Million-scale Near-duplicate Video Retrieval System*

Yang Cai[†‡], Linjun Yang[‡], Wei Ping[§‡], Fei Wang[⊥],Tao Mei[‡], Xian-Sheng Hua[⊥],Shipeng Li[‡]
[†]Zhejiang University [‡]Microsoft Research Asia [⊥]Microsoft Bing [§]Tsinghua University
yangcai1988@gmail.com,{linjuny,feiw,tmei,xshua,spli}@microsoft.com,weiping.thu@gmail.com

ABSTRACT

In this paper, we present a novel near-duplicate video retrieval system serving one million web videos. To achieve both the effectiveness and efficiency, a visual word based approach is proposed, which quantizes each video frame into a word and represents the whole video as a bag of words. The system can respond to a query in 41ms with 78.4% MAP on average.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval tenance—*Search process, Selection process, Information filtering.*

General Terms

Algorithms, Experimentation, Performance

Keywords

Near-duplicate, video retrieval, large scale

1. INTRODUCTION

Near-duplicate video retrieval has been a hot research topic for the past decades, since it is highly useful for automatic video indexing, management, retrieval, rights protection, copy detection, etc. The recent rapid growth of the web videos, evidenced by the fact that on YouTube there are over 35 hours videos being uploaded every minute, imposes an urgent need for a practical large-scale near-duplicate video retrieval system, to facilitate users' consumption of the ever-increasing web videos.

While the decades' research on near-duplicate video retrieval witnessed a large amount of achievements, the practical large-scale near-duplicate video retrieval system is still not mature enough. Most of the existing methods require several seconds to minutes to respond to a normal video query on a database comprising tens of thousands of videos. The approach based on a compact video representation and an improved inverted file index, requires 17ms to query against a 50K video database [4]. However its scalability remains unknown when applied to a web-scale system, which comprises millions or even billions of web videos. The multiple feature hashing method [5] which costs 553ms to retrieve a dataset of 133K videos may still not meet the web search requirement.

*This work was performed when Yang Cai and Wei Ping were visiting Microsoft Research Asia as research interns.

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.
ACM 978-1-4503-0616-4/11/11.

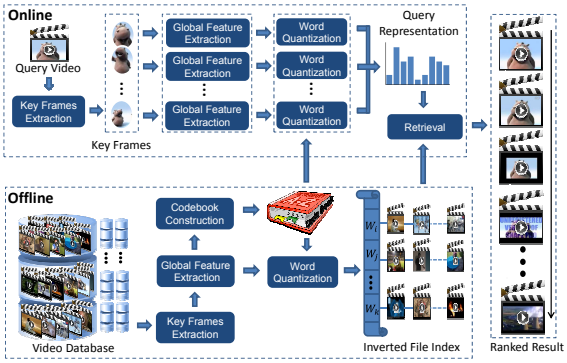

Figure 1: The framework of our proposed system.

To make a further step towards the large-scale near-duplicate web video retrieval, we construct a one million video database by crawling from the web, and based on that, propose a million-scale near-duplicate video retrieval system. To the best of our knowledge, this is the first near-duplicate video retrieval system, serving one million web videos.

The increasing scale imposes a big challenge to the retrieval system in order to achieve a reasonably high precision with a low time cost. To this end, we propose a visual word based approach by generalizing the video signature in [4]. Different from previous methods which construct visual words on each interest points detected in frames, we quantize each frame into a word, to achieve a fast and compact video representation. Specifically, the global features are extracted from each frame and then quantized into a so-called "word". Then the inverted file index is naturally adopted to index these words. While simple, our system demonstrates that it is both effective and efficient for most web video near-duplicates. The system can respond to a query in 41ms on average with a MAP(Mean Average Precision) of 78.4%.

2. THE PROPOSED APPROACH

To develop a scalable near-duplicate video retrieval system, we first need to design a compact video representation, also called video fingerprint. As suggested by [4], representing a video by a bag of visual words extracted from key frames seems to be a promising approach and can achieve a reasonable performance in a moderate scale. We implemented this method and applied it on our collected one

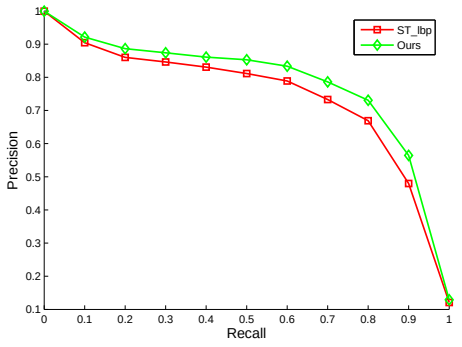

Figure 2: The performance comparison of ST_lbp [4] and ours in terms of precision-recall.

