News Video Search with Fuzzy Event Clustering using High-level Features

Shi-Yong Neo, Yantao Zheng, Tat-Seng Chua

School of Computing, National University of Singapore
3 Science Dr 2, Singapore 117543
(65) 6516-4426

{neoshiyo, yantaozheng, chuats}@comp.nus.edu.sg

Qi Tian

Institute for Infocomm Research (I2R)
21 Heng Mui Keng Terrace, Singapore 119613
(65) 6874-7588

tian@i2r.a-star.edu.sg

ABSTRACT

Precise automated video search is gaining in importance as the amount of multimedia information is increasing at exponential rates. One of the drawbacks that make video retrieval difficult is the lack of available semantics. In this paper, we propose to supplement the semantic knowledge for retrieval by providing useful semantic clusters derived from event entities present in the news video. These entities include the output from keywords derived from the automated speech recognition (ASR) and event-related High-level Features (HLF) extracted from the news video at the pseudo story level. Fuzzy clustering is then carried out to group similar stories together to form semantic clusters. The retrieval system utilizes these clusters to refine the re-ranking process in the Pseudo Relevance Feedback (PRF) step. Initial experiments performed on video search task using the TRECVID 2005 dataset show that the proposed approach can improve the search performance significantly.

Categories and Subject Descriptors

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

General Terms

Design, Experimentation

Keywords

Video Retrieval, Event-based Clustering

1. INTRODUCTION

Efficient automated multimedia information retrieval (MIR) is becoming increasingly important especially with the up surging amount of multimedia data. Users increasingly desire to search on complex semantic queries such as: "Show me the video clips depicting George Bush entering a car" or "Find shots of buildings covered in flood water" used in TRECVID 2005 [1]. However, current state-of-the-art video retrieval systems still face many challenges due to the lack of valuable semantics. To alleviate this problem, we propose the mining of semantic clusters by performing an unsupervised clustering on the multimodal event entities present in news video. This is similar to Topic Detection and Tracking (TDT) [2] which has been well researched in the pure text document domain. The leverage of topic/event structures from news video TDT will provide excellent partial semantic toward many queries as news videos are generally the multimedia depictions of events. Moreover, given the clusters of similar stories, retrieval can be carried out more effectively as relevant shots for semantic queries tend to be semantically similar.

MM'06, October 23--27, 2006, Santa Barbara, California, USA.

However, due to errors such as sentence boundary misalignment, missing and misrecognized words and language translation mismatch in ASR, the use of only textual information from ASR is not sufficient to obtain good semantic clustering. [3] shows that these errors can significantly degrade the performance of clustering. It is therefore necessary to supplement the erroneous ASR with additional features in order to obtain a clustering result that is more representative and descriptive of the underlying distribution of news events/stories. In this paper, we integrate event-related high-level features (HLFs) to provide the additional context and knowledge about the events in news story. The event-related HLFs are relevant visual features that can contribute to describing the event or what is happening in the shot. For example: if there are shots containing the HLF "fire", it could strongly indicate stories on topics like "forest fire", "fire breakout", "explosion", etc. However, it is also known that not every HLF present in the news video can be helpful. For example, the detected high-level semantic concepts of "*walking people*" or "*moving cars*" in a financial news report are definitely not useful. Therefore, it is necessary to analyze the structure of news video to determine which HLF are relevant to the news story. We heuristically compute the importance of HLF to news video based on factors such as: a) the detection confidence of the HLF, and b) the type of video news in which the HLF appears. With the selected HLFs, we can then derive and utilize the visual and lexical similarity measures between news stories by computing the implicit semantic relationships and similarities between HLFs.

The feature space of news events is constructed by concatenating the features of event keywords from ASR and the HLFs. An unsupervised clustering method is employed to discover the topical and event structures of the various news video stories. The fuzzy clustering is utilized to cluster the news stories into groups associated with membership values to their member news events. This characteristic of fuzzy clustering can address the phenomenon that news event clusters can be overlapping and not exclusive. In the implementation, we adopt the fuzzy c-means technique [4] to perform the news event mining. We observe large improvement in MAP when these semantic clusters are integrated into our previous video search system [5] with Pseudo Relevance Feedback (PRF).

