 Face Detection, Recognition and Shot Genre Detection

Our face detection is based on a cascade of boosted classifiers using an extended set of Haar-like features, which is an extension of Viola’s fast face detector. The face recognition algorithm we used is based on 2DHMM which is mainly use for anchor person detection. The detection of faces and recognizing them helps to provide person-related shot genres which is useful in providing partial semantic for news video. In particular, anchor person shots are also very useful in story boundary detection.

The shot-genres that are in this system includes: Political, Scientific, Entertainment, Sports, Weather, Financial, Disaster, Commercial and General which is detected by using our text classifier (Chua et al, 2005).

2.1.6 Story Boundary

Story segmentation is critical in retrieval especially at answer generation or summarization. If the segmentation is inaccurate, there will be a mixture of different stories summarized into one clip and will cause confusion to the searcher. In our implementation, we make use of the story boundary detection result provided by IBM-Columbia (Chang et al, 2005). We enhance the segmentation output by utilizing anchor-person shots for second level segmentation (Neo et al, 2006). The underlying reason is that we prefer over- rather than under-segmentation as the latter tends to cause the clusters to overlap more frequently. Furthermore, the analysis of various shorter segments within a long story is crucial to better understanding of the content of the main story.

2.2 News Video Events and Event Clustering

Almost every news video story is a description of an event. Typical event will have location, time of occurrence, people involved and a detailed description of what has happened. Using the story boundaries, we extract sets of semantically meaningful terms as well as visual features which denote the video events

2.2.1 Extracting Text Event Entities

Location is the most reliable features in the analysis of the event since they are seldom misrecognized or wrongly translated. Person names on the other hand are far more vulnerable to errors as they are non-vocabularies. There is however a need to identify the actual event location since there may be a number of locations mentioned in the news video story. We first group the various locations accordingly to their spatial relationship. For example, if “Iraq, Baghdad” are both mentioned in the story, “Baghdad” will be selected as event location as Baghdad is the capital of Iraq. Intuitively, a more specific place mentioned in the ASR suggests that it is likely that the event occurs there. For cases where there are multiple cities or countries involve, we choose the location which appears at the first. The reasoning is that this is often a standard practice in news report to state the location of happening at the beginning of the news. The time entity of the event is taken by default to be the date of the news video when no other date information is found in the story. However, cue terms such as “yesterday”, “two days ago” might signify that the event happens earlier. A list of predefined key-terms which helps to depict events like “fire, explosion, flood, war, fighting, murder…" is also use to extract keywords which can provide description to the news video story. The list which consists of approximate 700 words is automatically generated by gathering news snippets of the relevant period of the news video corpus. From past experiences, the person or organization which is involved in the event usually plays the most significant part in terms of the event identification. It is almost true that an event changes once the related people are all different. For this reason, we choose to include all the person names mentioned in the story by allowing an event entity to contain several key persons. However, it is also evidence that non-English names tend to be misrecognised as the pronunciation might not be what is assumed which therefore lead to a lost of crucial information. For MT texts, these errors are even more evident.

2.2.2 Inducing Text Extra Entities from External Resources

In order to recover missing key context which could be important, we employ external news resources from WIKIPEDIA NEWS as well as from our text news corpus database gathered in the same period as the TRECVID corpus. A total of several hundreds of news articles are available each day in the news article corpus. Morphological analysis is first performed on the retrieved documents to obtain the Part-of-Speech (POS). These POS-tagged documents are passed to the generic NE extractor module to obtain various NE types such as: Person Name, Location and Time.

2.2.3 HLF Event Entities

In order to select only the event-related HLF or HLFs which are useful using clustering, we adopt a selection scheme which bias against non-event-significant HLFs. For example, the semantic concepts of objects and people tend to be subjects of the news story, while semantic concepts of "place" and "scene/event" categories tend to provide a background or related incidents to the news story topic. In particular, various type of news can have different HLFs which may be important. For a disaster news report, the place and scene HLFs are the crucial elements to disaster-type news. While for a political news, the people and venue are more important. Table 1 illustrates the association of various shot-context types and the relevant categories of HLFs. Only HLFs relevant to the particular type of news story will be considered as the event-HLF and use in the clustering.

