# LARGE-SCALE NEAR-DUPLICATE WEB VIDEO SEARCH: CHALLENGE AND OPPORTUNITY

*Wan-Lei Zhao, Song Tan and Chong-Wah Ngo*

Department of Computer Science
City University of Hong Kong

## ABSTRACT

The massive amount of near-duplicate and duplicate web videos has presented both challenge and opportunity to multimedia computing. On one hand, browsing videos on Internet becomes highly inefficient for the need to repeatedly fast-forward videos of similar content. On the other hand, the tremendous amount of somewhat duplicate content also makes some traditionally difficult vision tasks become simple and easy. For example, annotating pictures can be as simple as recycling the tags of Internet images retrieved from image search engines. Such tasks, of either to eliminate or to recycle near-duplicates, can usually be achieved by the nearest neighbor search of videos from Internet. The fundamental problem lies on the scalability of a search technique, in face of the intractable volume of videos which keep rolling on the web. In this paper, we investigate scalability of several well-known features including color signature and visual keywords for web-based retrieval. Indexing these features based on embedding technique for scalable retrieval is also presented. On an Internet video dataset of more than 700 hours collected during years 2006 to 2008, we show some preliminary insights to the challenge of scalable retrieval.

## 1. INTRODUCTION

The number of web videos has exploded with the proliferation of digital video-capturing devices and the popularity of social media in Web 2.0. A rough statistic, as indicated in [1], shows that more than $65,000$ videos have been uploaded on video sharing web site YouTube daily. It is believed that this number is still increasing with fast speed. Among those uploaded videos, many of them are partially or fully duplicate to existing ones. Especially for hot topics, it becomes inevitable that the same video has been uploaded repeatedly.

Table 1 shows the amount of near-duplicate Internet videos from a brief survey of 13 topics. These videos were collected from YouTube, Google and Yahoo!. The first time we collected the videos was in year 2006 [1], where these topics were popularly viewed during that period. We crawled the videos of same topics recently during Dec 2008 and noticed that there still exists large amount of near-duplicates. When comparing these two collections of videos as indicated in Table 1, topic likes "I will survive Jesus" still has high percentage of near-duplicates (72.1%) for videos uploaded during Dec 2006 to Dec 2008. Based on our statistics, the average percentage of partially and fully duplicate videos for 13 topics is $14\%$. This number evidences the continuous rolling of near-duplicate videos for popular topics over years.

In this paper, we investigate scalable near-duplicate retrieval using the global (color signature) and local (visual keyword) features. Both features are popularly used in the current literature. It is our primary intention to see how global and local features react separately to scalability. Color signature is used in many content-based search tasks including near-duplicate retrieval. The signature is fast to retrieve and tolerant to slight change of video content. We also study feature embedding technique [2], which maps color feature into high-dimensional space in a way that specially designed indexing structure can be utilized for high speed retrieval. Visual keywords (VK) is generated based on a dictionary of local keypoint clusters [3]. Due to the consideration of local variations, VK becomes popular for near-duplicate search for its tolerance to geometric and photometric transformations.

## 2. RELATED WORK

Existing works on near-duplicate retrieval can be broadly grouped into two categories. One category aims for rapid retrieval and thus global features derived from color and ordinal signature [4, 5] are popularly employed. These approaches are highly suitable for identifying near identical videos. For videos which are partially duplicated, either spatially or temporally, global features are known to be less robust. The second category addresses the robustness issue by employing local keypoint features [6]. Keypoints are salient local patches detected over different scales. Its effectiveness have been demonstrated in [7, 8, 9], where video copies with considerable changes in background, color and lighting can still be successfully identified.

Keypoint-based approaches can be further subdivided into

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 119508).

