Video fingerprinting for copy identification: from research to industry applications

Jian Lu
Vobile, Inc.

A preprint from
Proceedings of SPIE - Media Forensics and Security XI, Vol. 7254, January 2009.

Video fingerprinting for copy identification: from research to industry applications

Jian Lu*
Vobile, Inc.

ABSTRACT

Research that began a decade ago in video copy detection has developed into a technology known as "video fingerprinting". Today, video fingerprinting is an essential and enabling tool adopted by the industry for video content identification and management in online video distribution. This paper provides a comprehensive review of video fingerprinting technology and its applications in identifying, tracking, and managing copyrighted content on the Internet. The review includes a survey on video fingerprinting algorithms and some fundamental design considerations, such as robustness, discriminability, and compactness. It also discusses fingerprint matching algorithms, including complexity analysis, and approximation and optimization for fast fingerprint matching. On the application side, it provides an overview of a number of industry-driven applications that rely on video fingerprinting. Examples are given based on real-world systems and workflows to demonstrate applications in detecting and managing copyrighted content, and in monitoring and tracking video distribution on the Internet.

Keywords: Video fingerprinting, video copy detection, video sharing, UGC

1. INTRODUCTION

In February 1999, a few graduate students at Stanford University wrote a technical report entitled, "Finding pirated video sequences on the Internet".[1] The work contained in that technical report eventually became part of two distinguished Ph.D. dissertations by Shivakumar[2] and Indyk[3], respectively. And the technique that was used for identifying pirated videos developed into a technology known as "video fingerprinting".

To be sure, Shivakumar and Indyk's work was not the only one on video fingerprinting and copy identification. There were quite a number of related publications by other researchers around the same time in late 1990s and early 2000s. [4-9] However, Shivakumar and Indyk's work was particularly notable in two respects. First, it was the first to use video fingerprinting in identifying unauthorized copyrighted videos on the Internet. Years later, such application became the driving force in the development and deployment of video fingerprinting technology in the era of YouTube. Secondly, it was the first to use Locality Sensitive Hashing (LSH) in fingerprint matching. To date, the LSH algorithm remains the state-of-the-art in similarity search in high dimensions; new techniques for fingerprint matching are often compared to LSH.

Research and development activities in video fingerprinting subsided for a period of time after the burst of the first Internet bubble in early 2000s. In the last few years, however, the proliferation of online video in the peer-to-peer (P2P) and user-generated-content (UGC) networks has brought renewed interest in video fingerprinting technology for solving the copyright violation problems. At issue is unauthorized distribution of copyrighted video content in the P2P and UGC networks. In recent high-profile legal cases such as MGM v. Grokster and Viacom v. YouTube, the plaintiffs argued that the sites that host video content or a search index of video content should proactively police their sites by identifying and filtering out copyright-infringing content. The enabling technology recommended by the plaintiffs for identification and filtering of copyright-infringing content at large-scale is fingerprinting.

In late 2006, the Motion Picture Association of America (MPAA) and the Motion Picture Laboratories (MovieLabs) initiated a Content Recognition Systems Study that focused specifically on evaluating fingerprinting technologies for video identification. The Study lasted more than 6 months and was joined by 12 participants including large corporations, start-up companies, and a university. Today, all of major Hollywood film and TV studios have adopted

* jian@vobileinc.com; phone 1 408 217-5010; fax 1 408 212-8300; www.vobileinc.com

video fingerprinting technology. In practical applications, video fingerprinting is used in identifying, monitoring, tracking, as well as filtering of unauthorized, illegal, or offensive content. Researchers and practitioners are also exploring other applications of video fingerprinting, such as video asset management, contextual advertising, and content-based video search.

Despite the resurgence in research activities in video fingerprinting and its adoption by the industry, there has not been a comprehensive review about video fingerprinting technology and its applications. This paper attempts to fill that gap by providing a review of both research work and real-world applications. Before getting down to the details, it is helpful to review and clarify on related terminology.

1.1 Related Terminology

A *video fingerprint* is an identifier that is extracted from a piece of video content. The process of extracting a fingerprint from the video content is referred to as *fingerprinting* the video or *video fingerprinting*. There is an obvious analogy to human fingerprint and fingerprinting. Just like human fingerprint that can uniquely identify a human being, video fingerprint can uniquely identify a piece of video content. The analogy extends to the process of subject identification by fingerprint: first, known fingerprints must be stored in a database; then, a subject's fingerprint is queried against the database for match.

In a broad sense, the term *video fingerprinting* has been used to refer to the technology encompassing algorithms, systems, and workflows that use video fingerprint for video identification. It should be evident from the context if the term is used to refer to the fingerprinting process or more broadly the technology.

In research literature, video fingerprinting and fingerprint-based video identification are also commonly known as *video copy detection* or more generally *content-based copy detection* (CBCD). Indeed, copy detection is the application that motivated development of video fingerprinting. Here, "copy" has a quite broad meaning. It could be a small segment cut from the original, lasting only a few seconds, and possibly embedded in a long edit or "mash-up". It could be transformed into different formats, codecs, resolutions, frame rates, and bitrates. And it could be modified and distorted by scaling, cropping, frame dropping, and overlay of text and graphics.

Another term that is used to describe video fingerprinting is *robust video hashing*. It comes from the observation that conventional cryptographic hashing such as MD5 is fragile and sensitive to even a single bit change in the content. The idea is to design hashing schemes that are robust to distortions that do not change our perception of the video content. For this reason, it is sometimes also called *perceptual hashing*. However, the use of "hashing" can be confusing because of the additional security requirements that are often imposed on hash functions. For example, one desired property for a hash function is uniform distribution of hash values in order to minimize collisions, the incident that has two different entities hashed into the same point in the hash space. Yet for video fingerprinting, it can be ideal to have different versions (can be infinite in number) of the same video content hashed into the same point in the fingerprint space. In such observation, "robust hashing" sounds like a self-conflicting proposition.

Lastly, it is worth noting that the word fingerprinting has also been used in the research literature of watermarking to describe the process of adding an identifier (watermark) to the content. To date, the industry has an unambiguous view of what it calls watermarking and fingerprinting. When an identifier or signature is *added* to the content and thereby changing the content, it is watermarking; when an identifier or signature is *extracted* from the content without changing the content, it is fingerprinting.

1.2 Organization of Sections

The rest of this paper is organized as follows. Section 2 reviews research work on video fingerprinting and fingerprint matching algorithms and designs. First, a set of desired properties and common metrics for fingerprinting algorithms are introduced. Then, fingerprinting algorithms are grouped into several categories and reviewed based on their use of spatial, temporal, color, and transform-domain signatures. For fingerprint matching, a complexity analysis is given for exhaustive fingerprint search; general strategies for reducing complexities are discussed. A few existing and new algorithms for fast approximate fingerprint matching are reviewed. The last part of Section 2 contains the author's observations and remarks on the video fingerprinting research. Section 3 provides an overview of a few industry-driven applications that rely on video fingerprinting. Examples are given based on real-world systems and workflows to demonstrate applications in identifying and managing copyrighted content, and in monitoring and tracking video distribution on the Internet. Finally, a few promising new applications are previewed. Section 4 concludes the paper with a summary.

