BJTU TRECVID 2006 Video Retrieval System

Shikui Wei, Yao Zhao, Zhenfeng Zhu, Nan Liu , Yufeng Zhao, Liang Zhang, Fang Wang
Institute of Information Science, Beijing Jiaotong University, Beijing, 100044, China
E-mail: shkwei@gmail.com

ABSTRACT

In this paper, we describe our experiments of search task for TRECVID 2006. This year we participated in the automatic video search subtask and the interactive video search subtask, and submitted two runs to NIST, one for the automatic search (runid: F_A_1_JTU_FA_1_1), one for the interactive search(runid: I_B_2_JTU_I_1_1). The paper presents the approaches as well as the results analysis in details. In an effort to obtain an initial answer set and to provide an entrance for user interaction, a fully automatic video search engine is developed, which depends mainly on the transcribed speech text and some natural language processing techniques. After that, a Co-EM based interactive search scheme, which utilizes unlabeled data by explicitly splitting the feature space into two approximately independent views, is proposed to mitigate the burden and enhance the search performance at the same time. In automatic video search task, we were able to achieve the median or better performance for most of the topics and even achieve the best average precision for one topic. For the interactive search task, our approach gives high scores to the true relevant results by only labeling a very small set of positives.

1. INTRODUCTION

In video retrieval field, the key issue is how to bridge the semantic gap between low level feature and high level concept. As a kind of solutions, the interactive video search techniques can alleviate this problem to some extend by real-time intervention of users during the search process. In recent years, more and more researchers and organization focus on the research of the interactive video search. With development of research, a variety of interactive search schemes [1, 2,14,15,16,21,23] were presented. In fact, most interactive search schemes construct classifiers directly from unimodal features by using only labeled data, and then combine multiple unimodal classification scores into a final multimodal score. The faults of this kind of methods are that those unimodal classifiers can't contribute to each other during the process of training and require much more labeled samples.

To address above problems, we proposed a Co-EM based interactive search scheme in this paper. The virtue of proposed approach is its ability to automatically find positive examples from past answer set unlabeled. Moreover, both classifiers can contribute to each other during the training phase.

This paper is organized as follows. First, a fully automatic video search engine is presented to provide a search entrance for users in section 2. And then, proposed interactive video search system and the CO-EM algorithm are

discussed in section 3. In section 4, the experiment setting and results are addressed. Finally, we give the conclusion and discuss our future work in section 5.

2. AUTOMATIC VIDEO SEARCH ENGINE

In advance of the interactive search, an initial answer set must be obtained to provide entrance for user interaction. In fact, there isn't a generally accepted method to start the video search. As a kind of effective methods, example content-based retrieval approaches, are generally used in many video retrieval systems [15,18,22]. But, in a real world search process, users usually don't have some right video examples handy to formulate query intention, thus it is unreasonable to expect users to provide query examples for search system as stated in [1, 21]. In contrast to visual-based approach, text-based search approaches [1,12, 14,15,16,21,22] are more straightforward and natural manner for user to start a search task. After given a query topic text in everyday English by users, the text-based system can return a series of relevant video sequences. In this paper, for the purpose of obtaining an initial answer set for interaction, a text-based search engine is constructed in advance. The following two figures, figure 1 and figure 2, show the sketch of index building and text retrieval. We will discuss each part in detail.