**Table 1**. Amount of near-duplicates (ND) for videos uploaded to Internet during December 2006 to 2008 from a brief survey of 13 topics. The 3rd column shows the number of uploaded videos before Dec 2006; 4th column shows the number uploaded videos between Dec 2006 and Dec 2008; 5th and 6th columns show the amount and percentage of near-duplicates when comparing 3rd and 4th columns.

| ID | Topic | # | # | ND | % |
|---|---|---|---|---|---|
| 1 | The lion sleep tonight | 792 | 395 | 77 | 19.5% |
| 2 | Fold shirt | 436 | 355 | 46 | 13% |
| 3 | Cat massage | 344 | 433 | 10 | 2.3% |
| 4 | Ok go here it goes again | 396 | 255 | 9 | 3.5% |
| 5 | Real life Simpsons | 365 | 304 | 16 | 5.3% |
| 6 | Napoleon dynamite | 881 | 326 | 35 | 10.7% |
| 7 | I will survive Jesus | 416 | 326 | 235 | 72.1% |
| 8 | Korean karaoke | 205 | 350 | 10 | 2.9% |
| 9 | Panic at the disco I write sins not tragedies | 647 | 375 | 34 | 9.1% |
| 10 | Changes Tupac | 194 | 446 | 25 | 5.6% |
| 11 | Afternoon delight | 449 | 426 | 28 | 6.6% |
| 12 | Numa Gary | 422 | 375 | 63 | 16.8% |
| 13 | Shakira hips don't lie | 1322 | 342 | 49 | 14.3% |

different categories: trajectory-based [8, 10], keyframe-based matching [9] and visual keywords based [7]. Trajectory-based approaches track keypoints temporally along the video sequence, which forms bag of trajectories summarizing the moving pattern of keypoints. Such representation offers two advantages: facilitates the localization of near-duplicate segments, and supports high-speed online retrieval as demonstrated in [8, 11]. Nevertheless, trajectory extraction is generally a very expensive offline processing due to the need for extracting keypoints over frames. The fact that trajectory feature is sensitive to camera motion also makes it only applicable for exact duplicate (or copy) retrieval. Keyframe-based matching, while not able to characterize temporal content, is shown to exhibit excellent performance for near-duplicate image/video detection [9]. Nevertheless, even matching keypoints across two keyframes is already considered expensive. Such approaches in general is difficult to be scaled up for online retrieval.

Visual keyword (VK) based techniques, which quantize keypoints and perform matching based on bin-to-bin comparison, are accelerated version of direct keypoint matching. VK has been paid more attention recently for its ability in trading off speed and retrieval effectiveness. Different from trajectory-based approaches, VK is appropriate for both near-duplicate and copy retrieval. The recent work in [7] has demonstrated excellent performances of VK for video copy detection. A major weakness of VK is information loss during keypoint quantization. Several approaches have been proposed for addressing this problem, including hamming embedding [12], soft-weighting [13], and post-processing using weak geometry consistency [12].

## 3. SCALABLE NEAR DUPLICATE RETRIEVAL

### 3.1. Global Signature

Given a keyframe equally partitioned into $5\times5$ grids, The first three color moments in $Lab$ color space are extracted from each grid. By concatenating 25 moment vectors one after another, we obtain a 225 color moment feature vector. A *video signature* (VS) is defined as a 225-dimensional vector averaged over all keyframes in the video. Generally, signature VS is not sparse, as a result, indexing structure such as inverted file cannot be applied to speed-up the retrieval. Given videos $V_i$ and $V_j$, the similarity between their signatures $VS_i$ and $VS_j$ is measured based on Euclidean distance:

$$R(V_i, V_j) = d(VS_i, VS_j) = \sqrt{\sum_{k=1}^{m}(s_k^i - s_k^j)} \qquad (1)$$

where $m = 225$ and $s_k$ denotes the *k*th component of a video signature.

