J Sign Process Syst (2010) 59:189–199
DOI 10.1007/s11265-008-0287-2

A Cooperative Learning Scheme for Interactive Video Search

Shikui Wei · Yao Zhao · Zhenfeng Zhu · Nan Liu

Received: 14 May 2008 / Revised: 15 July 2008 / Accepted: 23 September 2008 / Published online: 1 November 2008

Abstract The main idea of an interactive search is to gradually improve search quality of retrieval system via user interaction. While a large amount of work has been made in the past, most of the existing approaches typically require labeling effort for updating the query model. Unfortunately, it is time-consuming and tedious to label a large number of training examples. We aim to develop a novel text-driven cooperative learning scheme, which can offer users a quite natural query fashion and alleviate significantly the burden on users without compromising search performance. Starting with an advanced text-driven video search engine, a multi-view cooperative training strategy is proposed for learning from feedback data a refined ranking function. The main merit of proposed framework is its ability in mining training samples automatically from previous answer set and implicitly combining multiple modalities for effectively learning users' query intent. Evaluation on TRECVID' 06 video corpus shows that the proposed scheme with few training seeds achieves a comparable performance with classic interactive schemes.

Keywords Video · Search · Interactive · Cooperative learning

S. Wei (✉) · Y. Zhao · Z. Zhu · N. Liu
Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
e-mail: shkwei@gmail.com

Y. Zhao
e-mail: yzhao@bjtu.edu.cn

Z. Zhu
e-mail: zhfzhu@bjtu.edu.cn

N. Liu
e-mail: mysnowdays@126.com

1 Introduction

Video archives are cropping up everywhere now with the popularity of video capture devices, which greatly motivates researchers to find various approaches for naturally and effectively searching interesting information from such enormous multimedia resources. While various retrieval models have been developed for responding to the users' query intent, most of them investigate video search patterns by implicitly or explicitly measuring similarity between the query and the database shots in some low-level video feature spaces [1]. However, the similarity is not always consistent with human perception due to the limitation of current image/video understanding techniques. That is, semantic gap exists between low-level features and high-level semantics. Therefore, determining how to nicely model user preference and effectively bridge the semantic gap is a key issue in the multimedia search area.

As a kind of working way for alleviating this problem to some extent, the interactive technique, which combines users' preference into search process by real-time users' intervention, has attracted more attention in recent years. While numerous interactive models have been proposed in previous work [2–8], most of them formalize the interaction process as learning a retrieval function from only labeled data, namely, supervised learning approaches. Most closely related work is the scheme made by C. Snoek et al. [5], which first queries a topic by selecting some similar concept interfaces from a fixed set of total 106 query interfaces, and then learns a new retrieval model from feedback information using one-class SVM method. Although this kind of paradigm has achieved quite good search performance, it requires labeling a large number of training samples manually. Unfortunately, no users are willing to spend too much time for it in a real-world search

scenario, so it is crucial for a search engine to simplify the tedious labeling task.

For mitigating users' burden on labeling, various methods have been employed in previous literature. As an extreme case, pseudo-relevance feedback (PRF) technique is developed for automating the interactive process. PRF-based methods assume that the top-ranked documents in returned result list are relevant to query and are used to automatically refine the search process [9]. For instance, co-retrieval algorithm [10] treats the top-ranked results as positive examples and others as negative ones. From these noisy training samples, a re-trained retrieval model is then built using an Adaboost based ensemble learning method. Although this kind of methods frees users from the time-consuming labeling effort, user preference is ignored completely. Considering the variety of user's subjectivity on video content, it is necessary to involve user preference as well as to minimize user burden. To do this, an SOTM-based automatic machine interaction scheme is proposed in [11]. This method minimizes human involvement by employing a recursive approach based on the self-organizing tree map (SOTM).

Likewise, we aim to present an interactive search scheme that can nicely model user preference with less effort on labeling samples. To do this, a cooperative training framework is proposed for learning the ranking function, in which each shot is individually represented by multiple independent feature views. An important point of difference from previous work is that this scheme learns users' preference in a semi-supervised fashion. Given only few positive samples (conveying users' query preference) as training seeds, it can automatically find out more shots with high possibility being positive from currently returned shots and then separately update the training sets on different feature views. By doing this, the users are disentangled from the time-consuming task of labeling training samples, and multiple learners built for different feature views can also contribute to each other during the training phase.

As an inevitable step, all of the interactive search systems require providing a natural search entrance for user interaction. Yet, there is no a generally accepted way to start such a video search process. As a kind of effective method, example-based retrieval pattern is generally used in many video retrieval systems [12–14]. Unfortunately, users usually do not have proper video examples at hand for formulating their query intent in a real world search scenario, and it is also unreasonable to expect users to provide query examples for search system [5, 6]. In contrast to the example-based approach, text-based search fashion, which is generally applied to commercial web search engines, is considered a good fashion for satisfying users' query practice [2, 5–7, 12, 15]. Unexceptionally, an advanced text-based search engine, which provides a natural fashion for triggering a search process, is also presented in this paper to return an initial ranked shot list for each query topic.

The rest of this paper is organized as follows. First, a fully automatic and text-based video search engine is presented in Section 2, leading to the problem analysis in Section 3. Section 4 then focuses on the discussion of proposed interactive learning framework. In Section 5, some experimental results and analysis are illustrated in detail. We finally give the conclusion and discuss our future work in Section 6.

