Content-based Video Signatures based on Projections of Difference Images

Regunathan Radhakrishnan and Claus Bauer
Dolby Laboratories Inc
100 Potrero Ave, San Francisco, CA
Email: {regu.r,cb}@dolby.com

***Abstract*— We propose a novel video signature extraction method based on projections of difference images between consecutive video frames. The difference images are projected onto random basis vectors to create a low dimensional bitstream representation of the active content (moving regions) between two video frames. A sequence of these signatures serves to identify the underlying video content in a robust manner. Our experimental results show that the proposed video signature is robust to most common signal processing operations on video content such as compression, resolution scaling, brightness scaling.**

I. Introduction

The goal of content-based video signature extraction is to obtain a compact bitstream representation of the underlying video content that is robust to various signal processing operations. Past work on video signature extraction can be broadly classified under two approaches. The first approach derives signatures from a subset of frames selected from the video sequence either by shot detection followed by key frame extraction[2] or by random sampling[3]. One of the main drawbacks with this approach is that the extracted signatures cannot identify a portion of the video clip that is shorter than the original shot. The method requires the same subset of frames to be selected from a video shot for signature extraction. The second approach is to apply robust image hashing techniques such as [4],[5] to individual video frames. Since this method extracts fingerprints from individual video frames it can identify smaller sequences of video than the first method can. Our proposed scheme follows the second approach. However, unlike prior approaches, our method works with difference images. The use of difference images leads to a derivation of fingerprint bits from active moving regions of the video frame alone. This, in turn, translates to a smaller video fingerprint length for our proposed method than for previous methods which try to capture all the information in an individual video frame. A coarse representation of the difference image of consecutive video frames is used to extract signature bits. We use the robust visual hash proposed in [1] to create a low dimensional bitstream representation of the difference images in the video sequence. The extracted signature serves as an identifier for the underlying video content and is invariant to most signal processing operations.

II. Proposed Video Signature Extraction

The proposed signature extraction consists of the following two steps: (i) Video feature extraction (ii) Computation of the signature bits from extracted features.

A. Video Feature Extraction

The goal of the video feature extraction block is to obtain a set of robust features that are invariant to certain signal processing operations on the video. In this section, we describe a set of such robust features extracted from a coarse representation of the absolute values of intensity differences between consecutive frames. This set of features is chosen such that they survive a variety of processing of the video, including compression, color space conversion, intensity adjustment, addition of computer generated graphic objects and format conversions.

Figure 1 illustrates the proposed video feature extraction approach. An absolute difference image is computed between adjacent video frames (frame(n) and frame(n+1)). This step ensures that we only extract signature bits based on active moving regions between video frames. We verified that the amount of motion between two adjacent frames can be as small as in a talking head sequence and still our method can identify the sequence. It will, however, fail if the video sequence is of a static graphic image and there is no activity. The next step is to downsample and crop the absolute difference image. This step ensures that the extracted features are invariant to interlaced and progressive video formats and addition of graphics and letterboxes on the corners of the individual frames. Then, the cropped absolute difference image is tiled horizontally and vertically. Finally, the intensities of the absolute difference image within each tile are summed up to obtain a coarse absolute difference image. Let us represent the cropped absolute difference image by Δ. We obtain a coarse representation (Q_v) of Δ by averaging pixel intensities in image blocks of size $W_x \times W_y$ such that $K \times W_x = 120$ and $L \times W_y = 160$. Q_v is of size $(K \times L)$

$$Q_v(k,l) = \frac{1}{W_x * W_y} \sum_{i=(k-1)W_x}^{kW_x} \sum_{j=(l-1)W_y}^{lW_y} \Delta(i,j)$$

$$k = 1, 2...K; l = 1, 2...L$$

Here i and j represent the indices for the horizontal and vertical dimensions for the absolute difference image Δ. k and l represent the indices of the sub-blocks of the absolute difference image Δ. This coarse representation (Q_v) helps us

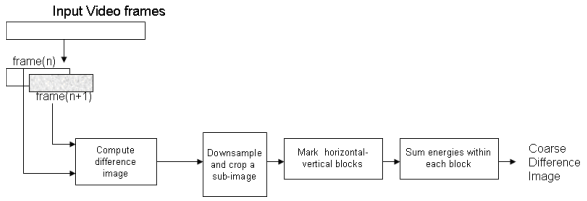

Fig. 1. Video Feature Extraction

to achieve robustness by allowing for certain variations within a block while preserving the average intensity within a block.

