FEATURE BASED FUSION APPROACH FOR VIDEO SEARCH

Ameesha Reddy[1], B. Sivaiah[2], Rajani Badi[3], Venkateshwararaju Ramaraju[4]
[1]Department of CSE, CMR College of Engineering and Technology, Hyderbad, Andhra Pradesh, India.
[2]Associate Professor, [3]Associate Professor, & [4]Professor,
Department of CSE, CMR College of Engg. and Technology, Hyderbad, A. P., India

ABSTRACT
The main objective of every search engine is to satisfy the users' informational need. So we need to consider the behavior of user when we are designing a search engine. Based on some analysis, users are not patient enough to look at the entire result set that is given by search engine. Hence it is very crucial to achieve more accuracy on the documents that are ranked on top than improving the performance of search on the whole result set. Even though there are various methods in order to boost the performance of video search, they pay less attention in achieving high accuracy on the top ranked documents. In this paper, we are presenting a flexible and effective reranking method, called Cross-Reference (CR) Reranking to enhance the performance of retrieval of videos. In order to get more accuracy on the results of the top ranked documents CR-Reranking uses cross-reference method to combine or fuse various features. First, the initial result is reranked at the cluster level separately, based on multiple features. Further all the ranked clusters from different features are used cooperatively to get the final result set with high relevant documents ranked at the top.

KEYWORDS *- Clustering, Content Based Video Retrieval, multimedia databases, Ranking.*

I. INTRODUCTION

Today most of the people are opting Data mining domain as their research field. Specifically, Content-based video retrieval (CBVR) has drawn lot of attention from many people in recent years.The current approaches for video search are limited to text based search, which processes the keyword queries against the tokens of the text that are associated with the video, such as speech transcripts, closed captions, and recognized video text. However, such textual information may not necessarily be useful for image or video sets. The use of other features such as content of the image, audio, face detection and high-level concept detection has shown that it improves the video search than the text based video search systems. Video contains several types of audio and visual information which are difficult to extract and combine in common video retrieval. In video search, the quality of the video is of at most important to the user. The search for videos in web is extremely challenging for the web search engines. The capabilities of video retrieval of conventional search engines were limited when they were devoid of the capacity to understand contents of the media. There is abundant space to enhance the traditional techniques in the field of video search. Due to the growing profusion of digital video contents, competent techniques for analysis, indexing, and retrieval of videos which are based on their contents have gained more significance. Previously many retrieval models have been developed in order to improve the quality of video search. But the search procedure implemented by most of these models are based on the similarity between the query that is given by the user and the shots in the database based on some low level features [1]. Hence the result set to the query of the user is obtained only based on some of the low level features. However, mostly this similarity is not consistent with the human expectation due to the drawbacks of current existing

techniques to understand image and video. Because of this there is a lot of semantic gap. And this semantic gap will keep increasing linearly as the data set in the search engine increases with time. Thereby this leads to rapid deterioration of search performance. But the final aim of every search engine is to improve the search performance by retrieving the top ranked relevant documents to satisfy the users' need who are rarely patient to look at the entire result set.
The rest of the paper is organized as follows: The following is about the existing and related work of video search engines. Section 2 brushes about the basic idea of CR Reranking and the steps used in the method. Section 3 elaborates the proposed CR-Reranking scheme. In Section 4, experimental results and performance analysis is given. In Section 5 and 6, we discuss the conclusion and future work.

1.1 Related Work

To improve the retrieval performance of video search engines many methods have been proposed. The main idea of Relevance Feedback(RF) [2] is to take the results that are initially returned from a given query and to use information about whether those results are relevant to perform a new query or not. The modification of the search process to improve the accuracy by using information obtained from the prior relevant documents. Below Fig1 depicts how exactly Relevance Feedback works. Pseudo Relevance Feedback (PRF) [3] which is also known as Blind Relevance Feedback, provides a method for automatic local analysis. It automates the manual part of relevance feedback, so that the user gets improved retrieval performance without an extended interaction. The method is to do normal search to find an initial set of most relevant documents, then to assume that the top "k" ranked documents are relevant, and finally to do relevance feedback. The procedure is:

1. Take the results returned by initial query as relevant results (only top k with k being between 10 to 3).
2. Select top 10-20 terms from these documents.
3. Expand query by adding these terms to query, and then match the returned documents for this query and finally return the most relevant documents.

