Video Retrieval using High Level Features: Exploiting Query Matching and Confidence-based Weighting

Shi-Yong Neo*, Jin Zhao, Min-Yen Kan, and Tat-Seng Chua

Department of Computer Science, School of Computing,
National University of Singapore, Singapore, 117543
{neoshiyo,zhaojin,kanmy,chuats}@comp.nus.edu.sg

Abstract. Recent research in video retrieval has focused on automated, high-level feature indexing on shots or frames. One important application of such indexing is to support precise video retrieval. We report on extensions of this semantic indexing on news video retrieval. First, we utilize extensive query analysis to relate various high-level features and query terms by matching the textual description and context in a time-dependent manner. Second, we introduce a framework to effectively fuse the relation weights with the detectors' confidence scores. This results in individual high level features that are weighted on a per-query basis. Tests on the TRECVID 2005 dataset show that the above two enhancements yield significant improvement in performance over a corresponding state-of-the-art video retrieval baseline.

1 Introduction

News video retrieval systems often perform retrieval based solely on automatic speech recognition (ASR) results on the video's audio. This is because ASR, while not fully accurate, is reliable and largely indicative of the topic of videos. Such a transformation of the video retrieval problem into a text-based one has been shown to be effective [1].

To further increase the accuracy and resolution of video retrieval requires analysis and modeling of the video and audio content. The community has investigated this in part by developing specialized detectors that detect and index certain **H**igh-**L**evel **F**eatures (HLFs; e.g., presence of cars, faces and buildings). As such, research retrieval systems incorporate both standard text-based information (from ASR and/or closed captions) with results from an inventory of detectors designed to capture HLFs. In order to carry out a large-scale retrieval of video in a real time environment, most features have to be extracted and preprocessed during offline indexing. In the current state of the art, systems cannot detect and index (or even conceptualize) every possible useful high-level semantic feature. Therefore, it is necessary to carry out inference on a limited set of detectable HLFs that cover and support a wide range of queries. Thus we focus on using only ASR and the HLFs to support news video search.

We offer two extensions to this basic framework that enhance the contributions of HLFs, based on two observations. First, we note that many HLFs have a natural textual description (e.g., "car", "face") that have not been widely utilized for retrieval. We show

* Contact author, supported by the Singapore Millennium Foundation (SMF).

how to match such feature descriptions with the user's textual query to enhance retrieval performance in a time-dependent manner. We approach this by employing morphological analysis followed by selective expansion using the WordNet [2] lexical database on both the feature descriptions and the user's query. The stronger the match between the descriptions and the query, the more important the HLF is to the query. However as queries are often time-sensitive (featuring new personas, corporations each day, using only the static information in WordNet is not enough. Thus we further employ the use of comparable news articles within the same period of time to further build and expand word-based relationships. Crucially different from previous work that only employs lexical expansion, our method fuses both static lexical information with dynamic correlation by calculating time-dependent mutual information [3].

Secondly, as HLF detectors vary greatly in performance, it is necessary to consider their accuracies in the fusion process. Currently, retrieval systems have used the output of such batteries of detectors "as-is", without considering the confidence of individual detectors. For example, detectors for faces are fairly robust, whereas detectors for cars and animals are not. We introduce a performance-weighted framework which accounts for this phenomenon. Different from previous work, it evaluates the accuracy of individual high-level detectors during training/validation and utilizes probability of correct detection in feature weighting during testing.

We have validated our approach on the TRECVID 2005 dataset [4] and queries. Our experimental results show that the appropriate use of HLFs in retrieval outperforms text-based systems and improves results on a representative state-of-the-art multimodal retrieval systems.

2 Use of High-level Features in Video Retrieval

Starting from text-based search, video retrieval has incorporated the use of low-level video features (e.g., color, motion, volume) and, more recently, high level features for specific objects or phenomenon (e.g., cars, fire, and applause). To create such high level features, recent work has taken a machine learning approach, where each HLF detector is trained against an annotated corpus of video clips [4, 5]. A well-known example is the LSCOM set, which contains approximately 1000 concepts which can be used for video annotation. In TRECVID 2005, the LSCOM-lite set (a LSCOM subset of 39 interesting concepts) have been selected and tagged to provide training examples of approximately 50,000 shots or 70 hours of video. The detectors trained using these examples introduce useful and partial semantics to retrieval systems.

