A Cooperative Learning Strategy for Interactive Video Search

Shikui Wei[1]*, Zhenfeng Zhu*[1,2]*, Yao Zhao*[1]*, Nan Liu*[1]

[1]Institute of Information Science, Beijing Jiaotong University, Beijing, 100044, P.R.China
[2]National Key Laboratory on Machine Perception, Beijing University, Beijing 100871, P.R.China
E-mail: shkwei@gmail.com

***Abstract*—**The goal of this paper is to develop a learning strategy for interactive video search that can effectively mitigate the burden on users without decreasing search performance. Taking SVM as underlying learner, a cooperative training strategy is proposed for learning a ranking function, in which semi-supervised learning procedure is started with a combination of a few positive training seeds and a relative large set of unlabeled data. The main merit of the proposed framework is its ability to mine automatically training samples from previous answer set and to refine gradually ranking model during cooperative training phase. In addition, as an extension of the proposed framework, multiple modalities can be potentially combined for effectively learning user's query intention. Following the guideline of TRECVID' 06 video search task, we validate the effectiveness of our proposed method.

***Keywords*— cooperative, learning, interactive, video, retrieval, SVM**

I. Introduction

In the video retrieval field, the key issue is how to effectively bridge semantic gap between low level feature and high level concept. As a kind of solutions, the interactive search techniques can alleviate this problem to some extent via real-time user intervention. In recent years more and more researchers and organizations focus on the research of interactive video search. TRECVID [13], a well-known community in the video retrieval field, also lists the interactive search as a subtask of video search. The latest interactive techniques formalize the interaction process as learning a retrieval function from training samples labeled by users [1,2,3]. Most closely related is the interactive framework proposed by C. Snoek et al. [1]. They treat the interactive search as a combination of querying a lexicon-based search engine and learning a retrieval model from feedback data. In particular, with a lexicon-based search engine the user can obtain an initial ranking by selecting a query topic from a set of total 106 query interfaces. After a certain number of training examples are labeled from the initial answer set by user, one-class SVM is exploited to learn a new retrieval model from them. In their scheme, the returned results from multiple query interfaces can also be integrated into a unified ranking.

While many efforts have been made upon machine learning methodology for the interactive video search, most approaches are based on supervised learning approaches and require labeling a large number of samples via user intervention. Unfortunately, no users are willing to spend too much time labeling data. To address this problem, some attempts have been made to simplify the labeling task. M.Y. Chen et al. [4] put a pool-based active learning method into the domain of interactive video search. The main idea is to narrow the range of answer set needed to be labeled by extending next training set based on the past answer set. However, the approach can only show its efficiency after a minimum of two feedback processes.

As mentioned above, previous work focuses mainly on learning a ranking function using solely labeled data, namely supervised learning. Although the paradigm of utilizing supervised learning methodology has achieved quite good performance, it requires much more labeled samples. In addition, all learners, constructed for various modalities, are independent during training process.

To deal with those problems above, we present a cooperative training framework for learning the ranking function in a semi-supervised learning fashion. In this framework, we take SVM algorithm as the underlying classifier. After given a few positive samples as the training seeds, the proposed learning scheme can automatically find out additional positive examples from current answer set and update the training set on each view iteratively. Moreover, multiple learners can also contribute to each other during the training phase.

II. Problem Analysis

When we design an interactive search system, two important factors must be taken into account. First, users are less willing to spend too much time labeling data in a real world search scenario. Hence, it is crucial to alleviate the burden on users without decreasing the search quality of system. Second, users are usually interested in a very small set of relevant shots. Therefore, it is necessary to have high accuracy on top returned shots after user intervention. Before achieving the goals above, we first analyze the feasibility of proposed scheme. As a matter of fact, any interactive search systems require an initial answer set to provide entrance for user interaction. Unexceptionally, a text-based search engine, which is based on a powerful program package named Lemur toolkit [11], is developed to return an initial ranking of 1000 shots. However, it is necessary to analyze the quality of initial search before designing an interactive scheme. NIST TRECVID provides 24 search topics for all participants to test the performance of their retrieval system. We make statistics on the average numbers of relevant shots over these 24 topics at different depths. Without loss of generality, we plot the statistics results on our text retrieval system (BJTU) and all 76

1-4244-0983-7/07/$25.00 ©2007 IEEE

runs, respectively, as shown in Figure 1. The approximate likeness of bins indicates that our text-retrieval system is representative.

