VIDEO COPY DETECTION BASED ON HEVC INTRA CODING FEATURES

Kai-Wen Liang, Yi-Ching Chen, Zong-Yi Chen, and Pao-Chi Chang
Department of Communication, National Central University
{kwliang, ycchen, zychen}@vaplab.ce.ncu.edu.tw, and pcchang@ce.ncu.edu.tw

Abstract-- This work utilizes the coding information in HEVC for video copy detection. Both directional modes and residual coefficients of the I-frames are employed as the texture features for matching. These features are robust against different quantization parameters and different frame sizes. The accuracy is comparable with traditional pixel domain approaches.

I. INTRODUCTION

Thanks to the tremendous growth of multimedia technologies and rapid development of network environment, the amount of the multimedia data is dramatically increased. Accordingly, duplicates of original media files with various bitrate, quality, and frame size are widely spread on the internet. These duplicates may lead to the problems of digital right management (DRM). Thus, copy detection technique which can automatically detect the duplicates is getting important.

The primary concept of content-based video copy detection (CBVCD) is that "the media itself equals to the watermarking," which exhibits an alternative to the watermarking for persistent identification of media content without additional information [1]. CBVCD first extracts the visual words in frames as the signature to represent the unique characteristics of contents. The copy detection is then performed by comparing the signatures of videos.

Traditionally, visual features are extracted from raw pixel data; referring to pixel domain (PXD) approaches. However, videos distributed and stored are almost necessarily compressed by coding techniques such as MPEG, H.264, or HEVC, which means an extra decompression process is needed for PXD approach. A more efficient approach is to operate the detection procedures in the compression domain (CPD) [2]; that is, extracting features from the encoded bitstream without fully decoding the video frames.

Numerous studies related to the CPD feature extraction on videos and images have been conducted in recent years. In MPEG4 I-frame coding as well as JPEG coding, discrete cosine transform (DCT) is applied to image blocks. The intensity and the shape information in a block can be represented by the DC [3] and AC coefficients respectively. Therefore, the features extracted and generated from the DCT coefficients have been applied to the copy detection. Zhang et al. used the AC coefficients to obtain edge information in video frames [4]. Li et al. analyzed the texture according to the amount of non-zero AC coefficients in one block of I-frames [5]. Some works such as [6] and [7] also addressed the importance of efficient feature generation and utilization.

Predictive I-frame coding is an innovation in H.264 where the DCT is applied to the residual signal instead of the original frame. Ali et al. [8] counted the number of intra 4x4 MBs within the I-frame and the number of skip MBs within the B-frame and P-frame, and then compared each other to generate the rank matrix as a signature for copy detection.

Practically, a duplicate of video may not have the same quality or resolution with the original. In the literature, the methods that count the number of prediction modes in I-frame, P- frame, or B-frame, and the number of non-zero AC coefficients [5] are easily affected by the change of quantization parameters (QP). The accuracy of the detection results will significantly decreases when they are applied to the same content with different QPs. In consequence, finding QP invariant feature is a very important issue in the compression domain approach.

High efficient video coding (HEVC) is the newest video standard which has outstanding compression performance. It provides finer predicted direction and hierarchical partition size than H.264. This study proposes a QP invariant video copy detection approach in the HEVC compression domain. It extracts the features, the directional modes and coefficients of residual, from an I-frame, which is periodically inserted in the video sequence and can be independently decoded without other reference frames, fitting the efficiency requirement in this work. The refinement procedure is proposed to overcome the variations of multiple QPs and different frame sizes.

The rest of this paper is organized as follows. Section II describes the background on CPD approach and intra prediction in HEVC. Section III presents the proposed scheme for copy detection. Section IV describes the experimental results and Section V concludes this work.

II. BACKGROUND

This section first explains the idea of the compression domain approach and why it is more efficient than the pixel domain approach. Then, the method of intra frame prediction in HEVC is briefly reviewed.