Table 4: List of Shot-type and their important HLFs

Shot context type	Categories of related HLF
Political	People, places
Entertainment	Scene/event
Scientific	People, scenes/events
Unclassified/ General	Objects
Financial	People, scenes/events
Weather/ Commercial	-
Disaster	scenes/events, places
Sports	People, actions

2.2.4 Event Clustering

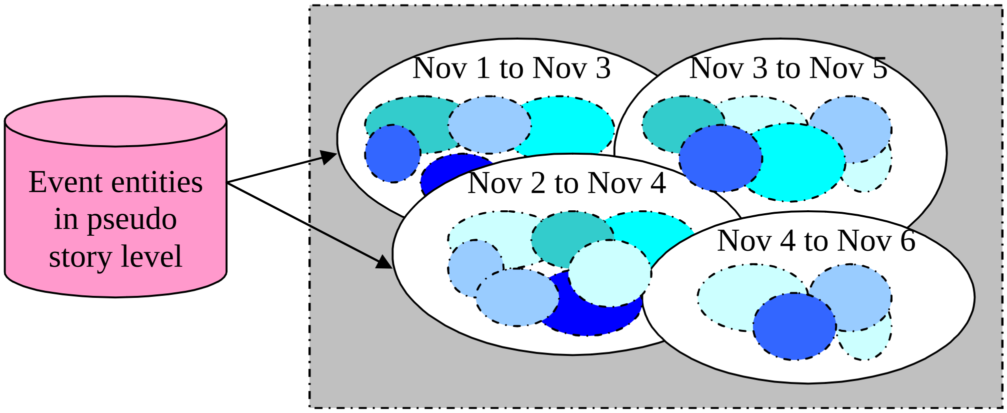
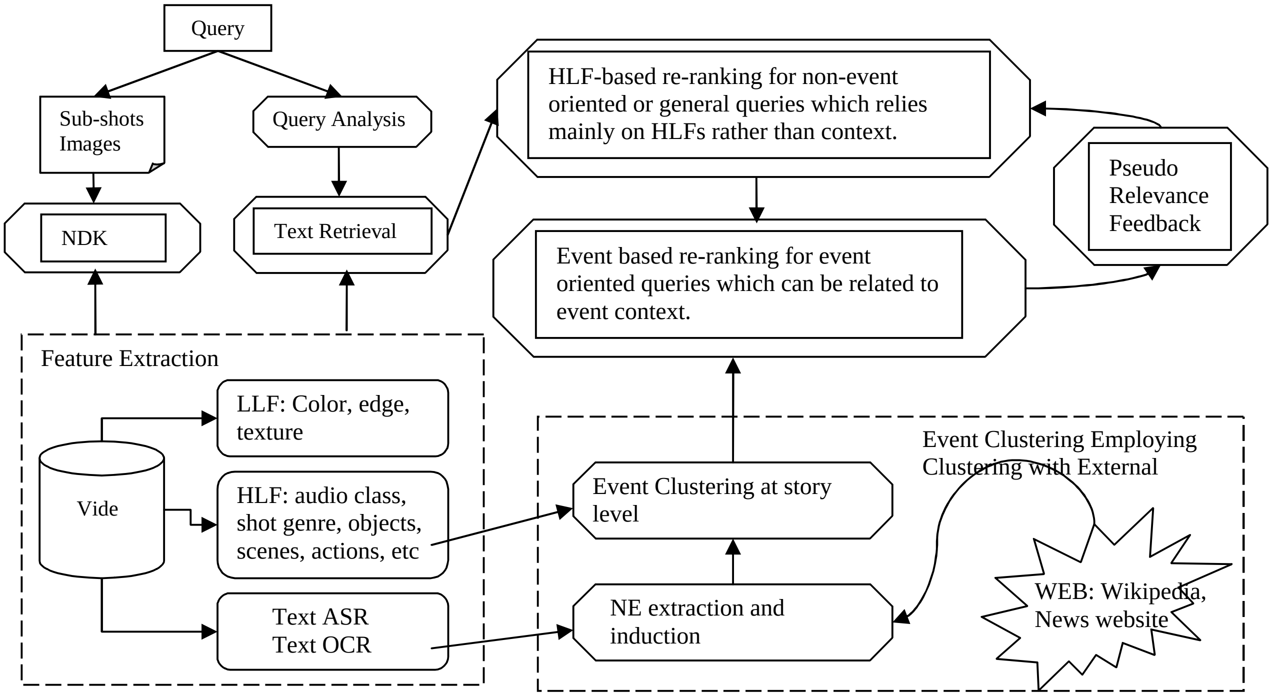


Figure 4: Time-dependent Event Clustering

We make use of a sliding window shown in Figure 3 to pre-group news stories. The grouping takes into consideration of the event-time rather than the video time-stamp. Such grouping allows a more robust clustering result as it is able to better model rare events. In an entire month of news, there can be a number of single-event news which does not belong to any large clusters. A generic clustering over the whole news corpus will induce many mistakes due to the single-event news. Therefore, it is necessary to make use of the news video properties in a time-dynamically way to achieve a good clustering performance. We employ a simple cosine similarity function over agglomerative clustering algorithm to groups the news video.

