ACCURACY AND STABILITY IMPROVEMENT OF TOMOGRAPHY VIDEO SIGNATURES

POSSOS, Sebastian. KALVA, Hari
Department of Computer Science and Engineering, Florida Atlantic University, Boca Raton, FL 33431
Email: spossos@fau.edu. Hari.kalva@fau.edu.

ABSTRACT

Video signature techniques based on tomography images address the problem of video identification. This method relies on temporal segmentation and sampling strategies to build and determine the unique elements that form the signature. In this paper an extension for these methods is presented; first an implementation for high contrast sub-sampling and up-sampling filters is used to increase tomography edge stability, then a robust temporal video segmentation system is used to replace the original method applied to determine shot changes more accurately, and finally a new feature extraction method, derived from the previously proposed sampling pattern, was implemented and tested, resulting in a highly distinctive set of signature elements. The video set for the test was recorded from live broadcast TV, using as queries 20 different commercials, with the objective of analyzing the audit proficiency of the signatures for new applications. The performance obtain over the version presented in [3] is significant with 99.58% of Recall and 99.35% of Precision prediction..

Keywords— video tomography, video signature, shot detection, independence, sampling, interpolation, histogram.

1. INTRODUCTION

The sheer volume of today's media distribution options through formal and informal channels (for example through websites like YouTube) make video identification a key, taking into account that copyright laws are continuously challenged and tracing and controlling of video content is becoming a necessity. Research areas like copy detection, multimedia indexing and content retrieval are gaining importance to a point where MPEG has called for proposals for video signature standardizing [1].

A proposal based on video tomography signatures is presented in [3]. There, a two component signature is proposed consisting of a shot signature, which is globally unique, and a locally unique frame signature. [3] Stands out for exploring the uniqueness and independence aspects of video tomography signatures. This paper shows that the stability and accuracy of tomographic signatures can be further improved through several techniques which include image normalization through the use of sub- and up-sampling filters, a improved shot detection algorithm as well as a new feature extraction method used over each shot segment. We applied this solution for ad tracking in videos. This is achieved through an experiment that involved querying for 20 different commercials on a recorded live TV broadcast and comparing the previous and current results.

These results are preceded by a background section and an overview of the enhancements implemented in the original video tomography signature algorithm. Finally, to wrap up conclusions are presented and future work is suggested.

2. RELATED WORK

2.1. Video Identification

Among the ways to identify video, watermark based techniques and content based techniques are well known [2].

Digital watermarking embeds a watermark in the video to identify its source. It's first proposed in [4] as a means for video identification, but its principal drawback is the large amount of "un-watermarked" content already being distributed.

Content based identification, on the other hand, uses the content of the video to compute a unique signature based on various video features. Several video identification system surveys are presented in [9] and [10].

A solution for copy detection in streaming videos is proposed in [11]. There, the similarity between video sequences is measured; the similarity measurement is done by comparing a composition of frame fingerprints extracted from each video individual frames. Also [12] and [13] base their technique on key frame analysis. In [12] a clustering technique is proposed where key frames for each cluster of the query video are taken, then a search based on key frame distribution is performed to determine similarity regions in the target videos. [13] uses local features, extracting key frames and matching them certain features against a database.

Many of the content based video identification methods use video signatures generated from individual frame content, like those previously mentioned; the downside of this methods is that they add a large amount of overhead and complexity, especially in long duration videos, as they require feature extraction and comparison on a frame basis.

The Tomography video signatures propose an alternative to this problem.

2.1.1. Tomography Video Signatures

In video tomography, a sample consisiting on a single line is extracted from each frame of the Y component of a video. These single lines are sequentially transposed to create a new tomography image. Figure 1 shows the process of generating a tomography image. This image is then processed through a Canny edge detector which obtains the edges to reveal patterns with high spatio-temporal correlation [3].