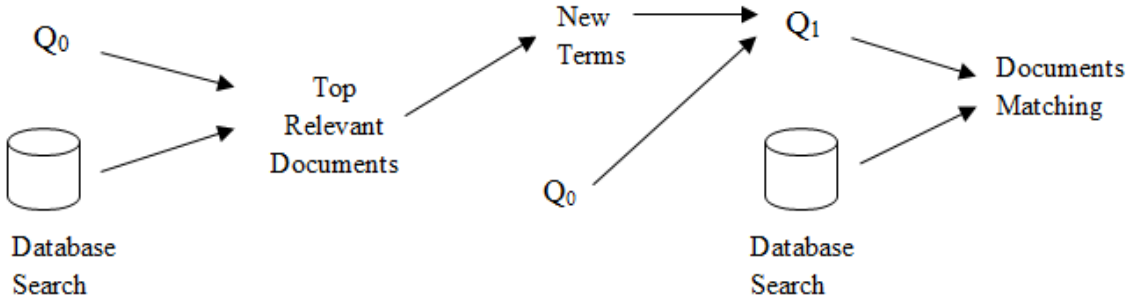

Figure 1: Relevance Feedback

Although both RF- and PRF-based methods have achieved precision improvement on the entire result list by returning more relevant shots, no mechanism guarantees that these relevant shots will be top positioned. Now-a-days, the metasearch strategy [4], [5], which is originally put forward in the field of information retrieval, is imported to CBVR for improving the effectiveness of video search. Meta search engine is a search tool that will send user requests to many search engines and/or databases and further it will aggregate the results into a single list or displays them according to their source. Different search engines retrieve many of the same relevant documents. The combination of the returned lists is performed by simply giving higher ranks to the documents that are present simultaneously in multiple result lists. Metasearch engines will enable the users to access multiple search engines simultaneously by entering their search criteria only once. As a kind of feature based fusion method, metasearch can simultaneously leverage multiple ranked lists from several search engines based on various features. However, the problem with metasearch is that it is usually hard to expect users to provide query examples with feature representations. In addition, it is not easy in practice to get access to multiple search engines based on different features.

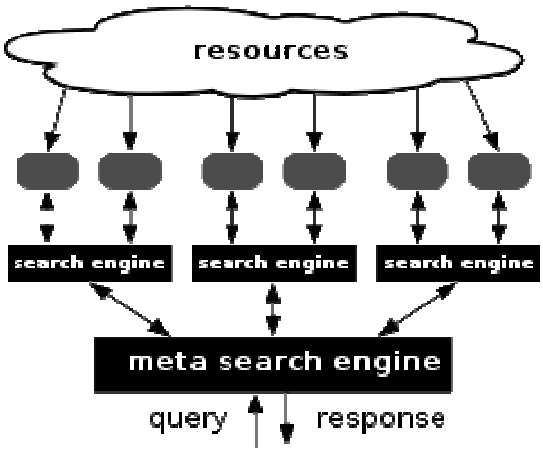

Figure 2: Architecture of Metasearch engine

II. BASIC IDEA OF CR-RERANKING

In this paper, we introduce a reranking method, called Cross Reference Reranking. This method combines multiple features in the manner of cross reference. The core idea of CR-Reranking lies in understanding the semantics or the meaning of content of video based on different features. Basically, this idea is derived from the multiview learning strategy [6], which is a semi supervised method in machine learning. In multiview learning, it will first partition the available attributes into disjointed subsets (or views), and then cooperatively uses the information from various views to learn the target model. Its theoretical foundation depends on the assumption that different views are compatible and uncorrelated. In our context, the assumption means that various features should be comparable in effectiveness and independent of each other. Multiview strategy has been successfully applied to various research fields, such as concept detection [7]. However, here, this strategy is utilized for retrieving the most relevant shots in the initial search results, which is different from its original role. CR-Reranking method contains three main stages:

1. Clustering the initial results based on each feature view.
2. Determining the rank of each cluster that is formed based on the query.
3. Fusing the ranked clusters into new result set using Cross Reference method.