2.3 Retrieval Framework



Query
Sub-shots Images
Query Analysis
NDK
Text Retrieval
HLF-based re-ranking for non-event oriented or general queries which relies mainly on HLFs rather than context.
Pseudo Relevance Feedback
Event based re-ranking for event oriented queries which can be related to event context.
Feature Extraction
LLF: Color, edge, texture
Vide
HLF: audio class, shot genre, objects, scenes, actions, etc
Text ASR
Text OCR
Event Clustering Employing Clustering with External
Event Clustering at story level
NE extraction and induction
WEB: Wikipedia, News website

Figure 5: Overall Retrieval Framework

In the automated search where there is no user feedback in the process, it is necessary to combine various modalities appropriately to obtain good results. Query analysis therefore plays an important part by “understanding” the user’s underlying intention as well as initialing what multimodal features to consider. Subsequently, the retrieval will then be carried out based on the query-content.

2.3.1 Query Analysis

The five main query-content which is used in our system are the keywords, query-type, query-class, query-events and query-HLF. Images and video which is supplied together with the query is also processed. Keywords from the query provide the necessary initial context for retrieval. Keywords usually contain the topic and subject which pose the context basis of the query. The query-type and query-class (Chua et al, 2005) suggest the various types of shots-genre to be considered as well as if the query is directed at an event, person, scene or object. Query-type refers to the type of story context which is relevant to the query. The query-class on the other hand provides a visual descriptor to the type of shots required. Using the above examples: like “*Find shots containing a goal post*” requires a sports scene while “*Find shots of Condoleezza Rice*” is looking for a person’s face. We derived seven useful classes in news video*: anchor-person, live-reporting-person, live-reporting-non-person, sports scene, weather maps, financial charts and text scene*. A query can only be map to a single class. In general, it can be seen that the query-type and the query-class are very similar to the shot context and shot visual genre.

2.3.1.1 Query-events

The query-event suggests the relevancy of the query to the individual time-dependent clusters. We first formulate expanded query *Q’* by using query expansion on a parallel set of text articles to obtain high mutual information (MI) [20] words. In addition, sample video shots which forms part of the query will be processed to detect the presence of any HLFs. If a certain HLF is detected within these sample shots, this could mean that the particular HLF is essential.

2.3.1.2 Query-HLFs

We match the various available HLFs to query by applying morphological analysis on the query and the HLF description followed by selective expansion using the WordNet lexical database. The stronger the match between the descriptions and the query, the more important the HLF is to the query. In addition, we further employ the use of comparable news articles within the same period of time to further build and expand word-based relationships. WordNet gloss sometimes provides visual information about an object – its shape, color, nature and texture; whereas the latter only provides direct relations (e.g., *aircraft* & *airplane*; *fire* & *explosion*). For example, the word *boat* can not be related to *water* by virtue of any relationship link in WordNet, but by its gloss – *"a small vessel for travel on water."*

$$Sim\_Lex(Q_j, HLF_k) = (\sum_{t_q \in Q_j} \sum_{t_f \in HLF_k} \mathrm{Re}\,snik(t_q, t_f) / (|Q_j| \times |HLF_k|))$$