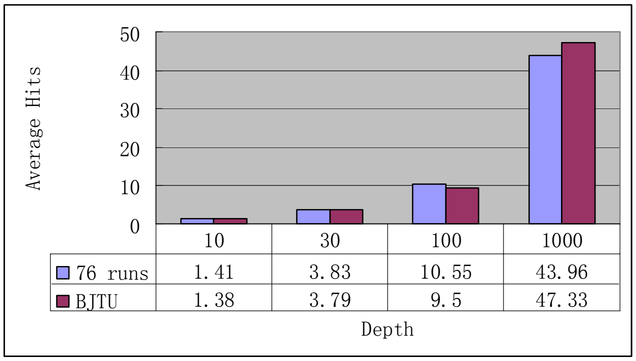

Figure 1. Average numbers of relevant shots after X shots have been retrieved

Analyzing the data in Figure 1, while there are a lot of relevant shots at a large depth (e.g. depth =1000), these shots scatter over the whole result set and the relevant shots are scarce in the top-ranked shots. These make it time-consuming to label a large number of positive examples. Therefore it is essential to develop an approach that can automatically mine training samples, given a few training seeds.

III. COOPERATIVE LEARNING FOR INTERACTION

The main idea of the proposed approach is to automatically mine training samples from initial answer set so as to alleviate the burden on users and more effectively learn user's query intention. The general framework of the proposed scheme is illustrated in Figure 2. We will describe each component in more detail.

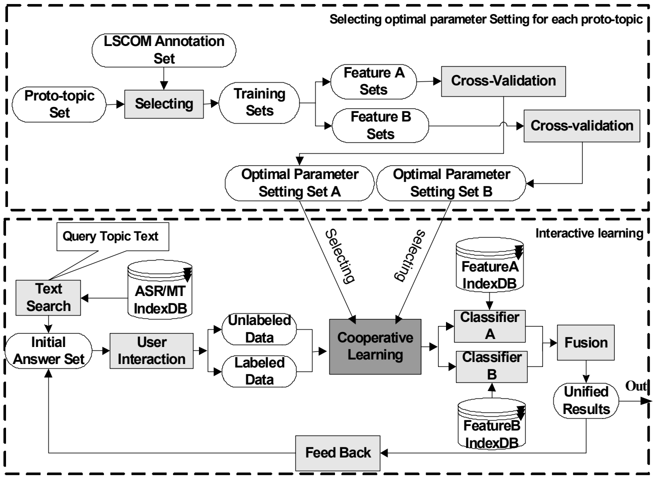

Figure 2. The framework of proposed interactive video search

A. Cooperative Learning Scheme

The core problem is how to construct a learning strategy for modeling user's query intention with a few labeled samples. For this purpose, a multi-view cooperative learning strategy is explored to automatically mine positive examples from initial search results after user labels a few relevant shots. The idea of multi-view learning was first suggested in [7,8], and R. Yan et al.[5] applied it to concept detection of video shot. In our case, multi-view strategy is extended to interactive search application for the purpose of alleviating user labeling burden. An important difference between our scheme and traditional Co-Training is how to exchange labeled samples between classifiers on different views. Table 1 shows overall flowchart of the proposed learning strategy.

TABLE I. COOPERATIVE LEARNING SCHEME

```
Inputs an initial answer set $R_0$, the number of feedback $M$ and the number of iteration T
for i = 1 to $M$
1). $R_{i-1}=P_{i-1} \cup U_{i-1}$, $P_{i-1,A}=P_{i-1,B}=P_{i-1}$
2). Selects negative data set $N_{i-1}$ randomly
3). For j = 1 to T
    $C_{i,A}^{j}$ = TrainSVM ($P_{i-1,A}$, $N_{i-1}$, A)
    $C_{i,B}^{j}$ = TrainSVM($P_{i-1,B}$, $N_{i-1}$, B)
    Updates $P_{i-1,A}$ using the output of $C_{i,B}^{j}$ on $U_{i-1}$
    Updates $P_{i-1,B}$ using the output of $C_{i,A}^{j}$ on $U_{i-1}$
4). Outputs $C_{i,A}^{T}$, $C_{i,B}^{T}$
5). $R_i = \{ F_i(D) = \alpha C_{i,A}^{T}(D) + \beta C_{i,B}^{T}(D) \}$
Output $C_{M,A}^{T}$, $C_{M,B}^{T}$
```