B. Robust Hash

This block takes as input the matrix Q_v, and generates the signature by generating K hash bits. We use a robust hash function for this purpose as small perturbations in the video features caused by signal processing operations such as compression, filtering etc would not change the hash bits drastically. The Robust hash function also serves to reduce the bit-rate of the signature stream. Let us represent the dimensions of the matrix Q_v by $(M \times N)$. By using a robust hash, instead of sending $(M \times N)$ values we only send K bits. A robust hash function is unlike a regular cryptographic hash function. A cryptographic hash function changes its output for every single bit change in the input. However, we would like our hash output to change slowly with small changes in features. This would enable us to allow for certain signal processing operations on the content which do not change the content but only slightly disturb the features. We use one such robust hash function proposed in [1] for generating the hash bits from the feature matrix, Q_v. We generate K random matrices each with the same dimensions as the matrix, $Q_v(M \times N)$. The matrix entries are uniformly distributed random variables in $[0, 1]$. The state of the random number generator is set based on a key. Let us denote these random matrices by $P_1, P_2, ...P_K$ each of dimension $(M \times N)$. We compute the mean of matrix P_i and subtract it from each matrix element in P_i (i goes from 1 to K). Then, the matrix Q_v is projected onto these K random vectors as shown below:

$$H_k = \sum_{i=1}^{M} \sum_{j=1}^{N} Q_v(i,j) * P_k(i,j) \quad (1)$$

Here H_k is the projection of the matrix Q onto the random vector P_k. Using the median of these projections (H_k, $k = 1, 2K$) as a threshold, we generate K hash bits for the matrix Q_v. We generate a hash bit '1' for k^{th} hash bit if the projection H_k is greater than the threshold. Otherwise, we generate a hash bit of '0'.

III. Experimental Results

A. Test Content

The test content used for the performance assessment of the proposed video signature extraction consisted of 36 one-minute original video clips of a variety of content types in various formats. The video formats include standard broadcast formats such as Standard Definition (SD) 480/30i, High Definition 1080/30i and 720/60p (HD). Each of the test clips was processed in a variety of ways to simulate the processing in modern broadcasting and post production facilities. We considered the following processing operations: Brightness modification by 10%, Median noise reduction, Addition of random film grain type noise, MPEG compression and decompression at 2,4,8 and 12Mbps, Down conversion from HD to SD, Logo and Graphic insertions. Overall, we created 432 test cases from the original 36 one-minute video sequences.

B. Parameter Settings

The signature extraction parameters were empirically selected and set as described below. The size of coarse difference image was set to be 8×9 ($K = 8, L = 9$). The number of random vectors for the hash was chosen to be 36. This means for every pair of frames we create a signature of length 36 bits.

C. Sensitivity of the Proposed Signature

Before we show the robustness of the proposed signature, we first illustrate how sensitive the signature is to the underlying content. Let us assume we perform a MPEG compression attack on certain content which causes x% of the original signature bits to be flipped. Now, when comparing the signatures of two different video files we would like the percentage of bits that flip to be much larger than x%. Then, we can say that the proposed signature is sensitive enough to be used as a content identifier and is robust to MPEG compression attack. To verify the sensitivity of the proposed signature, we perform the following for two scenarios. Under scenario 1, we compare the signatures from two different video files (A and B). We compute the histogram of hamming distances between signature of A and signature of B. Under scenario 2, we compare the signature of A against the signature of some modified version (e.g MPEG Compression) of A. We compute the histogram of hamming distances between the signature of A and the signature of modified A. A robust signature would have a distribution that is heavily skewed near hamming distance value of zero under scenario 2. A signature that is also sensitive to underlying content, should show minimal overlap between the histograms under scenario 1 and scenario 2. In other words, the BER for comparison between signatures of A and B should be larger than the BER for comparison of signatures of A and modified A. Figure 2 shows the comparison of histogram of hamming distances for scenario 1 and MPEG compression (scenario 2) and it confirms that there is indeed minimum overlap between the two histograms as we expected. Note that most of the hamming distance values for MPEG compression attack is below 15 whereas the hamming distance for two different files is atleast 15. The overlap is small even for one of the severe modifications on content (Rotation by 3^o) as shown in Figure 3.

