A Light-weight Relevance Feedback Solution for Large Scale Content-Based Video Retrieval

Zimian Li[1], Ming Zhu[2]

[1] The Key Lab of Network Communication System & Control, The Chinese Academy of Sciences, The Key Lab of Network Communication System & Control, Anhui
Department of Automation, University of Science and Technology of China

[2] The Key Lab of Network Communication System & Control, The Chinese Academy of Sciences, The Key Lab of Network Communication System & Control, Anhui
Department of Automation, University of Science and Technology of China

Abstract

This paper addresses the problem of large scale content-based video retrieval with relevance feedback. We analyze the common methods which leverage local feature detectors to extract feature descriptors from video collections and perform multi-level matching after indexing and retrieval of feature vectors. Instead of learning similarity-preserving codes, we introduce the relevance feedback approach in a light-weight way. A relevance model is proposed to merge semantic similarity with the original distance matching at descriptor level. By learning several weights using canonical correlation analysis (CCA), the resulting candidate list of similar videos changes according to relevance feedback. Finally, we demonstrate the improvement of the proposed method by experiments on a standard real world dataset.

Keywords: *Content-based Video Retrieval, Relevance Feedback, CCA.*

1. Introduction

With the rapid growth of digital video content production on the web, content-based video retrieval (CBVR) has been receiving increasing attention over the last decade. In the computer vision and machine learning community, many approaches focus on multimedia information indexing and retrieval techniques. Compared with individual images, videos have much richer content and therefore need a more complicated structure to describe, index and retrieve.

Recently, different methods have been proposed for video structure analysis, including shot boundary detection, key frame extraction and scene segmentation. The general procedure of existing work can be summarized as three stages. First, using shot detection methods, videos are segmented into clips, which then represented by one or more key frames. Second, a set of high dimensional feature vectors are extracted by feature detector and descriptor. Finally, the similarity between videos is computed from the feature vectors under spatial and/or temporal sequence matching schemes, see [1] for a comprehensive review.

Unlike video copy detection (VCD) or near-duplicate video detection (NDD), content-based video retrieval searches for a more semantic sense of similarity, moreover, compared with content-based image retrieval (CBIR), some additional spatial/temporal information plays an important role in matching stage. So, how to measure the similarity and to perform nearest-neighbor search are the essential problems. In this paper, we leverage the common initial strategies and focus on the semantic retrieval with relevance feedback.

Approximate nearest neighbors (ANN) search methods are used to perform nearest neighbor search in large scale retrieval, especially for high dimensional datasets. One of the most popular techniques is Locality Sensitive Hashing (LSH), which was first introduced in [3]. LSH function families have the property that objects that are close to each other have a higher probability of colliding than objects that are far apart. For different distance measures, different LSH families have been proposed, e.g. LSH for p-norms based on p-stable distribution [4]. However, recent research [5] [6] shows that the Chi2 distance often leads to better results than Euclidean metric for image and video retrieval task, especially when histogram-based descriptors, such as SIFT and SURF, are used to describe the images and video frames. In this paper, we pursue the new LSH scheme fitted to the Chi2 distance, which was introduced by Gorisse [2] for approximate nearest neighbor search in high-dimensional spaces.

Another important question is how to extract and represent semantic information from video data. In the vision community, most recent approaches concern about learning similarity-preserving binary codes. Large scale image/video collections are represented in generally two

major steps: embedding and binarization. Regressing, classification and clustering techniques are merged with indexing and retrieval approaches to generate semantic structures, e.g. Principal component analysis (PCA) based method [7], Spectral Hashing [8], Kernelized LSH (KLSH) [9], Product Quantization (PQ) [10], Linear Discriminant Analysis based method LDAHash [11] and iterative quantization (ITQ) [12]. In all these approaches, compact binary codes are learned by training examples and the performance of similarity-preserving depends on the sample representativeness. What's more, they pay more attention to content-based image retrieval and less to video scenario. Compared with image retrieval, video retrieval has additional spatiotemporal characteristics and it's almost impossible to learn one compact code to represent a video clip totally.

In this paper, we follow the general procedure: leveraging local feature detectors to extract feature descriptors from video collections and perform multi-level matching after indexing and retrieval of feature vectors. Relevance feedback techniques based on canonical correlation analysis (CCA) are introduced to bridge the gap between semantic notions of search relevance and the low-level representation of video content. The main contribution is as follows:

- We analyze the indexing and retrieval stages in the common framework of content-based video retrieval.
- We leverage the state-of-the-art techniques in content representation, similarity measure selection and multi-level matching and merge them to work in an incremental way.
- We introduce a novel light-weight relevance feedback approach to refine the original resulting list.