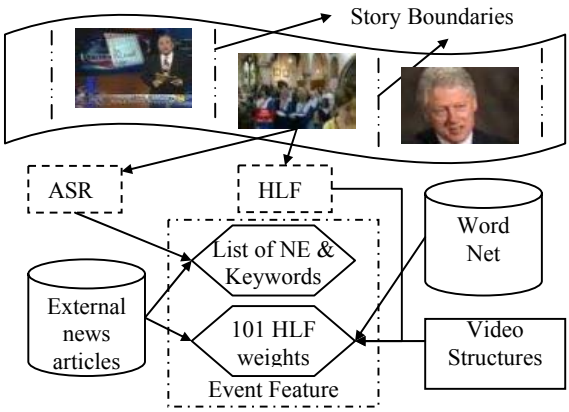

Figure 1: Extracting features for news event clustering

2. NEWS VIDEO EVENT FEATURES

The news video event feature space is the combination of event keywords derived from ASR and event related HLFs extracted from the multimodal analysis of news video stories. Figure 1 describes the framework for the integration of HLFs and ASR as event features for clustering.

2.1 Video Story Boundary Detection

Prior to constructing the feature space, the first task is to define a suitable basic unit for event representation as news video is continuous. Even though shot boundary detection rates are excellent, we choose to use pseudo story segments as they provide more coherent textual and visual semantics. In our implementation, we make use of the story boundary detection result provided by IBM-Columbia [6]. We enhance the segmentation output by utilizing anchor-person shots for second level segmentation [7]. 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 Event Features from Automated Speech Recognition (ASR)

The recognition errors from ASR and the translation errors from Machine Translation (MT) can significantly degrade the performance of semantic clustering. Therefore, we choose to utilize ASR transcripts using the bag-of-word assumption. The bag-of-word assumption allows us to treat each news story as a holistic unit, enabling us to avoid some ASR issues such as the sentence boundary, sentence truncation and grammatical/word alignment errors caused by machine translation.

We extract the news event feature from ASR using the following steps. First, we obtain the corresponding ASR for each story based on the video story boundaries generated above. Second, we extract the important keywords and Name Entities (NE) in the ASR transcripts to form the event entities representing the video segment. These keywords and NEs include people, organization, location, time of the event, and a list of predefined key terms in the event story. We make use of the rule-based extraction technique used in [8] to automatically extract these terms. Third, we employ relevant external news articles from the same period to add highly correlated NEs to the feature vector of the news story. These additional NEs are selected based on the Mutual Information (MI) overlap between the person, location and time entities extracted from the news video story and the online news [9]. [9] shows that the locations and time NEs in ASR are seldom misrecognized or wrongly translated even for spoken documents; however, the person names are more vulnerable to errors as they are non-vocabularies. The use of NE list harvested from parallel news articles has been shown to be effective in recovering missing person names in the ASR transcripts.

2.3 Event Features from High-level Features

To supplement the ASR text, we utilize the 101 high-level features extracted from news video shots by MediaMilll [10]. These HLFs are detected using a machine learning approach, where each HLF detector is trained against an annotated corpus of video clips. The mean average precision (MAP) performance of HLF detection is about 0.3 as reported in TRECVID 2005. The high-level features from MediaMilll [10] can be classified into the following categories: objects (such as *aircraft*, *boat*, and *car*), people (such as *crowd*, *female*, *police*, and *prisoners*), places (such as *beach*, *government building*, and *houses*), scenes/events (*horse racing*, *natural disaster*, *people marching*) and actions (*walking & running*, *cycling*).

2.3.1 Selection and Weighting of Event-related HLF

As a news story may consist of a number of detected HLFs, we first determine which HLFs are event-related by analyzing the structure and grammar of news videos. In a standard news video setting, the script writers construct the speech transcripts for the anchor-person to narrate based on the "real-life" scenes or shots [11]. The video speech provides the explanation or content of the story while the visual counterpart displays the real-life events. Very often, besides these real-life events shots, additional shots (such as led-in/out, special graphic effects, etc.) are padded to the news video. Therefore, it is necessary to identify the shots which are not related to the news event. We make use of the shot-genre and shot-type detector [12] [7] to isolate various redundant shot-genre like led-in, led-out, commercials, anchor person, special effects shots, etc in the news segment. The remaining shots are used as representative shots of the video event. A limitation is imposed on the number of continuous live-reporting series used. For example, if there are more than 2 live-reporting series in the video story, only the first 2 series will be selected.