Specifically, after an answer set R_i is obtained, the user then labels a small set P_i of positive examples as training seeds and leaves the others as the unlabeled data set U_i .. The negative data set N_i is just selected from database randomly. During the training phase for each feedback, two learners are trained separately on each view of P_i and N_i iteratively. This process is formulated as follow:

$$C_{i,v}^{j} = \text{TrainSVM}\ (P_{i,v}, N_i, \text{v}) \qquad (1)$$

where, v ∈ $\{A, B\}$ denotes the feature view, i is the i^{th} feedback process, j is the j^{th} iteration, $C_{i,v}$ is the classifier on view v, $P_{i,v}$ is the sample set for training classifier on view v.

As an important step, selecting reliable training samples from output of the other classifier on U_i have a direct impact on the final learning performance. Here, training sets on individual views are updated separately, instead of retaining a common training set for all classifiers as Co-training does. For instance, using the label information of $C_{i,A}$ on U_i, the most likely positives of U_i are added only into the $P_{i,B}$ .. However, our sample exchanging strategy across different views is also different with so-called Co-EM algorithm which trains one classifier using directly the assigned labels from the other classifier on U_i.

To fuse the outputs from two view classifiers, we use linear weighted score to integrate the search results from two learners, which is defined as follow:

$$F\,(D) = \alpha C_A(D) + \beta C_B(D) \qquad (2)$$

where D denotes dataset, $C_A(D)$ stands for the returned ranking on view A, $C_B(D)$ indicates the returned ranking on view B, α and β are constants, usually $\alpha \le \beta$.

B. Optimal Parameter Selection

In this scheme, SVM with RBF kernel function [12] is employed as underlying learner for the cooperative learning. As a matter of fact, the parameter setting for SVM significantly influences classification performance of video information [1]. Unfortunately, it is difficult to know in advance which setting is optimal for a specified query topic [6]. To address the problem, we proposed a simple but effective method here. Specially, a series of representative query topics are selected as proto-topic set first, and then a parameter setting is obtained separately on each view for each proto-topic using cross-validation and grid-search methods against training set. 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. The upper box in Figure 3 shows the procedure. Note that each query topic is treated separately as a proto-topic in our case due to the focus of this scheme on learning strategy. We leave automatic topic mapping from query topics to proto-topics for future studies.

IV. Experiments

To construct the experiments, we employ the NIST TRECVID'06 benchmark, which is composed of approximately 343 hours of MPEG-1 broadcast news video, 169 hours for TRECVID'05 dataset viewed as training set in TRECVID'06, 174 hours as test set. Together with this corpus, the LSCOM workshop [9] provided the ground truth of annotation for the TRECVID'05 development set, and Fraunhofer Institute [10] provided the master shot reference for all data as well. In addition, the automatic speech recognition (ASR) output and machine translation (MT, Chinese/Arabic->English) output are distributed with this corpus by NIST.

A. Experiment Setup

In our experiment, 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.

In the video retrieval, 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 or feature A, and the other is the text vector or feature B.

Concerning visual descriptor, we employ a color histogram with 36 dimensions in HSV space for each keyframe. To obtain text descriptor, we first select 78 concepts from concept ontology of LSCOM to build a proto-concept set. Those proto-concepts are carefully selected so as to cover broad categories from generic concepts to specific objects. After that, a training set of 40 shots with corresponding speech transcript text is then chosen for each proto-concept against the annotation ground truth. Finally for each shot a 78-D text vector can be constructed by individually measuring similarity between the shot and the training sets of proto-concepts.

B. Performance Evaluation

For verifying the search performance of proposed approach, the 24 search topics of TRECVID'06 are employed and a set of 1000 total shots is returned for each topic.

In this paper, our aim is to develop an algorithm which can effectively alleviate the burden on users by labeling only a small set of positives, and give high accuracy on top-ranked shots. Hence we use the precision at 9 document cutoff values to evaluate the effectiveness of this interactive scheme.. The precision is computed after a given number of documents have been retrieved, which reflects the actual system performance at different depths. Note that the precision after X documents, here, is the precision average over all of 24 topics.