The rest of this paper is organized as follows. In Section 2, we present the framework of content-based video retrieval with relevance feedback. Section 3 presents the structure to index SURF descriptors using Locality-Sensitive Hashing under Chi2 distance. Section 4 introduces the light-weight relevance feedback solution using Canonical Correlation Analysis (CCA). Section 5 gives the experimental results and performance analysis of our proposed algorithm. Finally we conclude this paper and give some future work in Section 6.

2. Content-Based Video Retrieval Framework

Figure 1 illustrates the framework of content-based video detection with relevance feedback. The processing consists of three parts: Indexing, Retrieval and Relevance Feedback. Indexing videos are processed by shot detection, key-frame and local feature extraction (we use SURF descriptors in this paper) to generate a set of 64-dimensional feature vectors. Then a video database is built using an indexing structure. In the retrieval parts, the same local features extraction is performed. By retrieving in the database a candidate result set is generated and then multi-level matching methods are applied to get the final similar video result list.

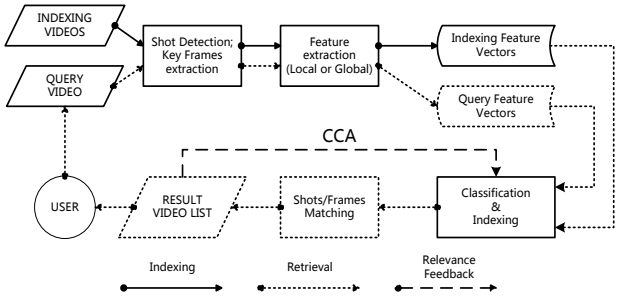

Fig. 1 Framework of CBVR with relevance feedback.

We focus on the indexing structure and relevance feedback techniques in retrieval for large-scale video collections. Different strategies have been proposed for local feature based near-duplicate video detection following the above framework. As to content-based video retrieval, which searches not the copy but the semantic similar ones, after indexing and retrieval in descriptor level, voting-based multi-level matching is not enough. Also, instead of leveraging users' feedback to learn a similarity measure or perform classification/clustering, we "delay" the semantic learning to the shots/frames matching stage.

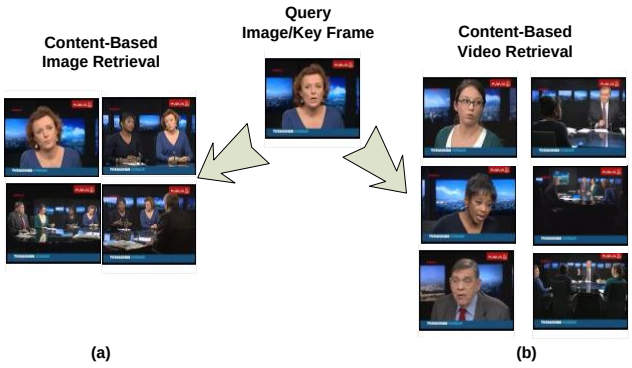

Fig. 2 Difference between CBIR and CBVR

Figure 2 shows the difference of content analysis between image retrieval and video retrieval. In (a) content-based image retrieval, where the query image contains all the semantic information, the results are more intuitive (mainly dealing with some transformations). However, in (b)

content-based video retrieval, where query video consist of many key frames(query images), for one query image, it's necessary to learn/find/create relations with other images extracted from the similar video and make up the semantic lost in representing video with image sequences.

3. Indexing SURF descriptors using Locality-Sensitive Hashing under Chi2 distance

SURF (Speeded-Up Robust Features) detector and descriptor [2] is based on calculating approximate Hessian response for image points and is efficiently implemented on the basis of integral images. SURF is proved to be equal or superior to performance and significantly better computational efficiency in comparison with other local feature methods, such as SIFT, PCA-SIFT. In this paper, we use 64-dimensional SURF descriptors as the feature vectors for key frames extracted from video collections. LSH (Locality Sensitive Hashing) is introduced in [3] for approximate nearest neighbors search in high dimensions. LSH function families have the property that objects that are close to each other have a higher probability of colliding than objects that are far apart. For different distance measures, different LSH families have been proposed. We consider the LSH for chi2 distance [2] since SURF descriptors are designed to be histogram-based descriptors measured by the Euclidean distance. We briefly describe the indexing structure in our scenario and show the characteristics of results after feature vector retrieving. In the basic LSH scheme, a query point is hashed into several buckets in different hash tables to retrieve all points in these buckets, then the distances to each point is computed using the chi2 distance (1):

$$\chi^2(x,y) = \sqrt{\sum_{i=1}^{d} \frac{(x_i - y_i)^2}{x_i + y_i}} \quad (1)$$

