

Statistical Summarization of Content Features for Fast Near-duplicate Video Detection

ABSTRACT

This paper outlines a system for detecting near-duplicate videos based on a novel statistical summarization of content features for each clip. It captures the dominating content and content changing trends of a video, so this representation is very compact and effective. Unlike traditional frame-to-frame comparisons that involve quadratic computational complexity, the similarity measure of our method is only linear in dimensionality of feature space and independent of video length. To further improve the search efficiency for very large video databases, an effective indexing structure is deployed to significantly reduce the number of videos for comparison. This demo shows that our system can accurately find near-duplicates from a large collection of tens of thousands of video clips extremely fast.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—*Abstracting methods, Indexing methods*

General Terms

Design, Algorithms, Experimentation

1. INTRODUCTION

A massive influx of video clips on the Internet has been dubbed as a new phenomenon which has a profound impact on both the Internet and other forms of media. Video clips are short clips in video format predominantly found on the Internet and expressing a *single moment of significance* [1]. The widespread popularity of video clips has evolved into clip culture. Different from traditional TV programs and full movies, video clips are mostly less than 15 minutes, referring to an Internet activity of sharing and viewing a short video. Sources for video clips are various, typically including music videos, TV commercials, news and sporting events and movie trailers. Traditional long videos can also be segmented into clips, each of which represents a scene or story.

Near-duplicate video clips are short videos that visually similar or nearly identical to each other, but may appear differently due to various changes introduced during capturing stage (camera view point or setting, lighting condition, background, foreground, etc.), transformations (frame format, rate, resolution, shift, contrast, brightness, saturation, crop, blur, sharpen, etc.), and editing operations (frame adding, dropping, re-ordering or content modification). Near-duplicate video detection is an important research problem with a wide range of applications such as TV broadcast monitoring, video copyright enforcement, content-based video clustering and annotation, video database purge, cross-modal divergence detection, etc. For example, when a company contracts TV

stations for its commercials, it often asks a market survey company to monitor whether the commercials are actually broadcasted as contracted (when and how). Some other companies may also approach such market survey companies to seek information about how their competitors market their commercials. While the same commercial is given to all TV stations for broadcasting, it can eventually appear with some variations, such as station-specific parameters, reception and recording errors, and inserts of different products or local contact information. Thus, the ‘same’ commercials broadcasted by different TV stations at different time could be near-duplicates.

Some preliminary works on detecting near-duplicate videos focus on extracting some video signatures, such as key-frame or point of interest. The frame pairwise comparisons, such as using Edit distance [2, 3] as the similarity measure, though could be somehow robust, the time complexity is quadratic in video length. Therefore, most of them are tested and workable on small video collections only. With the increasingly widespread popularity of online digital videos, real-time detection from large databases becomes strongly desired. In this demo system, effective and compact video representations are summarized from frame content features for accurate and fast similarity comparison. Moreover, we deploy an effective indexing structure called Bi-Distance Transformation (BDT) which utilizes the power of a pair of optimal reference points to reduce search space. We first introduce the proposed video representation, similarity measure and indexing issues, and then briefly describe the features of our system, which shows detecting near-duplicate videos from a large video database can be performed extremely fast and very accurately.

2. SYSTEM OVERVIEW

2.1 Algorithm

Through image feature extraction, each video is represented by a sequence of frame feature vectors. Given the understanding that a video clip often shows a moment of significance, from human perception, different dominating visual content may express different significance. Meanwhile, different content changes may also suggest different meanings. Inspired by this, we propose a single video representation model to capture the dominating content and content changing trends of each clip by statistically exploiting the tendencies of frame vector point distribution. Given a video $X = \{x_1, x_2, \dots, x_n\}$ where x_i is a d -dimensional feature vector, it can be summarized by $(O, \Phi_1, \dots, \Phi_d)$, where O is the mean for all x_i , and Φ_1, \dots, Φ_d are d Bounded Principle Components (BPCs), each of which is a line segment that shows the direction of large variance bounded by two furthestmost projections on Φ_i . In other words, BPCs indicate the ranges of content feature dispersions along certain orientations in vector space. Independent of frame number n , this model only records an origin and d BPCs to represent a video. Therefore, it actually summarized each video by $(d + 1)$ d -dimensional vectors.