In the first stage, clustering is performed by considering the multiple features individually, on the initial result set that is obtained based on the users query. Retrieval of initial result set is based on Pseudo Relevance Feedback. Now that clusters have been formed on the initial result set, they have to be ranked. In second stage of CR-Reranking, ranking of each cluster in accordance to the users query is done. Then in the last stage, we assume that the shot with high relevance should be the one that simultaneously exists in multiple high-ranked clusters obtained from different features. Based on this assumption, the shots that are high relevant can be retrieved using the cross-reference strategy and then they can be on top of the result list. As a result, the accuracy on the top-ranked documents is given more consideration. Because the "unequal overlap property" is employed implicitly, this fusion strategy is similar to the metasearch methods to a certain extent. However, our crossreference strategy differs in two ways from metasearch. The first difference is that, instead of combining multiple ranked lists from different search engines, we integrate multiple reordered variants of the same result list obtained from only one text-based video search engine. The second one is that, instead of fusing multiple lists at the shot level, we first coarsely rank each list at the cluster level, and then integrate all the resulting clusters hierarchically.

2.1 Algorithm

Step 1: Initial Result is taken and it is processed in two distinct feature views, i.e. feature view A and feature view B.

Step 2: In each feature view they are ranked in ascending order based on Euclidean distance
Let $A=\{A_1, A_2, \ldots.., A_{10}\}$

$$md(A_i, A \backslash A_i) = \min_{A_j \in A \backslash A_i} \{d(A_i, A_j)\},$$

d(.,.) is the Euclidean Distance, *md* is the smallest distancepossible.

Step 3: In each feature view all the results are first clustered into three clusters and then they are mapped
into three predefined rank levels ie., High, Medium and Low based on their relevance to the query.

Step 4: All the ranked clusters, from different features are hierarchically fused using cross reference strategy.

Step 5: Two ranked cluster sets can be integrated into a unique ranked subset list using the rule:

$$Rank(A_{high} \cap B_{medium}) > Rank(A_{medium} \cap B_{low})$$

If(high+medium) < (medium+low)

A_{high},A_{medium} are the clusters of feature view A andB_{medium} , B_{low} are the clusters of feature view B.

Step 6: When (high +medium) = (medium + high), we use the Hausdorff distance as follows:

$$Rank(A_{high} \cap B_{medium}) > Rank(A_{medium} \cap B_{high}),$$

When $hd(E, A_{high} \cap B_{medium}) < hd(E, A_{medium} \cap B_{high})$

where E is the query relevant set.

Step 6: Thus a final result set is formed and accuracy is achieved on the top ranked results.

III. FEATURE BASED RERANKING SCHEME

3.1 Overview

The overall framework of CR-Reranking is illustrated in Fig3, where the initial result list of videos is obtained according to text-based search scores. This initial result becomes the input for our proposed model. It is processed individually in two distinct feature views, i.e., feature view A and B. In each feature view clustering is performed. And we obtain three clusters in feature view A and feature view B. Then these clusters are ranked as High, Medium and Low, according to their relevance to the query. In the Fig3, we consider red as High ranked cluster, blue as Medium ranked cluster and finally yellow as Low ranked cluster. Finally, a unique and improved result set is formed by hierarchically combining all the ranked clusters from two different views. Note that only two features are considered here, however, the system can be easily extended to more features.