A. Compression domain approach

As illustrated in Fig. 1, the path with solid arrows represents the detection process in the PXD. The process requires full decoding and feature extraction to obtain features for matching or decision. It is interesting to note that in order to improve the coding efficiency many operations in modern video encoders are similar to those in content analysis while the objectives are different. The coding information, such as the prediction modes and transformed coefficients in a bitstream, contains the content information of a frame. The detection process in the CPD, as described by the dotted

978-1-4799-7543-3/15/$31.00 ©2015 IEEE

arrows in Fig. 1, generates the features by extracting and refining the coding information directly. In this case, only partial decoding is required to return the binary code back into the corresponding values, thus consumes less memory and computational complexity than those in the PXD.

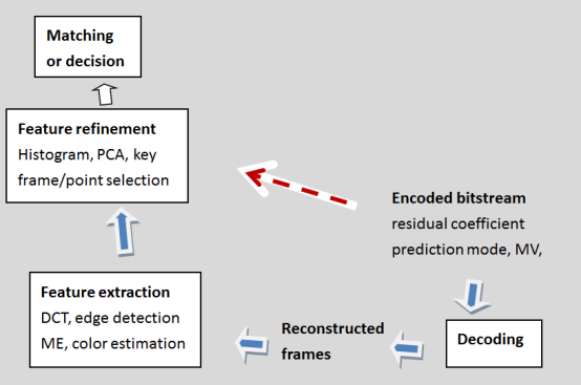


Fig. 1. The detection process in the pixel domain (solid arrow) and the compression domain (dotted arrow).

B. Intra prediction in HEVC

Video frames are compressed based on the block-base coding scheme, in which each frame is divided into several non-overlap blocks and then predictive coding is applied to each block. The residual r instead of the original block is coded and transmitted in order to reduce the data amount:

$$r = f - p \tag{1}$$

where f is the original block; p is the prediction block. In intra frame coding, the prediction block is predicted from the neighboring reconstructed pixels with a prediction direction. The residual signal is coded by DCT transform and quantization and then it is accompanied by additional side information to form the final bitstream. The side information includes the partition modes and the direction modes both indicating how the prediction process is performed. Fig. 2 shows the available partitions and direction modes in HEVC [9].

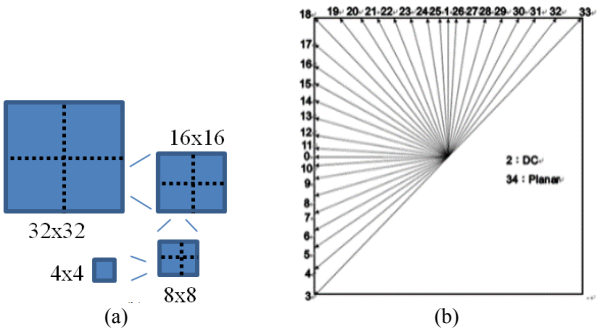


Fig. 2. Available partition and directional modes in HEVC (a) hierarchical partitions (b) fine directions.

In HEVC video coding, a rate-distortion optimization (RDO) function is used to tradeoff distortion E and bitrate B. The prediction direction θ can be determined through minimizing the cost function:

$$C = E(\theta) + \lambda(\mathrm{QP})B(\theta) \tag{2}$$

where λ is Lagrange parameter, which is a function of QP.

III. Proposed Method

This section describes the proposed method that is based on the HEVC I-frame coding. Not only the intra prediction modes but also the residual coefficients are extracted from the bit stream to generate features for detection. The characteristic of the modes and residuals are analyzed, and then two dissimilarity criterions are defined as the indication for copy detection.

A. Directional mode

In equation (2), the cost function usually reaches its minimum value when the prediction direction matches the edge direction of the original block. Therefore, the directional mode determined by the video encoder can sufficiently represent the content, and is utilized as a good texture feature.

The difference between two predicted directions of two images is used as the first score for matching. Since an angle or its complement angle can equivalently represent the difference of two directions for matching, the smaller one that is limited within 90° is used to represent the angle dissimilarity:

$$D^A = \frac{1}{I}\left(\sum_{i=1}^{I} \min\left(\left|\theta_i^Q - \theta_i^D\right|, 180^\circ - \left|\theta_i^Q - \theta_i^D\right|\right)\right) \tag{3}$$

where θ_i^Q and θ_i^D are the predicted directions of collocated partitions in query image and database, respectively, I is the number of 4x4 blocks in one frame.