Given videos X and Y , their difference can be estimated with



Figure 1: Demo System Snapshots.

the translation, rotation and scaling operations needed to match their corresponding compact summaries ($O^X, \Phi_1^X, \dots, \Phi_d^X$) and ($O^Y, \Phi_1^Y, \dots, \Phi_d^Y$), where d is the numbers of BPCs. A *translation* allows one to move its origin to another position. A *rotation* defines an angle which specifies the amount to rotate a BPC to match its counterpart of another. A *scaling* operation can stretch or shrink a BPC to have the same length of another. In vector space, the difference of two vectors is given by the length of their subtraction. Therefore, the overall difference can be measured by

$$\|O^X - O^Y\| + \sum_{i=1}^d \|\Phi_i^X - \Phi_i^Y\|.$$

Each clip can be globally summarized by a single point so this representation reduces video complexity greatly. Meanwhile, its similarity measure is linear in dimensionality of feature space and independent of video length, so it improves the efficiency for fast search. However, as the video collection becomes very large, exhaustive scan is still undesirable. To further improve the efficiency, we extend the optimal one-dimensional transformation method accommodated with B^+ -tree introduced in [4] for indexing the compact summaries. The one-dimensional transformation of a database point P can be simply achieved with a mapping function $D(P, R)$ which computes the distance between P and the selected reference R . The derived one-dimensional distance values for all points are then indexed by a B^+ -tree. Given a search radius r and a query Q , a range search $[D(Q, R) - r, D(Q, R) + r]$ in B^+ -tree is performed so the points with distances in the range are considered for actual distance computations. As shown in [4], either the furthestmost projection R_1 or R_2 in Figure 2 which bounds the bi-directional first BPC could be an optimal reference point. We propose a two-dimensional transformation method called Bi-Distance Transformation (BDT) to further reduce the number of candidate accesses. The intuition of BDT is that two furthestmost projections R_1 and R_2 are far away from each other. Points that are close to one optimal reference point will be far from the other. Given a query point Q and a search radius r , the chance for P of satisfying both $D(Q, R_1) - r \leq D(P, R_1) \leq D(Q, R_1) + r$ and $D(Q, R_2) - r \leq D(P, R_2) \leq D(Q, R_2) + r$ simultaneously is much less than that of satisfying either $D(Q, R_1) - r \leq D(P, R_1) \leq D(Q, R_1) + r$ or $D(Q, R_2) - r \leq D(P, R_2) \leq D(Q, R_2) + r$. Therefore, the search space of using both R_1 and R_2 together will be much smaller than that of using a single R_1 or R_2 , i.e., less candidate videos will be accessed with BDT.

2.2 User Interface

This demo system is implemented with JSP. The database consists of more than 30,000 video clips (from a few seconds to minutes) that are segmented from free-to-air TV programmes and pre-processed offline. The content features are extracted for each frame in 8-, 16-, 32- and 64-dimensional RGB color space and HSV color space, respectively.

The main functionalities of our system are shown in Figure 1. Among the returned results of near-duplicate video clips, the detailed differences of any two can be displayed clearly by brows-

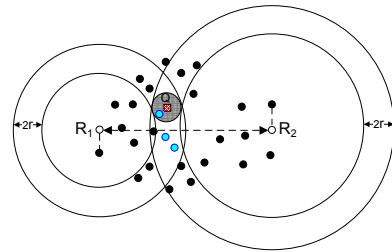


Figure 2: Search Space Comparison of A Single Optimal Reference Point and A Pair of Optimal Reference Points.

ing their individual key-frames at same time stamps. The system provides convenient ways to play two videos simultaneously for dynamic comparison as well. We also incorporate ViTri [4] and Edit distance [3] methods in our system, and a user can tick the methods to be used with different feature spaces and dimensionalities. From the search results of different methods together with their query response time, the search effectiveness and efficiency can be compared with these state-of-the-arts. From the large scale experiments, it is observed that the proposed method delivers most accurate detection. Compared with ViTri and Edit distance which responds in minutes and hours respectively due to their quadratic time complexity, our novel approach achieves real-time response only in milliseconds.

3. SUMMARY

In this demo, we describe a system for detecting near-duplicate video clips by statistically summarizing content features. This online system can search the near-duplicates of query video clips, with an easy-to-use tool for users to compare their differences. The effectiveness and efficiency of our method outperforms existing approaches and can be tested at the demonstration.

4. ACKNOWLEDGEMENTS

This work was funded in part the European Union Sixth Framework Programme (FP6) through the integrated project Pharos (IST-2006-045035).

5. REFERENCES

- [1] YouTube. <http://en.wikipedia.org/wiki/YouTube>, May 2007.
- [2] D. A. Adjeroh, M.-C. Lee, and I. King. A distance measure for video sequences. *Computer Vision and Image Understanding*, 75(1-2):25–45, 1999.
- [3] L. Chen, M. T. Özsu, and V. Oria. Robust and fast similarity search for moving object trajectories. In *SIGMOD Conference*, pages 491–502, 2005.
- [4] H. T. Shen, B. C. Ooi, X. Zhou, and Z. Huang. Towards effective indexing for very large video sequence database. In *SIGMOD Conference*, pages 730–741, 2005.