For each data point v, k independent hash functions of the form (2) are considered, where a is a d-dimensional vector whose elements are chosen independently from the Normal distribution and b is chosen uniformly from $[0, W]$.

$$h_{a,b}(p) = \frac{\sqrt{\frac{8a \cdot p}{W^2} + 1} - 1}{2} + b \quad (2)$$

Each hash function maps a d-dimensional data point onto the set of integers and the final result is a vector of length k of the form (3).

$$g(v) = \left(h_{a_1 b_1}(v), ..., h_{a_k b_k}(v)\right) \quad (3)$$

Thus, all points in dataset are hashed into buckets labeled by a k-dimensional vector in a hash table. To ensure the accuracy of similarity search, l independent hash tables are generated to construct the LSH indexing structure. Then, in the retrieval step, one can determine near-neighbors by hashing the query point l times to l buckets in l different hash table and retrieving elements stored in buckets containing that point.

Then the candidate feature vector set is used for retrieval on key-frame level and video shot level. Some recent approaches introduced some spatio-temporal matching and sequence matching approaches to further get the final list of similar video results. Note that in the framework based on local descriptor, the effectiveness of retrieving similar key frame (image) is insufficient for retrieving similar video. The improvements on indexing structure can only boost the retrieval efficiency. As mentioned above, similar videos have much more "semantic means" than similar images and the accuracy is defined fuzzy and depends mainly on user's opinions. We leverage the relevance feedback to learn an adaptive similarity score function, which works as a similarity measure in content-based video retrieval.

4. Relevance Feedback using Canonical Correlation Analysis

For each feature vector in each query video, a candidate set is retrieved from the indexing database. Each vector in the candidate set is associated with respective key frame and video shot. As in our scenario, the number of feature descriptors extracted from each key frame is about two hundreds, it's not computationally efficient to perform relevance feedback in descriptor level. We design a similarity function with a correlation matrix to project to semantic space in key frame level. We pick up key frames from the similar videos chosen by users, which are cached during Chi2-LSH retrieval. We represent key frames from the candidate videos as (4) and key frames from feedback video as (5):

$$F_i = \{f_i\}^n \quad (4)$$

$$F_f = \{f_f\}^n \quad (5)$$

where f is calculated by a voting method from below.

$$f_i = \sum_{N_{descriptor}} \sum_{L} w_b \cdot N_{matching} \quad (6)$$

$$f_f = \sum_{N_{descriptor}} \sum_{L} w_b \cdot N_{matching} \quad (7)$$

$N_{descriptor}$ is the number of descriptors extracted from the query frame, $N_{matching}$ is the number of matching descriptors between the two frames in one bucket and w_b is the weight of the corresponding bucket. We use the weight w_b to reduce the impact of large buckets, in which

many descriptors of the same frame match to one query descriptor of the query frame. Then we need a correlation matrix C.

$$C = \{c_{i,j} \mid i = 1,...,m;\ j = 1,...,n\} \quad (8)$$

The matrix C satisfies $F_f = CF_i$. The matrix could be learned by a supervised dimensionality reduction method to capture the result structure in semantic space.

We solve the problem by using the Canonical Correlation Analysis (CCA), which has proven to be an effective tool for extracting a common latent space from two views in a semi-supervised way. The goal of our approach is to find projection directions w_k and u_k for candidate key frame set and relevance key frame set to maximize the correlation between the projected $F_i w_k$ and $F_f u_k$. The problem is represented as:

$$\max C(w_k, u_k) = w_k^T F_i^T F_f u_k$$

subjected to:

$$w_k^T F_i^T F_i w_k = 1, u_k^T F_f^T F_f u_k \quad (9)$$

Solving the above optimal problem can use the generalized eigenvalue solution [15]. Once we get the canonical variables w_k and u_k, the optimal direction to project candidate set to relevance set is determined. Then we obtain the modified retrieval results using the similarity score function as:

$$score_c = \frac{\sum_{N_{frame}} f_m}{N_{frame}} \quad (10)$$

where $f_m \in F_m = CF_i$, N_{frame} is the number of key frames of the query video. An indexing video is considered to be a similar video of the query video if $score_c$ exceed a threshold S_t. The selection of S_t requires a trade-off between recall and precision during retrieval and is data dependent. Finally, the videos with first several highest scores are returned as the results of similar videos with relevance feedback considered.

