Volume 3, Issue 6, June 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Generic Framework for Video Search Using Feature Extraction and Annotation

Sunil Parihar*, Kshitij Pathak
MIT Ujjain, RGPV
Ujjain, India

Abstract—The continuous growth of computer technologies and Internet has made multimedia retrieval an active search area. Particularly searching for a video from massive video database effectively and accurately has become a challenging task. In this paper we provide a generic framework to find videos from a large database using feature extraction and annotations. Firstly video frames are extracted from the video and then low level feature of frames are extracted to store in video database. Secondly all the valid and related keywords and description are associated to the video and store them into database. On the basis of these feature vectors, tags and keywords that are stored in database, the proposed system retrieves the relevant videos from the database. The proposed system composed on the idea of content based video retrieval, low level feature extraction, search by text and search by example. Experimental results show that proposed method perform better than most of the video searching and reranking techniques in terms of both effectiveness and efficiency.

Keywords— Video search, color feature, content based video retrieval, feature extraction and annotation.

I. INTRODUCTION

As the demand of multimedia contents increasing rapidly over the Internet there is a need for multimedia library that can store different kind of multimedia contents. There are various popular multimedia libraries are available over the Internet. Retrieving correct information from these massive multimedia libraries has become an active research area. Videos can be searched by the semantics and text that are associated with it such as text annotation, descriptions, closed captions, or video optical character recognition text. Content based multimedia information retrieval methods are useful when text annotation and description are missing or incomplete. Moreover content based information retrieval methods can improve multimedia retrieval accuracy even when text annotations and descriptions are present by providing additional insight into the multimedia collections. Adding multiple modalities such as image features, audio, face detection, and high level concept detection can improve text based video search systems [1], [2]. There are numerous video searching methods, multimodal architectures and reranking schemes have been proposed to improve the performance of the search engines but they either time consuming or do not fulfill the user's information need.

The purpose of any search engine is to fulfill user's need and to provide accurate search results to the user. The behavior of user should also be considered while designing any search engine. Earlier work [3], [4] shows that users are highly interested in visiting top ranked documents they hardly check the entire result list. Thus it is more important to provide high accuracy on the top ranked documents. All existing techniques for video search and reranking schemes either pay less attention on retrieving most relevant documents on top of the result list or encounter difficulties in practical application. So there is a need for new framework which can deal with these problems. In this paper we propose, design and implement a searching technique that not only improves the search performance but also fetch the most relevant documents.

II. RELATED WORK

There have been number of methods presented to improve the retrieval performance of video search engines. Earlier works [5], [6], [7], [8], [9], [10], which are mainly deals with relevance feedback strategy to improve the retrieval effectiveness. Relevance feedback technique is an interactive technique which improves the initial search result but take much time in labeling for updating the query model. Pseudo relevance feedback [11], [12] is another tool for improving initial text search results in both text and video retrieval. Pseudo relevance feedback assumes that a significant fraction of top ranked documents are relevant and uses them to build a model for reranking the search result set. Relevance feedback and Pseudo relevance feedback techniques improve the retrieval performance of video search but do not promise that these relevant shots will be relevant and top positioned. Metasearch strategy [13], [14] is also used to improve the performance of video search. Meta-search is mainly based on the unequal overlap property in which result obtained by several search engines is combined to form a single list in an optimum way and assign a highest rank to the document which are simultaneously occur in multiple result list. But the main problem associated with metasearch is that in multimodal representations it is very difficult for a user to provide query examples and also it is very hard to access multiple search engines for the same query example.

Recently the reranking methods are used to improve the video search performance. In reranking the initial search results are rearranged so that highly relevance documents can be reached on the top of the result list. Usually reranking methods are used in web search for ranking the web pages [15], [16]. In multimedia search community reranking technique is successfully implemented in IB-Reranking [17], and CR-Reranking [18]. IB-Reranking uses bottleneck principle to reorder the initial search result and it works only for single feature space while CR-Reranking based on cross reference strategy to hierarchically combine all ranked clusters from various modalities and provides better performance over IB-Reranking. In CR-Reranking firstly initial search results are clustered separately in diversified feature space then ranking the clusters with relevance to the query, and then hierarchically fusing all the ranked clusters using a cross-eference strategy. This overall process of CR-Reranking makes it more time consuming.