### 3.2. LSH Embedded Global Feature

Instead of generating color signature, one can employ LSH embedding (LSH-E) [2] to embed the color moments into a long sparse feature vector. The idea of LSHE is as follows. Given hash function group, $\mathcal{H}_i = \{h_1(v), ..., h_i(v), ..., h_B(v)\}$, where $h_i(v)$ accepts a $D$ dimensional feature $v$. LSH-E produces a binary code with length of $B$. As a result, given a family of hash function groups $\mathcal{G} = \{\mathcal{H}_1, ..., \mathcal{H}_i, ..., \mathcal{H}_L\}$, $L$ binary codes can be generated for one feature vector. By viewing one hash code as a histogram bin, color moment vector extracted from each grid is then hashed to the bins where its hash codes locate. Ultimately, LSH-E obtains a histogram in high-dimensional space (embedded vector) for summarizing the moment vectors extracted from the keyframes of a video sequence. The atomic hash function $h_i(v)$, $v \in R^D$, is defined as

$$h_i(v) = \begin{cases} 0 & \text{if } \rho \cdot v < 0 \\ 1 & \text{if } \rho \cdot v \geq 0 \end{cases} \qquad (2)$$

where $\rho$ is a random *D*-dimensional vector sampled from the unit hyper-sphere $\{\rho \in R^D \mid ||\rho||_2 = 1\}$. The dimension of the embedded vector is defined by $2^B * L$. As discussed in [2], $B$ and $L$ must be specified particularly for different kinds of feature input. Generally, larger input dimension $D$ requires larger $B$ as well as $L$. With this sparse vector representation, inverted file index can be employed to support fast retrieval. The similarity between two videos is evaluated by cosine distance measure.

### 3.3. Visual Keyword

To label keypoints with visual keywords, we learn a visual dictionary of 10,000 keywords. The dictionary is generated by clustering 1,185,698 keypoints extracted from 3,000

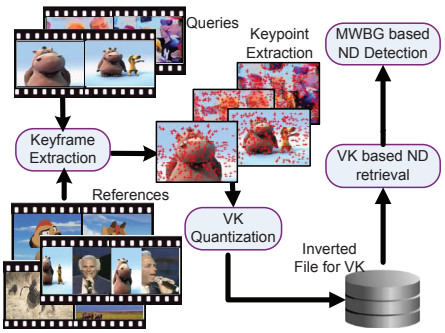


**Fig. 1**. Retrieval Framework using visual keywords.

keyframes randomly sampled from the dataset of [1]. We employ DoG [6] for keypoint detection and PSIFT [9] for feature description. Figure 1 shows the retrieval framework using visual keywords. Basically keyframes are first extracted. The detected keypoints in a keyframe is then quantized and labeled with keywords. Eventually, each keyframe is represented as a histogram of words with length equivalent to the size of dictionary. The feature vector is expected to be sparse and thus inverted file is employed as the index structure to support fast retrieval.

Different from global signature and LSH-E, we do not simply average or accumulate the bags of visual keywords for all keyframes in a video. Given two videos $V_i$ and $V_j$, the keyframes in $V_i$ are used to retrieve similar keyframes in $V_j$ using inverted file index. This indeed forms a bipartite graph where each keyframe is a node. The edge between two nodes, from $V_i$ and $V_j$ respectively, is weighted by visual keyword similarity between them. Since one keyframe may match to multiple keyframes in another video, we use the classical algorithm MWBG (maximum bipartite graph matching) to enforce that there is at most one match for each keyframe. Since the number of keyframes in an Internet video (of normally no more than 10 minutes) is not excessive and manageable, MWBG matching can be efficiently conducted. The similarity between two videos is measured based on the number of ultimately matched keyframe pairs in two videos.