The IBM group used a fusion of low-level features and HLFs based on two learning techniques: Multi-example Content Based Retrieval (a k-NN variant) and support vector machines [6]. Their system automatically maps query text to HLF models. The weights are derived by co-occurrence statistics between ASR tokens and detected concepts as well as by their correlations.

[7] represented the text queries and subshots in an intermediate concept space which contains confidences for each of the 39 concepts. The subshots are represented by the outputs of the concept detectors for each concept, smoothed according to the frequencies of each concept and the reliability of each concept detector. The text queries are

mapped into the concept space by measuring the similarity between the query terms and the terms in the concept's description. This approach was applied to automatic, manual, and interactive searches, yielding high performance for the few topics which have high-performing correlated concepts.

The MediaMill group [8] also extended the LSCOM-lite set by increasing the HLF pool to 101 features, some original as well as some recycled from the previous TRECVID tasks. Other top performing interactive retrieval systems from Informedia [9] and DCU [10] also show effective methods of integrating high level semantic features. One may conclude that even though the HLF detection accuracies are much lower than low level features, HLF have shown to be more useful for semantic queries.

In this work, we use a set of 25 HLFs for news video retrieval. Our primary reason for choosing this set is that the corresponding detectors are readily available and have been trained previously on both the TRECVID 2004 and 2005 HLF task. In addition, they have shown to be useful in retrieval in previous work [11, 12]. These 25 features are targeted towards identifying the video genre, objects, backgrounds and actions, as shown in Figure 1. The HLF task requires system to return ranklists of maximum 2000 shots for each HLF. Our system achieves a mean average precision (MAP) of 0.22. In order to maximize the detection accuracy, we combine the best available HLF detection results from various participating groups. We only select ranklists which have a MAP $\geq$ .2 and above (including IBM's HLF detector set [6], which has a .33 MAP). The score of shot S_c containing HLF_k is calculated using the following equation:

$$Score(S_c|HLF_k) = \alpha \sum_j Contains(S_c) + (1-\alpha) \sum_j \frac{maxPos - Pos(S_c)}{maxPos} \quad (1)$$

where $Contains()$ is an indicator function that checks whether a shot is present on the ranklist 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 ranklist. The resulting ranked list achieves a MAP of 0.38.

- Genres: anchorPerson, $commercial, politics, \underline{sports}, weather, financial$
- Objects: $face, \underline{fire}, \underline{explosion}, \underline{car}, \underline{U.S.flag}, boat, aircraft, \underline{map}, \underline{buildingExterior}, \underline{prisoner}$
- Scene: $\underline{waterscape}, \underline{mountain}, sky, outdoor, indoor, disaster, vegetation$
- Action: $\underline{peopleWalking}, peopleInCrowd$

Fig. 1. High level features used by our system. The ten underlined features indicate the required features from the TRECVID HLFs; italicized features come from LSCOM-lite.

However, having a well-trained, accurate set of HLF detectors is not sufficient for precise retrieval. This is because each HLF detector models a specific phenomenon, and which detectors are useful for particular queries varies greatly. Correctly determining and matching detectors to queries is therefore a critical task. Past systems have done this matching manually or using simple automated methods by unsupervised clustering or

simple expansion using dictionaries. In this work, we leverage the textual descriptions of the HLF set for time-dependent matching and also incorporate the confidence of the detectors in our fusion process. This is illustrated in Figure 2 which shows the placement of both of these modules in our processing framework for large-scale news video retrieval. We describe this two-fold approach in the following two sections.

Note that in the remainder of the paper, system parameters (normally indicated by lowercase Greek letters) that are introduced have all have been optimized by either manual tuning or learned from training.