2 Text-Based Video Search Engine

The main idea of text-based video search approach is to convert video retrieval into text document search [16]. In particular, given a query topic text in everyday English by users, the text-based system can return a ranked shot list by matching the query text with the text documents associated with video shots. As indicated, two key aspects of this issue have to be addressed: (1) how to properly map a video shot into a text document; (2) how to correctly rank these text documents after giving a query text. The first problem involves some effective text processing techniques used widely in the Natural Language Processing (NLP) area. We will show how those techniques construct a text document for each shot to effectively represent the semantic content of the shot after giving some certain text features about the shot, such as speech transcripts, closed captions, and video OCR text. In our study, speech transcript is solely imported to construct the text documents of shots. To address the second aspect, some statistical language models used in the information retrieval (IR) filed can be explored to match the query text and the candidate text documents of shots. Figures 1 and 2 demonstrate the sketches of mapping and ranking procedures, respectively. The detailed information will be described in the following two subsections.

2.1 Mapping

The speech transcripts of each shot are extracted from audio track using automatic speech recognition (ASR) techniques. For the non-English shots, they are processed further for translating their ASR output into English text exploiting machine translation (MT) approaches.

The purpose of mapping is to build a concise and effective text representation for each shot from its speech transcripts. Intuitively, it is a feasible way that the transcribed speech text of each shot is directly used to form a text document mapping to the shot. Nevertheless, the formed text document involves much non-relevant information such as function words and is also not robust

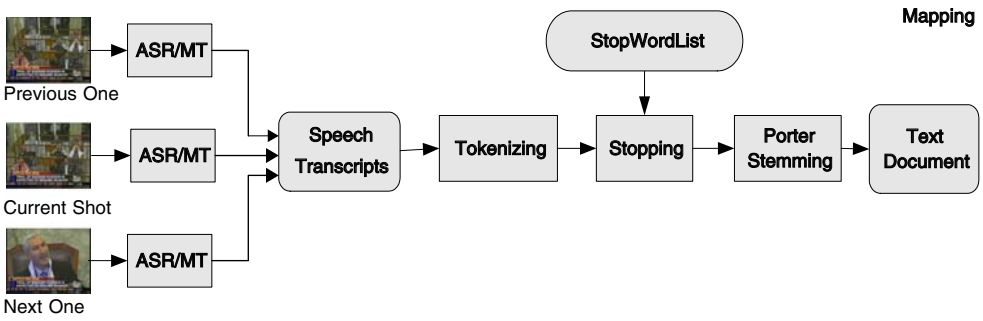

Figure 1 Sketch of constructing a text document for current shot from speech transcripts.

for text matching due to inner property of scoring functions. Therefore, it is necessary to process the speech transcripts before they are mapped to a shot. As shown in Fig. 1, this issue is addressed by exploring some text processing approaches. As an important preprocessing step, tokenizing technique, which splits sentences or paragraphs into individual words, is first employed to form a keyword list from the transcripts of shot since most of text processing techniques and scoring functions are directly or indirectly based on the keywords. Afterwards, function words such as "the", which are not meaningful for video shot content, are removed out from the keyword list according to a stopword list of total 460 terms, leading to a less noisy text representation. Moreover, since scoring functions commonly used in IR area are essentially the matching between query terms and key terms in the candidate documents, it is necessary to handle the semantic inconsistency problem of some words. For instance, although the terms "computation" and "computing" are of similar semantic interpretation, the scoring function will treat them as two completely irrelevant words due to their different morphological patterns. Therefore, stemming techniques, which reduce each word to its stem or root form, is further exploited to tackle the semantic inconsistency problem. In our work, Port stemming [17], a well-known stemming technique, is employed to process the keyword list.

In addition to text processing, each shot is extended by a shot length at the start and end points, considering the semantic similarity and the speech offset of adjacent shots in the same video archive. That is, the transcripts of the previous shot and the next shot are integrated into the transcripts of the current shot before text processing, as shown in Fig. 1. So far, a clear and robust text representation is built for each shot.

2.2 Ranking

The key issue in this phase is how to design a proper ranking function, which can make a fair judgment of the relevance between the query and the candidate shots and give a similarity score to each candidate shot. As shown in Fig. 2, the query text is treated as a naïve text document of a virtual shot, which is processed with the same way as the mapping procedure does. Using this processed query text document, a scoring function is then employed to rank the candidate shots by matching the query text document and the candidate text documents. In our work, KL-divergence retrieval model, an extension of query-likelihood approach [18], is used to score candidate documents after both the query document and the candidate documents are modeled using the same statistical language model presented in [18, 19].

3 Problem Analysis

When we design an interactive search system, two important factors must be taken into account. Firstly, users are less willing to spend too much time labeling data during the process of seeking needed information in a real world search scenario. Hence, it is crucial to alleviate the burden on users without decreasing the search quality of system. Secondly, users are usually more interested in a very small set of relevant shots. Therefore, it is vital to have more high accuracy on the top returned shots with less user intervention.

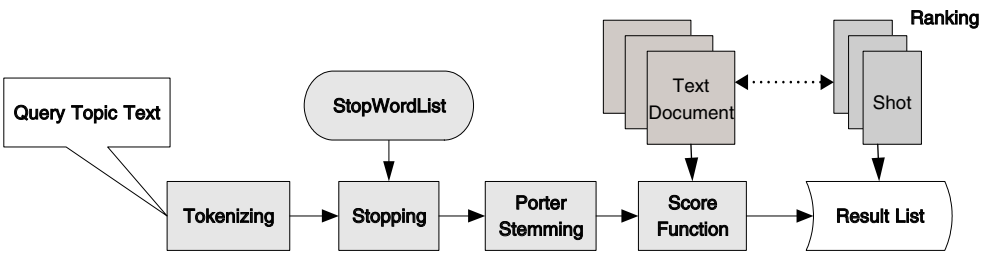

Figure 2 Framework of text-based video retrieval system.

Before giving our proposed interactive learning scheme, we first analyze why it is necessary and feasible to mitigate the burden on users during the interaction process. In fact, the problem can be revealed by analyzing the quality of text-based automatic video search engines. NIST TRECVID provides 24 search topics for all participants to test the search performance of their search systems. It also requires all its participants to return a ranking of 1,000 shots for each query topic and to submit at least one run (including 24 rankings, one ranking corresponds to one topic) for evaluation. Figure 3 illustrates the run score of proposed search method (dot) relative to the median and the best non-interpolated average precisions of all participants for TRECVID'06.

