CONDITIONAL ACCESS TO H.264/AVC VIDEO WITH DRIFT CONTROL

Enrico Magli, Marco Grangetto, Gabriella Olmo

CERCOM - Center for Multimedia Radio Communications
Dipartimento di Elettronica - Politecnico di Torino
Corso Duca degli Abruzzi 24 - 10129 Torino - Italy
Ph.: +39-011-5644195 - Fax: +39-011-5644099
firstname.lastname@polito.it

ABSTRACT

In this paper we address the problem of providing conditional access to video sequences, namely, to generate a low-quality video to be used as preview, which can be decoded at full quality if a decryption key is obtained. We propose and investigate the performance of two different techniques, based on smoothing and separate encoding in the compressed domain, and motion vector perturbation. We show that these techniques are able to provide conditional access to different quality levels of H.264/AVC video with very small rate overhead, and that their combination can provide different levels of security towards malicious attacks.

1. INTRODUCTION

Multimedia security and digital rights management are becoming very important issues in the deployment of multimedia content distribution systems. Recent encryption standards such as advanced encryption standard (AES) allow to protect data communications in a reliable way, providing a high degree of privacy with a reasonable computational cost. However, encrypting a complete compressed image or video file has a few disadvantages. The encryption algorithm must protect a potentially very high number of bits, which could result in excessive computational burden and/or power consumption on a battery-powered device. Moreover, multimedia compressed files typically exhibit well-defined structures that can be exploited in several ways, e.g. for quality-progressive scalability, transcoding, rate shaping, and so forth; however, these structures are not recognizable in the ciphertext, and hence become useless.

Several authors have proposed "selective encryption" systems, in which only the visually most significant transform coefficients are ciphered in order to prevent display of the image by an unauthorized user [1]. Encryption can be carried out at different stages of the compression process. Encrypting data before the entropy coder, or during the entropy coding stage, may result in a loss in coding efficiency due to the modified data statistics [3]. When employing an international standard, the protected file should also be syntax-compliant with the standard, so that, if a decoder attempts to decode a protected file it will generate a meaningless content, but will not crash due to syntax errors; this problem is typically incurred if the compressed file undergoes encryption.

Another interesting application, namely *conditional access*, has been proposed in [2]. In conditional access, a low-quality version of the image is left in the clear, and can be used to preview the multimedia content; the user can purchase a key, and then decode the content at full quality. Note that, unlike this paper, the term "conditional access" is also sometimes referred to the ability of a system to limit the access to the video in a "yes or no" fashion, without the option of a public low-quality version.

In this paper we address the problem of providing conditional access to video sequences. To our best knowledge, this is the first paper that addresses the conditional access problem for video, as opposed to images, using signal processing techniques instead of hardware devices such as smart cards or Java cards. We propose and evaluate two different techniques, respectively based on removal and separate encryption of DCT coefficients, and on random perturbation of the motion vectors. We investigate their flexibility, effectiveness, and compression loss, comparing them with baseline H.264/AVC with no conditional access, and with direct coding of encrypted the DCT coefficients. The techniques are embedded in the H.264/AVC video coding standard, which provides state-of-the-art compression performance; streaming of CIF video sequences is selected as target application.

2. BACKGROUND AND CRITICAL ISSUES

The concepts outlined in Sect. 1 have led to several image and video scramblers [4]. In [5] an arithmetic coder is used to carry out joint compression and encryption. However, its application to H.264/AVC has turned out to be difficult, because decoding encrypted data such as the motion vectors or the coding modes can lead to out-of-range values; if the decoder behavior for these values is not specified, the decoder will stop and exit instead of using a default value (e.g., a zero motion vector or an intra coding mode). The same problem is encountered if one attempts to selectively applying an external encryption scheme such as AES to parts of the compressed file. Moreover, in this latter case, even if the headers were left in the clear, the encrypted file would not be guaranteed to be compliant with the standard syntax because of potential marker emulations.

While selective encryption of image and video sequences has been addressed by several authors, conditional access has been dedicated less effort so far. As far as images are concerned, in [2] it has been proposed to scramble some bit-planes of high-frequency wavelet subbands after entropy coding. In [5] a similar approach is used, and some high-frequency subbands are encoded using the

This work was partially supported by the Italian Ministry of Education and Research under the PRIMO (Reconfigurable Platforms for Wideband Wireless Communications) grant.