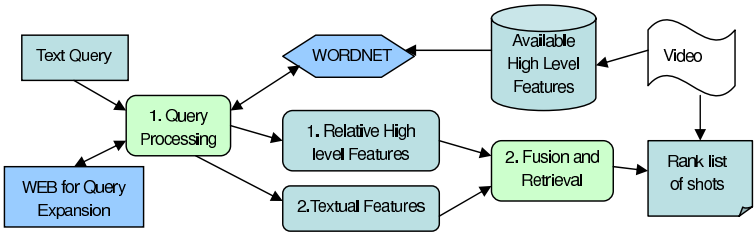

Fig. 2. Retrieval Framework

3 Query processing for HLF weighting

As user queries are usually short and contain insufficient context to perform a precise retrieval, we employ previous work on query expansion techniques using external resources [13] and query classification [14] during query processing to expand the user's original query (denoted Q_0) obtain an initial expanded query, Q_1.

WordNet has been a heavily utilized source of ontological lexical information in text retrieval. In text retrieval, systems relate terms by synonymy, hypernymy, hyponymy and overlap in definitions (gloss). We employ a technique close in spirit to the MediaMill group [8] to determine the match between a detector and a query. Both the short one or two word original description of the detector and the user's expanded query (HLF_0 and Q_1) are expanded using WordNet. Both pieces of data are first tagged for part-of-speech using a commercial product, and then closed-class words and words on a 400+ word video domain stopword list are removed.

WordNet expansion. Unlike previous work, we include terms from the WordNet gloss as we have found that the terms extracted from the gloss differs significantly from those extracted from a term's synonyms and the hypernym/hyponym hierarchy. The former 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."*

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).

A final matching phase is done to determine which high level features are most relevant to the query. To match HLF_1 to Q_2 we use the information-content metric of Resnik [15] (as was done in [16]), which equates similarity with the information content of the pair of words' most specific common ancestor: $Resnik(t_i, t_j) = IC(lcs(t_i, t_j))$ where $lcs(t_i, t_j)$ is the most deeply nested concept in the *is-a* hierarchy that subsumes both t_i and t_j). Here, we factor in the expanded term weights from the previous step.

$$Sim\_Lex(Q_j, HLF_k) = (\sum_{t_q \epsilon Q_j} \sum_{t_f \epsilon HLF_k} Resnik(t_q, t_f))/(|Q_j| \times |HLF_k|)) \quad (2)$$

After summing all such scores for each HLF, the k top scoring HLFs are taken with their weights and used in the final retrieval.

This framework would be fine for video in which the associated text information is aligned exactly to the clip. However, in professionally edited video, speech often comes before the corrsponding visual information. We therefore carry forward β seconds of speech of each preceding shot to its succeeding shot ($\beta = 12$, roughly equivalent to an average shot duration).

Time sensistive expansion. 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 systems relies solely on lexical links between words as processed from such dictionaries, they may return spurious results.

To overcome these problems, we sampled external (e.g., non-TRECVID) 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. Given a query, the system first retrieves the top relevant shots from the test set. We build a corpus of news articles centered on the timestamp of each retrieved shot. As illustrated in Figure 3, given a period $\delta = 3$ days, all the available news articles for these 3 consecutive days will be used for finding the MI (Mutual Information) of between the word $feature_1$ and $feature_2$. This score is then fused with $Lex\_Sim()$ from above to obtain the time-dependent similarity function $Lex\_Sim_t()$. Equation 3 gives the final, time-sensitive similarity measure.

$$Sim\_Lex_t(Q_j, HLF_k) = \gamma Sim\_Lex(Q_j, HLF_k) + (1 - \gamma) MI(Q_j, HLF_k|t) \quad (3)$$

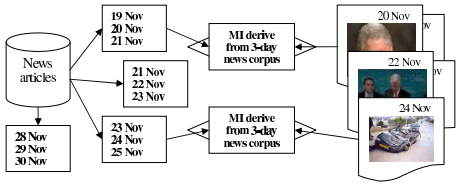

Fig. 3. Dynamic corpora generation for term correlation in video shots

4 Confidence fusion and retrieval

The retrieval step is a text-based retrieval scoring function enhanced with HLF confidence scoring. We use the text scoring function $Text(S_i)$ from [11]. This scoring function utilizes the query's class and other additional contextual information to retrieve the relevant documents. In this previous work, it was experimentally shown that irrelevant segments were eliminated while recall was maintained.