D. Robustness of proposed video signature

In this section, we present more results to show the robustness of the proposed video signature to various signal processing operations. Towards that end, we first extract the signature from the original video content and then compare it against the

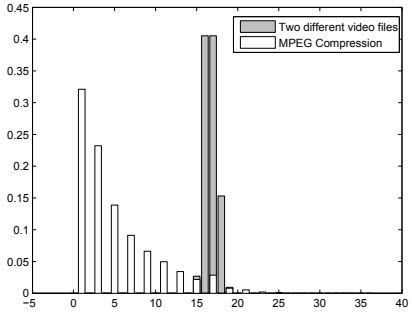

Fig. 2. Comparison between histograms for scenario 1 and scenario 2 (MPEG Compression)

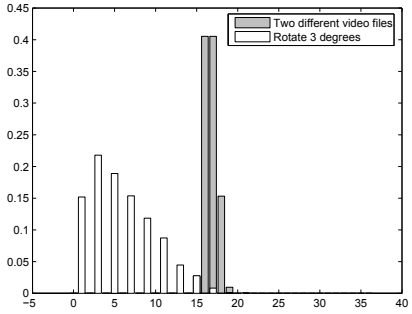

Fig. 3. Comparison between histograms for scenario 1 and scenario 2(Rotation by 3^o)

signature extracted from processed (attacked) video content using a hamming distance measure. Then, we compare the BER (Bit Error Rate) between signatures of original content and its modified content with the BER between signatures of two different video files. In the previous section, we showed that on an average 15 out of 36 bits flip, when comparing signatures of two different video files. This means that the average BER for this case is 41%. In the remainder of this section, we will compare this BER of 41% to the BER caused by various attacks to illustrate the robustness of the signature.

1) MPEG Video Compression: Tables I and II summarize the performance of the video signature for MPEG compression attack on HD sources and SD sources respectively. Note that MPEG compression at 8Mbps causes a BER of 15% which is smaller than the BER of 41%. Therefore, we can conclude that the proposed signature is robust to MPEG compression attack. The higher the bitrate used for compression the closer the compressed video is to the original. Consequently, fewer number of bits in the signature flip.

Figure 4 illustrates the effect of increasing the MPEG-compression bitrate from 500kbps to 4Mbps on the robustness of the signature for a SD source video.

2) Brightness scaling: Table III shows that the proposed signature is robust to brightness scaling by +/-10% as the maximum BER for this case (11.44%) is far less than the BER of 41%. Since the signature extraction is based on difference images, uniform scaling of the brightness across frames in

Attack	B	C	D
MPEG-8Mbps	463950	84438	0.1526
MPEG-12Mbps	445898	84672	0.1462

TABLE I
ROBUSTNESS OF PROPOSED VIDEO SIGNATURE FOR MPEG COMPRESSION ATTACK ON HIGH-DEFINITION SOURCES; B: $Num.BitErrors$; C: $Num.Frames$; D:$BER = \frac{BC}{(SigSize*C)}$, $SigSize = 36$

Attack	B	C	D
MPEG-2Mbps	174744	31170	0.1557
MPEG-4Mbps	160204	31329	0.1420

TABLE II
ROBUSTNESS OF PROPOSED VIDEO SIGNATURE FOR MPEG COMPRESSION ATTACK ON SD SOURCES;

video doesn't affect the signature bits as much as MPEG compression does. Also, note that brightness scaling by -10% causes slightly more number of bits to flip as some of the frames can turn out completely black after the intensity adjustment.

3) Down-Conversion and Up-Conversion: We perform the following attacks on HD content that occurs in practice. HD content is in 16:9 aspect ratio and when it is to be displayed on standard definition TV screen with 4:3 aspect ratio one of the following three is usually done:

- The 16:9 content is horizontally cropped to fit 4:3 aspect ratio. We refer to this down conversion as hcrop in Table IV.
- A letterbox is added to 16:9 content to fit 4:3 aspect ratio. There is no horizontal cropping or vertical stretching. We refer to this type of down conversion as Letterbox1 in Table IV.
- A letterbox of smaller size than in the case of Letterbox1

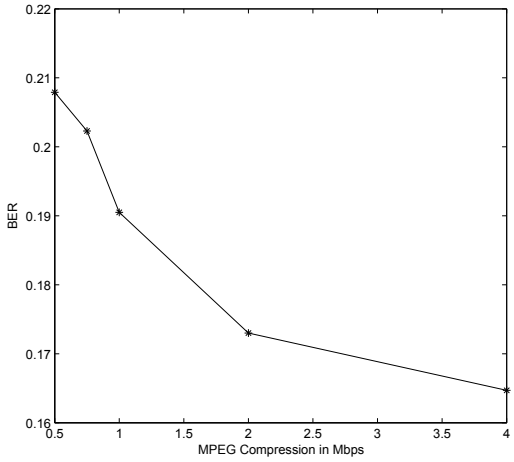

Fig. 4. MPEG-Compression Vs Bit Error Rate of the signature

Attack	B	C	D
Brightness Up 10%	567689	144233	0.1093
Brightness Down 10%	593478	144091	0.1144

TABLE III
ROBUSTNESS OF PROPOSED VIDEO SIGNATURE FOR BRIGHTNESS SCALING;