Considering the data in Fig. 3, our system takes a moderate achievement among all 76 fully automatic runs [20] from all participants. Furthermore, some statistics are taken on the average numbers of relevant shots over 24 topics at different return depths. Without loss of generality, the statistical data on our text retrieval run (BJTU) and on all 76 runs is shown together in Fig. 4. The blue bins show the average numbers of relevant shots at different depths, and the corresponding values are also given under the bins. For instance, the average number for 76 runs is 1.41 at depth 10. Likewise, the red bins indicate the BJTU case. The approximate likeness of bins indicates that our text-based video retrieval system is representative, so this system is impartial to serve as the entrance for interaction.

As shown in Fig. 4, there are quite a few relevant shots at some great depth (e.g. depth = 1,000, average number of relevant shots = 43.96), which means that there indeed exists a certain number of positive samples for learning. However, users have to scroll down the returned list far enough in order to label enough samples, which is time-consuming. In addition, Fig. 4 also indicates that few relevant shots appear in the most top-returned shots (e.g. depth = 10, average number of relevant shots = 1.38), so it is necessary to improve the precision of the most top ranked results.

Therefore it is essential to develop an approach that can automatically mine training samples given only few training seeds and pay more high precision on the top-ranked results.

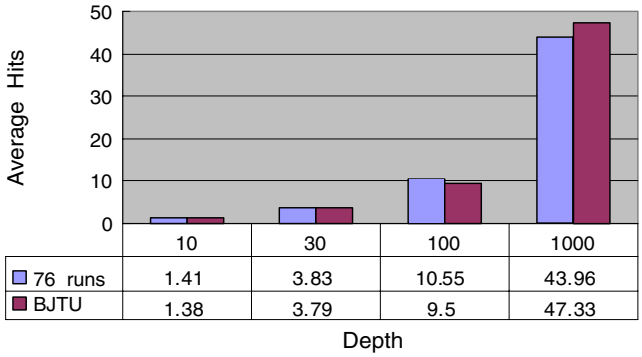

Figure 4 Average numbers of relevant shots at different depths for 76 runs and BJTU run.

4 Cooperative Learning for Interaction

The main contribution of the proposed interactive approach is to automatically mine positive training samples from the initial search results so as to alleviate the burden on users and more effectively learn user's query intent. The general framework of the proposed scheme is illustrated in Fig. 5. We will describe each component in more details below.

4.1 Cooperative Learning Scheme

We aim to develop a novel learning strategy that can effectively model users' query intent after users label a very small set of relevant shots from the initial search results. For this purpose, the multi-view learning method [21–23] is extended in this paper. The essential of the multi-view learning strategy is that each training example is represented explicitly in multiple distinct views. The improvement of learning quality on each independent view will be iteratively benefited from the other independent views, leading to a more reliable learning outcome finally. An important difference between our scheme and the tradition-

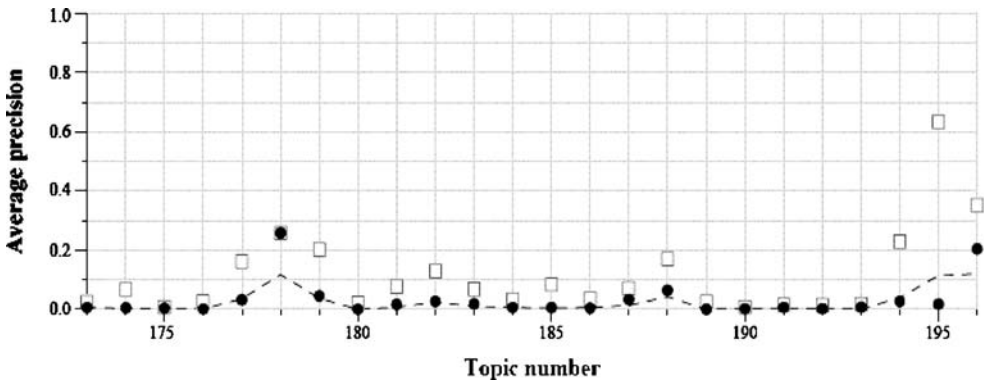

Figure 3 Run score (*dot*) versus median (*line*) versus best (*box*) by topic.

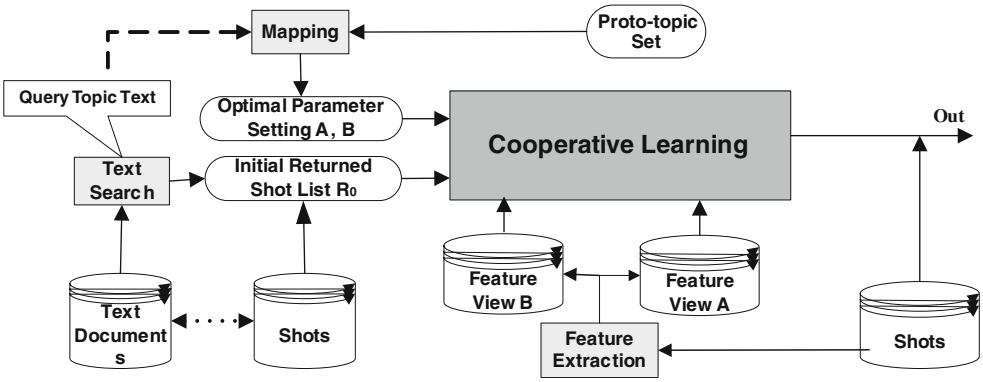

Figure 5 The framework of the interactive search system.

al multi-view learning approaches is the exchange strategy of labeled samples on the different views.