As the HLF detectors perform at different accuracy levels, we must consider their precision in retrieval. Using the available training samples, we obtain accuracies for each detector using 5-fold cross validation, in terms of mean average precision. The final score of shot S_i with respect to query Q is given below.

$$Score(S_i) = \zeta * Text(S_i) + (1-\zeta) * \sum_{HLF_k \epsilon S_i} Conf(HLF_k) \times Sim\_Lex_t(Q, HLF_k) \quad (4)$$

where $Conf(HLF_k)$ is estimated MAP of the $Detector_k$. The score for each shot will be computed based on the available textual features as well as the HLFs and their detection confidence with respect to the query.

5 Evaluation

The goal of our evaluations is to show the efficacy of both modules: HLF weighting and confidence-based fusion. For the weighting, we can measure how well our automatic weighting scheme agrees with the importance assigned to the HLFs by human subjects. For fusion, we measure the gain in retrieval performance when incorporating confidence in the HLF weighting scheme. We also measure the synergy when employing both modules together in the retrieval framework.

5.1 HLF weighting agreement with human subjects

We asked 12 paid volunteers to take a survey that assessed how they would weight HLFs in video retrieval. All participants were either university postgradute or undergraduate students and had not used textual descriptions to search for videos before. We selected 8 queries from past TRECVID queries that were representative of different semantic classes (e.g., "George Bush", "Basketball players on court" and "People entering and leaving buildings"), and asked the participants to first freely associate what types of

HLFs would be important in retrieving such video clips, and second, to assign a value on a scale from 1 (important) to 5 (unimportant) of the specific HLF inventory set used in our system (c.f., Figure 1) for the same 8 queries. 5 here refers to a strong positive correlation between the HLF and the query; 1, a negative correlation. In total, we gathered about 2400 judgments.

An analysis of the free association subtask shows that over 90% of the responses are concrete nouns, confirming earlier work that searchers focus on nouns as cues for retrieval. In addition, although only 5% of the features used in our experiments are mentioned explicitly in the free association task, some were later ranked as "Important" by participants in the second subtask. Calculating the interjudge agreement using Kappa (which varies from -1 (no agreement) to 1 (full correlation)), we found only a low agreement ranging from 0.2 to 0.4, which varied with the search task. The low agreement may be partially due to the inexperience of the participants in searching video. We also calculated the standard deviation of a feature's score for each search task, shown in Table 1, which showed results similar to [17].

Table 1. Importance ratings of most HLFs across all 8 search tasks. Blank cells indicate high standard deviations (above 0.7). Features sorted by standard deviation.

Feature (Avg. s.d.)	Search Task							
	Map	Tree	Office	Basketball	Ship	Hu Jintao	George Bush	Fire
Fire & Explosion (0.6)		1.4	1.4	1.3		1.3	1.4	5.0
Car (0.7)		1.4	1.2	1.3	1.4		4.4	4.0
Boat (0.8)	1.1	1.5	1.1	1.1	4.5	1.4		
Aircraft (0.9)		1.4	1.2	1.1	1.6			
Face (1.1)	1.2	1.3						
People Walking (1.2)					1.5			
Map (0.7)	4.8		1.4	1.1	1.2	1.5	1.6	1.4
People in Crowd (1.2)								
Sports (0.8)	1.2		1.4	4.7		1.5	1.5	
Weather (0.9)				1.4			1.1	1.4
Disaster (0.9)		1.5	1.4	1.2			1.7	4.5
Building Exterior (1.0)				1.1	1.2			
Waterscape (1.0)	1.5		1.3	1.1	4.1	1.7	1.5	1.3
Outdoor (1.0)	1.6		1.1		3.9			
Indoor (0.9)	1.2	1.2	4.5		1.5			1.5