We carried out the proposed interactive system by labeling only 5 positive examples, which is a quite small set. As shown in Figure 3, the average precision after X documents of interactive search is far higher than automatic text-based search within top 200 returned results, which indicates that the proposed approach do bring up the true relevant results in the initial ranking.

The last series of experiments are designed to compare the search quality of different interactive learning schemes. Consider that retrieval variability is dependent on both the ranking algorithm and the implement details, it is difficult to compare search performance across different interactive schemes. Hence, only some schemes available are compared with proposed interactive scheme under the same conditions. We describe those schemes in detail as follows:

Textual feature + SVM: textual feature is only extracted from shots of the training set to train the classifier and to rank the candidate documents separately.

Visual feature + SVM: color vectors of training shots are only 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.

To show the effectiveness of the proposed scheme, ten positive examples are manually labeled by user for individual supervised learning schemes, which are twice as large as the proposed method.

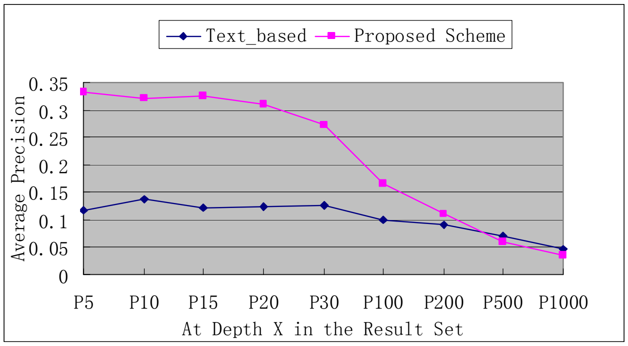

Figure 3. Interactive search VS. Automatic search at depth X in the result set

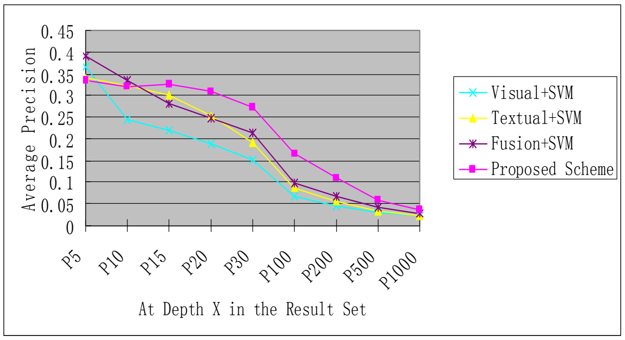

Figure 4. Performance comparison of different interactive learning schemes

To evaluate the rankings and provide a fair comparison, the same ground truth, generated by NIST when evaluating the search task, is used to judge if the result is relevant. The final evaluation results are shown in Figure 4. As we can see that the proposed scheme performs better than the others methods even if its training set is smaller than them, which suggests that the proposed scheme do mitigate the burden on users and enhance the final search quality at the same time. Figure 4 also demonstrates that the performance of textual feature based scheme is almost equal to the fusion scheme, which indicates the effectiveness of our proposed extraction scheme of textual feature.

V. Conclusions

In this paper, we developed an interactive video search scheme based on a cooperative learning strategy. This scheme utilizes the unlabeled data by explicitly splitting the feature space into two approximately independent views. The virtue of this approach is its ability to automatically mine positives from past unlabeled answer set. In addition, learners can contribute to each other by using the label information from different views. The experimental results show that our scheme works better than the supervised single-view algorithms and reduces a need for labeled data.

Acknowledgment

The This work was funded in part by NSFC (Nos. 60602030), 973 Program (No. 2006CB303104), Program for New Century Excellent Talents in University, Research Foundation of Beijing Jiaotong University (No. 2005SM013), Open Foundation from National Laboratory on Machine Perception and Open Foundation of National laboratory of Pattern Recognition.

References

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

[2] A.G. Hauptmann et al., “CMU Informedia’s TRECVID 2005 Skirmishes,” In *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID, Gaithersburg, USA, 2005.

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

[4] M.Y. Chen et al., “Putting Active Learning into Multimedia Applications: Dynamic Definition and Refinement of Concept Classifiers,” In *International Conference on Multimedia*, ACM, Singapore, pp.902-911, 2005.

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

[6] 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.

[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] 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.

[10] 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

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

[12] 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..

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