HLFs also have different level of significance in the video. 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. It is important to determine which categories of HLF play the leading roles for a particular shot-type. For example, the place and scene HLFs are the crucial elements to disaster-type news. Table 1 illustrates the association of various shot-types and the categories of HLFs. Only HLFs relevant to the particular type of news story will be considered.

Table 1: List of Shot-type and their important HLFs

Shot-type	**Categories of related HLF**
General (Political, Science, etc)	People, places, scenes/events
Unclassified	Objects
Financial	People, scenes/events
Weather	-
Disaster	scenes/events, places
Sports	People, actions

Since the same HLF may be detected in multiple shots in a news video segment with varying confidence, we take the highest detection score as the feature value of that HLF.

2.3.2 Inference of Similarity from HLF Semantics

The semantic relation (both lexical and visual) between the HLFs can also be leveraged to provide additional similarity measures during clustering. For example: the concept *police* and *prisoners* and *military* may be individual concepts but it is clear that there are intrinsic relationships among them. Likewise, it is also easy to see a shot that contain concept *waterscape_waterfront* is more probable to contain a *ship_boat* rather than a *car*. In this work, we use the high confidence HLF_h (detection confidence >= θ) to re-adjust the lower confidence HLF_l (detection confidence < θ) by considering their correlation with respect to time. In our previous work [12], a time-dependent "query to HLF" relationship is obtained by using WordNet [13] and relevant external news articles. We adopt the same principle to compute the semantic similarities between HLFs by using Eqn (1).

$$Sim_t(HLF_a, HLF_b) = \alpha Lex(HLF_a, HLF_b) + (1-\alpha) MI(HLF_a, HLF_b \mid t) \quad (1)$$

where $Lex(HLF_a, HLF_b) = \sum_{t_q \in HLF_a} \sum_{t_p \in HLF_b} \mathrm{Re}\, snik(t_q, t_p) / (|\, HLF_a \,| * |\, HLF_b \,|)$

and $MI(HLF_a, HLF_b|t)$ is the mutual information (MI) between them during the time period *t*. Specifically, the formula computes similarity based on the gloss as well as the hypernym/hyponym hierarchy from WordNet. The gloss supplements WordNet in the way that it sometimes provides visual information about an object – its shape, color, nature and texture; whereas the WordNet only provides direct relations (e.g., *aircraft* & *airplane*; *fire* & *explosion*). For example, the word *boat* is not related to *water* by virtue of any relationship link in WordNet, but is by its gloss – *"a small vessel for travel on water"*. The output of HLF similarity computation using Eqn 1 is the cross-matrix containing all the similarity scores between 2 HLFs with respect to date. For the lower confidence HLF_l, if its similarity score to a high confidence HLF_h in the same video story is greater than the predefined threshold β, the feature value of this HLF_l will be adjusted by taking the average value of the detection confidence of both HLF_l and the corresponding HLF_h. For example: if the value of HLF_h is 0.7 and HLF_l is 0.3, the feature value of HLF_l will be adjusted to (0.3+0.7)/2, which is 0.5. Such an arrangement will allow us to propagate weights from higher confidence HLFs to lower confidence but correlated HLFs in the same news story.

3. FUZZY SEMANTIC CLUSTERING

After constructing the feature space by fusing the features from keywords and NEs from ASR and HLFs at the story level, we perform the clustering of news stories belonging to similar events into homogeneous classes. The clustering can be either hard (crisp partition) like k-means clustering or fuzzy (soft partition with memberships) like fuzzy c-means clustering [14]. We employ fuzzy clustering as it allows the instances to be clustered into multiple classes/clusters, which makes it a natural match to the news event clustering task. This is because in news story clustering, a news story may belong to multiple story classes. For example, as Figure 2 shows, the news story of German train crash can be classified into both classes of *'Terrorism'* and *'Accident'* with different membership grades.