It should be noted that although the partition sizes could be different in each frame, the direction here is referred to the 4x4 partition; for example, there exist 4 repeated directions in an 8x8 partition. Moreover, if the resolution of two compared frames are different, the rescaling with an averaging filter is applied to the directions before the comparison. Accordingly, a threshold Th^A is used to determine whether the angle dissimilarity D^A points out a copy or not. While D^A closes to 0, it indicates that two frames are very similar.

B. Residual coefficients

In addition to the directional modes, quantized residual coefficients are also included in the bitstream as mentioned. According to the property that texture and non-smooth boundary which usually results in high prediction error, the magnitude of residual in spatial domain can indicate whether the block texture is massive or not. This feature can be used as the global texture structure. Applying Parseval energy theorem, the energy of the residual is equal to that after transform T:

$$\Sigma(r^2) = \Sigma(T(r)^2) \tag{4}$$

where r is the residual. It is thus possible to estimate the residual variation directly from transmitted coefficients.

The residual is not only affected by the image content. The residual coefficients of the same content but different QPs would be much different in the coding scheme. The accuracy of image matching will substantially decrease in this situation. To solve the problem caused by the variation of QP, the QPs of two frames are adjusted to be the same for good comparison results. Therefore, the residual coefficients of a frame video need to be inverse-quantized and then re-quantized to a coarse version. The definitions of these operations are:

$$X = X_{Level} \times Q_{step} \tag{5}$$
$$Y = round\left(X/Q'_{step}\right) \times Q'_{step} \tag{6}$$

where X_{Level} is the received coefficient index, Q_{step} is the corresponding quantization step, and Q'_{step} is quantization step of the coarse one. Both X and Y are residuals but with different quantization errors. Fig. 3 shows an example demonstrating the effect of re-quantization. The re-quantization procedure suppresses the impact of the QP variation when the residual frame with smaller QP becomes more similar to coarse one after it is re-quantized.

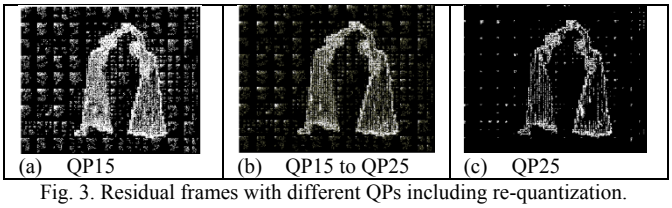
As mentioned above, we calculate the mean of residual powers σ_Y^2 after re-quantization. The frame is divided into K (9x11) patches with size NxN (16x16) and the mean of residual power for each patch is obtained as a low dimension feature, the corresponding formulas are shown as follows:

$$\sigma_Y^2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} Y(i,j)^2 \tag{7}$$
$$P_k = log(\alpha\sigma_Y^2 + 1) \tag{8}$$

where P_k is the mean of residual power in logarithmic scale and α is a constant. The dissimilarity for residual coefficients between query and database is defined as the normalized difference between residual powers in query P_k^Q and database P_k^D, which is represented as follows:

$$D^R = \begin{cases} \sum_{k=1}^{K} \frac{2\left|P_k^Q - P_k^D\right|}{(P_k^Q + P_k^D)} \\ 0, when\ P_k^Q = P_k^D = 0 \end{cases} \tag{9}$$

Similar to Section III A, a threshold Th^R is used to determine if the difference value D^R points out a copy or not.



(a) QP15 (b) QP15 to QP25 (c) QP25

Fig. 3. Residual frames with different QPs including re-quantization.