1-4244-0367-7/06/$20.00 ©2006 IEEE

randomized arithmetic coder. Both techniques are based on the fact that the encoding in JPEG 2000 can be done in a layered way, so that only some refinement layers need to be protected (in fact Part 8 of JPEG 2000, which addresses security, also supports encryption and access control). In the video coding case, this is made more difficult by the prediction feedback loop, since decoding an encrypted file without knowing the key will cause an uncontrolled error propagation.

Conditional access could be obtained by using a scalable video representation, and encrypting the enhancement layers while leaving the base layer in the clear. This approach is simpler from the system design standpoint, but has two drawbacks. The first one is that this technique would only work with video coders supporting the scalable modes, which are usually included only in advanced profiles for specific applications. The second one is that this approach has an inherent performance loss, i.e. the loss incurred by a quality-scalable coder with respect to a non scalable one, which can be significant [6]. Both the scalable and nonscalable approaches are viable; in this paper we pursue the non scalable one, as it does not impose any additional scalability requirement to the decoder, and can hence be applied more generally.

3. PROPOSED TECHNIQUES

One serious problem with scrambling techniques is that they require that the data are uncorrelated; this guarantees that no information can be gained about the encrypted data from the data that are left in the clear. However, transforms and prediction loops as commonly used in image and video compression are not perfect decorrelators; as a consequence, some care must be taken in order to avoid information leakage from non-ciphered data.

As for video sequences, besides header information, the compressed file contains the motion vectors (MV), coding modes, reference frame number, and DCT coefficients for each macroblock (MB) of each slice. Some of this types of data are not amenable to be used for conditional access. For example, the reference frame number would be relatively easy to estimate, given that, most of the times, the best prediction match occurs in the previous frame. Similarly, the coding modes are not so numerous, and are also amenable to attacks.

We believe that the most suitable data to be used to provide conditional access functionalities are the DCT coefficients and the MVs. Since the DCT coefficients provide a frequency-based, and hence hierarchical representation of the video content, they are amenable to differentiated encoding for content protection purposes. The MVs should be protected to the extent that they do not leak any information regarding the DCT coefficients, so that an attacker will not be able to use them for temporal or spatial error concealment.

Note that the algorithms presented in the following only operate on the luminance component of the video, because it contains most of the detail information about the sequence, which cannot be estimated from the chrominance components.

3.1. DCT coefficients

H.264/AVC follows the typical structure of motion-compensated video coding schemes, which is shown in Fig. 1. A temporal prediction loop is present, in which the prediction error is transformed and quantized; unlike other standards, H.264/AVC also has a prediction loop in the motion estimation stage, in that the MV for a MB can be predicted from other MVs of already encoded MBs in the same frame.

The protection is carried out for each MB separately, because this is the way the encoding is done in H.264/AVC. The standard employs a 4x4 DCT followed by uniform scalar quantization; we denote as d_i, with $i = 0, \ldots, 15$ the quantized values for a 4x4 block, ordered according to the zig-zag scan. Inside the prediction loop, the coefficients d_i are left unchanged, and all quantities used for rate-distortion optimization, e.g. the rate and distortion associated to a certain mode decision and whether a 4x4 block is nonzero, are computed using the unchanged d_i coefficients. This ensures that the encoder rate-distortion optimization produces exactly the same results as in case of no conditional access.

However, the d_i coefficients are modified prior to entropy coding, as shown in the shaded boxes in Fig. 1. The modification needs not involve all coefficients; rather, we decide to modify only a given number of least significant bit-planes of coefficients with index $i > i_{\min}$. This allows to accurately set the quality of the video thumbnail that all users are allowed to decode; a certain number of low-frequency coefficients are left unchanged, whereas some bit-planes of the high-frequency ones are processed to make the detail information unavailable to unauthorized users. The number of least significant bit-planes to be modified can differ according to whether the current MB is coded in intra or inter mode, and is denoted as L_I and L_P respectively.