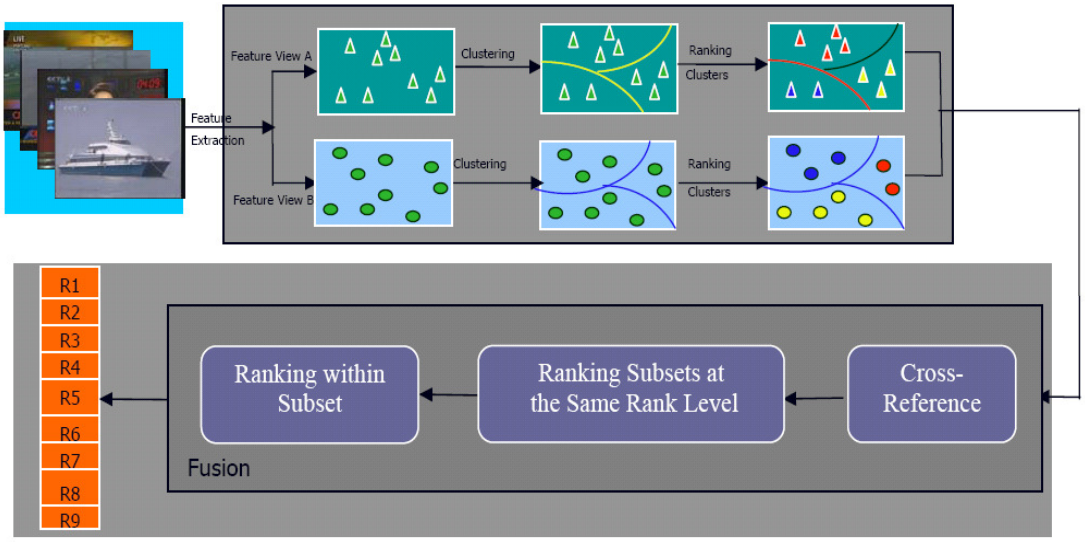

Figure 3: Architecture of the proposed Cross Reference Reranking method

3.2 Clustering in each Feature view

After extracting multiple features for each shot, we carry out clustering independently in these feature views. This provides a possibility for offering high accuracy on the top ranked documents. As a result, we can obtain a certain number of clusters from each feature view, which gives the way for implementing our cross reference strategy. An advanced Normalized Cuts (NCuts) clustering algorithm [8] is employed to cluster initial results. The experimental results show that NCuts clustering algorithm used in our scheme outperforms some other general clustering ones such as k-means.

3.3 Ranking within Cluster

After getting different clusters from each feature view, the next step in our proposed method is to rank them in accordance with their relevance to the query given by user. Some query relevant shots should be selected in advance to convey the intent of the query. Our selecting approach is also inspired by PRF method[3]. Hence top ranked initial results are considered as informative shots. Ranking is done in ascending order according to the following distance:

$$md(a_i, A\backslash a_i) = \min_{a_j \in A\backslash a_i}\{d(a_i, a_j)\}, \quad (1)$$

Here d(.,.) is the Euclidean distance and a_i and $a_j$are two different shots between which the Euclidean distance is calculated. The distance between relevant shots issmaller when compared to those distances between irrelevant shots or between relevant shots and irrelevant shots. Hence relevant shots are grouped together and irrelevant shots are scattered. Consider E as query relevant shot. 'K' shots with the smallest distances can be the most possible shots that convey the intent of the query.To measure the relevance between shot sets, we employ the modified Hausdorff distance [9], which is defined as follows:

$$hd(E, C) = mean_{e \in E}\{\min_{c \in C}\{d(e, c)\}\}, \quad (2)$$

Here E is the query-relevant set and C can be a cluster or any shot set. Note that *hd(E,C)* is a directed Hausdorff distance from E to C. Following (2), we can assign corresponding ranks to the clusters in each feature view. Here we have considered three ranks for the clusters in each feature view. The three ranks are High, Medium and Low.

3.4 Cross-Reference-Based Fusion Strategy

The final goal of our proposed method is to get high accuracy on the top ranked documents, by using improved reranking on the initial results. Thereby in order to move in the direction of this goal, we hierarchically fuse all the clusters that are ranked in different feature views in the previous step. We are using cross reference method to fuse all the clusters of one feature view with the clusters of another feature view.