In our context, multi-view means that the same original shot is explicitly and individually represented in multiple approximately independent feature spaces such as text feature space and visual feature space. In our study, only two views are adopted, which can be easily extended to more views.

Table 1 describes the overall flowchart of proposed learning strategy. Specifically, for feedback i, a returned shot list R_i is presented to user, the user then labels a very small set P_i of positive examples as training seeds and leaves directly the others as the unlabeled data set U_i. Our goal is to iteratively mine more positive examples from only the set U_i. Notice that the negative data set N_i is only selected from database randomly and fixed during the whole learning procedure. Note that, when i is equal to zero, all the processes are carried out on initial search list.

As described in Table 1, each learner is built separately on a specific view of P_i and N_i iteratively, which can be formulated as follow:

$$C^j_{i+1,v} = TrainSVM\left(P_{i,v}, N_i, v\right) \quad (1)$$

where, $v \in \{A, B\}$ denotes a specific feature view, i is the ith feedback process, j is the jth iteration, $C_{i+1,v}$ is the classifier on view v, $P_{i,v}$ is the positive sample set for training the classifier on view v.

After unlabeled set U_i is labeled individually by two classifiers, the selection of more reliable training samples has a direct impact on the final learning performance. In our scheme, the training set of positive samples on one view classifier is updated separately by importing the label information annotated by the other view classifier on U_i, instead of retaining a common training set for both classifiers as Co-training does. For instance, using the label information of $C_{i,A}$ on U_i, the most likely positives in U_i are added into the $P_{i,B}$. Indeed, our sample exchange strategy across different views is also different from so-called Co-EM algorithm [22] which trains one classifier using directly the assigned labels from the other classifier on U_i. Notice that the number of feedback M is determined by the user's satisfactory responding to the returned search results, which regulates the end point.

Table 1 Cooperative learning scheme.

Inputs an initial returned shot list R_0, the number of feedback M and the number of iteration T
For i=1 to M
$R_{i-1} = P_{i-1} \bigcup U_{i-1}, P_{i-1,A} = P_{i-1,B} = P_{i-1}$
Selects negative data set N_{i-1} randomly
For j=1 to T
$C^j_{i,A} = \text{TrainSVM}\left(P_{i-1,A}, N_{i-1}, \text{A}\right)$
$C^j_{i,B} = \text{TrainSVM}\left(P_{i-1,B}, N_{i-1}, \text{B}\right)$
Updates $P_{i-1,A}$ using the output of $C^j_{i,B}$ on U_{i-1}
Updates $P_{i-1,B}$ using the output of $C^j_{i,A}$ on U_{i-1}
Outputs $C^T_{i,A}$, $C^T_{i,B}$
$R_i = \left\{F_i(D) = \alpha C^T_{i,A}(D) + \beta C^T_{i,B}(D)\right\}$
Output $C^T_{M,A}$, $C^T_{M,B} R_M$

4.2 Feature Extraction

In the video search scenario, since video shot is referred as the final unit needed to be returned, the feature extraction is based on the shot unit. In our scheme, each shot is represented using two approximately independent feature views, one is the visual information, and the other is the ASR/MT text associated with shot. Color histogram presented in [24] is exploited as the visual descriptor, denoted as feature view A, which is actually extracted from key frame of shot picked by Fraunhofer Institute [25]. For text descriptor, an effective extraction scheme is presented for constructing a 78-D feature vector to represent the text feature view B. As demonstrated in Fig. 6, after N concepts are selected as proto-concepts from concept ontology of LSCOM [26], a training set of 40 shots with corresponding text documents are chosen for each proto-concept according to the annotation ground truth [26]. Here, N is fixed to 78. Finally, for each shot, a 78-D text vector can be constructed by measuring similarity between the text document of the shot and each training set of proto-concepts using the scoring function described in Subsection 2.2.

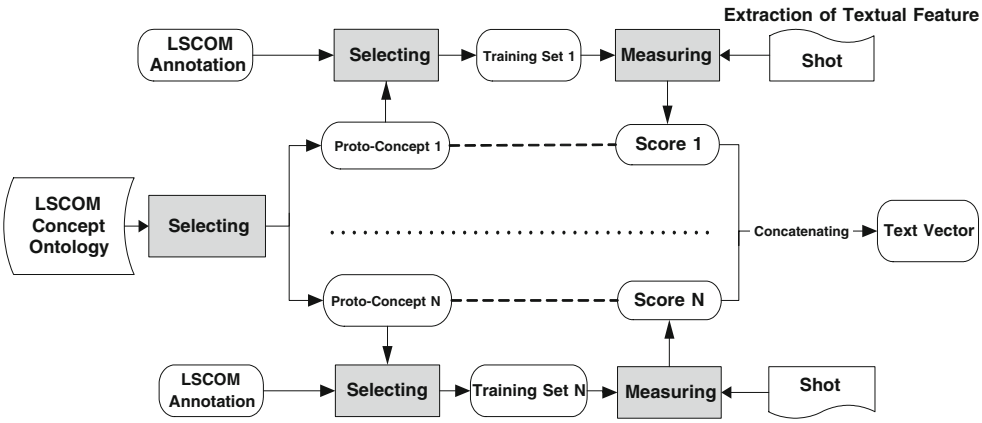

Figure 6 The extraction scheme of textual feature.

4.3 Optimal Parameters Selection

In this scheme, SVM with RBF kernel function [27] is employed as the underlying learner for the cooperative learning due to its powerful ability of learning a model from a small set of labeled samples. Instead of assigning only labels, an extended SVM classifier provided in [28] is utilized to predict the class probability information of samples. Although SVM has been proven to be an eminent classification tool, optimal parameter setting for SVM, which is difficult to know in advance for a specified query category [15], significantly influences the classification performance of video information [5]. Besides, the effect of different views should also be taken into consideration in our context during the parameter selection procedure.