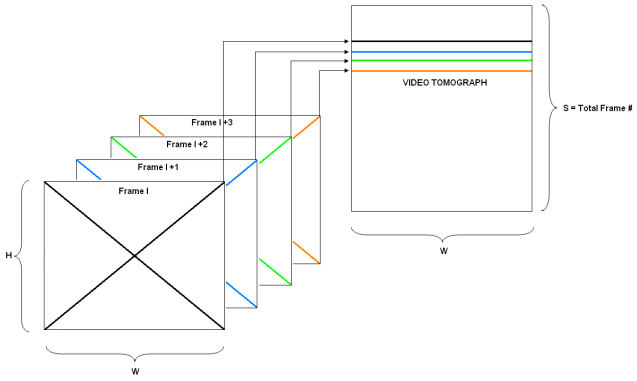

Figure 1. Video tomography extraction for 1 of 6 components

First, the video is scaled to a resolution of 360x240. Then, three composite tomography images are generated from six different sample patterns; two upper diagonals (from the upper corners to the middle), two lower diagonals (from the middle to the lower corners) and two regular diagonals (joining opposite corners). Both diagonals of each set are superimposed and a composite signature image is created using the OR operation.

The amount of level changes (edges) on these three composites are counted on 8 predefined vertical and 8 predefined horizontal lines, which are evenly distributed along the edge tomography. This produces 16 counts on the horizontal-vertical composite and 16 edge counts on the diagonal composite. These are then combined to form a 24 short integer signature for each shot.

Signatures comparison is achieved by finding the minimum Euclidean Distance between the points in a 48 byte 48-dimensional space:

$$D = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} \quad (1)$$

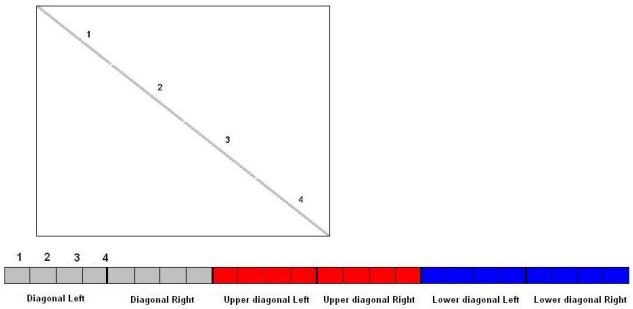

Figure 2. Frame tomography signature

By using this same technique, extraction of frame tomography signatures are also viable and can be performed from the same tomography image used to generate the shot signature; after retrieving the tomography lines of a frame, then the signature is obtained by dividing these lines into 4 segments and counting the edges among each of them. The result is a 48 byte 24-dimensional signature as shown in Figure 2.

2.2. Shot Detection

A shot in video is defined as a collection of consecutive frames that belong to the same camera, meaning that the correlation between those images is very high, been able to detect these segments is of crucial importance in several applications, like video coding where accurate P distance and I frame decision can result on a big difference in compression performance.

Among the basic methods for shot detection, pixel difference techniques and Histograms are worth mentioning. Both of them rely on finding differences between frames to judge if a frame boundary has been found.

2.2.1. Pixel difference

The pixel difference methods are one of the most straightforward approaches to find shot changes. They calculate a value, which represents the overall change in pixel intensities in the frame [5].

[6] makes use the sum of absolute pixel intensity differences between two frames as a frame difference. [7] and [8], on the other hand, calculate the number of pixels that change their value more than a predefined threshold.

In any case, the total sum of these pixel differences are compared against a scene change threshold to determine if a shot change has been found.

The downside of these techniques is their sensitivity to noise and camera motion as well as to lighting changes (for example a flashlight) [5].

2.2.2. Histogram

A histogram shows the distribution of pixel values in a frame. The most straightforward method in histogram comparison is the simple summation of the pair wise bin

differences [5]. Examples of these can be found in [6] and [8]. Linear combinations with different weights of the bin-wise differences can also be used, if some gray levels or colors are considered of greater importance [5].
Histogram techniques are very robust against noise and object movement. However, the histogram describes only the distribution of color or gray scale values. It doesn't take into account any spatial information of the image. In the same manner, small, visually significant image regions may not produce strong peaks in the histograms and hence can be neglected in the frame comparison [5].