III. PROPOSED WORK

In this section we introduce our proposed video search framework and Algorithms. Figure 1 shows the overall framework for proposed video search solution which can be divided into two modules training module and searching module.

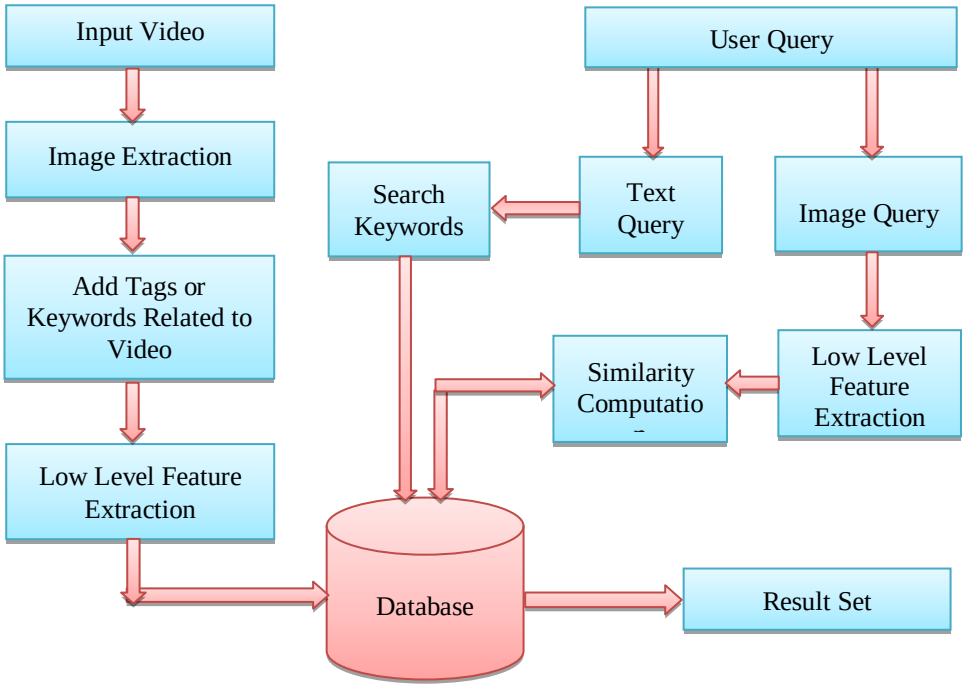

Fig. 1 Proposed Video Search Framework

A. Feature Extraction

The image is consist of colours and can be viewed as a color distribution in terms of probability theory. In probability theory, a probability distribution can be exclusively characterized by its moments. Thus the moments can be used to characterize the color distribution. In this work, color features from the images are extracted which are nothing but moments of the color distribution and the proposed system uses them to describe the image for image matching and retrieval [19].

The three color moments namely mean, standard deviation, and skewness have been proved to be efficient and effective in representing color distribution of images [19]. A color can be defined by three or more values, as in RGB model in which color is characterized by three channels. Color moments are calculated for each of these channels in an image. Therefore an image is characterized by 9 color moments, three color moments for each three color channels. Thus the i^{th} color channel at the j^{th} image pixel can be defined as p_{ij}. The three color moments can then be defined as:

1. Mean:

$$E_i = \frac{1}{N}\sum_{j=1}^{N} P_{ij} \quad \ldots (1)$$

2. Standard Deviation

$$\sigma_i = \sqrt{\left(\frac{1}{N}\sum_{j=1}^{N}\left(P_{ij} - E_i\right)^2\right)} \quad \ldots (2)$$

3. Skewness

$$S_i = \sqrt[3]{\left(\frac{1}{N}\sum_{j=1}^{N}\left(P_{ij} - E_i\right)^3\right)} \quad \ldots (3)$$