Here, a simple but effective method is proposed for dealing with the parameter selection problem. Instead of selecting a global optimal parameter setting, the selecting procedures on two feature views are individually consulted after assigning a specific concept category. Therefore, our parameter selection procedure is dependent on the topic category and feature view. Specifically, a number of representative query topics are first selected to construct a proto-topic set, and then a training set is selected for each proto-topic against LSCOM Annotation [26]. After two feature views of the training set are extracted for each proto-topic, optical parameter settings on two views are obtained separately in distinct feature spaces of the same training set using cross-validation and grid-search. Figure 7 illustrates the procedure of a given proto-topic.

When a new query topic comes in, it is first mapped into one of proto-topics, and then the optimal parameter settings corresponding to this proto-topic are chosen as parameter settings of the new topic. Note that the mapping process here is carried out in an interactive fashion, and we will do some research on automatically mapping in the future.

4.4 Multi-View Fusion over Search Results

As described in Table 1, a fusion method is required to integrate the search results from two view classifiers. As two main fusion strategies, the early fusion and the late fusion were generally adopted for integrating multiple modalities in most video indexing systems [2, 5, 7, 12, 29, 30]. In addition, the latest fusion strategy can also be considered from the view of query classes, namely query-class-dependent and query-class-independent models [15, 31]. In this paper, we don't focus on the development of

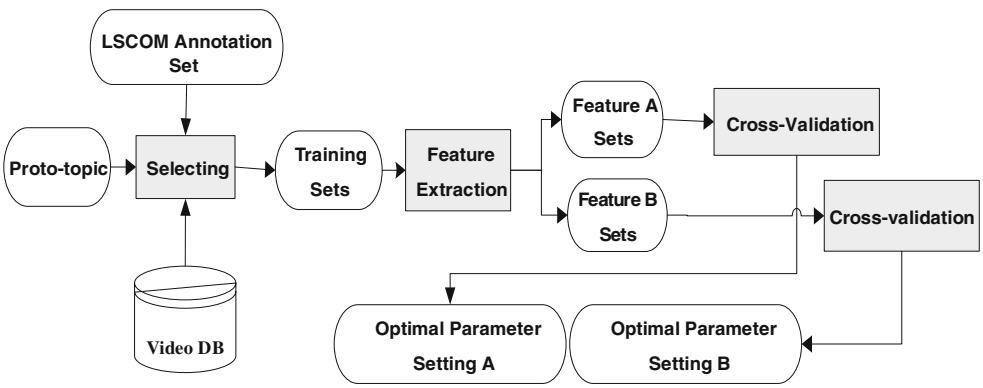

Figure 7 Selecting individually optimal parameter settings for one proto-topic on two views.

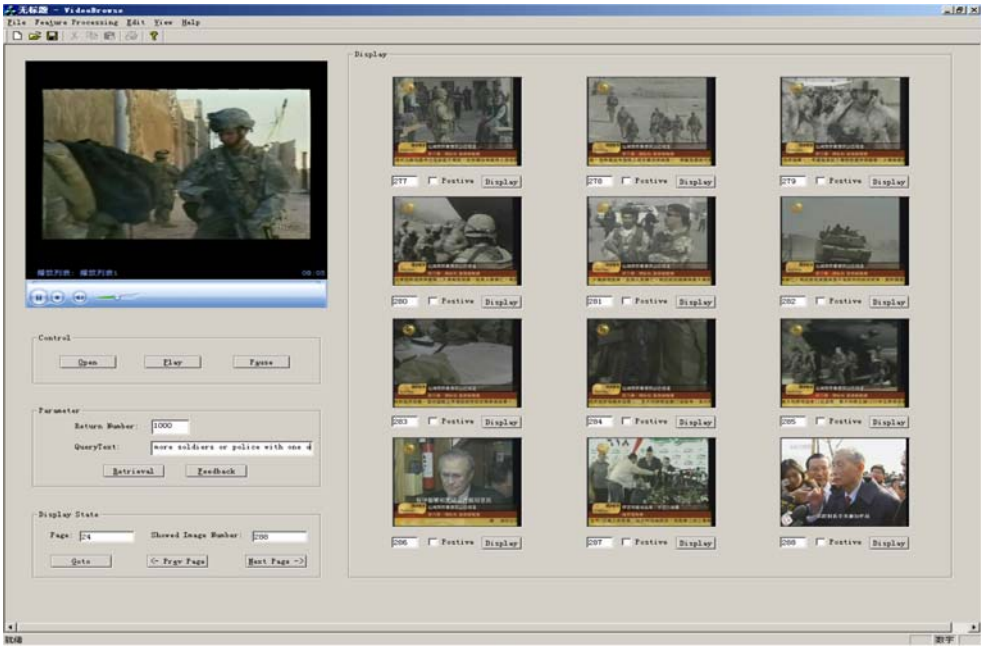

Figure 8 The interface of the proposed interactive search system.

new fusion strategies. Rather, a simple linear average weighted score [2, 13], generally used in multimedia area, is employed to integrate the search results, which is defined as follow:

$$F(D) = \alpha C_A(D) + \beta C_B(D) \tag{2}$$

where D denotes dataset, $C_A(D)$ and $C_B(D)$ stand for the returned ranking scores on view A and B, respectively, α and β are constants, which can be obtained experientially, usually $\alpha \leq \beta$.

5 Experimental Results and Analysis

We employ the NIST TRECVID'06 benchmark to evaluate the performance of our proposed interactive search scheme, which is composed of approximately 343h of MPEG-1 broadcast news video, 169h for TRECVID'05 dataset viewed as training set in TRECVID'06, 174h as test set. Together with this corpus, the LSCOM workshop [26] provided the ground truth of annotation for the TRECVID'05 development set, and Fraunhofer Institute [25] provided master shot reference for all data as well. In our experiments, the TRECVID'05 development data set with annotation information is employed to build training set for searching optimal SVM parameters. The test set for TRECVID'06 is adopted to answer the query topic and evaluate the search performance.