2.3. Image filtering

A digital image is a finite discreet representation of a picture, where each element or pixel stores a sampled value of an original image, and as the number of samples or pixels increases, the more accurate the representation of the original picture is. Sometimes in the capturing process some artifacts can appear in the scanned image, like moiré patterns which are the result of the superposition of two grid-like images [19], in case of a digital camera the grid can be the internal sensor array that captures the image, and the second could be a displayed picture on a TV screen. To remove these distortions digital filters have to be used, a digital filter is a system that can operate on the discreet or sampled space and boost or diminish defined frequency components on a given signal.

2.3.1. Image downscaling filters

To decrease the size of a digital image, removal of distributed pixel lines is done, but because of this resample process the Nyquist criterion is not going to be met because of the existence of high frequency components in the image, to eliminate or decrease the effect of this components a low pass filter should be applied first, before down-sampling the image, after that the filtered image would look blurry because no sharp changes (high frequency components) are not going to be present, then with the re-sampling procedure, the new smaller image will represent accurately the original image.[20]

2.3.2. Image up-scaling filters

To increase the size of a digital image, an interpolation filter is needed, the purpose of this filter is to increase the size of an image by generating intermediate pixels on defined locations [20], opposed to the downscaling filters the intention of this filter is to increase the contrast of the image, because when the image is up-sampled, by simply copying a neighbor pixel, the image will look pixelated, to solve this a interpolation filter should be applied. The interpolation filter also includes a low pass filter which is responsible for the smooth transition between pixels, but then the filter should also take into account the surrounding pixels to replicate the behavior and sharpen the details.

3. PROPOSED APPROACH

In order to achieve a stability and accuracy improvement two main techniques were applied: shot detection and frame size normalization. Both of them are discussed in the following subsections.

3.1. Shot detection

During the signature comparison experiments, the shot detection upgrade proved to be a decisive factor in signature accuracy and repeatability.
As a first approach, the crater distance method was explored as a shot detection technique. By taking three consecutive frames at a time, the Euclidean distance between frame signatures was measured and a pattern, specifically a depression pattern (meaning a high – low – high distance value), was searched. If this pattern complied with a predefined threshold value then a shot was declared.
As tomography based method, it incurs in little processing overhead. However, in this paper a new approach is explored which reduces complexity greatly. This technique takes a DC frame to start with, which reduces the size of the image to a $1/16^{th}$; the difference of the DC frame values of two consecutive frames is then calculated and a histogram created. An analysis of this histogram for three consecutive frames is used to reveal shot boundaries according to predefined thresholds.
One added value found from the analysis of this histogram is the possibility to catalog transitions like fade in or fade out from a solid color, and detection of sudden light changes, like the one present when a flash occurs, or from an explosion. The detection of this type of effects increases the shot detection reliability, because the most common false positive statistic reported by shot detection systems comes from abrupt luminance changes, introducing a large variation in almost all frame pixels, causing a large increment on the calculated SAD.

3.2. Frame size normalization

Frame size normalization to 320x240 images is achieved by using two type of sampling filters as opposed to previous pixel sampling.
The first filter type is an anti-aliasing filter used when frame reduction is needed. Aliasing artifacts appear when a signal is resampled to a lower resolution, if no anti-aliasing filter is used, some high frequency components from the original signal can "alias" to un-existing lower frequency components resulting on an erroneous image behavior. If an anti-aliasing filter is used, before the re-sampling procedure is done, a low pass filter is applied to the original image, then the high frequency components are attenuated, causing that the reduced signal power of those undesired components do not add noticeable interference to the final image.

The second filter type is an interpolation filter used to scale up the image in case the current video frame is smaller than the normalization size. The purpose of this filter is to generate the intermediate pixels taking into account the existing correlation from the original small image, keeping a smooth behavior between them. In H.264 standard [18], an interpolation filter is used for sub pixel motion estimation, the generation of those sub pixels can be used to increase the size of the original image, taking into account the high level of correlation of those pixels and the original image. The purpose is that the current block can be compared against fractional pixels, taking advantage of curve fitting properties of the interpolation filter, allowing a better motion match and a reduction on the SAD values.