2. RESEARCH IN VIDEO FINGERPRINTING

This section reviews research work on video fingerprinting, including algorithms and designs for video fingerprinting and fingerprint matching. Before considering specific algorithms and designs, it is helpful to examine what we aim to achieve with video fingerprinting and how to measure the effectiveness of designs and implementations.

2.1 Properties and Metrics

Ideally, a design of a video fingerprint should have the following characteristics that hold true for a large corpus of video content of diverse types.

Robust. A video fingerprint should stay largely invariant for the same video content under various types of processing, transformations and manipulations, such as format conversion, transcoding, and content editing.

Discriminating. The video fingerprints for different video content should be distinctly different.

Compact. A video fingerprint should be minuscule in data size, comparing to the data size of the original video content.

Low complexity. The algorithm for extracting video fingerprints should have low computational complexity so that a video fingerprint can be computed fast.

Efficient for matching and search. Although there are generic algorithms that treat all fingerprints as a string of bits in matching and search, a good design of video fingerprint should facilitate approximation and optimization to improve the efficiency in matching and search.

Because video fingerprints are generally not perfectly identical for different versions of the same content, fingerprint matching is not a simple table lookup in the database. Instead, it is a similarity search problem. Typically, a distance metric is defined to quantify the similarity between two video fingerprints being compared. Commonly used distance metrics include Manhattan (L1) distance and Euclidean (L2) distance, where a normalized L1 or L2 distance provides a good measure of similarity. When a video fingerprint consists of binary signatures, Hamming distance is often used, and a normalized Hamming distance or Bit Error Rate (BER) provides a good measure of similarity.

In most applications of video identification, having a quantified measure of similarity is not sufficient. An explicit judgment of whether two videos contain the same content is required. Thus, the effectiveness of a video fingerprint design and implementation can be measured by the rate of correct returns to fingerprint queries. A pair of commonly used measures is *precision* and *recall* rates that are defined as follows:

$$P_r\,(\%) = \frac{N_{tp}}{N_p} \times 100 \tag{1}$$

and

$$R_e\,(\%) = \frac{N_{tp}}{N_{ep}} \times 100 \tag{2}$$

where P_r is precision rate, R_e is recall rate, N_{tp} is number of true positives or correct matches, N_p is total number of positives or matches, and N_{ep} is number of expected positives or matches.

Corresponding to the desired properties of video fingerprint, precision rate is a measure of discriminability, and recall rate is a measure of robustness. Another pair of related measures is *false positive* (FP) and *false negative* (FN) rates that are defined as follows:

$$R_{fp}\,(\%) = \frac{N_{fp}}{N_{en}} \times 100 \tag{3}$$

and

$$R_{fn}\,(\%) = \frac{N_{fn}}{N_{ep}} \times 100 \tag{4}$$

where R_{fp} and R_{fn} are false positive and false negative rates, respectively; N_{fp} is number of false positives, N_{fn} is number of false negatives, N_{ep} is number of expected positives as previously defined, and N_{en} is number of expected negatives. N_{ep} and N_{en} add up to the total number of fingerprint queries in the test, denoted by N_q.

The above definitions for false positive and false negative rates assume the knowledge of expected positive and negative numbers, N_{ep} and N_{en}. This is usually true for controlled tests. For a real-world running system that receives a large volume of queries, it is hard to know or verify the expected positive and negative numbers. Therefore, the operating R_{fp} and R_{fn} of a fingerprinting system are often computed by replacing N_{en} in (3) and N_{ep} in (4) with N_q, the total number of fingerprint queries.

2.2 Video Fingerprinting Algorithms

2.2.1 Overview of Video Signatures

In its raw form, a video fingerprint is just a string of bits that represent the "signatures" of the video data. Different designs vary in the type of signatures that are chosen to characterize the video data, and the way to compute them. Almost all of the video signatures that have been proposed to date can be classified into four types: spatial, temporal, color, and transform-domain. In some designs, different types of signatures are combined to form video fingerprints.

A spatial signature characterizes spatial features of a video frame and is computed independent of other frames. Examples of spatial features include luminance patterns, differential luminance or gradient patterns, and edges. A temporal signature describes temporal features of a video and is computed between two frames in the temporal direction. Examples of temporal features include frame difference measures, motion vector patterns, and shot durations. A color signature captures color characteristics of a video frame and is computed in a color space such as RGB or YUV. Many color signatures are an abstraction of patterns in the color histogram. A transform-domain signature is computed from coefficients of an image or video transform such as a DCT or wavelet transform. Transform-domain signatures provide a different characterization and representation of some spatial and/or temporal features in the transform domain.

In addition to various types of video signatures, video fingerprints differ in granularity that is the smallest unit of video that a video signature characterizes. Spatial granularity can vary from entire video frame to subdivided blocks to points of interest in a frame. Temporal granularity can be key frames only, group of frames, downsampled single frames, or every single frame.

Different granularities of video fingerprint provide a tradeoff between discriminability, robustness, and compactness. For example, by dividing a video frame into multiple blocks and computing temporal and color signatures for each block, we gain finer spatial granularity or resolution in temporal and color signatures at the cost of additional storage.

2.2.2 Spatial Signatures

A class of spatial signatures is designed to characterize luminance patterns in a video frame. In such designs, a video image is first converted to the YUV color space; the luminance (Y) component is kept, and the chrominance components (U, V) are discarded. The luminance image is further subdivided into a fixed-sized grid of blocks (e.g., a 4x4 grid of blocks) independent of frame resolutions, as shown in Figure 1(a).

Note that unlike in image and video compression where a frame subdivision is by fixed-sized blocks (e.g., 8x8 blocks), here the frame subdivision is designed to produce a fixed-sized grid of blocks. The subdivision of a video frame serves two purposes. First, it leads up to block-based signatures that are robust to changes in pixel values; second, it produces a compact and fixed-sized frame fingerprint consisting of fixed number of block signatures.

One popular block-based luminance signature is based on ordinal ranking. It was designed by Bhat and Nayar[7] for image identification and first used by Mohan[8] in video fingerprinting and matching. The simplicity of ordinal ranking is illustrated in Figure 1(b)-(c). After frame subdivision, the average pixel value for each block is computed, and an abstraction follows by ranking the blocks by their average pixel values. The rank of each block in ordinal position is assigned to the block as its signature. Video fingerprints based on ordinal signatures have been studied and experimented extensively.[9-11] They were found to be more robust than some temporal and color signatures.[9] They are also compact in size: for a frame subdivision containing M blocks, the required number of bits for a frame fingerprint is $M * ceiling(\log_2 M)$.

(a)

74	128	46	133
62	78	58	145
60	82	116	157
87	70	214	167

(b)

5	10	0	11
3	6	1	12
2	7	9	13
8	4	15	14

(c)

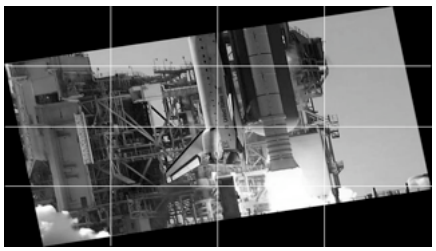
Figure 1: Ordinal ranking - an example of spatial signatures. (a) Color-converted and subdivided luminance frame; (b) average pixel values of blocks; (c) ordinal ranks of blocks.