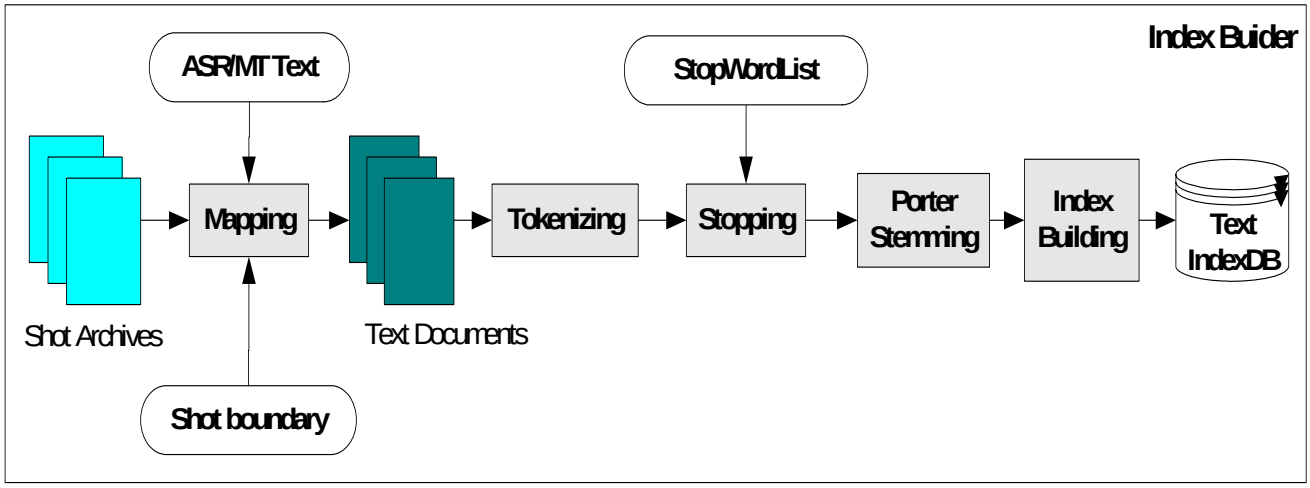

Fig1. A sketch of indexing speech transcripts of shots

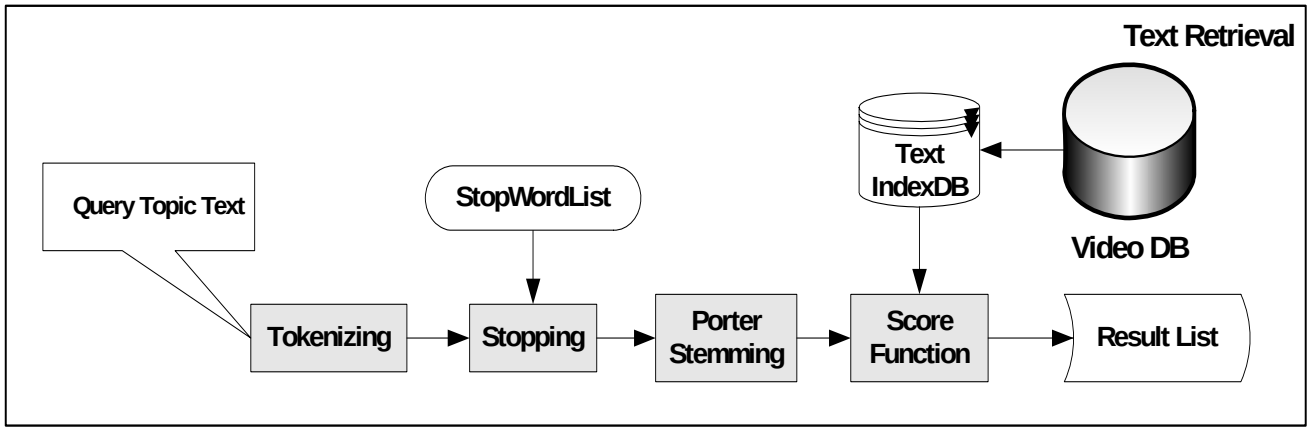

Fig2. A framework of text-based video retrieval system

2.1. Index Building

In this system, the text information related to video is first extracted from the audio track of video by using the automatic speech recognition (ASR) techniques. Furthermore, non-English sources are translated into English text by machine translation approaches. In our experiments, the video corpus together with ASR/MT text is provided by NIST. For building text index database of shots, the speech transcript is first mapped to corresponding shot based on the shot boundary information provided by NIST. And then, each text string corresponding to one shot is tokenized, and then a keyword list is formed. After that, functional words such as “the”, which are not meaningful for video shot, are removed out against a stopword list of total 460 terms. In addition, in the information retrieval application, the essence is the comparison between query terms and index terms (important words or phrases in candidate documents). Unfortunately, the words occurring in query text and in documents perhaps have similar semantic interpretations but morphological variants. For instance, pairs of terms like “computation” and “computing” are of similar semantic interpretations but different morphological manners, the system will consider them as completely different words. To address the problem, stemming techniques are used to process each word by reducing a word to its stem or root form. Here port stemming [20], a well-known stemming technique, is employed to process the keyword list in our experiments. After a series of processing by using natural language processing (NLP) techniques, an index builder is exploited to construct terms representation and document representation for each shot, and then an inverted index is built.