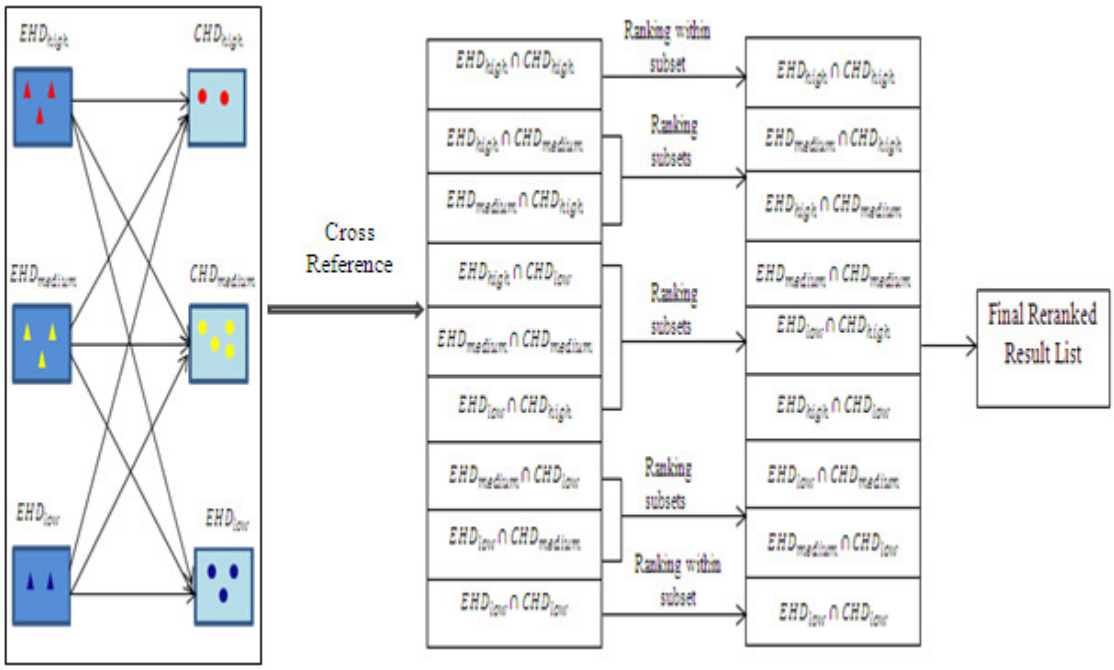

Figure 4: Proposed fusion Strategy.

Above Fig4 illustrates the schematic diagram of our fusion method with three rank levels (ie., High, Medium, Low). Our fusion approach is composed of three main components as shown in Fig4:

1. Combining the ranked clusters of one feature view with another using cross reference method.
2. Ranking the subsets with the same rank level.
3. Ranking the shots within the same subset.

In our proposed method we have taken two feature views as Color Histogram (CHD) and Edge Histogram (EHD). Some informative frames from the video are taken and stored in database. The CHD and EHD values for the stored informative frames are calculated using any feature extracting tool. The values obtained are also stored in database along with the informative shots. Note that the rank levels are denoted numerically in the following formulae for the convenience of expression. The rank levels High, Median, and Low in Fig. 4 are equivalent to the rank levels 1, 2, and 3, respectively. Here we assume that, a shot has got high rank if it is present in multiple high ranked clusters from different feature view simultaneously. Based on this assumption, we put forward a cross reference method to hierarchically combine all the ranked clusters. This gives a coarsely ranked subset list. Let $\{EHD_1, EHD_2,\ldots, EHD_N\}$ be the set of ranked clusters from feature view EHD and $\{CHD_1, CHD_2,\ldots,CHD_N\}$ be the set of ranked clusters from feature view CHD. Let *Rank* be the operation of measuring the rank level of a cluster or shot. In each ranked clusters the shots are arranged from high rank level to low rank level in ascending order of their subscripts, that is, $Rank(EHD_i)$ is greater than $Rank(EHD_{i+1})$. Then two ranked cluster sets can be integrated into a unique and coarsely ranked subset list based on the following rule:

$$Rank(EHD_i \cap CHD_j > Rank(EHD_m \cap CHD_n),$$
$$if(i+j) < (m+n), i, j, m, n = 1, \ldots., N, \quad (3)$$

Where N is the number of clusters, and $EHD_i \cap CHD_j$is the intersection of the shots $EHD_i$and CHD_j.