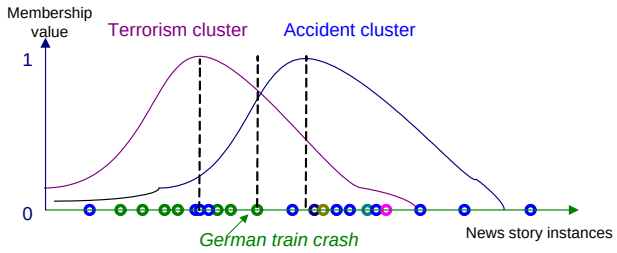

Figure 2: An example of multiple classes of news stories with different membership values.

There are a number of well-established objective function based fuzzy clustering techniques, like fuzzy c-means (FCM), Gustafson-Kessel algorithm (GK), and fuzzy c-varieties algorithm (FCV) [15]. In our implementation, we adopt the fuzzy c-means algorithm. The fuzzy c-means clustering partitions $X=\{x_1,.., x_i, ...,x_N\} \subset \chi$ into clusters $P=\{p_1, p_2, ...,p_C\}$ by means of a membership matrix $U = [u_{ij}] \subset [0,1]^{|P|\times|X|}$, where x_i is the news story instance *i* containing the ASR and HLF feature values, *N* is the number of news video stories, *C* is the number of clusters and u_{ij} is the membership in which x_i belong to cluster p_j. The fuzzy partition is carried out through an iterative updating of membership matrix *U* and cluster centroids c_j by minimizing a objective function *J*. The detailed procedures are defined as below:

Step 1: Randomly initialize the membership matrix *U* that has constraints $\sum_{j=1}^{C} u_{ij} = 1, \forall i = 1,..., N$

Step 2: Calculate centroids $c_j = \frac{\sum_{i=1}^{N} u_{ij}^{\ m} x_i}{\sum_{i=1}^{N} u_{ij}^{\ m}}$, where m ϵ [1,∞] is a weighting exponent.

Step 3: Compute the objective function (dissimilarity) between the cluster centroids and the news story instance using $J(U, c_1, c_2, ..., c_C) = \sum_{j=1}^{C} J_j = \sum_{j=1}^{C} \sum_{i=1}^{N} u_{ij}^{\ m} d_{ij}^{\ 2}$, where d_{ij} is the distance between centroid c_j and news story instance x_i. Stop the algorithms, if *error_function* < ε, where *error_function*= $\| U(n+1) - U(n) \|$, ε is the termination threshold between 0 to 1, and $n+1$ is the current step number.

Step 4: Compute new membership value $u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}}$.

Go to Step 2.

In order to improve the performance of clustering, FCM is run several times with different initial centroids.

4. REFINEMENT OF VIDEO SEARCH

With the semantic clusters obtained above, we aim to improve the performance of our previous video search system [5] by incorporating the cluster information into the re-ranking of result shots in the Pseudo Relevance Feedback (PRF) step. The PRF process is done based on the top 15 relevant shots retrieved. From previous evaluation results in TRECVID 2005, 14 out of the 24 queries have more than 8 correct shots in these top 15 retrieved shots, which suggest relatively accurate top-retrieved shots. The PRF in [5] makes use of the features weights in the top 15 shots to

perform the re-ranking by re-adjusting the linear fusion function. In this paper, we add an additional dimension to the linear fusion function at shot level to compute the scores of the retrieved shots by considering the event semantic clusters. The shot score for hard clusters (k-means) is defined in Eqn (2) and fuzzy clusters (fuzzy c-means used) in Eqn (3).

$$Score(s_{ik} \in CL_j) = Score(CL_j) = \frac{count(CL_j \cap Top)}{15} \quad (2)$$

$$Score(s_{ik} \in CL_j) = \max\left(\frac{\sum u_{ij}, \forall_{i:x_i \in (CL_j \cap Top)}}{15} \times u_{ij}\right), j = 1..C \quad (3)$$

where s_{ik} is the k^{th} shot of news story i, CL_j is the cluster j containing story i, and u_{ij} is the membership value of story i belonging to C_j where *Top* are the pseudo stories containing the top 15 returned shots. This cluster score will be the representative score for all the $shot_i$ in the $cluster_n$.