## 4. EXPERIMENT

### 4.1. Dataset

The Internet video dataset shown in Table 1 is used in our experiment. The dataset consists of two sets collected respectively during 2006 and 2008. The first set is originally from [1] and consists of 24 search topics. We use it as reference set. There are $12,790$ web videos with a total length of $735.4$ hours. A total of $391,952$ keyframes are available together with the reference set. The second set consists of $5036$ videos for 13 out of 24 topics (4th column of Table 1). These videos were crawled from YouTube where we collected up to the top 500 videos of each topic. We extracted $10$ keyframes per shot, resulting in a total of $243,471$ keyframes. We use all

**Table 2**. Time for querying 5036 videos against 12,790 web videos (excluding time for keyframe and feature extraction).

| | G-SIG | LSH-E | VK |
|---|---|---|---|
| Time costs | 53 sec | 34.5 sec | 811.6 min |

the $5036$ videos in second set as testing queries. There is no overlap between reference and testing sets. The former consists of videos uploaded to Internet before Dec 2006, while videos in the latter set are uploaded after Dec 2006.

To generate the ground-truth, two assessors were asked to briefly browse through the videos in reference set and compare to testing queries. This procedure, though involves only the comparison of partial keyframe sets extracted from videos, already took weeks of manual effort. Since the ground-truth is expected to be incomplete, we further pooled the search lists retrieved by three approaches in Section 3, and manually labeled the results in details to generate final ground-truth.

### 4.2. Scalability and Speed

We compare three approaches presented in Section 3: global signature (G-SIG), signature with LSH embedding (LSH-E) and visual keywords (VK). The experiment is done with a PC of 3G memory. Table 2 shows the speed efficiency for querying $5036$ Internet videos. G-SIG and LSH-E are extremely efficient, compared to VK, where both approaches complete all queries within 1 minute. LSH-E, despite is with higher dimension than G-SIG, is faster than G-SIG for the employment of inverted file. While LSH-E and VK both use inverted index, VK exhibits much slower speed than LSH-E. The main reason is due to the trade-off between online keypoint quantization and size of visual vocabulary. When visual vocabulary is large, the online time spent for quantizing the keypoints of queries is considerably significant. On the other hand, when the size is small, the inverted index becomes less sparse which slows down the retrieval speed. In practice, we find that it is difficult to trade-off these two factors. LSH-E, in comparison, does not require vocabulary as reference and is capable of embedding features of similar content into the same hash codes. As a result, inverted file is sparse and this indeed saves significant cost in both storage and computation.

The time cost in Table 2 does not include keyframe and feature extraction, which are also part of online processing and can take up significant portion of time. In our experiment, $40.5$ hours are spent for extracting keyframes from $5036$ videos of approximately $243$ hours. Another $4.2$ hours for extracting G-SIG, $4.4$ hours for LSH-E, and $68$ hours for keypoints. For a 3-minutes query with 67 keyframes, G-SIG takes about $34.1$ seconds, LSH-E takes $34.1$ seconds and VK takes $107.3$ seconds for the whole search procedure. All approaches perform in real-time on our dataset.

**Table 3**. Performance of near-duplicate detection (left) and retrieval (right).

| | Accuracy |
|---|---|
| G-SIG | 0.639 |
| LSH-E | 0.634 |
| VK | 0.680 |
| VK∩LSH-E | 1.000 |

| | Precision | Recall |
|---|---|---|
| G-SIG | 0.608 | 0.577 |
| LSH-E | 0.628 | 0.636 |
| VK | 0.831 | 0.653 |
| VK∩LSH-E | 0.874 | 0.412 |

### 4.3. Near-duplicate Detection and Retrieval

We conduct two experiments to test the retrieval effectiveness. The first experiment tests the accuracy of identifying queries with near-duplicate videos in reference set. The second experiments test the retrieval rate of queries in terms of precision and recall. The first experiment involves all $5036$ queries, while the second involves $637$ queries which are considered near-duplicate to at least one video in the reference set. Table 3 show the performance of different approaches. We also test the late fusion of LSH-E and VK (LSH-E $\cap$VK) by intersecting their search lists. In the table, accuracy refers to the percentage of queries being correctly identified as near-duplicate version to the videos in reference set. Recall refers to the percentage of near-duplicate videos being correctly retrieved compared to ground-truth near-duplicates. Precision refers to the percentages of correctly retrieved videos compared to the total retrieved videos.