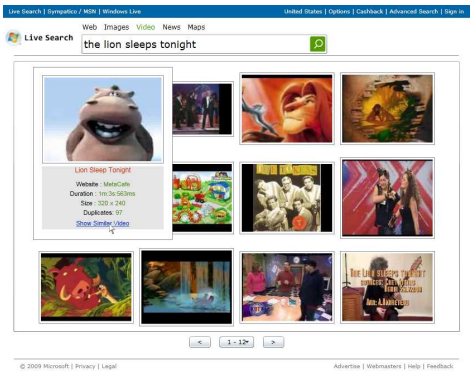

Figure 3: A snapshot of our proposed system.

million video database. However, the retrieval time for one query is still more than 300ms and a lot of false alarm videos are included in the retrieval results for some queries. There are two-fold reasons. First, the ordinal feature based video representation lacks discrimination and will lead to more false alarms in the large-scale setting. Second, due to the heuristics inherent in the quantization method in [4], the words distribute non-evenly and a small portion of words appear frequently in a large amount of videos. Hence, the inverted file index cannot speed up the retrieval too much and the retrieval time is nearly linear to the size of the database.

In this demo, we extend the visual word based approach in [4] by incorporating more complex and discriminative frame features. Instead of extracting ordinal relations from frames and encoding the ordinal relations based on heuristic, we extract more discriminative global features based on proven image processing techniques, e.g., color correlogram [1]. Then the high-dimensional frame features are quantized into words using Kmeans clustering algorithm. In this demo, a scalable Kmeans clustering algorithm [2] and the Kmeans tree search algorithm [3] are used to speed up the codebook construction and quantization. Compared to the previous methods, especially the bag of visual words on local features, the proposed approach is easier to scale up. Due to the content redundancy across frames, even though the frame representation is simple, aggregating them for an entire video will greatly improve the discrimination. Compared to [4], the application of color correlogram and Kmeans quantization can lead to a more discriminative video representation. Besides, with a large codebook being constructed, the induced video fingerprint can be made sparse and can therefore make the inverted file index more efficient.

Fig.1 illustrates the flowchart of our proposed system. Uniform sampling is employed to extract key frames every second from videos. After extracting the color correlogram features from the extracted key frames for all the one million videos in our database, we constructed 100K visual vocabulary by clustering on 1M randomly sampled key frames. The inverted file index is adopted to index the visual words and the *cosine* retrieval model with *tf-idf* weighting is employed to compute the relevance score between the query video and videos in the database.

3. SYSTEM DESCRIPTION

The system is implemented in a client-server architecture. The client is implemented using Silverlight 3.0 and the server side is implemented using Asp.Net 3.5 and C++.

In our system, a text-based video search engine is implemented based on Lemur[1]. Users can first input a text query and then select one of the returned videos in the text-based search results as query to retrieve near-duplicates. Fig.3 shows a snapshot of our demo system.

The video database comprises about one million videos crawled from the Web as well as 12K videos from CC_WEB_VIDEO dataset[2], which finally leads to 1,148,050 videos in our database.

4. EXPERIMENTAL RESULTS

We compared our method with LBP-based Spatiotemporal (ST_lbp) method [4], using the query sets in CC_WEB_VIDEO and the extended ground-truth. As shown in Fig.2, our proposed method outperforms ST_lbp in terms of precision-recall. Moreover, the retrieval time for one query is only 41ms, which is 9.4 times faster than ST_lbp (386ms). This demonstrates that our proposed approach is highly applicable for web-scale near-duplicate video retrieval, in terms of both precision and time cost.

5. CONCLUSION

In this demo we proposed a million-scale near-duplicate video retrieval system, which achieves both satisfactory precision and high efficiency by incorporating discriminative low-level frame features with visual word quantization techniques. Specifically, one query can be responded in only 41ms and the MAP can arrive at 78.4%, on average.

6. REFERENCES

[1] J. Huang, S. Ravi Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Spatial color indexing and applications. *IJCV*, 1999.

[2] D. Li, L. Yang, X.-S. Hua, and H.-J. Zhang. Large-scale robust visual codebook construction. In *ACM MM*, 2010.

[3] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISSAPP*, 2009.

[4] L. Shang, L. Yang, F. Wang, K.-P. Chan, and X.-S. Hua. Real-time large scale near-duplicate video retrieval. In *ACM MM*, 2010.

[5] J. Song, Y. Yang, Z. Huang, H. T. Shen, and R. Hong. Multiple feature hashing for real-time large scale near-duplicate video retrieval. In *ACM MM*, 2011.

[1] http://www.lemurproject.org/

[2] http://vireo.cs.cityu.edu.hk/webvideo/