5. Experiments

To evaluate the semantic effectiveness and robustness of our proposed algorithm, we conducted experiments using the MUSCLE-VCD benchmark [16], which is an evaluation set of the TRECVID 2008. The dataset consists of 101 videos with a combined length of about 100 hours. We divided the indexing videos into about 600 parts and the provided 15 query videos into about 100 parts. Then we performed 60 different similar video searches. Compared with the original process, we manually labeled a certain number of video parts from the candidate set as positive examples for relevance feedback. We modified the open source project E2LSH [4] provided by Alexandr Andoni to process locality sensitive hashing under chi2 distance. Experiments on precision/recall and percentage of relevance are made to illustrate the performance of our proposed method.

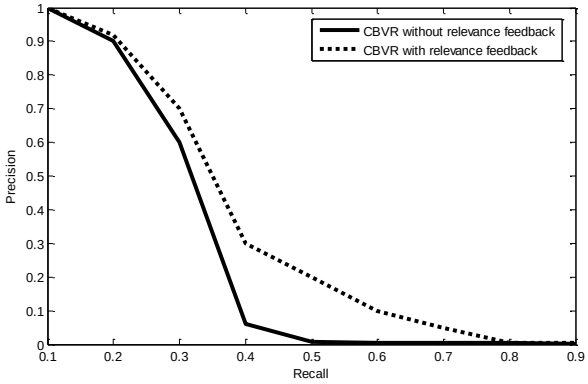

Fig.3 Precision-Recall Curve for CBVR with/without Relevance Feedback

Figure 3 shows the Precision-Recall curves on two different methods. Dot line is the CBVR with relevance feedback and plain line is the original CBVR. We can tell from figure 3 that the novel relevance feedback approach we propose can improve the recall percent for given precision. The recall percent increases 5% when the precision is below 50% and reaches up to 16.7% when the precision is above 50%. We can also see that the CBVR with relevance feedback can achieve higher precision for same recall percent. We can see that with a few similar videos labeled, when the recall percent increases, the precision decreases slower with the relevance feedback. The reason is that the original process ignores some of the similar videos only by distance calculating and locality sensitive hashing. The relevance feedback approach proposed in this paper works as an incremental tool to perform query expansion and boosts the precision in the same recall rate. The relevance feedback improves the efficiency of the retrieval.

Table 1 shows the how the feedback ratio of the candidate set affects the precision and recall percent in our experiments. We can see obviously that the relevance feedback approach could achieve high precision percent while improving recall percent by providing specific ratio of feedback information. The reason is that the artificial semantic information from the feedback of the users complements the semantic similarity loss caused by the

distance calculation fully based on the feature vectors. However, when the feedback ratio exceeds certain threshold such as 30%, the precision percent decreases substantially even the recall percent is still able to maintain increasing. This result shows that too much feedback information from the users reduces the weighting of the distances in multi-level matching and causes the fact that the video retrieval is actually fully depend on the correlation matrix learned from samples provided by the users. We make sure the feedback ratio is below 30% in our experiments in order to maximize the efficiency of relevance feedback approach instead of violating the principle of contend-based video retrieval.

Table 1: Feedback Ratio and Precision-Recall Percent

Feedback Ratio	***Precision***	***Recall***
10%	93%	13.2%
20%	92.7%	15.4%
30%	89.1%	23.4%
40%	55.8%	35.0%
50%	30.9%	40.5%
60%	34.2%	52.3%

6. Conclusion

In this paper, we leverage local feature detectors to extract feature descriptors from video collections and perform multi-level matching after indexing and retrieval of feature vectors using the state-of-the-art techniques in content representation, similarity measure selection. We introduce a novel light-weight relevance feedback approach based on canonical correlation analysis (CCA) to bridge the gap between semantic notions of search relevance and the low-level representation of video content. Experimental results on real world demonstrate the precision gains of our proposed method.

Acknowledgments

We thank Alexandr Andoni for providing his E2LSH binary and the anonymous reviews for their constructive comments and suggestions which greatly improve the quality of this paper. This research was supported by Network Video Communication and Control Project in Sensing China Program of Chinese Academy of Science under Grant XDA06030900.

References

[1] Hu, W. and Xie, N. and Li, L. and Zeng, X. and Maybank, S., "A Survey on Visual Content-Based Video Indexing and Retrieval", Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 41, no. 6, 2011, pp.797-819.