The expanded terms (Q_2, HLF_1) are then empirically weighted based on an approximate distance from the original terms *(Q_1, HLF_0)*. Expansion terms obtained from synonymy, hyponymy and gloss, where terms obtained from the gloss have a lower weight (due to noise words in the definition). We further enhance the formula by considering time as lexical similarity as computed from static dictionaries may not always be most suitable for news, especially because of news' transient nature. Aside from helping to increase to link named entities to common words, it refines the relations between words already linked by WordNet. For example, although the concept *fire* and *explosion* are associated in WordNet, in news stories the relationship can vary. A chemical factory explosion story is likely to have both terms highly correlated, but a story on forest fires is unlikely to have the *explosion* concept. Similarly, *car*, *boat* and *aircraft* are related in WordNet as means of transportation, but searches for any of the three usually should not return shots of the other two objects. Thus when system relies solely on lexical links between words as processed from such dictionaries, they may return spurious results. To overcome these problems, we sampled external sources of news to model the dynamic weighting of similarity between HLFs across time. We use the external news articles to calculate the co-occurrence of $feature_1$ and $feature_2$ with respect to time. The relationship between fire and explosion is thus modified according to their co-occurrence in the external articles. If no news articles directly relate explosion and fire during a certain time period t, the link weight between *explosion* and *fire* is reduced accordingly. This score is then fused with *Lex_Sim()* from above to obtain the time-dependent similarity function $Lex\_Sim_t()$. The equation below gives the final, time-sensitive similarity measure.

$$Lex\_Sim_t(Q_j, HLF_k) = \gamma \cdot Sim\_Lex(Q_j, HLF_k) + (1-\gamma) \cdot MI(Q_j, HLF_k \mid t)$$

More details can be found at (Neo et al, 2006)

2.3.2 Retrieval

During retrieval, we obtain relevant segments of news in the following 3 stages: 1) pseudo story-level retrieval; 2) multimodal shot-level re-ranking; and 3) pseudo relevance feedback based on top return results.

2.3.2.1 Story Retrieval

Previous experiments in (Chua et al, 2005) focus on using text features and the query-type for story retrieval. It was experimentally shown that irrelevant segments were eliminated while recall was maintained. In this work, we extend the idea by matching the query to event-clusters in the time-dependent clustering framework to further enhance the recall. Time information plays an important role in this clustering retrieval because we need to ensure good clusters and minimize noise. The story retrieval scoring function is:

$$\begin{aligned} score\_story(Q, PS) &= \alpha \cdot text(q_{text}, ps_{text}) \\ &\quad + \beta \cdot \max\{u_{PS,P} \cdot Sim\_cluster_t(Q', P \mid PS \in P)\} \\ &\quad + \gamma \cdot type(q_{type}, ps_{type}) \end{aligned}$$

where *α*, *β*, *γ* are variables *(α+β+γ=1)*, *text()* is the *TI.IDF* score, max{} choose best result of by computing the membership *u* of *PS* and cluster *P* with respect to all clusters, *type()* suggest whether the query-type of *Q* coincide with the context genre of *PS*.

2.3.2.2 *Shot level retrieval*

After obtained the pseudo segments, we can re-rank the shots based on the induced query-class, the query-HLF and the NDK. We make use of the HLF detection confidences (from Section 3.2.1) as a measurement of the confidence during fusion. The scoring function *score_shot(Q, S_i)* is used to re-rank the shots in the pseudo stories segments.

$$\begin{aligned} score\_shot(Q,S) = {} & \alpha_c \cdot score\_story(Q, PS \mid S \in PS) \\ & + \beta_c \cdot \sum_{HLF_k \in S} \left[Conf(HLF_k) \times Sim\_Lex_t(q, HLF_k) \right] \\ & + \gamma_c \cdot class(q_{class}, S_{class}) + \delta_c \cdot NDK(q_{images}, S) \end{aligned}$$

where α_c, β_c, γ_c, δ_c are variables for query-class *c*, $(\alpha_c+\beta_c+\gamma_c+\delta_c=1)$, *Conf()* is estimated confidence, *class()* suggest whether the query-class of *Q* coincide visual class of shot *S* and NDK(q_{images},*S*) suggest whether *S* is a near duplicate of an image in the set if Q_{images}.

2.4 Results and Discussions

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

Run 6. (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 5. (Query expansion text-only run). 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 6. The shots are returned in a similar manner as in Run 5.

Run 4. (HLF-only run). We match the query to high level features words. HLFs which overlap with the query-terms are automatically given same positive weights. Subsequently, we make use of the detection confidence describe in *Section 2.1.3* to rank the shots.

Run 3. (HLF-only run with inference). This run enhances Run4 by considering giving different importance to different HLF based on descriptive lexical matching and time-dependent mutual information.

Run 2. (Text + HLF + Visual feature rankings run) This run is based on the text run and fused with the visual feature. Each shot is indexed using 39-dimensional feature in the model space. Then we pseudo-relevance feedback to get the positive and negative samples for each query, and a LDF classifier based using the AUC maximization algorithm is trained for ranking the shots. The visual based ranking is further combined with the text run to get the final ranking. If the ranking list of text run has less than 1000 shots, we use the top-N visual ranking output to append it.

