VIDEO IDENTIFICATION USING VIDEO TOMOGRAPHY

Gustavo Leon, Hari Kalva, and Borko Furht
Department of Computer Science and Engineering, Florida Atlantic University, Boca Raton, FL, USA

ABSTRACT

Video identification or copy detection is a challenging problem and is becoming increasingly important with the growing popularity of online video services. The problem addressed in this paper is the identification of a given video clip in a given set of videos. For a given query video, the system returns all the instance of the video in the data set. This identification system uses video signatures based on video tomography. A robust and low complexity video signature is designed. The signatures are generated for video shots and not individual frames. This results in a compact signature of 64 bytes per video segment/ shot. The signatures are matched using simple Euclidean distance metric. Both signature generation and matching have a very low complexity. The system was evaluated using the dataset from video copy detection evaluations organized at the CIVR 2007. The results show that the proposed signatures outperform the results reported in the CIVR competition. The results also show that the signature is robust to common video transformations.

Index Terms — *Tomography, Motion Analysis, Scene change detection, Video copy detection, Video Signatures*.

1. INTRODUCTION

Video copy detection also referred to as video identification is an important problem that impacts applications such as online content distribution. The main problem addressed here is determining whether a given video clip belongs to a known set of videos. This is a challenging problem and the solutions fall into two classes 1) digital watermark based video identification and 2) content based video identification. Digital watermarking based solutions assume an embedded watermark that can be extracted anytime in order to determine the video source. Digital watermarking for video and images has been proposed as a solution for identification and tamper detection in video and images [1]. While digital watermarking can be useful in identifying video sources, they are not usually designed to address the problem of identifying unique clips from the same video source. Even if frame-unique watermarks are embedded, the biggest obstacle of using watermarking is the embedding of a robust watermark in the source. Another issue is that large collections of digital assets without watermarks already exist.

The drawbacks of digital watermarking are being addressed in an emerging area of research referred to as blind detection [2,3]. Blind detection based approaches, like digital watermarks, address the problem of tampering detection and source identification. Unlike watermarks, blind detection uses characteristics inherent to the video and capture devices to detect tampering and identify sources. Nonlinearity of capturing sources, lighting consistency, and camera response function are some of the features used in blind detection. Both digital watermarking and blind detection are more suitable for tamper detection and source identification and are not suitable for video copy detection or identification.

Content based copy detection has received increasing interest lately as this approach does not rely on any embedded watermarks and uses the content of the video to compute a unique signature based on various video features. A survey of content based video identification systems is presented in [4].

A content based identification system for identifying multiple instances of similar videos in a collection was presented in [5]. The system identifies videos captured from different angles and without any query input. Since the system is designed to identify similar videos this is not suitable for applications such as copy detection that require identification of a given clip in a data set. A solution for copy detection in streaming videos is presented in [6]. The authors use a video sequence similarity measure which is a composite of the frame fingerprints extracted for individual frames. Partial decoding of incoming video is performed and DC coefficients of key frame are used to extract and compute frame features.

A copy detection system based on the bag-of-words model of text retrieval is presented in [7]. This solution uses SIFT descriptors as words to create a SIFT histogram that is used in finding matches. The use of SIFT descriptors makes the system robust to transformations such as brightness variations. Each frame has a feature dimension of 1024 corresponding to the number of bins in the SIFT histogram. A clustering technique for copy detection was proposed in [8]. The authors extract key frames for each cluster of the query video and perform a key frame based search for similarity regions in the target videos. Similarity regions as short as 2x2 pixels are used leading to high complexity.

Most of these content based video identification methods operate with video signatures that are computed using

features extracted from individual frames. These frame based solutions tend to be complex as they require feature extraction and comparison on frame basis. Another common feature of these approaches is the use of key frames for temporal synchronization and subsequent video identification. Determining key frames either relies on underlying compression algorithms or requires additional computation to identify key frames. An important characteristic of video identification solutions is a robust and compact video signature that is computationally inexpensive to compute and compare.

In this paper we present a robust video identification system that uses spatio-temporal signatures based on video tomography (see Section 2. for an overview of video tomography). Video tomography captures the spatio-temporal changes in videos and is a measure of local and global motion in videos. The proposed video identification system is based on the hypothesis that the combination of local and global motion in a video clip can uniquely characterize and identify videos. The results of extensive evaluation show that the proposed system can identify videos with very high precision and recall rates.