One drawback of ordinal signatures has to do with the global ranking. The rank of each block is relative to all other blocks in the frame. This means local luminance variations such as a logo insertion that should change the rank of one block could actually change the ranks of multiple blocks or all blocks. Block-based differential luminance signatures such as those proposed[12-15] are more robust to local luminance variations while maintaining a compact representation similar in data size to the ordinal signatures. Oostveen et al[12] computed block difference in one spatial direction (horizontal) followed by a binary abstraction (greater than or not). Lee and Yoo[13-14] computed luminance gradients in each block and condensed them to the centroid (geometric average) of gradient orientations. Iwamoto et al[15] estimated luminance edges in 8 quantized directions for each block and kept the direction having maximum strength as an edge signature. It is worth noting that like ordinal ranking for block luminance patterns, differential block luminance patterns are quantized or abstracted to form differential signatures. Abstraction further increases robustness and reduces data size; it is key to all fingerprinting algorithms.

Block-based spatial signatures such as ordinal and differential signatures are susceptible to geometric transformations such as rotation, cropping, and scaling that changes aspect ratios. Figure 2 illustrates the mismatch of content in blocks between two transformed and the original frame images after rotation and cropping. This difficulty has motivated designs of spatial signatures that are resilient to geometric transformations. Many proposed algorithms employ a special image transform, such as polar Fourier Transform,[16] Radon Transform,[17-18, 22] or Singular Value Decomposition (SVD).[19] They have reportedly good resilience to affine transformations such as shift and rotation. However, they are still prone to cropping; see, e.g., Seo et al[17] for some experimental results. Unfortunately, in practical applications involving videos, cropping often accompanies shift, rotation, and scaling due to fixed video frame size, as shown in Figure 2. Additionally, most of the above techniques have high computational complexities that make them impractical for many applications of video fingerprinting in the real world.

Another approach that is fundamentally different from block-based designs is to compute spatial signatures around points of interest in a video frame. This approach is often combined with the use of key frames on which points of interest are computed. Examples of spatial signatures that are based on points of interest include those that use Harris points,[20-22] scale-invariant feature points,[23] and the Difference-of-Gaussian scale-space feature points.[24] Unlike block-based designs, spatial signatures based on points of interest lead to frame fingerprints of variable sizes, because the number of points of interest in a frame is content-dependent, and can be potentially very large. The variable number and configuration of points of interest in a frame necessitate an alternative way for similarity definition and search. The computational complexity of the above methods for extracting spatial signatures based on points of interest is significantly higher than that of block-based signatures.

(a)



(b)

(c)

Figure 2: Mismatch of content in blocks in a fixed-sized grid after rotation and cropping. (a) Color-converted and subdivided luminance frame; (b) rotation by 10 degrees; (c) rotation followed by cropping and expanding to full screen. Blocks in the center of the frame are less distorted than those on the edge.

2.2.3 Temporal Signatures

The groundbreaking work by Shivakumar[1-2] used temporal signatures. The design is quite straightforward. First, a video sequence is segmented into shots. Then, the duration of each shot is taken as a temporal signature, and the sequence of concatenated shot durations form the fingerprint of the video.

Related to shots are key frames where the content has abrupt changes, such as the boundary that separates two shots. Key frames are important anchors in a video sequence that are often used in video fingerprinting.[18, 20, 24-26] The locations of key frames in a video provide a natural temporal signature. However, not all designs make use of the temporal positions of key frames. For example, in Cheung and Zakhor's work,[26] the temporal information of key frames is discarded; the similarity between two video sequences is measured by the degree of match to a group of key frames. In such design, it is not possible to determine the location or offset of a match in the query and reference videos.

More commonly, temporal signatures are computed on adjacent frames in a video. Chen and Stentiford[27] proposed to use temporal ordinal signatures. Similar to the steps of deriving spatial ordinal signatures, a frame is subdivided into a fixed-sized grid of blocks, and the average pixel value of each block is computed. Unlike spatial ordinal ranking that ranks blocks in a frame, temporal ordinal ranking in Chen and Stentiford[27] ranks blocks in same spatial position across the frames in a temporal window; the temporal ordinal ranks of the blocks are used as temporal signatures.

With an approach similar to the way of computing spatial differential signatures, temporal differential signatures are explored. In such an approach,[11-12, 28] luminance difference between two adjacent frames or same-positioned blocks in two adjacent frames is computed, followed by an abstraction that quantizes the difference to 2 or 3 levels (i.e., greater, equal or less). Hampapur et al[9] estimated block motion vectors and quantized them to four orientations to form motion signatures. With a different approach to motion estimation based on tracking points of interest from frame to frame, Law-To et al[21] computed temporal trajectories of points of interest and abstracted their motion signatures by labels.

2.2.4 Color Signatures

Color signatures are among the first being used in video fingerprinting. Naphade el al[5] proposed a technique that was also experimented by Hampapur el al.[9] In this approach, a level-quantized histogram is computed for Y, U, and V components for each video frame. To reduce the resulting signature data, a polynomial approximation is used to model the pixel counts in each bin of the histograms along the temporal direction. A special distance metric based on histogram intersection is used as a similarity measure. Li et al[29] applied ordinal ranking to the bins of histogram based on the frequency counts of each bin. Hu[30] computed Alpha-trimmed average histogram on a group of subdivided frames and

quantized each color component to eight bins. In the latter two designs, the color signatures can be compared with common distance measures such as the Manhattan distance.

2.2.5 Transform-Domain Signatures

There are some designs that compute video signatures in a transform domain. In many cases, the choice of an image (2D) or video (3D) transform is motivated by some invariance properties of the transform. For example, Swaminathan et al[16] used polar Fourier Transform to compute an image hash that is resilient to rotation and translation. By a similar motivation, Radon Transform[17-18, 22] and Singular Value Decomposition[19, 31] were explored to generate transform-domain signatures that are robust to geometric distortions.