Run 1. (Run2 + Event Model). This runs make use of the suggested features fused in our proposed event-based framework.

Table 3: Evaluation Results in MAP

	Run1	Run2	Run3	Run4	Run5	Run6
MAP	0.075	0.067	0.051	0.043	0.042	0.039

The pure text runs like Run5 and Run6 yield bad results as text alone is not sufficient to support the precise shot retrieval required by TRECVID queries. Observation shows that some texts translated from other languages may include meaningless words or phrases that are useless during retrieval. In Run5, we notice that the usage query expansion to derive additional context and words improves the performance by about 5%. This improvement is lower compared to previous years (about 20% improvement for both TRECVID 2004 2005). One of the reasons which could suggest this phenomenon is the shift from context oriented queries to visual oriented queries this year. In particular, query expansion is more favorable to context oriented queries. The result of Run4 which only uses HLF through simple keyword matching is higher than results of both text runs further confirm our conjecture. Run3 which uses the inference of descriptive lexical matching and time-dependent mutual information produce significant improvement over Run4. For Run 2 and Run 1, we took advantage of all the features using the specification in Run3 and Run5. The only difference between both runs is that Run1 makes use of our event-based model while Run2 make use of our feature re-ranking model. Though Run 2 demonstrates the effectiveness of multimodal fusion, it is only able to achieve a MAP of 0.067. In Run 1, we are able to increase the performance by about 12% to 0.075.

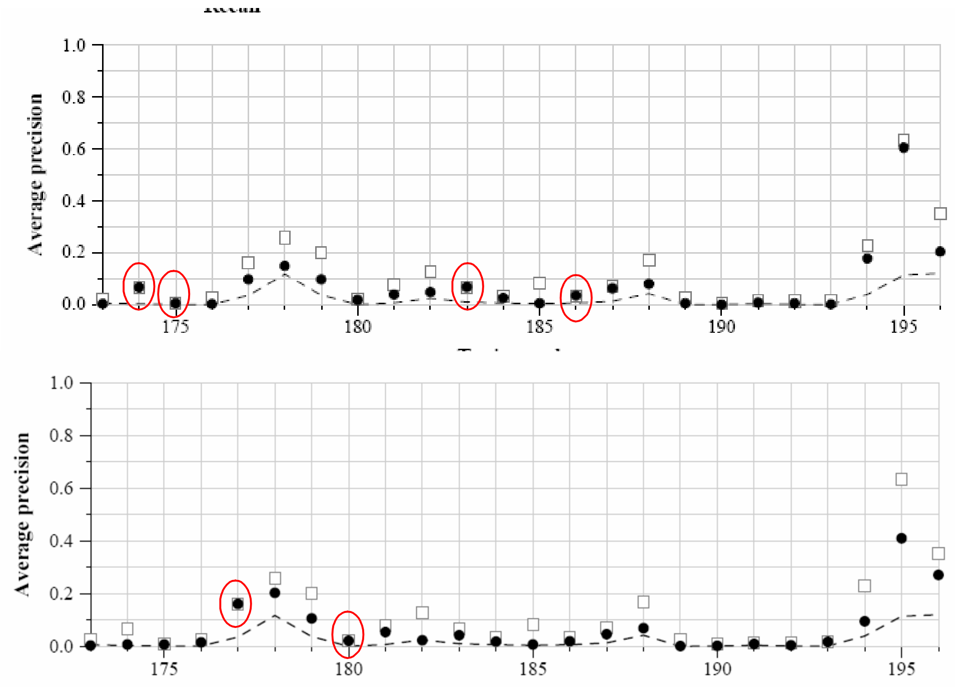


Figure 6: Best Performing runs: Run1 and Run2

A breakdown of our system performance shows that it is able to obtain best average precision for 6 of the queries, with also 10 other queries well above the median performance. An analysis of our system performance shows that our retrieval system is able to outperform other systems in Name entity or event-related queries, as well as produce similar comparative results in other classes of queries.

References

S-Y. Neo, J. Zhao, M-Y. Kan, T-S. Chua "Video Retrieval Using High-level features: Exploiting Query-matching and Confidence-based Weighting" CIVR 2006, Arizona, USA, July 2006.