The use of this type of filter can introduce a new issue because if the generated pixel behavior is too soft, then the enlarged image is going to look blurrier or out of focus; this would interfere with the edge detection process, resulting on a different tomography image, resulting on extraction of a signature with not enough similarity to the original. To face this matter, the selection process for the interpolation filter has to take into account the contrast level that it can add to the image, to facilitate the correct feature extraction at the generation of the signatures.

The usage of these filters makes the system robust to size transformations and to changes in brightness, because it emphasizes the spatial segmentation and maintains the correlation between pixel values, also avoids alias and noise introduced in the sampling process. For the down-sampling process a 5 tab version of the Lancsoz2 filter was selected, while for the up-sampling, the selection was a 6 tab version of the Lancsoz2 filter.

These upgrades to the system resulted in an increase in the stability of the tomography image, as consequence of a more robust canny edge detection which now has more correlated pixels as input.

3.3. New feature extraction

12 new methods for feature extraction were designed and tested to determine the best option that could outperform the original signature. These new methods followed the idea to count pixel transitions from shot segments, but with several modifications. Instead of counting through predefined lines [3], now the shot image is divided into 16 blocks, for each block the following characteristics are measured: Amount of white pixels inside the block, amount of horizontal and amount of vertical transitions from black to white. Then 3 tomography line combinations where used, full star pattern, right Z and left Z, and large X.

Full star pattern refers to the use the original pattern and compositions to generate the tomography images, as described on [3].

Right Z refers to the use of 3 diagonals, top right to middle left, top right to bottom left, and middle right to bottom left, to generate 3 different tomography images.

Left Z refers to the use of the opposing diagonals of Right Z, generating also 3 different tomography images.

Large X refers to the use of the two large diagonals to generate 2 tomography images, this last signature result in a smaller signature.

From the performance analysis of each new signature, a combination of full star pattern, and count of horizontal and vertical transition result in the best configuration for the solution. In this case the new signature has a size of 96 bytes, double from original.

4. PERFORMANCE EVALUATION

To determine if the new signatures design was able to outperform the original design, a similar scenario to the one presented on [3] was used, the master database consisted on 350 videos of 3 minutes each, and the query database had 100 videos, each query video belonged to one video from the master database. Then an exhaustive search is performed by comparing each query signature against the master database, and finally depending on the comparison threshold is determined if there is a match and if it is true on what position. A tolerance of +- 1 second has been determined for the correct time location of the query sequence in their matching master video.

Table 1: Performance Summary

Description	**2 Sec**	**5 Sec**	**10 Sec**
Total instances	35'000	35'000	35'000
Independent clip pairs	34'900	34'900	34'900
Identified as Independent	34'806	34'750	34'422
True Positive (tp)	34'805	34'748	34'421
False Positive (fp)	1	2	1
False negative (fn)	5	52	379
True negative (fp)	99	98	99
Identified as dependent	94	150	478
Recall: tp/(tp+fn)	99.99%	99.85%	98.91%
Precision: tp/(tp+fp)	100%	100%	100%
Prediction Precision	99.60%	98.68%	99.79%

Table 2: Performance Comparison

	Recall		
Original	99.16%	98.94%	96.26%
New	99.99%	99.85%	98.91%
	Precision		
Original	100%	100%	100%
New	100%	100%	100%
	Precision Prediction		
Original	94.44%	94.44%	96.30%
New	99.60%	98.68%	99.79%

A second test has been done for the proposed application using a HD 24 hour video recorded from a public channel broadcast on May 23. 20 different commercials are selected from it, signature generation is carried out on the original

video and the 20 advertisements then the matching procedure is performed.

For the comparison, the Euclidean distance is calculated between the first shot signature from the selected commercial and each shot detected after a hardcut from the 24h recording.