B. Training Module

Training module first required to provide input to the system in video file format which is processed using the EMF media library to extract intermediate video frame from the entire length of video. Now the system extract feature vector for each image and then store it into database. The annotation and tags are also associated with each feature vector of the image. Algorithm 1 describes the overall working of training module.

```
//Vid = Input video for training.
//Video = Video directory.
// n = Number of images extracted from Vid.
//Temp = Temporary directory
//tagTable = Contains all the relevant tag and keyword //associated with Vid
//K = Keywords and tags associated with videos
//E = Color Mean, σ = Standard deviation, S = Skewness
//tblData = Contains the entire feature vector for videos.

STEP 1: Insert Vid into system.
            Video←Vid
STEP 2: Extract images from Vid
            Temp←n
STEP 3: Add all related keywords and tags with Vid.
            tagTable←K
STEP 4: For i=1 to n do
            a) Calculate E for each channel
            b) Calculate σ for each channel
            c) Calculate S for each channel
            d) Store E, σ, and S in tblData
STEP 5: Exit
```

Algorithm 1: Algorithm for training module

C. Searching Module

In second module searching for the trained data sets is provided, where user is able to search any video using textual annotation as well as by the image query. If user inserts a text or keyword then database has been searched for related keywords and text and all the videos that are related to text or keywords has been retrieved and placed in result list. If user insert an image as a query then first all the feature vectors of the image has been extracted and compare with the feature vectors that are stored in the database and the relevant videos associated with matched feature vectors has been retrieved. Here the relationship is established using text, image and video by which system perform the faster and most relevant results are retrieved on the top of the result list. Algorithm 2 describes the overall working of searching module.

```
//QT = Query Type.
//Fn = Total number of tuples in tblData
//Tn = Total number of tuples in tagTable
//tagTable = Contains all the relevant tag and keyword //associated with Vid
//K = Keywords and tags associated with videos
//E = Color Mean, σ = Standard deviation, S = Skewness
//tblData = Contains the entire feature vector for videos.
STEP 1: Insert Query into system.
STEP 2: If QT = images
        {
            a) Calculate E for each channel
            b) Calculate σ for each channel
            c) Calculate S for each channel
            For i=1 to Fn
            {
                R ←Compare E, σ, S with F[i]
                If R>= Threshold
                {
                    Add video associated with F[i] to result list
```



```
                                    }
                            }
                    }
        STEP 3: If QT=Text
                    {
                            For i = 1 to Tn
                            {
                                    R ← compare QT with T[i]
                                    If R<= Threshold
                                    {
                                            Add video associated with T[i] to result list
                                     }
                            }
                    }
STEP 4: Exit
```

Algorithm 2: Algorithm for searching module

IV. EXPERIMENTS

For experiments on our proposed video search framework we applied our training algorithm on 100 video shots then we used TRACVID'06 suggested 24 query topics as an input to our searching algorithm and evaluated the results. We have then applied same 100 videos and 24 query topics on existing algorithms and evaluated the results.

A. Evaluation Criteria

For the performance evaluation, TRECVID suggests a number of criteria [20]. In this work three of them are employed for evaluation, including precision at different depths of result list (Prec_D), non-interpolated average precision (AP), and mean average precision (MAP). Here D as the depth where precision is computed. Let S be the total number of returned shots and Ri the number of true relevant shots in the top-i returned results. Then, these evaluation criteria can be defined as follows:

$$Prec\_D\ (T_n) = \frac{1}{D}\sum_{i=1}^{D} F_i \qquad \text{.... (4)}$$

$$AP(T_n) = \frac{1}{R}\sum_{i=1}^{S}\left(\frac{R_i}{i}.\ F_i\right) \qquad \text{.... (5)}$$

$$MAP = \frac{1}{N}\sum_{n=1}^{N} AP(T_n) \qquad \text{.... (6)}$$

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 [18].

B. Result Analysis

After implementation of video search framework , on the basis of proposed algorithms result analysis are required to see whether the proposed algorithms produces better results than previous algorithms. Figure 2 summarizes the evaluation results of different methods. Compared with existing method, proposed method achieves higher accuracy on the top ranked shots. As shown in figure 2 the proposed method not only performs better than existing methods at different depth of the result list but also improved overall performance (MAP) of the system.