2.2. Text-based video retrieval

The basis idea of text-based video retrieval is that the system returns a set of relevant video shots after the user specified a query topic text. As shown in figure 2, the query string is processed with the same manner as indexing procedure. The key issue in text retrieval is the score function, which compares query representation with each shot representation in the index database and gives a similarity score. KL-divergence retrieval model, an extension of the query-likelihood approach [24], is used to score candidate shots in this scheme. In essence, it scores a candidate shot document by computing the KL-divergence between the query representation and the document representation.

2.3 Implementation

A powerful program package named Lemur toolkit [6] is used to build the text-based search engine. The Lemur toolkit implements a series of language and retrieval models so as to facilitate research in language modeling and information retrieval. The NLP techniques used above also are implemented in the toolkit. In fact, our text-based video retrieval system is based on the package.

3. CO-EM BASED INTERACTIVE SCHEME

The aim of our method is to alleviate the burden on user and more effectively learn user’s query intention. The system architecture of our approach includes two main components, one for selecting optimal parameter settings for each proto-topic and one for interactive search based on Co-EM algorithm. The general framework of our proposed scheme is illustrated in figure 3. We will describe each component in more details.

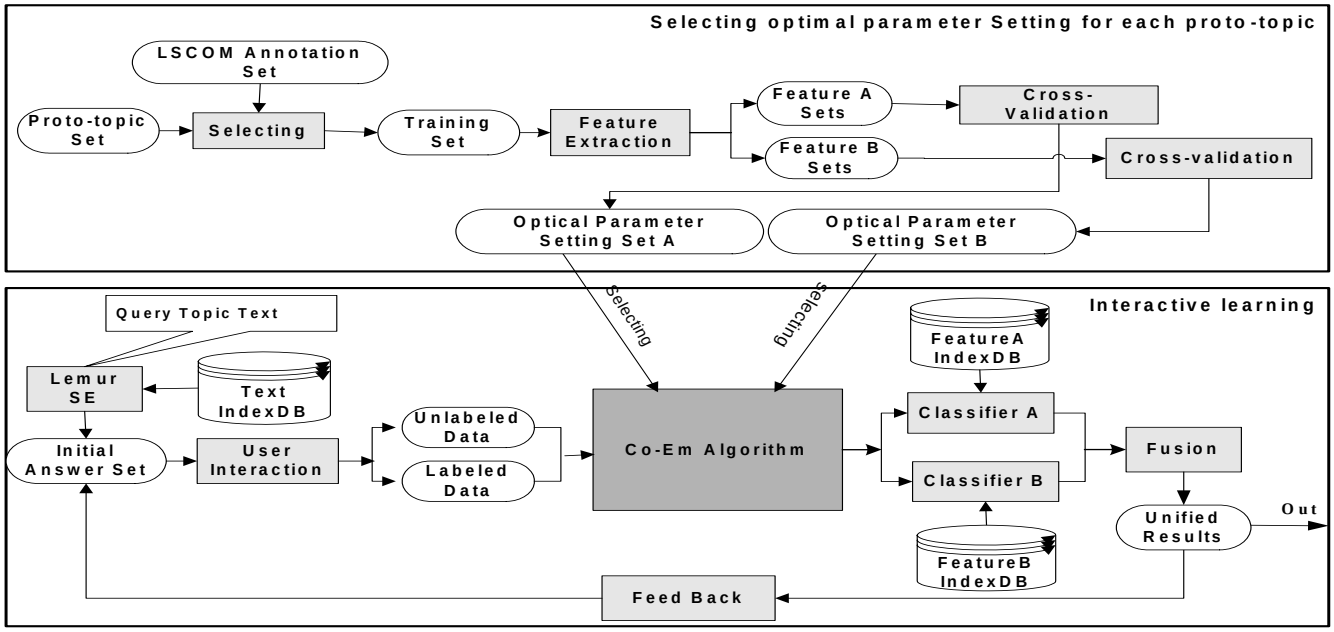

Fig.3 The framework of system architecture for interactive video search, which contains parameters selecting component and interactive learning component.

3.1. Interactive Learning

The core of the interactive learning in this scheme is the Co-EM algorithm. Co-EM algorithm [8, 9] is a hybrid method combining multi-view learning with probabilistic EM approach. As a powerful manner exploiting unlabeled data, multi-view algorithms leverage unlabeled data by explicitly splitting available features into independent and compatible multiple views. Nigam.K et.al demonstrated that multi-view algorithms outperform single-view methods [8] when the feature space can be divided into independent and compatible multiple views. Table 1 shows the flowchart of the Co-EM algorithm.