In another study, Coskun and Sankur[32] used 3D Discrete Cosine Transform (DCT) on a group of 64 frames and hashed the resulting coefficients into binary signatures with median-based thresholding. Their work was later extended[33] to also use a 3D Random Bases Transform (RBT) followed by hashing. So far, the DCT and RBT based signatures have been tested with a small set of short video clips.[33]

2.3 Fingerprint Matching

2.3.1 Exhaustive Search

As introduced in Section 2.1, fingerprint matching is a similarity search problem where the degree of similarity is quantified by certain distance metric. More specifically, for a given video fingerprint of a target video clip, fingerprint matching amounts to finding in a reference video fingerprint database the closest match or matches to the target video fingerprint. In some applications, it suffices to return a list of match candidates ranked by their similarity scores or distance values; in many other applications, however, an explicit judgment of match or no-match has to be made. In either case, the heavy load of computation lies in computing the distance values or similarity scores between the target video fingerprint and reference video fingerprints in the database. In a brute-force approach, this comes down to computing and comparing the distance from the target video fingerprint to each and every reference video fingerprint in the reference database, and finding the one(s) that have the shortest distance. Because a target video usually has different length than the ones in the reference database, and can match to any part of a reference video, comparing the target video fingerprint with each reference video fingerprint involves evaluating every offset position for the best alignment between the two video fingerprints that gives the smallest distance. To sum up, the time complexity of a brute-force fingerprint matching by exhaustive search can be expressed as follows:

$$\textit{Time complexity} = O(k*N) \tag{5}$$

$$= O(k*l*M) \tag{6}$$

where k is the length of the target video fingerprint, N is the total length of video fingerprints in the reference database, l is the average length of the video fingerprints in the reference database, and M is the total number of video fingerprints in the reference database; $N = l*M$.

Considering that the length of a given video is finite and not growing, the dominating factor of complexity in fingerprint matching is clearly the reference database size which can be represented either by the total length of reference video fingerprints, N, or the total number of reference video fingerprints, M. More specifically, the time complexity of fingerprint matching by exhaustive search is linear of the reference database size. Since in practice the reference database size can be very large and continues to grow, fingerprint matching by exhaustive search is clearly not scalable for practical applications. Fortunately, exhaustive search is rarely necessary in practice. In most cases, a well-designed approximate and fast search can find the same best match as an exhaustive search does in a tiny fraction of time required for the exhaustive search.

2.3.2 Approximate and Fast Search

A well-known algorithm for approximate similarity search is called Locality Sensitive Hashing (LSH). It was first introduced by Indyk and Motwani[34] and refined by Gionis el al.[35] Although LSH has been widely used in many applications that involve similarity search, video fingerprint matching was among the first applications in which LSH was used.[1-2] Since then, other researchers have explored LSH in fingerprint matching along with their video fingerprinting algorithms.[30, 36]

The LSH algorithm was conceived for solving the approximate Nearest Neighbor Search (NNS) problem in high dimensions; fingerprint matching can be formulated as an NNS problem. Consider a d-dimensional vector space $\boldsymbol{P}$ where

a distance metric $D(\boldsymbol{x},\boldsymbol{y})$ is defined. For a query point $\boldsymbol{q}$, if the nearest point $\boldsymbol{p}$ exists in $\boldsymbol{P}$ such that $D(\boldsymbol{p},\boldsymbol{q}) = r$, the so-called ε- Nearest Neighbor Search (ε-NNS) seeks to find a point $\boldsymbol{p}'$ in $\boldsymbol{P}$ such that $D(\boldsymbol{p}',\boldsymbol{q}) \leq r\,(1+\varepsilon)$ for any $\varepsilon > 0$. For $\boldsymbol{P}$ containing N points, it was shown[35] that with LSH the approximate nearest neighbor $\boldsymbol{p}'$ can be found with high probability in sublinear time of N, more specifically,

$$\textit{Time complexity} = O(\, d*N^{\,1/(1+\varepsilon)}\,) \tag{7}$$

Behind the mathematical rigor of LSH is an intuitive geometric reasoning: if two points in a d-dimensional space are close to each other, then, their projections onto lower dimensional spaces are very likely to be close as well. An important part of LSH design is to devise hash functions such that the points that are close to the query point will be hashed into the same bucket with high probability. Then, a linear, exhaustive search may be used to find the closest point(s) to the query point in the bucket that often contains a much smaller number of candidates than the original search space. In Gionis et al[35], the hash functions being chosen are random projections from high dimensions to a lower dimensional space. Recently, Baluja and Covell[37] took a different approach to hashing for reducing the search space. Instead of designing deterministic hash functions, they used machine learning techniques and training data to devise a hashing system that adapts to the identification task and data. This results in a more compact hash bucket that contains significantly fewer candidates that may need to be compared with a linear search, thus boosting the speed of fingerprint matching. So far, this "learning to hash" technique has been applied to audio fingerprint matching with excellent results;[37] it would be interesting to see how it works for video fingerprint matching.

One of the benefits of using training data for machines to learn to hash is to forgo an explicit definition of similarity that can be hard to define precisely for video content and its fingerprints. This is in contrast to LSH in metric space where a distance metric measuring similarity needs to be defined explicitly. In another approach that does without using distance-based similarity measures and deterministic hash functions, Joly et al[38] proposed to use a statistical similarity search based on probabilistic models of common distortion vectors. Similar to hashing, a statistical similarity search maps the full search space into a small bucket of candidates by probabilistic filtering. In Joly et al,[38] probabilistic models for common distortion vectors associated with Harris spatial signatures[20] were used and a substantial speed-up in fingerprint query was achieved over exhaustive search with little loss in accuracy.

From optimization point of view, a system for fingerprint matching can be divided into two parts; each part can employ some kind of approximation that trades possibly a little loss in accuracy for speed in a fingerprint query. The first part of approximation is to reduce the search space. More specifically, for a fingerprint query to a reference database containing M fingerprint records, we seek to map the reference database to a bucket of size B that is smaller than M. A refined search including possibly exhaustive search may be performed in the resulting bucket for the query fingerprint. Ideally, we would like the bucket to contain only the most likely candidates for match, and the mapping from the full search space to the bucket to be super fast. Besides the hashing and probabilistic mapping methods that have been reviewed above, some heuristic techniques can be also very effective for reducing search space. For example, Oostveen el al[12] attempted to reduce search space by selecting only reference candidates that contain identical anchor fingerprint blocks that are present in the query fingerprint.

It is possible that after search space reduction, the resulting bucket size B remains large, though it is smaller than M. In this situation, a refined search in the bucket of candidates can be quite costly by itself. Thus, the second part of approximation in fingerprint matching aims to reduce the cost of a linear search in the bucket. For systems that seek to match a sequence of frame fingerprints based on a distance measure, several approximation techniques are often used. One of these techniques is greedy search: if a portion of the query fingerprint is matched to some references, subsequent search for match is directed at the part that immediately follows the matched portion in both the query fingerprint and the references. Another technique is "early exit" that aborts the comparison with a reference if an intermediate value of distance measure is already above the threshold for no-match. Yet another technique is to use downsampled frames in distance calculation.

All of the approximation techniques except "early exit" can incur a loss in search accuracy. Nonetheless, real-world data suggests that with a good video signature design, a speedup in several orders of magnitude can be achieved using these approximation techniques with a negligible loss in search accuracy.

2.4 Remarks

2.4.1 Which One to Use?

With so many designs of video signatures and associated video fingerprinting algorithms, a natural question is: which one is the best? The answer is that there is no absolute best. Some video signatures are robust against certain types of distortions in video content but vulnerable to other types of distortions; other video signatures may be the other way around. Nonetheless, judging by the criteria outlined in Section 2.1, namely, robustness, discriminability, compactness, low complexity, and efficiency for search, the overall category winner appears to be spatial signatures, particularly block-based spatial signatures. Temporal and color signatures, while useful in improving discriminability, tend to fall short in robustness in comparison to spatial signatures. This observation is supported by experimental results reported in research literature, industry evaluation tests, as well as the success of some commercial systems deployed in the real world.