The key contribution of this paper is a robust video identification system based on video tomography. The proposed solution is low complexity as it uses a 64-byte signature for a shot or group of frames. The system has very low memory and computational requirements and is independent of video compression algorithms. This system can be easily implemented as a part of commonly available video players.

The rest of the paper is organized as follows: Section 2 summarizes video signature design and extraction, Section 3 presents performance evaluation, and conclusions are presented in Section 4.

2. VIDEO SIGNATURE DESIGN

2.1 Video Tomography

The proposed method of video identification is based on video tomography. Video tomography was first presented in ACM Multimedia '94 by Akutsu and Tonomura for camera work identification in movies [9]. Since then this approach has been primarily explored for summarization and camera work detection in movies [10]. The images of video tomographs in [9] and [10] reminded us of flow patterns of ridges in human fingerprints and thus began our exploration of video tomography for identification. The initial thought was to exploit the work done in fingerprint analysis to extract signatures from video tomographs. During the course of development we discovered the simple and elegant structure in video tomographs and developed a video signature based on easily computable features. Our experiments verify that these video signatures are robust and uniquely identify video shots. This approach is robust to transformations such as re-compression and is independent of the compression algorithms used.

Video tomography is the process of generating tomography images for a given video shot. A tomography image is composed by taking a fixed line from each of the frames in a shot and arranging them from top to bottom to create an image. Figure 1 illustrates the concept for a video shot of S frames. The figure shows horizontal tomography image, TH, created at height HT from the top-edge of the frame and a vertical tomography image, TV, created at position WT from the left-edge of the frame. The height of the tomography images is equal to the number of frames in a shot. Other line patterns can be used in addition to the vertical and horizontal tomography patterns shown in Figure 1; e.g., left and right diagonal patterns and any other arbitrary patterns.

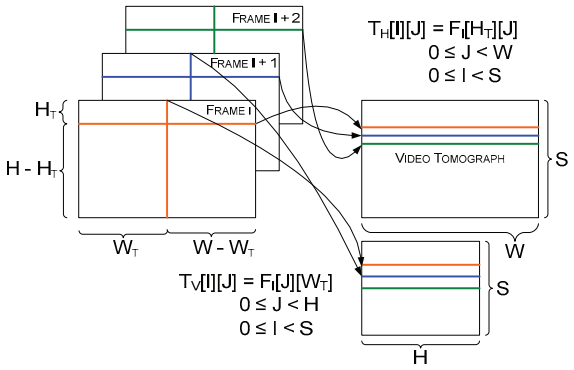

Figure 1. Video Tomography for a video shot of S frames with dimensions WxH

2.1 Signature Generation Using Video Tomographs

The image obtained using the composition process shown in Figure 1 captures the spatio-temporal changes in the video. The position of the scan line (HT or WT) strongly affects the information captured in the video tomograph. When scan lines are close to the edge (e.g., HT < H/5) the tomograph is likely to cut across background as most of the action in movies is at the center of the frame. Any motion in a tomograph that mainly cuts a static background would be primarily due to camera motion. On the other hand, with scan lines close to the center (e.g., HT = H/2) the tomography is likely to cut across background as well as foreground objects and the information in the tomograph is a measure spatiotemporal activity that is a combination of local and global motion. For video identification, capturing the interactions between global and local motion are critical and scan lines at the center of the frame are used.

Horizontal and vertical tomography for a 180 frame shot from the movie Shrek is shown in Figure 2. The tomographic images are created using only the luminance component; this has the side effect of making the system robust to color variations. The horizontal and vertical

patterns (figs. 2.d and 2.f) were combined using a pixel wise OR operation to create a composite image (fig. 2.b). The pattern edge images are aligned at the center to compose the image. A second composite image was created by combining the left and right diagonal patterns. The two composite images thus created form the basis for the video signatures.