Table.1 The Flowchart of Co-EM Algorithm

Input a small set of labeled data (D_l) and a large number of unlabeled data (D_u)

Co-EM algorithm

1. Build classifier A using feature split A of D_l
2. Label D_u probabilistically from view A using classifier A.
3. Loop until the number of iterations is met
   a) Build Classifier B using D_u with Classifier A's labels and D_l, and then probabilistically relabel D_u from view B
   b) Build Classifier A using D_u with classifier B's labels and D_l, and then probabilistically relabel D_u from view A

Output Two final classifiers, A and B, for predicting the probability estimates of shots in library

---

Multi-view learning has been successfully applied to many application domains, such as web page classification, named entity detection, etc. The co-training algorithm [7], a well-known multi-view learning algorithm, is first introduced into the domain of video concept detection by Yan.R et.al [3]. The experiments show that multi-view algorithms work well for classifying multimedia information.

Compared to the co-training method [7], the Co-EM algorithm iteratively uses the unlabeled data, this is, it probabilistically labels all the data for each round, instead of labeling a few samples per round. According to the analysis of K. Nigam et al [8], the Co-EM algorithm performs better than the Co-training in many applications.

After initial answer set is obtained, the user labels a small set of positive examples from them, and leaves the other results as unlabeled set. Together with above data, negative examples randomly selected from library are sent to the Co-EM classifier.

3.2 Feature Extraction

In video search, shots are referred as the final unit needed to be searched, for which some feature combinations are generally considered to make a characterization. But in our scheme, each shot is represented using two approximately independent feature views, one is the visual information of key frame, and one is the corresponding ASR/MT text. The feature extraction of both views is based on the shot. We will not elaborate here on the feature extraction approaches, which are also challenging and well-studied task in the domain of multimedia search. Rather, we exploit the color histogram as the visual descriptor or feature A, and a text vector as the textual descriptor or feature B.

Concerning visual descriptor, we employ a color histogram with 36 dimensions in HSV space for each keyframe, which represents the global color distribution of the keyframe. The more detailed description is presented in cited reference [19].

To obtain text descriptor, we adopt the similar manner as introduced in [1] which forms a vector feature for visual information, to construct text vector feature. Figure 4 illustrates the procedure of text feature extraction.