Attack	B	C	D
Letterbox1	482668	84546	0.1585
Letterbox2	441208	84396	0.1452
hcrop	437200	84693	0.1433

TABLE IV
ROBUSTNESS OF PROPOSED VIDEO SIGNATURE FOR DOWN CONVERSION FROM HD TO SD RESOLUTION;

Attack	B	C	D	PSNR(dB)
noise1	5568	3996	0.0387	44.49
noise2	6820	3996	0.0474	38.5
noise3	9610	3996	0.0668	34.62
noise4	12184	3996	0.0846	32.4

TABLE V
ROBUSTNESS OF PROPOSED VIDEO SIGNATURE AGAINST NOISE;

Attack	B	C	D
Rotation by 1^o	11300	1799	0.1744
Rotation by 3^o	12146	1799	0.1875

TABLE VI
ROBUSTNESS OF PROPOSED VIDEO SIGNATURE FOR ROTATION ATTACK;

is added to 16:9 content to fit 4:3 aspect ratio and there is also a small amount of horizontal cropping. This down conversion is a tradeoff between hcrop (more horizontal cropping) and Letterbox1 (no loss of content). We refer to this type of down conversion as Letterbox2 in Table IV.

For the aforementioned down conversion methods, we automatically detect the letter box and extract the smallest image region common to all of the down conversion methods and use that for signature extraction. Table IV shows the robustness of the signature against these down conversion attacks as the maximum BER is only 15.85% which is less than the BER of 41%.

We also converted SD content (4:3 aspect ratio) to HD content (16:9 aspect ratio) to study the robustness of the proposed signature for up-conversion process. Under this modification of the original signal, 14.25% of the signature bits were in error when compared with the signature of the original content. This error rate is similar to that for down conversion attacks.

4) Noise: We performed two noise attacks on the original content. In one attack, we employed a denoising algorithm on the original content and studied its effect on signature bits. We found that there were as many as 181466 bit errors in the signatures for a total number of frames equal to 88546 which is 5.69% Bit Error Rate (BER). The denoising attack has the smallest impact on the signature bits of all attacks we tried. In another attack, we varied amount of noise added to the original content and studied how the BER increases as a function of noise level. Table V shows the percentage of bits that flip for varying amount of Gaussian noise that is added to the content and BER is always less than 8.5%. Note that as PSNR goes down from 44.4dB to 32.4dB the BER increases from 3.87% to 8.46% as the amount of noise increases.

5) Rotation: We performed two rotation attacks on original content to study how robust the proposed signature is to geometric attacks. Since the signature extraction is based on a coarse representation of an image computed from a grid, geometric attacks which disturb the location of the features relative to the fixed grid would cause more bit errors than any other attack. Table VI shows the percentage of bits that flip for rotation attack and the BER of 18.75% for 3^o rotation is still less than 41%. Note that 1^o rotation has fewer bit flips than 3^o rotation. The rotation attack causes most number of bits to be flipped than any other attack as expected.

IV. CONCLUSION

We proposed a robust video signature extraction method based on projections of difference images onto random basis vectors. The extracted signatures were shown to be robust to various signal processing operations on video content such as MPEG compression, Spatial Scaling , Brightness Scaling etc. It was also shown to be sensitive to underlying content and hence can be used as a content identifier. However, the proposed signature is not robust to geometric attacks as is shown by the results for the rotation attack. Furthermore, the signature would not perform as well for frame-rate conversion attacks. We found that the signature is robust to +/-5% change in frame rate but for lower frame rates the difference image between consecutive video frames would not be the same as it was in the original video (especially so for high motion sequences). Our future work will focus on improving resilience to frame rate conversion, rotation and other geometric attacks.

ACKNOWLEDGMENT

The authors would like to thank Kent Terry of Dolby Labs for all the help with discussions, simulations and content acquisition.

REFERENCES

[1] J.Fridrich and M.Goljan, "Robust Hash Functions for Digital Watermarking", Proc. of ITCC, 2000.
[2] H.S. Chang, S. Sull and S.U. Lee, "Efficient video indexing scheme for content-based retrieval," in IEEE Trans. Circuits Syst. Video Technol., Dec 1999.
[3] S.S. Cheung and A. Zakhor, "Effcient Video Similarity Measurement with Video Signature", IEEE Transactions on CSVT,2003.
[4] Kozat, S.S. Venkatesan, R. Mihcak, M.K. , "Robust perceptual image hashing via matrix invariants", ICIP 2004.
[5] A.Swaminathan, Y.Mao and Min Wu, "Image Hashing Resilient To Geometric and Filtering Operations", MMSP 2004.