In some cases, the degree of agreement is high, especially when the search task mentions the feature directly (e.g., the basketball query mentioning "sports"). In fact, a trend of HLF rating stability was observed. Ratings for concrete nouns were most stable, followed by backgrounds and video categories, and with those describing actions being the most variable or unreliable. We also note that negative correlations (scores close to 1.0) are prominent in our dataset. We feel this is quite reasonable, as only a few HLFs are usually relevant per query. We have also computed the Kappa value between the HLF rankings from our system and the ones from the human judges. The value ranges between 0 to 0.25 with a mean value of 0.145. We believe this varying level of agreement is due to the fact that WordNet expansion works well for hypernym and hyponym relations, but less so for other relation types. As a result, the overall agreement is weak. We plan to look into the problem of how to enrich the WordNet so that it is capable of discovering other relations in the near future as an extension to this work.

5.2 Text Retrieval and Query Matching

We follow the evaluation standards in TRECVID 2005 automated search task. A maximum of 1,000 shots are returned for each query; performance is measured by MAP.

We modify our retrieval system [11] to perform the required text retrieval. The text-based retrieval engine uses the text query to retrieve pre-segmented passages of text in the (possibly machine translated) ASR transcripts. These segments correspond to phrase level video segments in the corpus. The video segment associated with the matched phrase and the segment immediately afterwards are retrieved as the retrieved results (because of the aforementioned time lag betweens speech and video). This text-based baseline system also incorporates query expansion using external news resources. The resulting text retrieval system achieves an MAP of 0.063 based on the TRECVID 2005 dataset and queries. In comparison, the top three performers in TRECVID 2005's search tasks yield MAP of 0.067, 0.062, 0.061 respectively (mostly based on common dataset), showing that our text baseline is competitive.

To test the effectiveness of our query matching techniques, we further compare the performance to this system [11] which uses heuristics to weight HLFs to individual queries. When HLFs are integrated into the text-only system [11], the jump in MAP is significant (from 0.063 to 0.104) and validates earlier reported work. To test the effectiveness of the various components in the query matching module, three runs have been carried out.

Table 2. The performance in MAP combining textual features and HLF in retrieval.

Technique of Using HLF + Text	**MAP (% Gain)**
Heuristics weighting by [11] (used as baseline)	0.104
Run1. Automated HLF query matching	0.106 (+1.9%)
Run2. Automated HLF query matching + Gloss (Eqn. 2)	0.110 (+5.8%)
Run3. Automated HLF query matching + Gloss + Temporal MI (Eqn. 3)	0.113 (+8.6%)

The table shows that the use of HLFs during fusion outperformed the text-based retrieval system by more than 50%. This is conclusive as textual feature alone are not reliable enough to pinpoint shots which are relevant to the query. Run1 and Run2 indicates that the use of WordNet glosses is positive as the performance increases from the MAP of 0.106 to 0.110. Run3, which uses all the components obtains a MAP 8.6% better than the baseline system. The main improvement comes from the sport and general queries. Queries which are directly or indirectly related to the available 25 HLFs benefit the most. This suggest that as more HLFs are added, a better performance can be obtained. The MAP performance is higher due to its re-ranking of relevant shots as it takes all 25 HLFs into consideration during fusion. This performance is also better than a similar evaluation run (MAP of 0.070) submitted by IBM [6] which uses only text and HLFs.

5.3 Confidence-based fusion and A/V Integration

For the confidence-based fusion, we carried out two more runs to investigate the effects of considering HLF detection accuracy in the retrieval. As Run1 to Run3 uses HLF detection result without considering the accuracy of the various HLF detectors (normal fusion), we added Run4 which applies confidence-based fusion as in Eqn. 4. Run5 is designed to investigate the overall performance of the system by integrating other A/V features including low level features from [11]. The fusion is done by modifying the query-dependent multimodal fusion function in [11] to accommodate Eqn. 4. Results of these experiments are reported in Table 3.

Table 3. Aggregate MAP of the runs. Percentages indicate performance gain over the baseline system.