Despite the motivation of using some special transforms for their resilience to geometric transformations, transform-domain signatures are not widely adopted in video fingerprinting in practice due to their computational complexity. On the other hand, by using some adaptive techniques in fingerprint matching, block-based spatial signatures that are known to be prone to geometric transformations can achieve good robustness against moderate geometric transformations, e.g., frame rotation by 10 degrees. One of commonly used adaptive techniques is to apply weighting in distance calculations in fingerprint matching. Generally speaking, simply weighting down the block differences towards the edges of a video frame is often helpful because content around the center of the video frame is better preserved in geometric transformations and less affected by logo and subtitle overlays. See, e.g., Figure 2 for a visual comparison of distortions on the center and edges of a video frame. Iwamoto et al[15] used a more sophisticated method in determining the weights in distance calculations.

Block-based spatial signatures are also compact and have low computational complexity. For many designs using spatial ordinal or differential signatures, the data size of a frame fingerprint is on the order of a few hundred bits, or less than 10 Kbps in data rate for video with frame rate at 30 fps. These fingerprints can be computed from a standard-definition video source in 1/10 of video playback time (or 10 times faster than real-time) on an off-the-shelf consumer-grade PC.

Because of their many advantages, spatial signatures particularly block-based spatial signatures are most widely used and studied in video fingerprinting. For designs that employ spatial signatures as the primary component of video fingerprint, temporal and color signatures are sometimes used as a secondary component to complement spatial signatures.

2.4.2 Temporal Structure from Spatial Signatures vs. Temporal Signatures

Many block-based spatial signatures are computed on each frame of the source video at certain frame rate, generating a sequence of time-stamped frame fingerprints. These frame fingerprints characterize not only spatial patterns in their corresponding video frames, but also temporal structure of the video. They provide a strong temporal constraint for a video being compared for match, increasing both discriminability and robustness of fingerprint matching. This assertion comes from an intuitive reasoning: if a single frame fingerprint is matched to a video, it may be by accident; if a number of consecutive frames are matched to a video in high degree, the chance of an accidental match decreases quickly as the number of consecutive frames increases. Indeed, many block-based spatial signatures correspond to an ultra-low resolution grid downsampled from the original frame resolution (e.g., 4x4 grid of blocks downsampled from a frame of 720x480); video sequence matching based on these low-resolution spatial signatures relies on a multitude of consecutive frame fingerprint matches to increase discriminability. On the other hand, robustness can also be enhanced with a multitude of consecutive frame fingerprints to smooth out a small number of mismatches due to distortions (e.g., frame drops) or local content changes (e.g., a fade or dissolve introduced by video editing).

The temporal structure that is imposed by a sequence of frame fingerprints with spatial signatures can be such a strong constraint in video identification that gives a nonessential role to separate temporal signatures such as the ones that characterizes the differential patterns in adjacent frames. Indeed, many proposed designs[8,10, 13-15] of video fingerprints do not include separate temporal signatures; they rely solely on spatial signatures to form a sequence of frame fingerprints. Like an I-frame only video sequence, a sequence of video fingerprints without inter-frame temporal signatures have some benefits in content editing and management; for example, it can be cut, split, and merged at any point without a need of re-computing or modifying temporal signatures on the boundaries. Nonetheless, temporal signatures can be a

good complement to spatial signatures in video identification. For example, video sequences containing many still frames (e.g., a slide show) can be characterized more effectively when temporal signatures are used.

2.4.3 Algorithms vs. Systems

The various approximation techniques used in fingerprint matching creates a complex pipeline where each stage can contribute to an increase in FN. Therefore, in an end-to-end video identification system including both fingerprinting and fingerprint matching, robustness depends on not only video signature and fingerprinting algorithms, but also a number of other factors in the system design. Specifically, reducing search space may introduce FNs. For example, a matching reference that does not fall into the bucket by hashing or mapping results in a FN. Additionally, the use of greedy search in computing alignment and frame downsampling in distance calculation can also introduce FNs.

To separate algorithms and systems, BER and normalized L1 or L2 distance can be used for pure video fingerprint evaluation and comparison, assuming a query fingerprint is perfectly aligned to the matching reference fingerprint, and no approximation is made in computing the distance or error rate. When recall or FN rate is reported, however, it should be understood that one is evaluating an end-to-end video identification system; the superiority of a video fingerprint measured by BER or other distance-based error rates may not translate into a superior video identification system. It is highly desirable that a video fingerprint design can facilitate and work well with various approximation techniques, because in the end, all practical systems must use some approximation techniques in fingerprint matching and what matters is the accuracy and speed of such systems.

2.4.4 Topics of Continuing Research

There is active research in new video signatures and fingerprinting algorithms as well as faster fingerprint matching algorithms and techniques. Designing an ultimate, versatile video signature is the Holy Grail. The author of this paper believes that it is more achievable and beneficial to develop a set of video signatures that are complementary to each other in enhancing robustness and discriminability. In fingerprint matching, there is a real need for continued advance in search algorithms because of the rapid growth in the size of video fingerprint database that is already in the order of tens of millions for UGC videos. Additionally, it would be highly useful to quantify the relationship between the loss in accuracy and gain in speed by various approximation techniques used in fingerprint matching.

3. APPLICATIONS OF VIDEO FINGERPRINTING

Video fingerprinting technology that can identify video content accurately, efficiently, and automatically has many practical applications. As was introduced in Section 1, the development of video fingerprinting technology has been driven largely by needs for finding copyrighted video content on the Internet. As the technology matures, other applications are also emerging. In this Section, we review a few industry applications that have been commercially deployed for copyright management in video distribution on the Internet. We also provide a brief overview for a few emerging applications that are being developed and experimented.

3.1 Video Content Registration

Before a video can be identified by its fingerprint, the video fingerprint must be extracted from at least one version of the same video content and ingested into a reference database. Typically, a master reference fingerprint database is centrally located while fingerprints are often collected and ingested from distributed locations. This is similar to the process of populating a human fingerprint database. Like a human fingerprint database, a video fingerprint database contains not only fingerprint data, but also information about or associated with the fingerprint data. Such information is of critical importance to applications that query a video fingerprint database.

One type of information that binds with video fingerprint is the so-called metadata that describes the video content and/or the particular instance of the video content from which the fingerprint is extracted. Currently, there is not yet a standardized metadata schema for video fingerprints, but commonly used content metadata includes title, ownership information, production and release dates, genres, etc. Instance metadata includes length, resolution, and frame rate of the video, codec and file format, etc. Because a wealth of video metadata already exists elsewhere, e.g., in an existing video asset management system, it is unnecessary to replicate a full set of metadata in a video fingerprint database. Most video fingerprint databases stores only a limited set of metadata for content identification purposes, and reference other unique video asset IDs (e.g., ISAN[39-40]) that link to sources of information outside the video fingerprinting system.