The rank levels of subsets cannot be compared using merely the above criteria if $(i + j)$ is equal to $(m + n)$, just like the intersections $(A_1 \cap B_2)$ and $(A_2 \cap B_1)$. To deal with this issue, we employ the method used in the cluster ranking step to order those subsets, which can be formulized as follows:

$$Rank(EHD_i \cap CHD_j > Rank(EHD_m \cap CHD_n),$$
$$if(i+j) = (m+n), hd\big(E, EHD_i \cap CHD_j\big) < hd(E, EHD_m \cap CHD_n) \quad (4)$$

Here $hd(.,.)$ is the Hausdorff distance which can be computed for any feature view.

Till now we have formed an ordered subset list. Although the ranks of shots in different subsets can be compared by the ranks of their corresponding subsets, we do not know which shot within the same subset is more relevant to the query. Hence, we need to find a method to order the shots within the same subset, i.e., ranking at the shot level. Here, the score or rank information of the initial ranking is used to order these shots. The ranking rule is defined as follows:

$$(d_m) > Rank(d_n), if(S_m > S_n), \quad (5)$$

Whered_m andd_n denote shots *m* and *n* within the same subset, respectively, S_m and S_n correspond to the scores orranks of shots *m* and *n*, respectively.

IV. RESULTS AND DISCUSSION

4.1 Data Set and Evaluation Criteria

We experimentally validate our reranking scheme on the NIST TRECVID'06 benchmark data set. The data set consists of approximately 343 hours of MPEG-1 broadcast news videos, which is divided into 169 hours of development videos and 174 hours of test videos. In all experiments, only the test data set is used for evaluation.

In scenarios of video search, a shot is considered as the fundamental unit. Hence, feature extraction is based on shots. For each video shot, two features are extracted: Edge Histogram (EHD) and Color Histogram (CHD). For the performance evaluation, TRECVID suggests a number of criteria [10]. Three of them are employed in our evaluation, including precision at different depths of result list (Prec_D), noninterpolated 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 true relevant shots in the top-*i* returned results. Then, these evaluation criteria can be defined as follows:

$$Prec_{D(T_n)} = \frac{1}{D}\sum_{i=1}^{D} F_i \quad (6)$$

$$AP(T_n) = \frac{1}{R}\sum_{i=1}^{S}(\frac{R_i}{i}.F_i) \quad (7)$$

$$MAP = \frac{1}{N}\sum_{n=1}^{N} AP(T_n) \quad (8)$$

where T_n is the nth query topic, F_i= 1 if the *i*th shot is relevant to the query and 0 otherwise, *R*stands for the total number of true relevant shots, and N denotes the number of query topics.

Prec_D is utilized to assess the precision at different depths of the 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 the query topics. Note that only the top-100 shots in the reranked result list are considered for computing both AP and MAP.

4.2 Text-only Baseline

The basic idea of text-based video search approach is to convert video retrieval into text document search. When a query text is given by users, the system returns a set of approximately relevant video shots by matching the text of the query with the text of the documents that are associated with the video shots. Using a fully automatic text-based video search engine, we can obtain an initial search list of 1,000 shots for each query topic. Reordering the initial list, our proposed reranking scheme leads to high accuracy on the top-ranked results.

4.3 Number of Clusters

In our case, the number of clusters is identical to the number of rank levels used in cluster ranking stage. Generally, varying cluster number should not have a significant impact on the reranking performance, as stated in [11]. However, the performance of proposed method is sensitive to the number of clusters due to the limitation of cluster ranking. As stated in Section 3.3, the clusters can only be coarsely ranked according to their similarity to a noisy query relevant shot set E. If the initial results are partitioned into too many clusters (or rank levels), the effect of noise will significantly violate the correctness of cluster ranking, which thereby deteriorates the reranking performance.