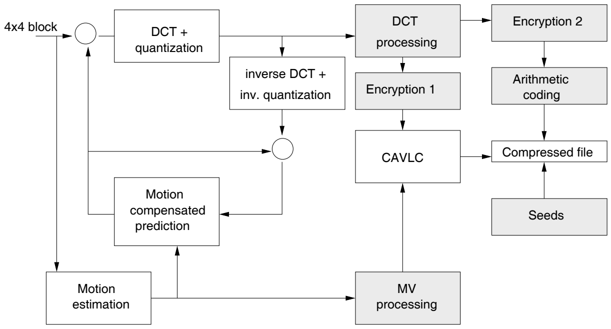

Fig. 1. Block diagram of the proposed conditional access scheme.

3.1.1. Benchmark: direct encryption of DCT coefficients

The technique we consider as benchmark is the direct encryption of DCT coefficients, and corresponds to the bottom arrow of the "DCT processing" box in Fig. 1. This technique employs a stream cipher to encrypt L_I (or L_P) least significant bit-planes of coefficients d_i with $i > i_{\min}$ before entropy coding. In particular, a pseudorandom number is generated and ex-ored with each bit to be encrypted. This kind of approach, i.e. encryption before entropy coding, is expected to provide some rate loss due to the more random statistical properties of the encrypted bits as opposed to the plaintext bits [3]; the advantage is that no external entropy coder is required (e.g. as opposed to Sect. 3.1.2).

The seed of the pseudorandom number generator must be communicated to the decoder, in order to allow it to decode the video sequence at full quality. Initializing the generator with the given seed, the decoder will reproduce the same sequence of numbers used by the encoder, and by ex-oring them with the received bits, it will be able to reconstruct the plaintext exactly. In this case, since the rate-distortion optimization has been made as in the case of no protection, the decoded video will be exactly equal to the unprotected video generated by an encoder with no conditional access features.

3.1.2. Cancellation and separate encryption of selected bit-planes

The second technique we propose aims at reducing the performance loss incurred because of coding encrypted data; this corresponds to the arrow on the right side of the "DCT processing" box in Fig. 1, with the "encryption 1" box being disabled. In particular, instead of encrypting the data before entropy coding, in this scheme we *remove* the data before entropy coding, and then code and encrypt them separately. In particular, we remove the L_I (or L_P) least significant bit-planes of coefficients d_i, with $i > i_{\min}$, by appropriate right-shift, and, at the same time, we save these bits in a separate auxiliary file. The auxiliary file is coded using an arithmetic coder, and the compressed file is encrypted using a stream cipher. If required, the resulting ciphertext can be embedded in the compressed H.264/AVC video, for example as a supplemental enhancement information (SEI) message. The rationale for this design is that, although we incur a rate overhead due to the auxiliary file, we also save some rate in the encoding of the DCT coefficients, because their energy is lower (and zero-runs potentially longer) after the bit-plane cancellation.

The decoder will need the key used for encrypting the auxiliary file, and will restore the cancelled bit-planes before inverse quantization.

3.2. Perturbation of motion vectors

Although processing the DCT coefficients provides enough flexibility in generating the desired degree of quality, an attacker could still apply error concealment techniques based on exact knowledge of the MVs of each MB, in order to improve the quality of the decoded video. To avoid this, it is useful to perturb the MVs in a controlled and reversible fashion, in such a way that there is no information leakage from the MVs regarding the DCT coefficients.

In particular, in the proposed scheme we modify each MV using the formula $V' = V + \mathrm{round}(\alpha S)$, where V and V' are the original and modified MVs respectively (represented as integers), S is a random number uniformly distributed between -1 and 1, α is a constant that sets the desired amount of modification to the MV, and $\mathrm{round}(\cdot)$ denotes rounding off to the nearest integer. The choice of S to be a random variable with zero mean ensures that the MV perturbation has zero mean. This is important because, unlike the high frequency DCT coefficients, the MVs do not have zero mean. Therefore, simple bit-plane removal would generate a quick propagation of the perturbation because of the MV prediction, leading to very poor and not controllable quality.

Note that, as done for the DCT coefficients, the rate-distortion optimization employs the unperturbed MV values, so as to provide exactly the same mode decisions and prediction errors as a regular encoder.