For the performance metric, TRECVID suggests a number of evaluation criteria [20]. Three of them are selected in our work, including precision at different depths of result list (Prec_D), non-interpolated average precision (AP), and mean average precision (MAP). We denote D as the depth where precision is computed. Let S be the total number of returned shots, and R_i the number of the true relevant shots in top-i returned results. Then, these metrics can be defined as below:

$$\Pr ec\_D(T_n) = \frac{1}{D}\sum_{i=1}^{D} F_i \tag{3}$$

$$AP(T_n) = \frac{1}{R}\sum_{i=1}^{S} \frac{R_i}{i} * F_i \tag{4}$$

$$MAP = \frac{1}{N}\sum_{n=1}^{N} AP(T_n) \tag{5}$$

where T_n is the *nth* query topic, F_i = 1 if the ith shot is relevant to query and 0 otherwise, R stands for the total

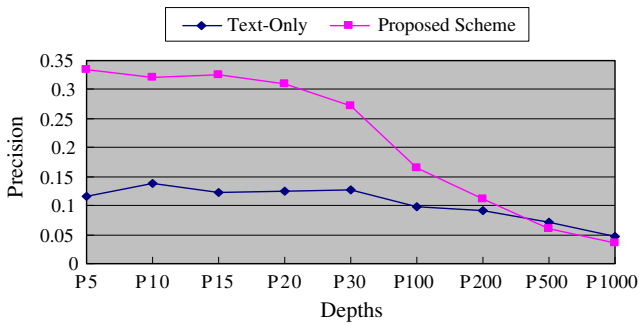

Figure 9 The proposed scheme VS. Automatic search at depth X in the result set.

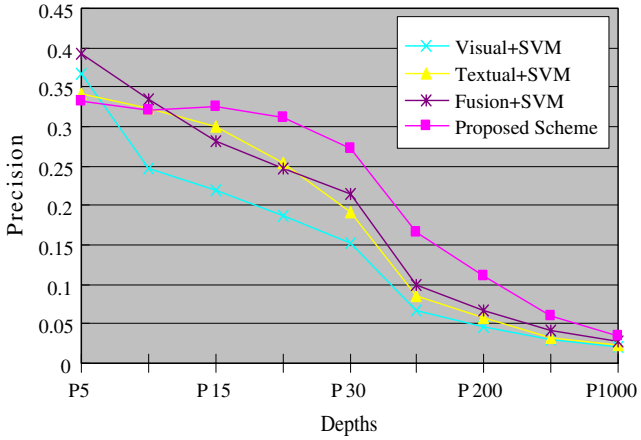

Figure 10 Performance comparison on different interactive learning schemes.

number of true relevant shots, and N denotes the number of query topics.

Prec_D is utilized to evaluate the precision at different depths of result list. AP shows the performance of a single query topic, which is sensitive to the entire ranking of documents. MAP summarizes the overall performance of a search system over all query topics. Note that only the top-100 shots in the result list are considered for computing both AP and MAP.

5.1 Experimental Setup

As an unavoidable aspect, interactive interface must be taken into account when designing an interactive video search system. Therefore, a video search system named VideoBrowse, which implements all of the algorithms mentioned above, is developed for providing the text-based search function and the natural interactive interface. Figure 8 shows its interface. Using this search system, we can easily label some interesting shots after obtaining an initial search list of total 1,000 shots for each query topic.

5.2 Evaluation on Different Methods

We participated in the automatic search and interactive search subtasks for TRECVID2006 and submitted two runs to NIST, one for the fully automatic search, and the other for the interactive search. The returned evaluation results imply the effectiveness of proposed search scheme. The detailed experimental analysis will be given below.

For automatic text-based video retrieval, because our purpose is to offer an entrance for user interaction, the more the number of relevant documents in the returned shot list, the better the performance of the search system. As described in Section 2, we exploited some natural language processing techniques with only the transcribed speech text to construct the text search engine. As shown in Fig. 3, the scores of our performance are slightly greater than or equal to the median value for most of the topics, and even one topic achieves the best performance. Considering that many other automatic systems combine multimodal features into search procedure, this result is relatively significant.

For the proposed interactive search scheme, our aim is to develop an algorithm which can not only alleviate the burden on users effectively by labeling only a very small set of positives but also give high accuracy on the top-ranked shots. Hence, the measurement of precision at 9 document cutoff values is suitable to evaluate the effectiveness of the proposed interactive scheme. The precision is computed after a given number of documents have been retrieved, which reflects the actual system performance at different return depths. Note that the precision at depth X, here, is the precision average over all of the 24 topics. We test the proposed interactive system by labeling only 5

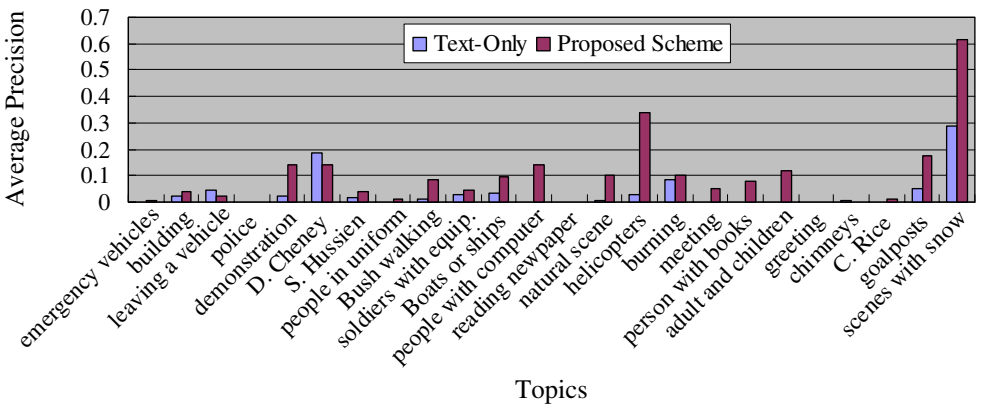

Figure 11 Performance of the proposed scheme and the text-only baseline across all 24 query topics of TRECVID 2006.

positive examples, which is a quite small set. As shown in Fig. 9, the average precision at depth X is far higher than automatic text-based search within top-200 returned results, which indicates that the proposed approach indeed brings up the true relevant results in the ranking (average 106% increase in AP).