The experimental results are shown in Table 1. Asexpected, increasing the number of clusters leads to worse performance, and the search quality is even worse than the text-only baseline when the cluster number is greater than 15. In the following experiments, the number of clusters is fixed to 3 unless noted otherwise.

Table 1. Comparison of Different Cluster Numbers

System	MAP	Gain
Text-only baseline	0.0333	0.0%
NCut+3	0.0454	36.3%
NCut+5	0.0443	33%
NCut+10	0.0378	13.5%
NCut+15	0.0328	-1.5%
NCut+30	0.0262	-21.3%

4.4 Performance Analysis on all Query Topics

In this section, we evaluate our proposed scheme on varied query topics. Fig. 5 illustrates the statistics on APs across 24 query topics used in TRECVID'06 evaluation. The results show that the proposed reranking scheme works well for named persons and named objects, such as "D. Cheney" and "Boats," as the search quality on these topics can benefit from the TEXT feature used in our scheme. Moreover, our approach is also suitable for some query topics that are of distinctive visual properties, such as "soccer goalposts" and "scenes with snow."

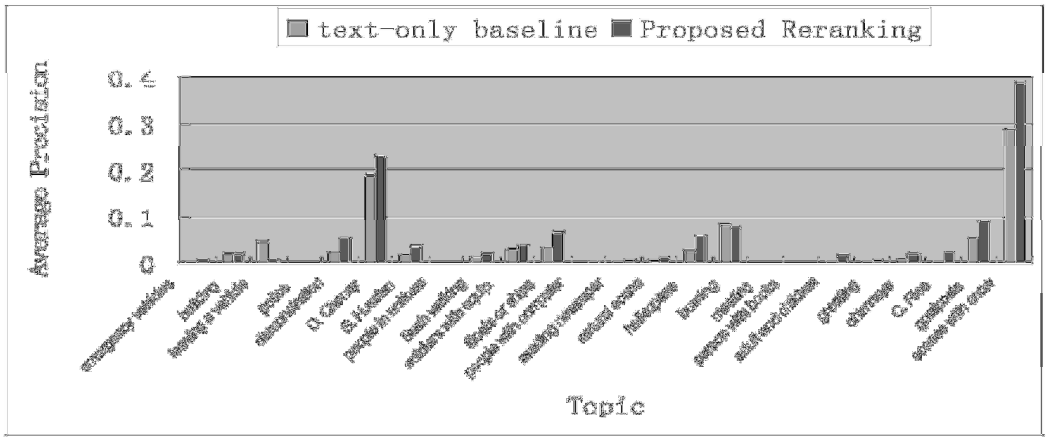

Figure 5: Performance of the proposed reranking scheme and the text-only baseline across all the 24 query topics of TRECVID 2006.

However, the search performance after reranking is even below the performance of text-only baseline for some topics with motion properties, like "leaving a vehicle." The reason is that features used in our scheme lack the capability to capture motion properties in video. Hence, new research fruits in precise representation of shot will provide much more room for performance improvement. In addition, our proposed method also fails in some query topics with very few relevant shots within the top-30 results, such as "meeting" and "people with uniform." It is because cluster ranking is based essentially on the relevant shots within the top-30 results. Incorrectly ranked clusters will deteriorate the reranking performance.

V. CONCLUSION

In this paper, we introduce a new reranking method called Cross-Reference Reranking. It combines the multiple features using cross reference strategy. First initial result set is obtained based on users query. Then clustering is performed on it by considering multiple features individually. Clusters thus obtained are mapped to predefined ranks, according to their relevance to the query. Cross-Reference strategy hierarchically fuses these ranked clusters from all feature spaces. Hence a unique and improved result set is obtained. The effectiveness of top ranked documents is improved as we have

considered multiple features. This improves the efficiency of video search engine and users can access the query relevant documents on the top ranked result set without searching the entire result set.