C. Combined dissimilarity score D^C

Finally, we combine the scores of the mode and residual. If one of the values is close to 0 or equal to 0, it has high possibility that the video is a copy. For this reason, two scores are multiplied to generate a final dissimilarity score, and then the score is incorporated with a threshold Th^C for detection:

$$D^C = D^A \times D^R \tag{10}$$

IV. Experiment Results

The evaluation of the detection accuracy and computational complexity are presented in this section. We test and analyze the impact of QP and frame size variations, which are the most serious issue for the CPD approach. Instead of using general copy detection benchmark, we specifically built up a dataset and set up a procedure to focus on the impact.

The dataset collects 110 CIF sequences with QP 25, including MPEG official test sequences and the clip from three commercial movies, Jack and Jill, One for the Money, and the Watch. Samples of the dataset are shown in Fig. 4. There are 10 test sequences, in which one half of these test sequences are the duplicates from the original dataset and the others are not included in the dataset. The test sequences were encoded with QP 15, 25, and 35 individually, and then compared with all of the clips in the dataset to evaluate the detection accuracy.

The test sequences were encoded with a different resolution QCIF to evaluate the detection accuracy against the variance of resolution. In this case, bi-cubic filter was used to rescale the frame size before encoding. HEVC reference software HM6.1 was used for the configuration of encoding. The MaxCUWidth was set to be 64 and MaxPartitionDepth was 4.

Fig. 4. Samples of the dataset.

First of all, the video sequences encoded by different QPs with the same frame size were tested based on the combined dissimilarity score D^C. The results with different threshold levels are shown in Table I. The terms Tp (true positives), Fp (false positives), and Fn (false negatives) as well as commonly adopted evaluation criterions were used to compare the results of the classifier under test with trusted external judgments, including precision, recall, accuracy and F1. As commonly defined, the precision equals to Tp/(Tp+Fp), recall equals to Tp/(Tp+Fn), and F1 score equals to 2Tp/(2Tp+Fp+Fn).

As shown in Table I, the experiment results show that the proposed method performed very well in terms of precision, which means the method can detect the copy with high accuracy while Fp indicating non-relevant detection is 0. Since the score of recall refers to the ratio of undetected copies, it shows that good recall results can be obtained by fine tuning the threshold Th^C.

TABLE I
EXPERIMENT RESULTS ON MIXED QPS WITH DIFFERENT Th^C

Th^C	Tp	Fp	Fn	Precision	Recall	Accuracy	F1
10	12	0	3	100	80	**99.55**	**88.89**
15	14	0	1	100	93.33	**99.85**	**96.55**
20	15	0	0	100	100	**100**	**100**

The test sequences including CIF and QCIF were compared with sequences in the database with only CIF format. The experiment results for test sequences with mixed frame sizes are shown in Table II. Furthermore, the results based on the directional modes and residual power are separately listed because the rescaling procedure may give different influences. The performance in Table II is only slightly decreased compared with in Table I. The residual power is more robust to frame size variation because the prediction mode may be changed due to the re-scaling while the residual pattern is more consistent.

TABLE II.
EXPERIMENT RESULTS ON MIXED FRAME SIZES BY DIFFERENT DISSIMILARITY SCORES D^A(ABOVE),D^R(MIDDLE), AND D^C (BELOW) WITH DIFFERENT THRESHOLD LEVELS

Th^A	Tp	Fp	Fn	Precision	Recall	Accuracy	F1
20	7	0	23	100	23.33	**96.52**	**37.83**
21	11	0	19	100	36.67	**97.12**	**53.66**
22	16	2	14	88.89	53.33	**97.58**	**66.66**

Th^R	Tp	Fp	Fn	Precision	Recall	Accuracy	F1
0.75	26	3	4	89.66	86.67	**99.1**	**88.14**
0.78	27	3	3	90	90	**99.23**	**90**
0.8	27	4	3	87.1	90	**99.1**	**88.53**
0.85	27	8	3	77.14	90	**98.59**	**83.08**

Th^C	Tp	Fp	Fn	Precision	Recall	Accuracy	F1
10	14	0	16	100	46.67	**97.58**	**63.64**
15	26	0	4	100	86.67	**99.39**	**92.86**
20	28	6	2	82.35	93.33	**98.79**	**87.5**