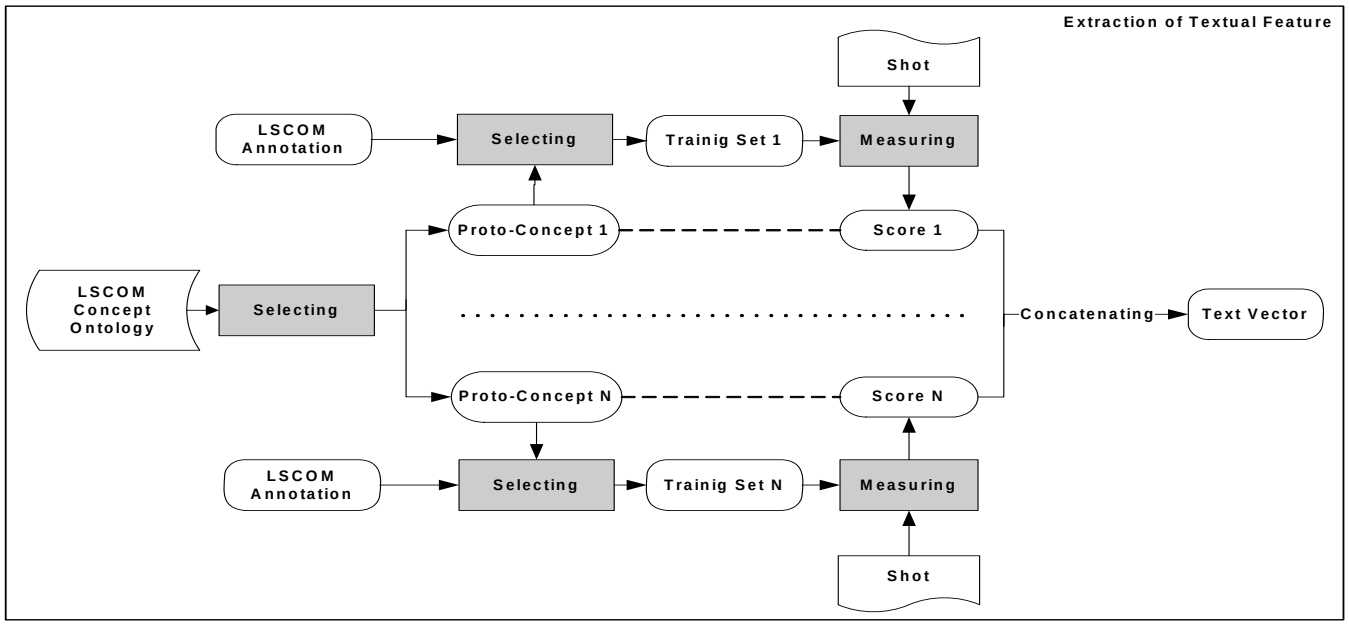

Fig.4 The scheme of textual feature extraction

In particular, we first select 78 proto-concepts from concept ontology of LSCOM [4]. And then a training set of 40 shots with corresponding speech transcript text is chosen for each proto-concept against the annotation ground truth [4].

Finally, for each shot a 78-D text vector can be constructed by measuring similarity between the shot and each training set. The similarity measurement, here, is the same score function used in text-base retrieval system.

3.3. Optimal Parameters Selection

In this scheme, Support Vector Machine (SVM) with RBF kernel function is employed as underlying learner for Co-EM learning. As an eminent classification tool, SVM has been proven to be a powerful way to tackle unbalanced data and learn from a small set of labeled samples. In our experiments, we utilize an extended SVM classifier provided in LIBSVM [11] to predict class probability information of samples, instead of only labels.

In fact, the parameter setting for the SVM significantly affects the classification performance of video information [1]. Generally it is difficult to know in advance which setting is optimal for a specified query topic. To focus on the problem, a subset from query topics used in TRECVID search task is selected as the proto-topic set, for which a training set is formed against LSCOM Annotation [4]. Moreover, two feature sets from training samples, A and B, can be extracted by the way discussed above. Finally, from each view, optimal parameter settings for every proto-topic are searched by using cross-validation and grid-search against training set, and two optimal parameter setting sets are then built. When a new query topic comes in, it is first mapped to one of proto-topic, and then two parameter settings corresponding to this proto-topic are chosen as the optimal parameter settings of the new topic for training Co-EM classifier. However, the idea is similar in spirit to [12], which performs similar fusion strategy for the same query class by clustering queries in performance space. But what the difference in our scheme is that the obtaining of optimal parameters for each query class is conducted in feature space.

3.3. Search Results Fusion

In the fusion phase, the search results r from two view classifiers are integrated into a unified and ranked result list. As two main fusion strategies, the early fusion and the late fusion are adopted for integrating multiple modalities in most video indexing systems [1, 2, 3, 15, 16,22]. In fact, the main difference between two integration strategies is how to construct the patterns used by classifiers from multiple modalities. The algorithms, which rely on early fusion, first extract feature pattern from each channel of modalities, and then concatenate feature patterns of all channels to form a unified feature pattern. In this manner, a single classifier is enough to carry out the task of video indexing. Compared to early fusion method, late fusion methods build a separate classifier for each feature pattern extracted from all modalities, instead of combining all features into a single pattern. Usually, the late fusion is less efficient due to the construction of a separate learner for each modality and have potential loss of correlation along modalities; however, for early fusion, although this manner avoid the faults of late fusion, it is difficult to construct the unified multimodal representation [1]. Generally speaking, the late fusion includes score normalization and score aggregation [16,18]. For score normalization, some common methods, such as the statistical normalization, can be utilized. For score aggregation, the general methods include weighted-average score aggregation method, non-weighted score fusion method. In fact, the latest fusion strategies [12,17] are considered from the view of query classes, namely query-class-dependent models and query-class-independent models. L. S. Kennedy et al [12] discussed this topic in more detail. In this paper, we don't focus on the selection of fusion strategies. Rather, a simple linear average weighted score is used to integrate the search results in our experiments, which is defined as follow:

$$F_s(r) = \sum_{i=1}^{M} \omega_i * s_i \quad (1)$$

where M is the number of classifiers of different views, s_i are the probability estimates of r being relevant, ω_i are the weights on classifiers. However, fusion strategy will impact significantly on the fusion result for different queries, we leave it for the feature work.

4. EXPERIMENTS

To construct the experiments and evaluate the performance of proposed algorithm, a text-based interactive video search system is developed. Figure 7 shows the interface of the system.

In our experiment, the TRECVID'05 development data set with annotation information is employed to build training set for searching optimal SVM parameters. In addition, the test set for TRECVID'06 is adopted to answer the query topic and evaluate the search performance.