Another type of information that is associated with video fingerprint is business rules. They are specified by content owners to determine what actions should be taken when an unauthorized copy of reference content is identified. Recently, MovieLabs published Content Recognition Rules (CRR)[41] that defines standard XML interfaces to communicate about business rules for identified unauthorized content. Putting a content identification system in the center, the CRR defines two interfaces. One interface is from content owners to a content identification system for specifying business rules; the other interface is from the content identification system to a caller of content identification service to communicate match results as well as the rules and actions specified by the content owner. The CRR provides a flexible framework for existing and anticipated business and application scenarios. For example, the content owner could specify if the UGC site should take it down or allow it to post, or could advertise on it when a copy of unauthorized content is identified on the site. Furthermore, these actions could be determined on different conditions and circumstances, such as how long the identified content is, and the geography of the site.

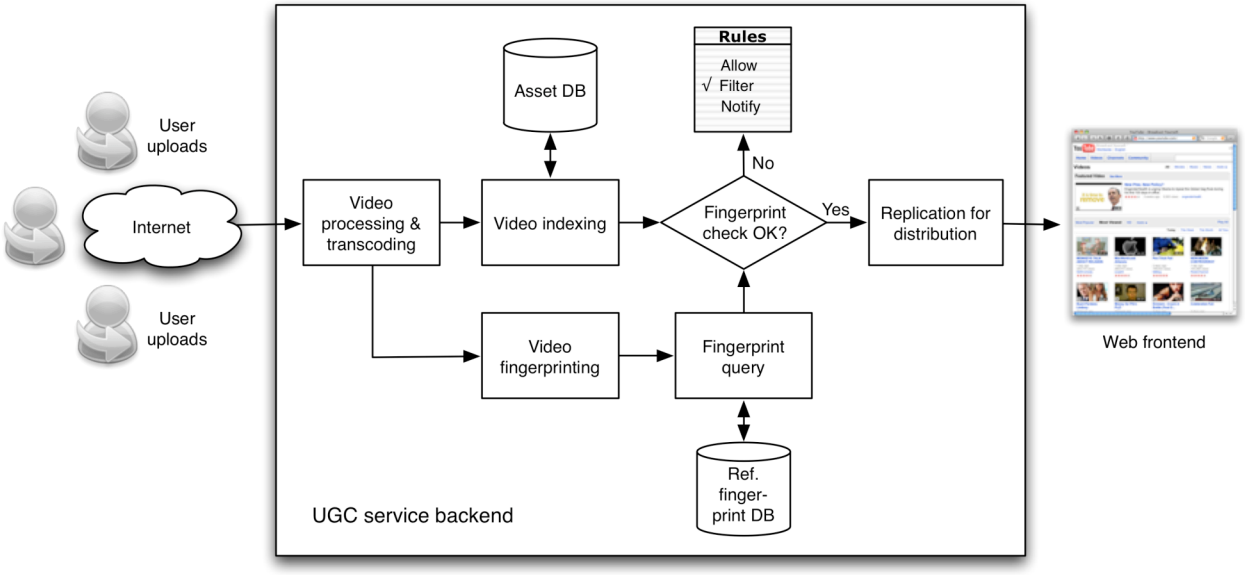


Figure 3: Integration of video content filtering in a UGC service backend.

3.2 Video Content Filtering

Content filtering has long been proposed as a solution to the piracy problem on the Internet. In 2001, Napster implemented a music content filter based on audio fingerprinting in its P2P network. Because of its availability and maturity, audio fingerprinting has also been used in video content filtering by identifying the associated audio tracks. Only in the last few years content filtering systems based on video fingerprinting have become available and been deployed. Today, video content filtering systems can be based on video fingerprinting, audio fingerprinting, or both, but they share basically the same architecture and workflow.

Figure 3 is a diagram illustrating a content filtering system based on video fingerprinting in the service backend of a UGC site. The filter is integrated in an automatic workflow for converting and publishing user uploaded video clips. Typically, a user uploaded video clip is processed and transcoded into a common format in site-specified settings (e.g., FLV in 320x240, 30 fps, 200 Kbps). The video is also indexed by its metadata for search and retrieval. Video fingerprinting and identification can be done before or after the transcoding. In the diagram of Figure 3, fingerprinting is done on transcoded video content. The resulting fingerprint is queried against a pre-populated reference fingerprint database for identification. The identification results are fed back to the publishing workflow. If the query fingerprint is matched to a copyrighted asset in the reference database, the corresponding video clip is taken down or handled differently according to the rules and actions specified by the copyright owner; otherwise, the video is replicated and published to the Web frontend.

The point of integration of content filters is an important design consideration for effective and efficient content filtering. Today, a top online video site such as YouTube has hundreds of thousands of video uploads each day; a second-tier site

also has tens of thousands. While these numbers are non-trivial for a content filtering system, they are a small fraction compared with the number of downloads that are in the order of a hundred millions each day for YouTube alone. So, in the UGC filtering above, the optimum point of integration is in the path of upload processing on the UGC service backend. In the P2P networks, however, the best point of integration of a content filter is less obvious. Existing designs and implementations put content filters in the P2P client software or on servers that host a directory or search index of P2P video content.

There are also content filtering systems that are designed to work in the packet network to identify and filter out unauthorized copyrighted video content in P2P file swaps. These systems are typically deployed at the gateway of an intranet, such as a college campus network, or a traffic aggregation point in a broadband access network operated by an ISP. Content identification in the packet network is more challenging because of the stricter requirement on low latency and high scalability. For this consideration, practical systems often adopt a hybrid design that combines content fingerprints with packet-level signatures for increased efficiency in content identification.

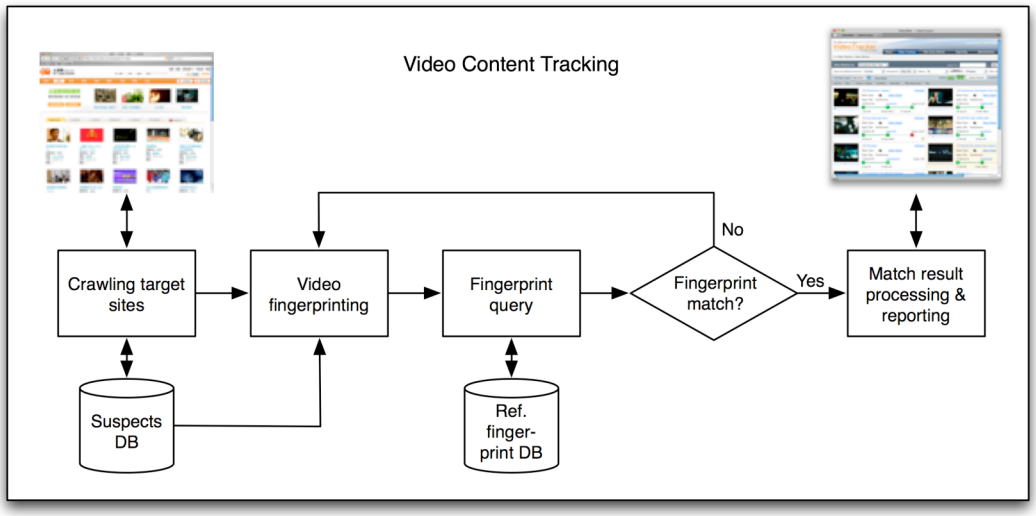


Figure 4: Workflow of a video content tracking system.