The last series of experiments is designed to compare the search quality of different interactive learning schemes. Ten positive examples, which are twice as large as the account of positives labeled for the proposed method, are manually labeled by user for testing these supervised learning schemes. The details of these supervised learning schemes are as follows.

Textual feature + SVM: textual feature of training shots is only utilized to train the classifier and rank the candidate documents.
Visual feature + SVM: color feature of training shots is solely employed to train the classifier and give a ranked result list.
Fusion + SVM: The results from two classifiers above are combined into a unified ranking by using linear average weighted method mentioned early. The main difference from the proposed scheme is that, instead of contributing to each other during training phrase, the two classifiers in Fusion+SVM scheme are built separately.

The same ground truth, which is generated by NIST for evaluating the search task and providing a fair comparison, is used to judge if the result is relevant. The final evaluation results are depicted in Fig. 10. As shown in Fig. 10, the proposed scheme performs better than the others methods even if its training set is smaller than others, which suggests that the proposed scheme indeed mitigates the burden on users and enhances the final search quality meanwhile. Also, Fig. 10 demonstrates that the performance of textual feature based scheme is almost equal to the fusion scheme, which implies the effectiveness of our proposed extraction scheme of textual feature.

5.3 Performance Analysis on All Query Topics

In this section, we evaluate our proposed scheme on varied query topics. Figure 11 illustrates the statistics on APs across 24 query topics used in TRECVID'06 evaluation.

The results show that the proposed scheme works well for named persons and named objects, such as "S. Hussien" and "Boats or Ships", as search quality on these topics can benefit from the textual feature used in our scheme. The only exception lies in the topic "D.Cheney", for which the performance after interaction is below the text-only baseline. One possible explanation is that text-only baseline itself has achieved considerable retrieval effectiveness on the top-ranked shots. Therefore, it is difficult to improve the performance further.

Moreover, our approach is also suitable for some query topics that are of distinctive visual properties, such as "soccer goalposts" and "scenes with snow". Similarly, prominent improvement of them is due to the usefulness of the visual feature.

On the other hand, the search performance after interaction is even below the text-only baseline for some topics with motion properties, like "leaving a vehicle". The reason is that features used in our scheme lack the ability to capture motion properties in video. Hence, new research fruits in precise representation of shot will provide much more room for performance improvement. Note that the performance improvement on the topic "Bush Walking" is due to the effectiveness of textual feature.

In addition, our proposed method also fails in some query topics with very few relevant shots in the initial list, such as "greeting" and "police". The reason is that no more positive examples can be found to train the new model, which thereby leads to the failure of performance improvement.

6 Conclusions

In this paper, we developed an interactive video search scheme based on an advanced text-based video search engine and a cooperative learning 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 in automatically finding positives from past unlabeled answer set. In addition, both learners can contribute to each other by using the label information from the other view, which indicates that multiple modalities are potentially fused 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 indeed reduces a need for the labeled data. In future work, we will develop some techniques to more effectively fuse the result list from different views.

Acknowledgments This work was supported in part by National Science Foundation of China (No. 60602030, No. 90604032), 973 Program (No. 2006B30314), 863 Program (No. 2007AA01Z175), PCSIRT (No. IRT0707), and Specialized Research Foundation of BJTU (No. 2005SM013, No. 2005SZ005).