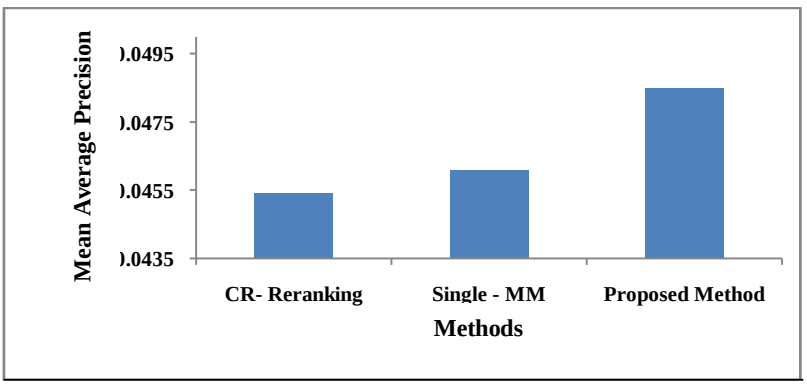

Fig 2: Performance comparison of existing and proposed methods

Now the proposed method is evaluated for different query topics. Figure 3 illustrates the statistics on averages precisions across 24 query topics used in TRECVID'06 evaluation. The results shows that proposed algorithm works well for named persons and object such as 'building', 'boats and ships', 'helicopters', 'D. Cheney' and 'C. Rice', because search quality on these topics can benefited from the text features used in proposed method. Furthermore proposed method is also appropriate for some query topics that are of distinctive visual properties, for example 'natural scene' and 'goalposts'. The Searching performance of proposed algorithm is overall better than the existing methods but for a few query topics which having moving properties, such as 'bush walking' and 'leaving vehicle', the performance is even below then CR-Reranking and text only baseline. The main cause behind this is that features used in proposed method having lacking the capability to capture motion properties in video. Moreover the proposed method also fails in searching very few query topics like 'people in uniform', 'people with computer' and 'person with book'.

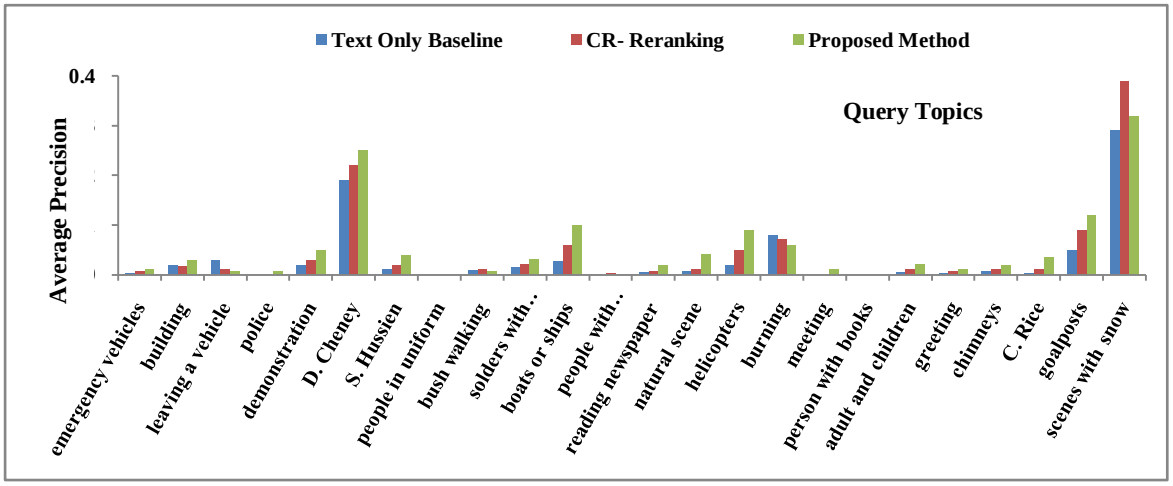

Fig 3: Performance comparison of proposed method with CR-Reranking and text-only baseline across all the 24 query topics of TRECVID 2006

V. CONCLUSIONS

In this paper we proposed a video search framework for retrieving the relevant videos from the database. The proposed method searches the videos on the basis of either text or image. The proposed work is divided into two modules training module and searching module. Training module performs video training on the basis of which searching module retrieve the videos from the database. The proposed method integrates feature-based and annotation-based retrieval approaches. In feature based approach the system extract color features from the image and compare them with the features that were already stored in the database. In annotation based approach the system search the text and tags from the database so that videos can be retrieved. The main advantage of proposed method is that it does not perform clustering, reranking and fusion on the result list as it is done in existing techniques. The proposed method searches the database and retrieves the relevant documents directly instead of working on retrieved result list. Experimental results show that searching effectiveness especially on top ranked results is improved significantly. Although overall performance has been improve significantly, the proposed method does not perform excellent for all query topics. In future the proposed method can be extended to improve the performance for all query topics by adding more motion and visual feature for large scale visual search.