The following two runs were submitted to NIST this year:

Runid: F_A_1_JTU_FA_1_1, fully automatic search depending mainly on the transcribed speech text and some natural language processing techniques

Runid: I_B_2_JTU_I_1_1, interactive search using textual vector, some visual features, and Co-EM algorithm.

For automatic text-based video retrieval, we exploited some natural language processing techniques with only the transcribed speech text. Figure 5 shows the performance relative to the median and the best scores. For most of the topics, the scores are not less than the median value, and even one topic achieves the best average precision. Considering that many other automatic systems combined multimodal features, this result is significant.

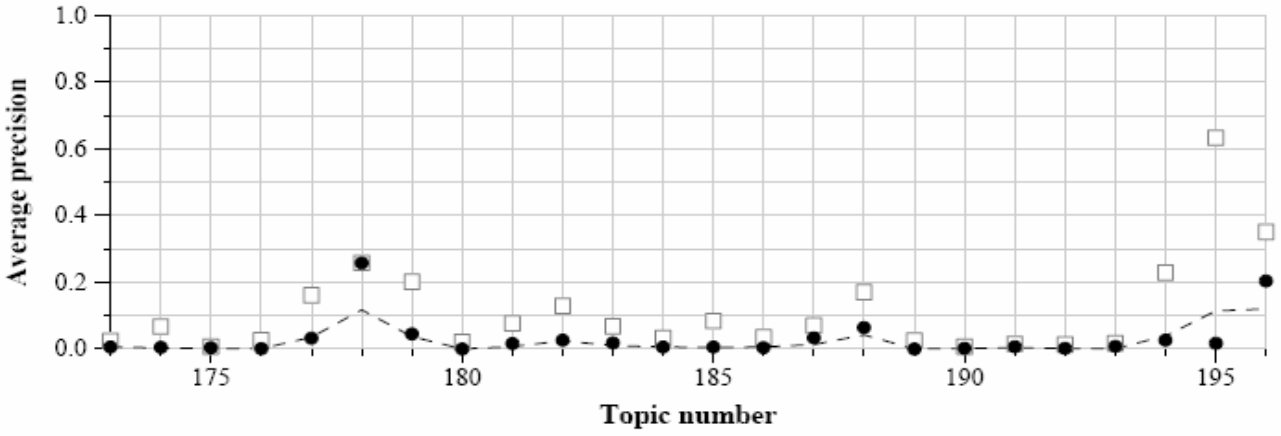

Fig. 5 Run score (dot) versus median (---) versus best (box) by topic

In interactive video search task, our aim is to develop an algorithm which can effectively alleviate the burden by labeling a small number of positives and give high scores to the true relevant results. Here we use the precision at X to evaluate the effectiveness of this interactive scheme. Note that the precision at X is the average precision at X over all of the topics. We carried out the proposed Co-EM based interactive system by labeling only 5 positive examples, which is a quite small set. As shown in Figure 6 , the average precision of interactive search is far higher than automatic search in the top 200 returned results, which indicates that the proposed approach do give high scores to the true relevant results.

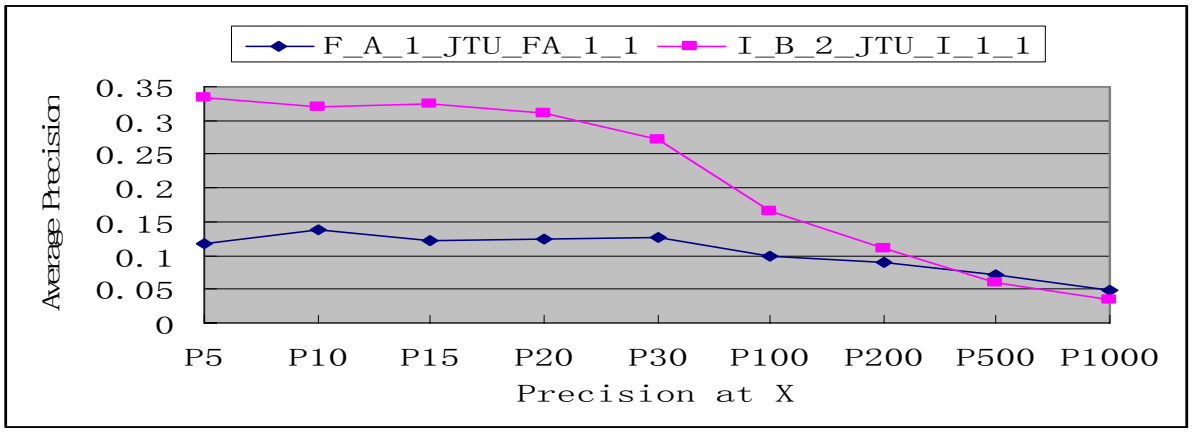

Fig.6 Interactive search average precision Vs. automatic search average precision at X returned documents