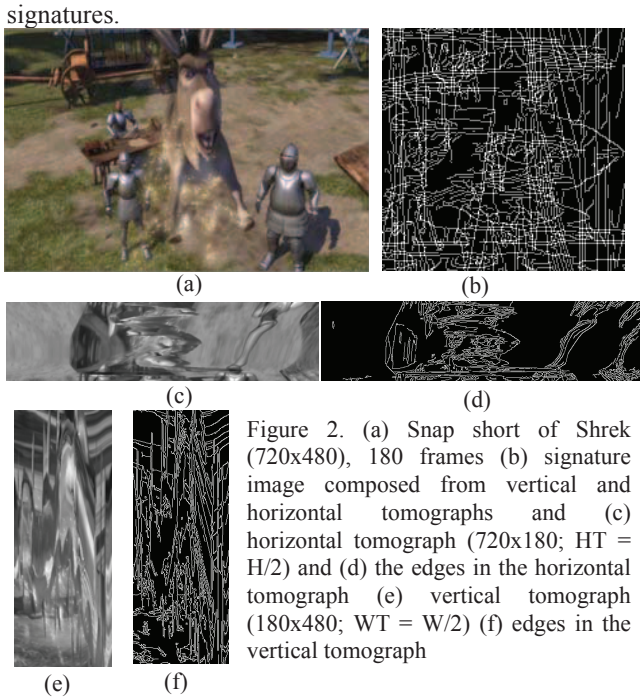

Figure 2. (a) Snap short of Shrek (720x480), 180 frames (b) signature image composed from vertical and horizontal tomographs and (c) horizontal tomograph (720x180; HT = H/2) and (d) the edges in the horizontal tomograph (e) vertical tomograph (180x480; WT = W/2) (f) edges in the vertical tomograph

The key component of the signature is the number of level changes at discrete points in the composite images. The level changes were measured along horizontal and vertical lines at predetermined points in composite images. The number of such points determines the complexity and length of a signature. Figure 3 shows eight horizontal and vertical positions used. At each of these positions on a tomograph edge image, the number of level changes is counted; i.e, the black to white transitions representing the number of edges crossed along the line. This count can be as high as half the width of an image and is stored as a 16 bit integer. The 16 counts on the horizontal-vertical composite and the other 16 edge counts on the diagonal composite form a 64 byte signature for each video clip. The signature size is always 64 bytes irrespective of the number of frames in a clip. Since signatures are not created for individual frames, this solutions results in a compact signature and the computational cost of finding a match is very low.

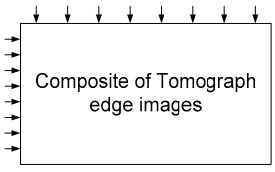

Figure 3. Level changes measured at eight equally spaced horizontal and vertical positions

2.1 Signature Generation Complexity

Generating the signatures for a video clip has relatively low complexity. The complexity is dominated by the complexity of edge detection in tomographic images. On a 2.4 GHz Intel Core 2 PC it takes about 100 milli seconds to generate a video signature for a 180 frame video clip with 720x480 resolution. At 30 frames per second, the complexity of signature generation is negligible and can be implemented in standard video player without sacrificing playback performance.

3. PERFORMANCE ANALYSIS

The performance of the proposed video signatures was evaluated by testing this on the data set used for content based copy detection at ACM International Conference on Image and Video Retrieval (CIVR 2007) [11]. The competition covered the following scenarios: 1) Transformed full-length movies with no post production and a possible decrease of quality (camcording i.e.) 2) Short segments on TV streams with possibly extensive large post-production transformations, and 3) Short videos on the Internet with various transformations (may be extracted from a TV stream). The teams which have taken part in the competition are: 1) IBM T.J. Watson Research Center, USA 2) ADVESTIGO, France, 3) City University of Hong Kong, Hong Kong, 4) Institute of Computing Technology, Chinese Academy of Sciences, and 5) Bilkent University RETINA Group, Turkey.

3.1 Test Database

The video data base had about 100 hours of video material that included web video clips, TV archives, and movies. The videos covered diverse set of programming including documentaries, movies, sports events, TV shows, and cartoons. The videos have different bitrates, different resolutions and different video format.

3.2 Query Construction

The queries were constructed using copies of videos in the database, the videos had length from 5 minutes to 1 hour. The queries were transformed versions of the original videos; transformations such as re-encoding with heavy compression, noised, slightly retouched, cropping, and screen recording. The task of the system to determine whether the query is in the database and if so, identify the matching video. The queries used are described in the first three columns of Table 1.

3.3 Experimental Results

Video signatures were generated for the entire database using the proposed approach. Signatures were also generated for each query, and the matches are detected using a distance between the signatures in the original and

the query. Euclidian distance in 16-dimensions (since the signature is composed of 16 short-integers) was used and the mean distance was calculated. Videos with a mean signature distance smaller than 2200 were identified as matches. The threshold was determined experimentally. The process of generating signatures for the queries was timed and is used as a performance metric.