[2] Gorisse, D. and Cord, M. and Precioso, F., "Locality-Sensitive Hashing for Chi2 Distance", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 34, no. 2, 2012, pp.402-409.

[3] Gionis, A. and Indyk, P. and Motwani, R., "Similarity search in high dimensions via hashing", Proceedings of the International Conference on Very Large Data Bases, 1999, pp.518-529.

[4] Datar, M. and Immorlica, N. and Indyk, P. and Mirrokni, V.S., "Locality-sensitive hashing scheme based on p-stable distributions", Proceedings of the twentieth annual symposium on Computational geometry, ACM, 2004, pp.253-262.

[5] Chapelle, O. and Haffner, P. and Vapnik, V.N., "Support vector machines for histogram-based image classification", Neural Networks, IEEE Transactions on, vol. 10, no. 5, 1999, pp.1055-1064.

[6] Gosselin, P.H. and Cord, M. and Philipp-Foliguet, S., "Combining visual dictionary, kernel-based similarity and learning strategy for image category retrieval", Computer Vision and Image Understanding, Elsevier, vol. 110, no. 3, 2008, pp.403-417.

[7] Gordo, A. and Perronnin, F., "Asymmetric distances for binary embeddings", Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 2011, pp.729-736.

[8] Weiss, Y. and Torralba, A. and Fergus, R., "Spectral hashing", NIPS, 2008.

[9] Kulis, B. and Grauman, K., "Kernelized locality-sensitive hashing for scalable image search", Computer Vision, 2009 IEEE 12th International Conference on, 2009, pp.2130-2137.

[10] Jégou, H. and Douze, M. and Schmid, C., "Product quantization for nearest neighbor search", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 33, no. 1, 2011, pp.117-128.

[11] Strecha, C. and Bronstein, A.M. and Bronstein, M.M. and Fua, P., "LDAHash: Improved matching with smaller descriptors", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 34, no. 1, 2012, pp.66-78.

[12] Gong, Y. and Lazebnik, S. and Gordo, A. and Perronnin, F., "Iterative Quantization: A Procrustean Approach to Learning Binary Codes for Large-scale Image Retrieval", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012.

[13] M. Yeh and K.-T Cheng, "Fast visual retrieval using accelerated sequence matching", Multimedia, IEEE Transactions on, vol. 13, no. 2, 2011, pp. 320-329.

[14] C. Chiu, H. Wang and C. Chen, "Fast min-hashing indexing and robust spatio-temporal matching for detection video copies", ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 6, no. 2, 2010, Article 10.

[15] Foster, D.P. and Kakade, S.M. and Zhang, T., "Multi-view dimensionality reduction via canonical correlation analysis", Technical Report, TR-2008-4, TTI-Chicago, 2008.

[16] Law-To, J. and Joly, A. and Boujemaa, N., "Muscle-VCD-2007: a live benchmark for video copy detection", Available: http://wwwrocq.inria.fr/imedia/civr-bench/, 2007.

[17] P. Haghani, S. Michel, and K. Aberer, "Distributed Similarity Search in High Dimensions Using Locality Sensitive Hashing", In Proceedings of the 12th International Conference on Extending Database Technology (EDBT 09), ACM, 2009, pp. 744-755.
[18] M. Bawa, T. Condie, P. Ganesan, "LSH forest: self-tuning indexes for similarity search", Proceedings of the 14th international conference on World Wide Web (WWW 05), ACM, 2005, pp. 651-660.
[19] Q. Lv, M. Josephson, Z. Wang, M. Charikar, K. Li, "Multi-probe LSH: efficient indexing for high-dimensional similarity search", Proceedings of the 33rd international conference on Very large databases (VLDB 07), ACM, 2007, pp. 950-961.

Zimian Li received the B.S. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2006. He is currently pursuing the Ph.D. degree in the School of Information Science and Technology of USTC. He already published three EI-indexed papers in international conferences and one in Chinese domestic journal. His research interests include self-healing multimedia system, multimedia communication.

Ming Zhu is currently a professor of University of Science and Technology of China. He received B.S., M.S. and Ph.D. degrees in Computer Science from University of Science and Technology of China in 1986, 1989 and 2001, respectively. He became an assistant professor in 1989 and a professor of USTC in 2004. He worked as a visiting scholar in Department of Computing, The Hong Kong Polytechnic University from 1997 to 1998. He is the Director of the Key Lab of Network Communication System & Control, Chinese Academy of Sciences and the Director of the Key Lab of Network Communication System & Control, Anhui. His research interests include intelligent software systems, data mining and network security.