3.3. Some comments about drift control

It is worth describing in some more detail the quality control mechanism on which the proposed scheme is based. We start with the first frame of a group of pictures (GOP), which is always an I frame; $i_{\min}$, L_I and L_P are constants and are not adjusted for each frame. The parameters $i_{\min}$ and L_I are selected so that the distortion (on top of the source coding one) incurred by a decoder that does not know the encryption key is the desired one, say D. If $L_P = 0$ the quality would be roughly constant also across the P frames of the GOP. This is due to the fact that the error introduced in a predicted MB in the first P frame is exactly the error D introduced in the reference MB. If another MB in the second P frame is predicted from this reconstructed MB, the same principle applies, i.e. the distortion with respect to the non modified DCT coefficients would still be D. This highlights that the proposed scheme adopts an effective form of drift control. The average distortion remains constant across different predicted frames, and, even more importantly, the "quality" of this distortion (i.e. its visual impact) is also constant, since it is always confined to the same high-frequency DCT coefficients, but does not propagate to the low-frequency ones.

From these remarks, it can be seen that, in order to provide conditional access functionalities, it is not needed to cancel or encrypt information in the MBs coded in inter mode. It is indeed possible to set L_P to a value larger than zero. This is useful to increase the security level, if it is feared that an attacker could try to estimate the missing information in the DCT coefficients of an intra-coded MB using the high-frequency information of the prediction error in an inter-coded MB. To avoid this attack, it is possible to employ a small value of L_P, so as to obtain a good balance between security and the additional overhead. Note that setting L_P to a value larger than zero still allows to achieve accurate quality control, because the error is kept confined to the same high-frequency DCT coefficients with index larger than $i_{\min}$.

A similar reasoning can be made for the MVs. The amount of error on the MVs can be kept under control by making sure that the perturbation has zero-mean, and the drift in the MV intra prediction loop turns out to be constant, similarly to the case of the DCT coefficients. However, it is difficult to employ the MV perturbation to control the decoded image quality, since even a moderate MV error could lead to discontinuities in the decoded image, with a significant negative impact on quality. Therefore, in this work we only perturb the MVs to the extent that they do not leak any information about the DCT coefficients.

4. EXPERIMENTAL RESULTS

We have tested the proposed techniques on CIF video sequences of 100 frames, using the baseline H.264/AVC encoder. The sequences have been coded at 10 frames per second, with GOPs of 15 frames (one I frame followed by P frames), which is suitable for a video streaming scenario. CAVLC has been used as entropy coder. For brevity, in the following we only report the results for the *Mobile* sequence.

In Tab. 1 we evaluate the performance of DCT coefficient cancellation and motion vector perturbation with various parameters. The original sequence, coded with no protection, has PSNR equal to 31.30 dB. The rightmost column reports the overhead incurred with respect to the no protection case. As can be seen, decreasing $i_{\min}$ decreases the quality obtained without the decryption key, since more DCT coefficients undergo bit-plane cancellation. For the same value of $i_{\min}$, L_I can be used to fine-tune the quality control. The incurred overhead is also quite small, typically within 1%. In isolated cases the proposed scheme is even slightly better than H.264/AVC; this is due to the fact that the auxiliary file is coded using a binary arithmetic coder, which is slightly more powerful that CAVLC. Obviously, the overhead increases as L_I

and $i_{\min}$ increase, since more information has been cancelled and has to be coded separately. As had been anticipated, when L_P is set to a value different from zero, there is no significant quality change, and the PSNR drops by only about 0.2 dB; there is however a overhead increase due to the need to code a significantly larger amount of information in the auxiliary file.

When MV perturbation is employed, i.e. $\alpha \neq 0$, the temporal prediction drift leads to larger errors, and hence lower PSNR; the overhead is also a bit higher because the MVs are noisier. However, thanks to the fact that the MV error has zero mean the visual quality is still suitable for a video preview.

Table 1. Performance of DCT coefficient cancellation and motion vector perturbation on the *Mobile* sequence.