G. Miller, "Wordnet: An on-line lexical database". International Journal of Lexicography (1995)

Shi-Yong Neo, Yantao Zheng, Tat-Seng Chua, Qi Tian "News Video Search with Fuzzy Event Clustering using High-level Features" In ACM MM 2006, Santa Barbara, USA, 23-27 October 2006.

Shi-Yong Neo, Tat-Seng Chua "Query-dependent Retrieval on News Video" In MMIR 2005, SIGIR 2005 workshop, Salvador, Brazil, 19 Aug 2005.

Yantao Zheng, Shi Yong Neo, Tat Seng Chua, Qi Tian, "Fast Near-duplicated Keyframe Detection in Large-scale Video Corpus for Video Search" IWAIT 2007, Thailand

Winston H. Hsu and Shih-Fu 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.

Tat-Seng Chua, Shi-Yong Neo, Ke-Ya Li, Gang Wang, Rui Shi, Ming Zhao and Huaxin Xu "TRECVID 2004 Search and Feature Extraction Task by NUS PRIS" In TRECVID 2004, NIST, Gaithersburg, Maryland, USA, 15-16 Nov 2004.

Tat-Seng Chua, Shi-Yong Neo, Hai-Kiat Goh, Ming Zhao, Yang Xiao, Gang Wang "TRECVID 2005 by NUS PRIS" In TRECVID 2005, NIST, Gaithersburg, Maryland, USA, 14-15 Nov 2005. [3] G. Miller, "Wordnet: An on-line lexical database". International Journal of Lexicography (1995)

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

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

Hauptmann, A.G., Christel, M., Concescu, R., Gao, J., Jin, Q., Lin, W.H., Pan, J.Y., Stevens, S.M., Yan, R., Yang, J., Zhang, Y.: CMU Informedia's TRECVID 2005 skirmishes. In: TRECVID, 2005. (2005)

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.

M. Zhao, S.Y. Neo, H. K. Goh, T. S. Chua, "Multi-Faceted Contextual Model for Person Identification in News Video" in MMM 2006, 7-10 Jan 2006 Beijing, China.

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.

Shi-Yong Neo, Hai-Kiat Goh, Tat-Seng Chua, "Multimodal Event-based Model for Retrieval of Multi-Lingual News Video" In International Workshop on Advance Image Technology (IWAIT), Okinawa, Japan, 9-10 Jan, 2006.

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.

D. Seidman, "Careers exploring in journalism", The Rosen Publishing Group, New York, 2000

J. C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York

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

Amir, A., Iyengar, G., Argillander, J., Campbell, M., Haubold, A., Ebadollahi, S., Kang, F., Naphade, M.R., Natsev, A.P., Smith, J.R., Tesic, J., Volkmer, T.: IBM research TRECVID- 2005 video retrieval system. In: TRECVID, 2005.

C. Snoek, C.G.M., van Gemert, J., Geusebroek, J.M., Huurnink, B., Koelma, D.C., Nguyen, G.P., de Rooij, O., Seinstra, F.J., Smeulders, A.W.M., Veenman, C.J., ,Worring, M.: The MediaMill TRECVID 2005 semantic video search engine. In: Proceedings of the 3rd TRECVID Workshop, NIST

Hauptmann, A.G., Christel, M., Concescu, R., Gao, J., Jin, Q., Lin, W.H., Pan, J.Y., Stevens, S.M., Yan, R., Yang, J., Zhang, Y.: CMU Informedia's TRECVID 2005 skirmishes. In: TRECVID, 2005.

Foley, C., Gurrin, C., Jones, G., Lee, H., McGivney, S., O'Connor, N.E., Sav, S., Smeaton, A.F.,Wilkins, P.: TRECVid 2005 experiments at dublin city university. In: TRECVID, 2005.

S. Gao & Q.B. Sun. "Classifier optimization for multimedia semantic concept detection", ICME'06.

S. Gao, X.L. Zhu & Q.B. Sun, "Exploiting concept association to boost multimedia semantic concept detection", submitted to ICASSP'07.

T. Joachims, Learning to Classify Text Using Support Vector Machines. Dissertation, Kluwer, 2002.

Hang Cui, Soft Matching for Question Answering, *PhD Thesis*, 2006.

Manning, Christopher D. and Hinrich Schütze. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts, 1999.

Yang Xiao, Tat-Seng Chua and Chin-Hui Lee, Fusion of Region and Image-based Techniques for Automatic Image Annotation, to be appeared in *The 13th International MultiMedia Modeling Conference* (MMM07), 2007.