The proposed method is compared with the related work that adopts the number of non-zero coefficients as a feature [5]. The results by using our dateset are shown in Table III, where Th^s is defined in [5]. The results show that the method [5] is seriously affected when the QP changes. The performance is significantly deteriorated because the number of non-zero coefficients is highly related to the QP in the encoding scheme. On the other hand, experimental results from a traditional method of PXD approach which is implemented by raw pixel subtraction between two video frames are also shown in Table III, where Th^p is its threshold. In comparison with the results in Table I, our methods achieve similar performance to the pixel domain approach. It suggests that the directional mode and residual power could efficiently represent the video content.

TABLE III.
EXPERIMENT RESULTS FOR THE ALGORITHM OF [5] (ABOVE) AND THE APPROACH IN PIXEL DOMAIN (BELOW)

Th^S	Tp	Fp	Fn	Precision	Recall	Accuracy	F1
0.31	7	98	8	6.67	46.67	**83.94**	**11.67**
0.32	7	61	8	10.29	46.67	**89.55**	**16.86**
0.33	6	19	9	24	40	**95.76**	**30**

Th^P	Tp	Fp	Fn	Precision	Recall	Accuracy	F1
20	13	0	2	100	86.67	**99.94**	**92.86**
21-112	15	0	0	100	100	**100**	**100**
113	15	2	0	88.24	100	**99.94**	**93.75**
114	15	3	0	83.33	100	**99.91**	**90.91**

Finally, when considering the computational complexity, the proposed method could save up to 49% and 77% decoding time in the all intra coding profile and random access profile respectively according to the complexity analysis on HEVC decoding [10]. This is reasonable because the entropy decoding is the only requirement in CPD approach.

V. CONCLUSION

This paper proposes a method to obtain the QP and resolution invariant signature by employing and refining the intra prediction mode and residual coefficients from HEVC bitstream. The experimental results show that the proposed method can achieve comparable performance with the pixel domain methods and reduce the resource consumption by taking the advantages of effective features from coding information.

REFERENCES

[1] Y. Min, X. Li, Y. Zhang, Y. Zhao, and H. Lian, "A New SIFT Key point Descriptor For Copy Detection," *International Congress on Image and Signal Processing (CISP)*, vol. 2, pp. 842-845, 2011.

[2] H. Wang, A. Divakaran, A. Vetro, S. Chang, and H. Sun, "Survey of Compressed-Domain Features Used in Audio-Visual Indexing and Analysis," *Journal of Visual Communication and Image Representation*, vol. 14, pp. 150-183, 2003.

[3] G. Schaefer, D. Edmundson, "DC Stream Based JPEG Compressed Domain Image Retrieval," *Lecture Notes in Computer Science*, vol. 7669, pp. 318–327, 2012.

[4] Z. Zhang and J. Zou, "Compressed video copy detection based on edge analysis," *Information and Automation (ICIA)*, pp. 2497-2501, June 2010.

[5] Z. Li and J. Chen, "Efficient compressed domain video copy detection," *International Conference on Management and Service Science*, pp. 1-4, Aug. 2010.

[6] Y. Tian, T. Huang, M. Jiang, and W. Gao, "Video Copy-Detection and Localization with a Scalable Cascading Framework," *IEEE Multimedia*, vol. 20, no. 3, pp. 72-86, July-Sept. 2013.

[7] R. Hu, B. Li, W. Hu, and J. Yang, "Spatio-temporal Features for Efficient Video Copy Detection," *Lecture Notes in Computer Science*, vol. 8261, pp. 120-127, 2013.

[8] M. Ali and E. Edirisinghe, "Efficient Spatiotemporal Matching For Video Copy Detection in H.264/AVC Video," *International Journal of Computer Applications,* vol. 41, no. 15, Mar. 2012.

[9] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE transaction on circuits and system for video technology*, vol. 22, no. 12, pp. 1649-1668, Dec. 2012.

[10] F. Bossen, et al., " HEVC complexity and implementation analysis, " *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1685-1696, Dec. 2012.