L_I	L_P	$i_{\min}$	α	PSNR (dB)	OVH %
1	0	12	0	27.16	0.33
2	0	12	0	26.60	0.40
1	0	10	0	23.20	0.23
2	0	10	0	21.86	0.93
1	0	8	0	21.22	-0.13
2	0	8	0	18.76	0.83
3	0	8	0	18.43	1.45
1	1	10	0	23.02	0.54
2	1	10	0	21.71	1.35
1	0	12	0.52	21.81	0.49
1	0	12	0.6	18.60	1.04
2	0	12	0.52	21.60	0.56
2	0	12	0.6	18.48	1.11

Tab. 2 contains results for direct encryption of DCT coefficients. It can be seen that the overhead is much higher than in the case of the proposed techniques, due to the fact that the modified bit-planes have poor correlation characteristics, and their encoding requires a lot of bits. This witnesses that the proposed techniques are quite effective as far as compression loss is concerned.

Table 2. Performance of direct encryption of DCT coefficients on the *Mobile* sequence.

L_I	L_P	$i_{\min}$	α	PSNR (dB)	OVH %
1	0	12	0	22.89	6.73
2	0	12	0	17.67	13.96
1	0	10	0	20.23	8.36
2	0	10	0	15.38	17.74

Fig. 2 shows an example of video sequence decoded without knowledge of the key. As can be seen, the considered techniques exhibit distinctive visual features. Cancellation of DCT coefficients leads to a very smooth image, with reasonable quality; direct encryption generates a lot of high frequency noise and visual artifacts. MV perturbation leads to another kind of artifacts, namely local image warping/distortions.

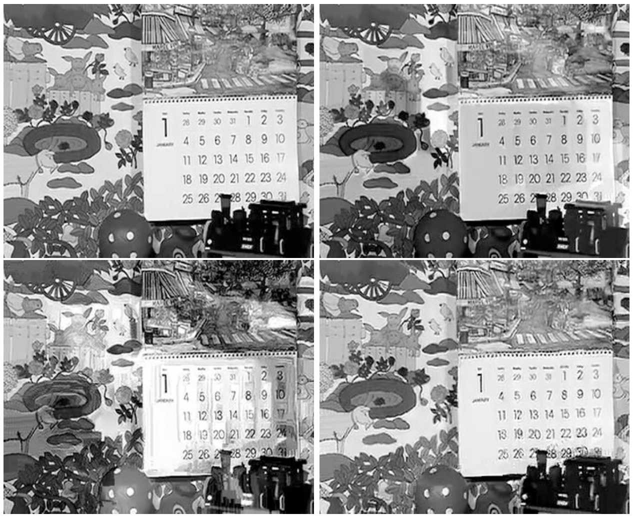

Fig. 2. Sequences decoded without knowledge of the key. Top left: Original with no protection (PSNR = 31.49 dB). Top right: cancellation of DCT coefficients (PSNR = 18.97 dB). Bottom left: direct encryption of DCT coefficients. Bottom right: MV perturbation. The PSNR for the last three images is equal to about 21.8 dB.

5. CONCLUSIONS

We have proposed two techniques that process the DCT coefficients and MVs inside an H.264/AVC video encoder in order to provide conditional access features. The proposed techniques allow to accurately select the quality achieved without knowledge of the key, whereas its knowledge allows to decode the sequence with no quality loss with respect to that produced by a standard encoder. The price to be paid is a small rate overhead, which is typically below 1 %.

6. REFERENCES

[1] H. Cheng, X. Li, “Partial encryption of compressed images and videos,” *IEEE Transactions on Signal Processing*, v. 48, n. 8, pp. 2439-2451, Aug. 2000.

[2] R. Grosbois, P. Gerbelot, T. Ebrahimi, “Authentication and access control in the JPEG 2000 compressed domain,” *Proc. of the SPIE 46th Annual Meeting*, USA, 2001.

[3] M. Wu, Y. Mao, “Communication-friendly encryption of multimedia,” *Proc. of IEEE MMSP 2002*.

[4] J. Wen, M. Severa, W. Zeng, M.H. Luttrell, W. Jin, “A format-compliant configurable encryption framework for access control of video,” *IEEE Trans. on Circuits and Systems for Video Technology*, v. 12, n. 6, pp. 545-557, Jun. 2002.

[5] M. Grangetto, E. Magli, G. Olmo, “Multimedia selective encryption by means of randomized arithmetic coding,” to appear in *IEEE Transactions on Multimedia*, 2006.

[6] L. Yang, F.C.M. Marins, T.R. Gardos, “Improving H.263+ scalability performance for very low bit rate applications,” *Proc. SPIE Vis. Comm. and Image Processing*, 1999.