4. CONCLUSIONS AND FUTURE WORKS

In this paper, we developed an interactive video search scheme based on Co-EM algorithm with SVM as underlying learner. This Scheme utilizes unlabeled data by explicitly splitting the feature space into two approximately independent views. The virtue of this approach is its ability to automatically find positives from past unlabeled answer set. In addition, both learners can contribute to each other by using the label information from another view, which indicates that multiple modalities are potentially combined during the training phase. We evaluate our approach against the TRECVID'06 benchmark. The experimental results show that our scheme works better than single-view algorithms and reduces a need for labeled data. In future work, we will develop some methods to more effectively fuse the result list from two classifiers of different views.

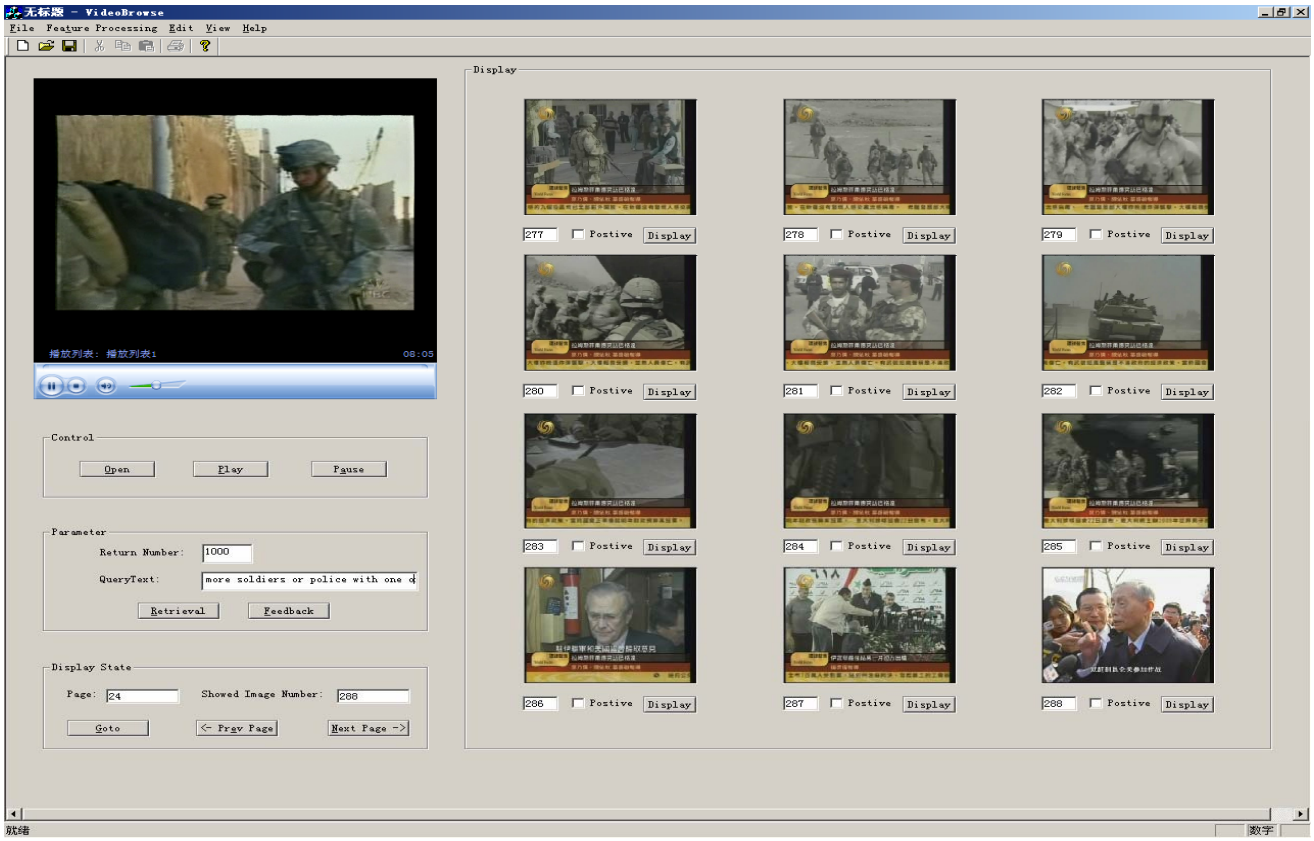

Fig.7. The interface of proposed interactive search system

REFERENCES

[1] C.G.M. Snoek et al., "The MediaMill TRECVID 2005 Semantic Video Search Engine," In *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID, Gaithersburg, USA, 2005.