In the experiments, G-SIG considers two videos as near-duplicate if their distance is below $1.0$. For LSH-E, the parameters are $B = 10$ and $L = 18$. Only if the similarity excesses $0.95$, two videos are regarded as near-duplicate. For VK, we require at least $25\%$ of keyframes being matched after employing MWBG. These thresholds are set manually with the aim to see the best possible performance of each approach. From Table 3, VK basically shows better performance than G-SIG and LSH-E in both detection and retrieval. G-SIG which is not designed to deal with partial duplicates shows the worst performance. LSH-E which considers keyframe-level information, in contrast, offers the higher recall for retrieval performance than G-SIG. When intersecting VK and LSH-E, we see some improvement for detection and retrieval precision. This result probably hints the potential of fusing VK and LSH-E for scalable search, which worth further research.

## 5. DISCUSSION AND FUTURE WORK

We have presented some preliminary insights to the scalability of three quite different features for large-scale retrieval. On an Internet video dataset of 700 hours, the good news is that real-time retrieval is achievable by all three approaches with reasonable detection and retrieval performance. The fact that only a small portion of near-duplicate videos uploaded to Internet shows significant variations, simpler approaches like G-SIG and LSH-E can already offer satisfactory performance. Similar observation was indeed also pointed out in our previous work [1] based on the dataset crawled during 2006. In our experiment, LSH-E seems offering better scalability by the fastest retrieval speed and minimal effort in feature extraction. VK, compared to LSH-E, is considerably slow for both online retrieval and feature extraction, despite showing the best detection rate. The trade-off between dictionary size and quantization speed become a bottleneck to scale-up VK based retrieval. LSH-E, while seems as a preferable choice, indeed suffers from the potential problem of huge memory consumption, which we do not have space to further elaborate in this paper. In short, when the amount of videos continue to scale up, there is no guarantee that the three tested approaches can still achieve real-time performance. Advanced high-dimensional indexing techniques which consider both storage and computation efficiency are highly demanded.

## 6. REFERENCES


[1] X. Wu, C.-W. Ngo, and A. G. Hauptmann, "Practical elimination of near-duplicates from web video search," in *ACM Multimedia*, 2007.

[2] W. Dong, Z. Wang, M. Charikar, and K. Li, "Efficiently matching sets of features with random histograms," in *ACM Multimedia*, 2008.

[3] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*.

[4] A. Hampapur and R. Bolle, "Comparison of sequence matching techniques for video copy detection," in *Conf. on Storage and Retrieval for Media Databases*, 2002.

[5] J. Yuan, L.-Y. Duan, Q. Tian, S. Ranganath, and C. Xu, "Fast and robust short video clip search for copy detection," in *Pacific Rim Conf. on Multimedia*, 2004.

[6] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, 2004.

[7] M. Douze, A. Gaidon, H. Jegou, M. Marszatke, and C. Schmid, "Inria-lear's video copy detection system," in *TREVCID*, 2008.

[8] J. Law-To, B. Olivier, V. Gouet-Brunet, and B. Nozha, "Robust voting algorithm based on labels of behavior for video copy detection," in *ACM Multimedia*, 2006.

[9] W-L. Zhao and C-W. Ngo, "Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection," *IEEE. Trans. on Image Processing*, 2009.

[10] D.-D. Le, X.-M. Wu, S.-N. Rajgure, J.-V. Gemert, and S.-I. Satoh, "National institute of informatics, japan at trecvid 2008," in *TRECVID*, 2008.

[11] A. Joly, J.-L.To, and N. Boujemaa, "Inria-imedia trecvid 2008: Video copy detection," in *TRECVID*, 2008.

[12] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *ECCV*, 2008.

[13] Y. G. Jiang and C. W. Ngo, "Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval," *CVIU*, 2009.