References

1. Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl. 2*, 1–19 (2006). TOMCCAP doi:10.1145/1126004.1126005
2. Amir, A., Argillander, J., Campbell, M., Haubold, A., Iyengar, G., Ebadollahi, S., et al: J. Teˇsi'c, and T. Volkmer, "IBM Research TRECVID-2005 Video Retrieval System," In TREC Video Retrieval Evaluation Online Proceedings, TRECVID, Gaithersburg, USA, 2005.
3. Chang, S. F., Hsu, W. H., Kennedy, L., Xie, L., Yanagawa, A., Zavesky, E., et al: Columbia University TRECVID-2005 video search and high-level feature extraction. In TREC Video Retrieval Evaluation Online Proceedings, TRECVID, Gaithersburg, USA, 2005.
4. Kacprzyk, J., & Zadrozny, S. (2005). Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. *Information Science*, *173*, 281–304. doi:10.1016/j.ins.2005.03.002.
5. Snoek, C. G. M., van Gemert, J. C., Geusebroek, J. M., Huurnink, B., Koelma, D. C., Nguyen, G. P., et al (2005) The MediaMill TRECVID 2005 semantic video search engine. In TREC Video Retrieval Evaluation Online Proceedings, TRECVID, Gaithersburg, USA.
6. Snoek, C., Worring, M., Koelma, D., & Smeulders, A. (2006). Learned lexicon-driven interactive video retrieval. In CIVR 2006, pp. 11–20.
7. Zhang, D. S., & Nunamaker, J. F. (2004). A natural language approach to content-based video indexing and retrieval for interactive E-learning. *IEEE Transaction on Multimedia*, *6*(3), 450–458.
8. Zhou, X. S., & Huang, T. S. (2002). Relevance feedback in content-based image retrieval: some recent advances. *Information Science*, *148*, 129–137. doi:10.1016/S0020-0255(02)00286-4.
9. Hsu, W. H., Kennedy, L. S., & Chang, S.-F. (2007). Reranking methods for visual search. *IEEE Transaction on Multimedia*, *14*, 14–22.
10. Yan, R., & Hauptmann, A. G. (2005). Co-retrieval: a boosted reranking approach for video retrieval. *IEE Proceedings Vision, Image and Signal Processing*, *152*, 888–895. doi:10.1049/ip-vis:20045188.
11. Muneesawang, P., & Guan, L. (2002). Automatic machine interactions for content-based image retrieval using a self-organizing tree map architecture. *IEEE Transactions on Neural Networks*, *13*(4), 821–834. doi:10.1109/TNN.2002.1021883.
12. Hauptmann, A. G., et al. (2005). CMU Informedia's TRECVID 2005 Skirmishes. In TREC video retrieval evaluation online proceedings, TRECVID, Gaithersburg, USA.
13. Natsev, A., Naphade, M. R., & Tesic, J. (2005). Learning the semantic of multimedia queries and concepts from a small number of examples. In International Conference on Multimedia, ACM, Singapore, pp. 598–607.
14. Yuan, J. H., Zheng, W. J., Chen, L., Ding, D. Y., Wang, D., Tong, Z. J., et al. (2005). Tsinghua University at TRECVID 2005. In TREC video retrieval evaluation online proceedings, TRECVID, Gaithersburg, USA.
15. Kennedy, L. S., Natsev, A., & Chang, S. F. (2005). Automatic discovery of query-class-dependent models for multimodal search. In International Conference on Multimedia, ACM, Singapore, pp. 882–891.
16. Hsu, W. H., Kennedy, L. S., & Chang, S.-F. (2006). Video search reranking via information bottleneck principle. In 14th annual ACM international conference on Multimedia, Santa Barbara, CA, USA, pp. 35–44.
17. Porter, M. (1980). An algorithm for suffix stripping. *Program*, *14* (3), 130–137.
18. Lafferty, J., & Zhai, C. (2001). Risk minimization and language modeling in information retrieval," In 24th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01).
19. The Lemur Toolkit for Language Modeling and Information Retrieval: URL:http://www.lemurproject.org.
20. TRECVID, TREC Video Retrieval Evaluation.: In http://www-nlpir.nist.gov/projects/trecvid.
21. Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In Proceedings of the Workshop on Computational Learning Theory, ACM, New York, USA, pp. 92–100.
22. Brefeld, U., & Scheffer, T. (2004). Co-EM support vector learning. In Proceedings of the twenty-first International Conference on Machine learning, Canada.
23. Nigam, K., & Ghani, R. (2000). Understanding the behavior of co-training. In Proceedings of the Workshop on Text Mining, ACM.
24. Su, H. J., Zhao, Y., & Yuan, B. Z. (2002). A new composite histogram integrating each bin's spatial distribution for image retrieval," In IEEE TENCON'02.
25. Petersohn, C. (2004). Fraunhofer HHI at TRECVID 2004: Shot Boundary Detection System", In TREC Video Retrieval Evaluation OnlineProceedings, TRECVID, URL: http://www.nlpir.nist.gov/projects/tvpubs/tvpapers04/fraunhofer.pdf
26. Lexicon Definitions, L.S.C.O.M.: and Annotations Version1.0, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #217-2006-3, March 2006.
27. Vapnik, V. (2000) The nature of statistical learning theory. Tsinghua University Press, Chinese Language Edition.
28. Chang, C. C., & Lin, C. J. (2001). LIBSVM: a library for support vector machines," 2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
29. Yan, R., & Naphade, M. (2005). Multi-modal video concept extraction using co-training. In International Conference on Multimedia and Expo, IEEE, pp. 514–517.
30. Snoek, C. G. M., & Worring, M. (2005). 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.
31. Chua, T.-S., Neo, S.-Y., Li, K.-Y., Wang, G., Shi, R., Zhao, M., et al (2004). TRECVID 2004 search and feature extraction task by NUS PRIS. In TREC Video Retrieval Evaluation Online Proceedings, TRECVID, Gaithersburg, USA.

Shikui Wei received the B.S. degree in electrical engineering from Hebei University, Baoding, China, in 2003, and the M.S. degrees in signal & information processing from Beijing Jiaotong University, Beijing, China, in 2005. He is now a Ph. D. candidate in signal & information processing at Beijing Jiaotong University. His research interests include computer vision, pattern recognition, multimedia analysis and retrieval.

Zhenfeng Zhu received the bachelor's degree and M. S. degree from the Wuhan University of Science and Engineering and Harbin Institute of Technology in 1996 and 2001, respectively, both in electromechanical engineering. After receiving the Ph. D. degree in Pattern Recognition and Intelligence System from Institute of Automatication, CAS, in 2005, he joined the Institute of Information Science of Beijing Jiaotong University. His main research interests include image and video understanding, pattern recognition and computer vision.

Yao Zhao received the bachelor's degree and M. S. degree from the Fuzhou University and Southeast University in 1989 and 1992, respectively, both in radio engineering, and the Ph. D. degree in signal and information processing from Beijing Jiaotong University in 1996. Since then he has been a faculty member in the Institute of Information Science of Beijing Jiaotong University. His main research interests include multimedia data compression, digital watermark, and content based multimedia information retrieval. He has been a visiting profession in Delft University of Technology from May 2001 to May 2002. Now he is the director of Institute of Information Science of Beijing Jiaotong University and member of IEEE.

Nan Liu was born in 1983 in China. He received his M. S. in biomedical engineering from Beijing Jiaotong University and is a Ph. D. student of Institute of Information Science in Beijing Jiaotong University now. His current research interests include pattern recognition, commercial detection, commercial analysis etc.