VI. FUTURE WORK

As analyzed previously, the proposed re ranking method is sensitive to the number of clusters due to the limitation of cluster ranking. In the future, it can be extended to develop a new method to adaptively choose cluster number for different feature views. In addition, new strategies can be investigated for selecting query-relevant shots, e.g., using pseudo negative samples to exclude irrelevant shots.

REFERENCES

[1]. M.S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-Based Multimedia Information Retrieval: State of the Art and Challenges," ACM Trans. Multimedia Computing, Comm., and Applications,vol. 2, pp. 1-19, 2006.
[2]. C.G.M. Snoek, J.C. van 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, and M. Worring, "The MediaMill TRECVID 2005 Semantic Video Search Engine," TREC Video Retrieval Evaluation Online Proc., 2005.
[3]. W.H. Hsu, L.S. Kennedy, and S.-F. Chang, "Reranking Methods for Visual Search," IEEE Trans. Multimedia, vol. 14, no. 3, pp. 14-22,July-Sept. 2007.
[4]. J.H. Lee, "Analyses of Multiple Evidence Combination," ACM SIGIR Forum, vol. 31, pp. 267-276, 1997.
[5]. J.A. Aslam and M. Montague, "Models for Metasearch," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 276-284, 2001.
[6]. A.Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," Proc. 11th Ann. Conf. Computational Learning Theory, pp. 92-100, 1998.
[7]. R. Yan and M. Naphade, "Multi-Modal Video Concept Extraction Using Co-Training," Proc. IEEE Int'l Conf. Multimedia and Expo, pp. 514-517, 2005.
[8]. J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888-905, Aug. 2000.
[9]. D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge, "Comparing Images Using the Hausdorff Distance," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 15, no. 9, pp. 850-863, Sept. 1993.
[10].TRECVID, "TREC Video Retrieval Evaluation," http://www-nlpir.nist.gov/projects/trecvid/, 2009.
[11]. W.H. Hsu, L.S.Kennedy, and S.-F. Chang, "Video Search Reranking via Information Bottleneck Principle," Proc. 14th Ann. Int'l Conf. Multimedia, pp.35-44, 2006.

AUTHORS

Ameesha Reddy was born in 1989 in India. She received Bachelor of Technology degree in Computer Science and Engineering from Jawaharlal Nehru Technological University in India in 2010. Now she is pursuing Master of Technology degree in Computer Science and Engineering from Jawaharlal Nehru Technological University in India.

Borra Sivaiah was born in 1979 in India. He received Bachelor of Technology degree in Computer Science and Engineering from Jawaharlal Nehru Technological University in India in 2002. And recevied Master of Technology degree in Computer Science and Engineering from Jawaharlal Nehru Technological University in 2010. He is currently working as Associate Professor in CMR College of Engineering and Technology in Hyderbad in India.

Rajani Badi was born in 1979 in India. She has received Master of Science degree in Mathematics from Osmania University in 2005 in India. She has received Master of Technology degree in Computer Science and Engineering from Sathyabama University in 2008 in India. She is currently working as Associate Professor in CMR College of Engineering and Technology in Hyderbad in India.

Venkateshwarla Rama Raju was born in India in 1965. Did his honors in electronics B.Sc. (Hon's) from the Osmania University college of science in 1984, Hyderabad, then he did post B.Tech in Computer Science &Engineering in 1988 from the Hyderabad central university(HCE), a M.Tech in artificial intelligence &

computer science in 1992 from Jawaharlal Nehru University (JNU) New Delhi. After spending time in industry both research & development and in Academics, he moved to England (UK), there he did his M.S in natural language processing & artificial intelligence from the university of Sheffield (1995), M.Phil in Biosignal processing and computational neuroscience from university of Leicester (1997) and a PhD in brain signal processing in 2009 from the Nizam's institute of medical sciences (NIMS, a university established under the state act 1989).He has more than 230 publications in different journals and conference proceedings, symposiums, colloquiums, etc.