5. EXPERIMENTS AND DISCUSSIONS

We carried out a series of experiments on TRECVID 2005 dataset to assess the improvement brought by the fuzzy semantic clusters. The automated search task in TRECVID 2005 consists of 24 queries directed at finding people, scenes, objects and actions in 80 hours of multilingual news video. A maximum of 1,000 shots are returned for each query and the performance is measured in terms of mean average precision (MAP). Five runs of experiments are conducted, as illustrated below:

- *Run0: Best run from NUS [5] with standard PRF (baseline)*
- *Run1: PRF with hard clusters using only ASR*
- *Run2: PRF with fuzzy clusters using only ASR*
- *Run3: PRF with hard clusters using only ASR and HLF*
- *Run4: PRF with fuzzy clusters using only ASR and HLF.*

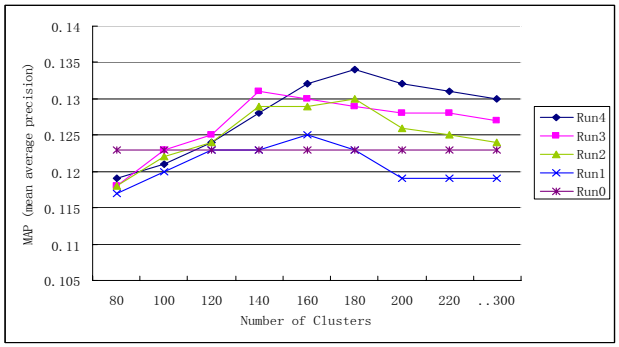

Figure 3: MAP performance against number of clusters

These experiments are repeated with varying number of clusters n. Figure 3 display the MAP performance of each run against the number of clusters. The baseline Run0 has a best MAP of 0.123. Run1 which only uses ASR has almost no improvement or performed worse except when using optimal number of clusters. This shows that the semantic clusters based on ASR-alone and hard clustering is not effective. Run3 and Run4 which uses the combination of HLF and ASR generally perform better than Run1 and Run2. This suggests that the appropriate use of HLF is helpful in semantic clustering of stories. The results also indicate a preference of fuzzy cluster over hard clusters. In particular, the best performing run Run4 which uses both ASR and HLF based on fuzzy clusters is able to achieve a best MAP of 0.134 based on a cluster size of 180. This improvement is significant over the baseline Run0. We also observe a trend that the hard-cluster runs (Run1 and Run3) achieve the best performance with a lower number of clusters (approx 140) while the fuzzy clusters peak at about 180 clusters. This could be due to the benefit of having membership score as they allow more flexibility. It is also clear that the number of clusters plays an important role in the effectiveness of clustering.

6. CONCLUSIONS

This paper presented a news video search framework that incorporates event-based fuzzy clustering. Our contribution is two-fold. First, we integrated high-level features (HLF) into the traditional text clustering approach. Second, we explored the use of an unsupervised fuzzy clustering model to provide semantic clusters to support news video search. Our initial results show that the proposed approach improves the search performance significantly.

This work is still in the initial stage. Future works are expected to address how to utilize highly relevant and confident HLFs to train event classifiers for effective event TDT. The results of this event TDT will be able to resolve more complex search queries in a robust and flexible way.

7. REFERENCES

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

[2] J Allan, R Papka and V Lavrenko: "On-Line New Event Detection and Tracking". SIGIR 1998: 37-45.

[3] J. McCarley, M. Franz. "Influence of speech recognition errors on topic detection". SIGIR 2000, pages 342-344, New York, NY, USA, 2000.

[4] J. C. Bezdek: "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York, 1981

[5] T. Chua, S. Neo, H. Goh, M. Zhao, Y. Xiao, G. Wang "TRECVID 2005 Search Task by NUS PRIS" In TRECVID 2005, NIST, Gaithersburg, Maryland, USA, 14-15 Nov 2005.

[6] 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, New York 2005.

[7] L. Chaisorn, T.-S Chua and C.-H. Lee. "The segmentation of news video into story units." ICME '02, Ischia, Italy, Jul 02

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

[9] S. Neo, H. Goh, T. Chua "Multimodal Event-based Model for Retrieval of Multi-Lingual News Video" In IWAIT 2006, Okinawa, Japan, 9-10 Jan 2006.

[10] C.G.M. Snoek, J. Gemert, J.M Geusebroek,., B. Huurnink, D.C. Koelma, G.P. Nguyen, O. de Rooij, F.J. Seinstra, A.W.M. Smeulders, C.J Veenman, M. Worring, "The mediamill trecvid 2005 semantic video search engine", TRECVID Workshop, NIST (2005)

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

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

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

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