Experiment	**Normal fusion**	**Confidence-based fusion (Eqn. 4)**
Run4. Text + HLF	0.113 (+8.6%)	0.117 (+12.6%)
Run5. Text + HLF + A/V features[11]	0.127 (+22.1%)	0.131 (+25.9%)

The result shows that the use of confidence-based fusion yields significant improvement over normal fusion. The confidence-based fusion Run4 achieves a MAP of 0.117. This performance is statistically comparable to top performing submissions. The run that incorporate the rest of the A/V features obtains the MAP of 0.127 and 0.131 respectively, which is better than the best published MAP of 0.123 in TRECVID 2005 automated search task. The bulk of improvement come from the general queries as they depend largely on the use of HLFs as evidence of relevancy. Person-oriented queries on the other hand have less significant improvement as textual features and video OCR still constitute the main score. As the confidence-based fusion and the automated HLF to query matching affect different parts of the retrieval system, they can be combined easily, producing largely independent gains on MAP.

6 Conclusion

As video analysis has advanced to building high-level semantic features from low-level ones, schemes that judisciously employ such HLFs are needed. We explore two distinct and complementary approaches to extend the current frameworks of such multimodal retrieval systems. We have investigated methods to automate and expand the matching of HLFs to user query terms. In particular, our query to HLF mapping methods examine 1) the use of dictionary definitions (WordNet's glosses) to help relate terms, and 2) time sensitive mutual information to make sure that the scores are sensitive to the timeframe and story distribution in the video corpus. Overall, our newly Text + HLF retrieval system is able to outperform baseline system and achieve similar results to top performing automated systems reported in TRECVID 2005. This framework is further tested by integrating other A/V features and the resulting performance is better than the best reported result.

References

1. Hauptmann, A., Chen, M.Y., Christel, M., Huang, C., Lin, W.H., Ng, T., Papernick, N., Velivelli, A., Yang, J., Yan, R., Yang, H., Wactlar, H.D.: Confounded expectations: Informedia at TRECVID 2004. In: TRECVID, 2004. (2004)
2. Miller, G.: Wordnet: An on-line lexical database. International Journal of Lexicography (1995)
3. Neo, S., Goh, H., Chua, T.: Multimodal event-based model for retrieval of multi-lingual news video. In: IWAIT. (2006)
4. Over, P., Ianeva, T.: TRECVID 2005: An introduction. In: TRECVID, 2005. (2005)
5. Smeaton, A.F., Kraaij, W., Over, P.: TRECVID - an overview. In: TRECVID, 2003. (2003)
6. 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. (2005)
7. Chang, S.F., Hsu, W., Kennedy, L., Xie, L., Yanagawa, A., Zavesky, E., Zhang, D.Q.: Columbia university TRECVID-2005 video search and high-level feature extraction. In: TRECVID, 2005. (2005)
8. 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 (2005)
9. 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)
10. 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. (2005)
11. Chua, T.S., Neo, S.Y., Goh, H.K., Zhao, M., Xiao, Y., Wang, G.: TRECVID 2005 by NUS PRIS. In: TRECVID 2005. (2005)
12. Chua, T.S., Neo, S.Y., Li, K., Wang, G., Shi, R., Zhao, M., Xu, H.: TRECVID 2004 search and feature extraction task by NUS PRIS. In: TRECVID 2004. (2004)
13. Yang, H., Chua, T.S., Wang, S., Koh, C.K.: Structured use of external knowledge for event-based open-domain question-answering. In: SIGIR 2003, Canada, Jul 2003. (2003)
14. Neo, S., Chua, T.: Query-dependent retrieval on news video. In: MMIR'05 workshop in SIGIR'05. (2005)
15. Resnik, P.: Semantic similarity in a taxonomy: An information- based measure and its applications to problems of ambiguity in natural language. Journal of Artificial Intelligence Research,11 (1999) 95–130
16. Kennedy, L.S., Natsev, A.P., Chang, S.F.: Automatic discover of query-class-dependent models for multimodal search. In: ACM Multimedia (MM '05). (2005) 882–891
17. Christel, M.G., Hauptmann, A.G.: The use and utility of high-level semantic features in video retrieval. In: Conf. on Image and Video Retrieval, Singapore (2005) 134–144