[2] M.Y. Chen, M. Christel, A. Hauptmann, H. Wactlar, "Putting Active Learning into Multimedia Applications: Dynamic Definition and Refinement of Concept Classifiers," In *International Conference on Multimedia*, ACM, Singapore, pp.902-911, 2005.

[3] R. Yan, M. Naphade, "Multi-Modal Video Concept Extraction Using Co-Training," In *International Conference on Multimedia and Expo*, IEEE, pp. 514-517, 2005.

[4] LSCOM Lexicon Definitions and Annotations Version1.0, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #217-2006-3, March 2006.

[5] C. Petersohn. "Fraunhofer HHI at TRECVID 2004: Shot Boundary Detection System", In *TREC Video Retrieval Evaluation OnlineProceedings*, TRECVID, 2004, URL: http://www.nlpir.nist.gov/projects/tvpubs/tvpapers04/fraunhofer.pdf

[6] The Lemur Toolkit for Language Modeling and Information Retrieval, URL:http://www.lemurproject.org.

[7] A. Blum, T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," In *Proceedings of the Workshop on Computational Learning Theory*, ACM, New York, USA, pp. 92-100, 1998.

[8] K. Nigam, R. Ghani, "Understanding the Behavior of Co-training," In *Proceedings of the Workshop on Text Mining*, ACM, 2000.

[9] U. Brefeld, T. Scheffer, "Co-EM Support Vector Learning," In *Proceedings of the twenty-first International Conference on Machine learning*, Canada, 2004.

[10] V. Vapnik, *"The Nature of Statistical Learning Theory,"* Tsinghua University Press, Chinese Language Edition, 2000.

[11] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," 2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[12] L.S Kennedy, A. Natsev, S.F. Chang, "Automatic Discovery of Query-Class-Dependent Models for Multimodal Search," In *International Conference on Multimedia*, ACM, Singapore, pp.882-891, 2005.

[13] C. G. M. Snoek, M. Worring, "Multimodal Video Indexing: A Review of the State-of-the- art," In *Multimedia Tools and Applications*, 2005 Springer Science + Business Media, Netherlands, pp. 5-35,2005.

[14] D.S. Zhang, J. F. Nunamaker, "A Natural Language Approach to Content-Based Video Indexing and Retrieval for Interactive E-Learing," In IEEE *Transaction on Multimedia*, vol. 6, no. 3, pp.450-458, 2004.

[15] A.G. Hauptmann et al, "CMU Informedia's TRECVID 2005 Skirmishes," In *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID, Gaithersburg, USA, 2005.

[16] A. Amir et al, "IBM Research TRECVID-2005 Video Retrieval System," In *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID, Gaithersburg, USA, 2005.

[17] T. S. Chua et al, "TRECVID 2004 search and feature extraction task by NUS PRIS," In *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID, Gaithersburg, USA, 2004.

[18] A. Natsev, M. R. Naphade, and J. Tesic, “Learning the semantic of multimedia queries and concepts from a small number of examples”, In *International Conference on Multimedia*, ACM, Singapore, pp.598-607, 2005.

[19] H. J. Su, Y. Zhao, and B. Z. Yuan, “A new composite histogram integrating each bin's spatial distribution for image retrieval,” In *IEEE TENCON’02*, 2002.

[20] M. Porter, “An algorithm for suffix stripping,” Program, pp: 14(3): 130-137, 1980.

[21] C. Snoek, M. Worring, D. Koelma, and A. Smeulders, “Learned Lexicon-Driven Interactive Video Retrieval,” In *CIVR 2006*, pp. 11-20, 2006.

[22] J. H. Yuan et al, “Tsinghua University at TRECVID 2005,” In *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID, Gaithersburg, USA, 2005.

[23] S. F. Chang, “Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction,” In *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID, Gaithersburg, USA, 2005.

[24] J. Lafferty and C. Zhai, “Risk minimization and language modeling in information retrieval,” In *24th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*, 2001.