3.3 Video Content Tracking

Owners of video content often want to know where their content is distributed on the Internet and how many people have watched it. Video content tracking is an application and service that serve this purpose. Figure 4 is a diagram of a video tracking system that consists of a web crawler, a video fingerprinting system, and a web interface. The web crawler serves to discover the "suspects". Crawling may be targeted to specific sites or specific content categories on a site, and may be guided by keywords. The video fingerprinting system serves to check and verify the "suspects". The reference fingerprint database can be very targeted; it may contain only the fingerprints of the video content being tracked. The web interface is used to report and update the tracking results. As the crawler continues to discover new "suspects", they will be verified and reported (if matched) in a continuous, 24x7, and non-stopping workflow. Figure 5 shows a screenshot of the VideoTracker™ system developed by Vobile, Inc.

To date, video tracking has been successfully deployed to track high-valued copyrighted video content, from Hollywood blockbuster releases to the 2008 Beijing Summer Olympic programs.[42-43] Besides copyright enforcement, video tracking has also been used to track various types of video content on the Internet, such as commercials and political campaign videos. One may observe that some of these tasks used to be done by humans before video fingerprinting technology became available. Indeed, the key value that video fingerprinting brings in these applications is enabling an automatic, low cost, and more accurate and efficient workflow. This has changed the way business is done. For example, automatic video tracking systems based on video fingerprinting can now track tens of thousands of titles simultaneously, comparing to no more than tens of titles previously by human-based tracking services. Better yet, the tracking results can now be updated instantly and continuously instead of daily or weekly reports.

Figure 5: VideoTracker™ - screenshot of a real-world video tracking system.

3.4 Other Applications

Broadcast monitoring is among the first applications of video fingerprinting. It monitors broadcast programs in various local markets to find out where and when a program is broadcast and for how many times. The gathered information is useful for rights owners to collect royalties for their content or for advertisers to audit the airing of their commercials in the paid time slots in a broadcast network.[29]

Contextual advertising based on video identification works in a way similar to Google's AdSense. While AdSense pairs ads with keywords indexed from Web pages or text derived from other media (e.g., audio tracks), fingerprint-based video identification can tell exactly what content is being consumed, thus providing a good context for serving relevant ads. When a content owner allows advertising on its content on UGC sites (e.g., specified in a CRR-compliant rule), it creates a new model for monetizing its content. Currently, the industry is experimenting with this approach that will hopefully lead to a new way for solving the piracy problem.

Video asset management using video fingerprints recognizes the fundamental role of video fingerprints: they are content-based IDs. There are many benefits of using video fingerprints as content-based IDs in an asset management system. Because a video fingerprint is computed from video content, it is a permanent ID that can always be regenerated. Copies, segments, and edits of the same video content can be easily identified and related to each other by their fingerprints. For example, Kasutani et al[44] developed an video archiving system that automatically detects and links video edits to the source video footages based on video identification.

Content-based video search is an area of active research. The notion of "query by video clip" was coined by Jain et al[45] a decade ago. The techniques described in their paper – using color, texture, and motion signatures in a query – are essentially a video fingerprint query. To date, video fingerprint queries are primarily for finding copies of the same content, partial or whole, transformed or unaltered. There have been attempts to explore video fingerprinting in broader video search for discovering similar but different video content.

4. CONCLUSION

Research in video fingerprinting has come a long way since it began a decade ago and developed into a technology that is adopted by the industry. Key areas of research include designs of video signatures, fingerprinting and fingerprint matching algorithms. Among the large number of designs, video signatures can be classified into spatial, temporal, color, and transform-domain signatures. Although none is perfect, spatial signatures are found to be the overall winner

in terms of robustness, discriminability, compactness, and computational complexity. Temporal and color signatures can provide enhanced discriminability when used together with spatial signatures. Fingerprint matching by exhaustive search has a linear time complexity with regard to the size of reference database. Fortunately, effective approximation techniques have been developed that provide a dramatic reduction in computational complexity, speeding up fingerprint queries by several orders of magnitude over an exhaustive search with a negligible loss in accuracy. This made it possible to build practical fingerprint matching systems that are scalable.

The adoption of video fingerprinting technology was accelerated in the last few years as the content industry responded to the increasing cases of copyright violations in the rapidly growing P2P and UGC networks. As such, major commercial applications of video fingerprinting to date are for identifying unauthorized distribution of copyrighted video content on the Internet, including video content filtering and tracking. Moving forward, researchers and practitioners are also exploring and experimenting other applications of video fingerprinting, including contextual advertising, video asset management, and content-based video search.

REFERENCES

[1] Indyk, P., Iyengar, G. and Shivakumar, N., "Finding pirated video sequences on the Internet," Tech. Rep., Stanford InfoLab, Stanford University, Feb. 1999.

[2] Shivakumar, N., "Detecting digital copyright violations on the Internet," Ph.D. Dissertation, Stanford University, Aug. 1999.

[3] Indyk, P., "High-dimensional computational geometry," Ph.D. Dissertation, Stanford University, Sep. 2000.

[4] Cheung, S.-C. and Zakhor, A., "Estimation of web video multiplicity," Proc. SPIE, Internet Imaging, vol. 3964, pp. 34-36, Jan. 2000.

[5] Naphade, M. R., Yeung, M. M. and Yeo, B.-L., "A novel scheme for fast and efficient video sequence matching using compact signatures," Proc. SPIE, Storage and Retrieval for Media Databases, vol. 3972, pp. 564-572, Jan. 2000

[6] Hampapur, A. and Bolle, R. M., "Comparison of distance measures for video copy detection," Proc. IEEE Int. Conf. Multimedia and Expo (ICME), pp. 188-192, Aug. 2001.

[7] Bhat, D. N. and Nayar, S. K., "Ordinal measures for image correspondence," IEEE Trans. Pattern Ana. Mach. Intell., vol. 20, no. 4, pp. 415-423, Apr. 1998.

[8] Mohan, R., "Video sequence matching," Proc. Int. Conf. Acoust., Speech and Signal Processing (ICASSP), vol. 6, pp. 3697-3700, Jan. 1998.

[9] Hampapur, A., Hyun, K.-H. and Bolle, R. M., "Comparison of sequence matching techniques for video copy detection," Proc. SPIE, Storage and Retrieval for Media Databases, vol. 4676, pp. 194-201, Jan. 2002.

[10] Hua, X.-S., Chen, X. and Zhang, H.-J., "Robust video signature based on ordinal measure," IEEE Int. Conf. Image Proc. (ICIP), vol. 1, pp. 685-688, Oct. 2004.

[11] Kim, C. and Vasudev B., "Spatiotemporal sequence matching for efficient video copy detection," IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 1, pp. 127-132, Jan. 2005.

[12] Oostveen, J., Kalker, T. and Haitsma, J., "Feature extraction and a database strategy for video fingerprinting," Proc. 5th Int. Conf. Recent Advance in Visual Information Systems, pp. 117-128, 2002.