Table 3: Advertisement audit application performance

Commercial index	# of appearances	# of correct detections	# of false detections
Commercial 1	4	4	0
Commercial 2	1	1	0
Commercial 3	1	1	0
Commercial 4	1	1	0
Commercial 5	2	1	1
Commercial 6	1	1	0
Commercial 8	1	1	0
Commercial 9	1	1	0
Commercial 10	11	11	0
Commercial 11	1	1	0
Commercial 12	1	0	1
Commercial 13	5	5	0
Commercial 14	1	1	0
Commercial 15	1	1	0
Commercial 16	1	1	0
Commercial 17	7	7	0
Commercial 18	1	1	0
Commercial 19	1	1	0
Commercial 20	1	1	0

5. CONCLUSIONS

This paper presents a development to the video identification method and a practical application of this system to solve a current industry challenge.

As seen on table 2, the performance of this new implementation for the signature generation has been upgraded, having an average improvement of 1.46% in Recall and 4.29% in Precision prediction. On table 3, this system proves to be an excellent method for advertisement or video content audit, and can be extended to monitor any kind of prerecorded material.

For future development, the implementation of an adaptive filter for up-scaling and down-scaling can lead to a improvement in signature uniqueness, taking into consideration that the current method has fixed set of filters, which work very well but could perform even better if automatic adaptation could be added.

6. REFERENCES

[1] MPEG Video Subgroup, "Updated Call for Proposals on Video Signature Tools," MPEG2008/N10155, October 2008, Busan, KR.

[2] G. Leon, "Content Identification using video tomography", M.Sc. Thesis, College of Engineering and Computer Science, Florida Atlantic University, August 2008

[3] S. Possos, A. Garcia; M. Mendolla, J. Schwartz, H. Kalva, Multimedia and Expo, 2009. ICME 2009. IEEE International Conference, June 28 2009-July 3 2009, Pages 698 – 701.

[4] 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.

[5] J. Korpi-Anttila, "Automatic color enhancement and scene change detection of digital video", Licentiate Thesis, Department of Automation and Systems Technology, Helsinky University of Technology, November 2002.

[6] A. Nagasaka, Y.Tanaka, "Automatic Video Indexing and Full-Video Search for Object Appearances", Visual Database Systems, II, Elsevier Science Publishers, 1992, pp. 113 – 127

[7] K. Otsuji, Y. Tonomura, Y. Ohba, "Video Browsing Using Brightness Data", Visual Communications and Image Processing, SPIE 1606, 1991, pp. 980 – 989

[8] H.J. Zhang, A. Kankanhalli, S.W. Smoliar, "Automatic Partitioning of Fullmotion Video", Multimedia Systems, Vol. 1, 1993, pp. 10 – 28

[9] X. Fang, Q. Sun, and Q. Tian, "Content-based video identification: a survey," Proceedings of the Information Technology: Research and Education, 2003. ITRE2003. pp. 50-54.

[10] 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.

[11] 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

[12] 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.

[13] G. Singh, M. Puri, J. Lubin, and H. Sawhney, "Content-Based Matching of Videos Using Local Spatio-temporal Fingerprints," Computer Vision – ACCV 2007, LNCS vol. 4844/2007, Nov. 2007, pp. 414-423.

[14] 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, pp. 349-356, 1994.

[15] Manjuntah, P. Salembier and T. Sikora "Introduction to MPEG-7: Multimedia content Description Interface", John Wiley and Sons, 2002.

[16] M. Bertini, A. Del Bimbo, W.Nunziaty, "Video Clip Matching Using MPEG-7 Descriptors and Edit Distance", Image and video retrieval, pp .133-142, 2006.

[17] B. Baker, "Anti-Aliasing, Analog Filters for Data Acquisition Systems", Application Note AN699 Microchip Technology, pp 4 – 5, 1999.

[18] ITU-T, "ITU-T Rec. H.264 Series H: Audiovisual and multimedia systems", Avanced video coding for generic audiovisual services, pp. 160 – 164, 03/2005.

[19] R. Chang, J. Sheu, C. Lin, H. Liu, "Analysis of CCD moirépattern for micro-range measurements using wavelet transform", Optics and Laser Technology, pp 43 – 47, 2003.

[20] K. Turkowsky, "Filters for Common Resampling Tasks", Filters for Common Resampling Tasks, Apple Computer, April 10 1990.