REFERENCES

[1] T. S. Chua et al., *TRECVID 2004 Search and Feature Extraction Task by NUS PRIS*, TREC Video Retrieval Evaluation Online Proc., 2004.

[2] S. F. Chang et al., *Columbia University TRECVID - 2006 Video Search and High-Level Feature Extraction*, TREC Video Retrieval Evaluation Online Proc., 2006.

[3] T. Joachims, *Öptimizing Search Engines Using Clickthrough Data*, Proc. ACM SIGKDD, pp. 133-142, 2002.

[4] F. C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, *Analysis of a Very Large Web Search Engine Query Log*, ACM SIGIR Forum, vol. 33, pp. 6-12, 1999.

[5] 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.

[6] A. Amir, J. Argillander, M. Campbell, A. Haubold, G. Iyengar, S. Ebadollahi, F. Kang, M. R. Naphade, A. Natsev, J. R. Smith, J. Te_si_c, and T. Volkmer, *IBM Research TRECVID-2005 Video Retrieval System*, TREC Video Retrieval Evaluation Online Proc., 2005.

[7] S. F. Chang, W. H. Hsu, L. Kennedy, L. Xie, A. Yanagawa, E. Zavesky, and D.-Q. Zhang, *Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction,* TREC Video Retrieval Evaluation Online Proc., 2005.

[8] A.G. Hauptmann, M. Christel, R. Concescu, J. Gao, Q. Jin, W.-H. Lin, J.-Y. Pan, S. M. Stevens, R. Yan, J. Yang, and Y. Zhang, *CMU Informedia's TRECVID 2005 Skirmishes,* TREC Video Retrieval Evaluation Online Proc., 2005.

[9] J. H. Yuan, W. J. Zheng, L. Chen, D.Y. Ding, D. Wang, Z. J. Tong, H. Y. Wang, J. Wu, J. M. Lin, and B. Zhang, *Tsinghua University at TRECVID 2005,* TREC Video Retrieval Evaluation Online Proc., 2005.

[10] S. K. Wei, Y. Zhao, Z. F. Zhu, N. Liu, Y. F. Zhao, L. Zhang, and F. Wang, *BJTU TRECVID 2006 Video Retrieval System,* TREC Video Retrieval Evaluation Online Proc., 2006.

[11] 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.

[12] A. Natsev, M. R. Naphade, and J. Tesic, *Learning the Semantics of Multimedia Queries and Concepts from a Small Number of Examples,* Proc. ACM Int'l Conf. Multimedia, ACM Press, 2005, pp. 598-607.

[13] J. H. Lee, *Analyses of Multiple Evidence Combination,* ACM SIGIR Forum, vol. 31, pp. 267-276, 1997.

[14] 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.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web,* Stanford Digital Libraries Working Paper, 1998.

[16] T. H. Haveliwala, *Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search,* IEEE Trans. Knowledge and Data Eng., vol. 15, no. 4, pp. 784-796, July/Aug. 2003.

[17] 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.

[18] Shikui Wei, Yao Zhao, Zhenfeng Zhu, and Nan Liu, *Multimodal fusion for Video Search Reranking,* IEEE Trans. Knowledge and Data Eng., vol. 22, no. 8, pp. 1191-1199, August 2010.

[19] M. Stricker and M. Orengo, *Similarity of color images,* In SPIE Conference on Storage and Retrieval for Image and Video Databases III, volume 2420, pages 381-392, Feb. 1995.

[20] TRECVID, *TREC Video Retrieval Evaluation*, http://www-nlpir.nist.gov/projects/trecvid/, 2009.