[13] Lee, S. and Yoo, C. D., "Video fingerprinting based on centroids of gradient orientations," Proc. Int. Conf. Acoust., Speech and Signal Processing (ICASSP), vol. 2, pp. 401-404, May 2006.

[14] Lee, S. and Yoo, C. D., "Robust video fingerprinting for content-based video identification," IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 7, pp. 983-988, Jul. 2008.

[15] Iwamoto, K., Kasutani, E. and Yamada, A., "Image signature robust to caption superimposition for video sequence identification," IEEE Int. Conf. Image Proc. (ICIP), pp. 3185-3188, Oct. 2006.

[16] Swaminathan, A., Mao, Y. and Wu, M., "Image hashing resilient to geometric and filtering operations," IEEE Workshop on Multimedia Signal Processing (MMSP), pp. 355-358, Sep. 2004.

[17] Seo, J. S., Haitsma, J., Kalker, T. and Yoo, C. D., "Affine transform resilient image fingerprinting," Proc. Int. Conf. Acoust., Speech and Signal Processing (ICASSP), vol. 3, pp. 61-64, Apr. 2003.

[18] De Roover, C., De Vleeschouwer, C., Lefèbvre, F. and Macq, B., "Robust video hashing based on radial projections of key frames," IEEE Trans. Signal Proc., vol. 53, no. 10, pp. 4020-4037, Oct. 2005.

[19] Radhakrishnan, R. and Bauer C., "Robust video fingerprints based on subspace embedding," Proc. Int. Conf. Acoust., Speech and Signal Processing (ICASSP), pp. 2245-2248, Apr. 2008.

[20] Joly, A., Frélicot, C. and Buisson, "Robust content-based video copy identification in a large reference database," Int. Conf. on Image and Video Retrieval (CIVR), pp. 414-424, 2003.
[21] Law-To, J., Buisson, O., Gouet-Brunetand, V. and Boujemma, N., "Robust voting algorithms based on labels of behavior for video copy detection," Proc. ACM Int. Conf. on Multimedia, pp. 835-844, 2006.
[22] Maani, E., Tsaftaris, S. A. and Katsaggelos, A. K., "Local feature extraction for video copy detection in a database," IEEE Int. Conf. Image Proc. (ICIP), pp. 1716-1719, Oct. 2008.
[23] Sarkar, A., Ghosh, P., Moxley, E. and Manjunath, B. S., "Video fingerprinting: features for duplicate and similar video detection and query-based video retrieval," Proc. SPIE, Multimedia Content Access: Algorithms and Systems, vol. 6820, Jan. 2008.
[24] Massoudi, A., Lefebvre, F., Demarty, C.-H., Oisel, L. and Chupeau, B., "A video fingerprint based on visual digest and local fingerprints," IEEE Int. Conf. Image Proc. (ICIP), pp. 2297-2300, Oct. 2006.
[25] Lee, S. and Yoo, C. D., "Robust video fingerprinting based on affine covariant regions," Proc. Int. Conf. Acoust., Speech and Signal Processing (ICASSP), pp. 1237-1240, Apr. 2008.
[26] Cheung, S.-C. S and Zakhor, A., "Efficient video similarity measurement with video signature," IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 1, pp. 59-74, Jan. 2003.
[27] Chen, L. and Stentiford, F. W. M., "Video sequence matching based on temporal ordinal measurement," Pattern Recognition Letters, vol. 29, no. 13, pp. 1824-1831, Oct. 2008.
[28] Radhakrishnan, R. and Bauer C., "Content-based video signatures based on projections of difference images," IEEE 9th Workshop on Multimedia Signal Processing (MMSP), pp. 341-344, Oct. 2007.
[29] Li, Y., Jin, J. S. and Zhou, X., "Matching commercial clips from TV streams using a unique, robust, and compact signature," Proc. Digital Imaging Computing: Techniques and Applications, pp. 266-272, Dec. 2005.
[30] Hu, S., "Efficient video retrieval by locality sensitive hashing," Proc. Int. Conf. Acoust., Speech and Signal Processing (ICASSP), vol. 2, pp. 449-452, Mar. 2005.
[31] Jeong, K.-M., Lee, J.-J. and Ha, Y.-H., "Video sequence matching using singular value decomposition," Proc. 3rd Int. Conf. Image Analysis and Recognition (ICIAR), pp. 426-435, Sep. 2006.
[32] Coskun, B. and Sankur, B., "Robust video hash extraction," Proc. European Conf. on Signal Processing (EUSIPCO), pp. 2295-2298, Sep. 2004.
[33] Coskun, B. and Sankur, B. and Memon, N., "Spatio-temporal transform based video hashing," IEEE Trans. Multimedia, vol. 8, no. 6, pp.1190-1208, Dec. 2006.
[34] Indyk, P. and Motwani, R., "Approximate nearest neighbor – towards removing the curse of dimensionality," Proc. 30th Symposium on Theory of Computing, pp. 604-613, 1998.
[35] Gionis, A., Indyk, P. and Motwani, R., "Similarity search in high dimensions via hashing," Proc. 25th Int. Conf. Very Large Data Bases (VLDB), pp. 518-529, 1999.
[36] Yang, Z., Ooi, W. T. and Sun, Q., "Hierarchical, non-uniform locality sensitive hashing and its application to video identification," Proc. IEEE Int. Conf. Multimedia and Expo (ICME), vol. 1, pp. 743-746, Jun. 2004.
[37] Baluja, S. and Covell, M., "Learning to hash: forgiving hash functions and applications," Data Mining and Knowledge Discovery, vol. 17, no. 3, pp. 402-430, Dec. 2008.
[38] Joly, A., Buisson, O. and Frélicot, C., "Content-based copy retrieval using distortion-based probabilistic similarity search," IEEE Trans. Multimedia, vol. 9, no. 2, pp. 293-306, Feb. 2007.
[39] International Standard Audiovisual Number (ISAN) – Part 1: Audiovisual work identifier, ISO 15706-1, 2002.
[40] International Standard Audiovisual Number (ISAN) – Part 2: Version identifier, ISO 15706-2, 2007.
[41] "Content Recognition Rules", MovieLabs TR-CRR1, 2008. http://www.movielabs.com/CRR/
[42] Burrows, P., "Video piracy's Olympic showdown," BusinessWeek, Jun. 9, 2008. http://www.businessweek.com/magazine/content/08_23/b4087073685542.htm
[43] Voigt, K., "Let the Games begin, online," CNN.com International, Aug. 6, 2008. http://edition.cnn.com/2008/TECH/08/06/db.olympicdigitalrights/
[44] Kasutani, E., Oami, R., Yamada, A., Sato, T. and Hirata, K., "Video material archive system for efficient video editing based on media identification," Proc. IEEE Int. Conf. Multimedia and Expo (ICME), vol. 1, pp. 727-730, Jun. 2004.
[45] Jain, A., Vailaya, A. and Xiong, W., "Query by video clip," Multimedia Systems, vol. 7, no. 5, pp. 369-384, Sep. 1999.