Table 1: Video Query Descriptions and Results

No.	Origin	Transformation	minimum distance	Origin	Match
1	movie27	color adjustment + Blur	449.26	27	1
2	not_in_db		empty	empty	1
3	movie8	Re-encoding + color adjustment + cropping	224.83	8	1
4	not_in_db		empty	empty	1
5	movie44	Re-encoding with strong compression	1614	44	1
6	movie76	frontal camcording + subtitles	1372	76	1
7	not_in_db		empty	empty	1
8	not_in_db		798	31	0
9	movie9	colors phase modification + color adjustment	139	9	1
10	movie21	non frontal camcording	1717	21	1
11	movie37	frontal camcording	1795	4	1
12	not_in_db		empty	empty	1
13	movie11	flip	226	11	1
14	movie17	resizing + subtitles	1810	17	1
15	movie68	resizing (longest video)	1003	68	1

Table 1 shows the result for the 15 query videos, the first column indicates query number, second column shows the original video the query came from, the third column correspond to the transformation used. Minimum distance indicates the numeric value of the minimum mean distance of all comparisons below the threshold of 2200, origin column is where the minimum value comes from, and the final column shows whether a match has been found. The entry 'not_in_db' indicates that the query is not in the database and the entry 'empty' indicates that no videos meeting the criteria were found. For the 15 queries used, only one video was misclassified because the minimum value was located under the threshold, the query was incorrectly matched to video 31.

Table 2: Performance Comparison

Team - run	Score	Search Time
Advestigo	***0,86***	64 min
Bilkent	n/a	n/a
Chinese academy of sciences - 1	0.46	41 min
Chinese academy of sciences - 2	0.53	14 min
City university of Hong Kong	0.66	45 min
IBM - 1	***0.86***	44 min
IBM - 2	0.73	68 min
IBM - 3	0.8	99 min
Video Tomography Signatures	***0.93***	***30 min***

Table 2 presents the performance against the results from the CIVT 2007 evaluations. The proposed approach has a has a precision of 14/15 = 0.93 and time spent in query signature generation and search is 30 minutes, well below the best performing run from IBM.

4. CONCLUSION

This paper presents a novel low complexity tool for video identification. The system uses video signatures based on video tomography. The signatures are designed for a group of frames (shots) and have low complexity for both creation and matching. The performance was evaluated by comparing with the CIVR 2007 Video Copy Detection results. The results show that the proposed approach outperforms the results from the 5 organizations reported at CIVR 2007. The results also show that the proposed approach is robust to transformations common in video copying and distribution. The proposed signature is compact and has low extraction complexity.

5. REFERENCES

[1] G. Doerr and J.-L. Dugelay, "A guide tour of video watermarking," Signal Processing: Image Communication, Volume 18, Issue 4, April 2003, Pages 263-282.

[2] T.T. Ng, S.F. Chang, C.Y. Lin, and Q. Sun, "Passive-blind Image Forensics," in Multimedia Security Technologies for Digital Rights, Elsevier (2006).

[3] W. Luo, Z. Qu, F. Pan, J. Huang, "A survey of passive technology for digital image forensics," Frontiers of Computer Science in China, Volume 1, Issue 2 , May 2007, pp. 166 – 179.

[4] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford, "Video copy detection: a comparative study," In Proceedings of the 6th ACM international Conference on Image and Video Retrieval, CIVR '07, pp. 371-378.

[5] T. Can, and P. Duygulu, "Searching for repeated video sequences," Proceedings of the international Workshop on Multimedia information Retrieval, MIR '07, pp. 207-216.

[6] Y. Yan, B.C.Ooi, and A. Zhou, "Continuous Content-Based Copy Detection over Streaming Videos," 24th IEEE International Conference on Data Engineering (ICDE) 2008

[7] C.-Y. Chiu, C.-C. Yang, and C-.S. Chen, "Efficient and Effective Video Copy Detection Based on Spatiotemporal Analysis," Ninth IEEE International Symposium on Multimedia, 2007, pp.202-209.

[8] N. Guil, J.M. Gonzalez-Linares, J.R. Cozar, and E.L. Zapata, "A Clustering Technique for Video Copy Detection," Pattern Recognition and Image Analysis, LNCS, Vol 4477/2007, pp. 451-458.

[9] A. Akutsu and Y. Tonomura, "Video tomography: An efficient method for camera work extraction and motion analysis," Proceedings of the 2nd international Conference on Multimedia, ACM Multimedia 94, 1994, pp. 349-356.

[10] A. Yoshitaka and Y. Deguchi, "Video Summarization based on Film Grammar," Proceedings of the IEEE 7th Workshop on Multimedia Signal Processing, Oct. 2005, pp.1-4.

[11] A. Joly, "Video Copy Detection Showcase," CIVR 2007, http://staff.science.uva.nl/~civr2007/videocopy.php.