EUSIPCO 2013 1569735865

CONFIDENTIALITY METRICS AND SMART SELECTIVE ENCRYPTION FOR HD H.264/AVC VIDEOS

Loïc Dubois [*], *William Puech* [†]

LIRMM Laboratory, UMR 5506 CNRS,
University of Montpellier II
161, rue Ada, 34095 Montpellier Cedex,
France.

Jacques Blanc-Talon[‡]

DGA, 7, rue des Mathurins,
92221 Bagneux Cedex,
France.

ABSTRACT

In the field of video protection, selective encryption (SE) is a scheme which ensures the visual security of a video by encrypting only a small part of the data. This paper presents a new SE algorithm for H.264/AVC videos in CAVLC mode. This algorithm controls the amount of encrypted alternative coefficients (AC) of the integer transform in the entropic encoder. The structural similarity (SSIM) is used to measure the visual confidentiality level of each video frame and to control the amount of encrypted AC. Moreover, a new psychovisual metric to measure the flickering is introduced, the so-called TSSIM. This method can be applied on *intra* and *inter* frame video sequences. Several experimental results show the efficiency of the proposed method.

Index Terms— Selective encryption, H.264/AVC, Psychovisual metrics, Flickering, Visual confidentiality.

1. INTRODUCTION

With the rapid evolution in digital media, growth of processing power and availability of network bandwidths, digital videos are commonplace and their numbers are rising exponentially. Consequently, archived and transmitted data must be protected because they can be easily copied and modified. Data security and network security are two common solutions to solve these issues. The first one protects a network of vulnerable data while the second ensures the safety of all data which can be shared on unsecured networks. Data protection to ensure the security is generally preferred to the network security thanks to a better optimization of processing time and data-size.

Furthermore, video data require compression in order to reduce the transmission time, and they need to be encrypted to ensuring their confidentiality. In video processing, full data encryption is rarely used because the processing time is twice that of the compression process. That is why Selective Encryption (SE) algorithms are usually recommended. SE algorithms aim to specify part of a video data bitstream to which the encryption algorithms are applied. This scheme guaranties visual confidentiality and protection without a data increase. Moreover, SE algorithms can lean on psychovisual metrics. Psychovisual metrics are mathematical tools which measure the perceptual quality of a processed image relative to its original. The goal of this combination is to optimize the encryption scheme: psychovisual metrics guarantee the quality of each encrypted frame during the encryption process.

This paper presents an analysis of video SE combined with similarity measures by taking into account the temporal aspect. Section 2 presents the H.264/AVC codec, the main previous work on SE of video and an overview of the actual psychovisual metrics. In Section 3, we present our approach and analysis in detail. In Section 4, experimental results are given and discussed. In Section 5, concluding remarks and prospects for about the proposed scheme are discussed.

2. STATE OF ART

2.1. Selective Encryption of H.264/AVC

H.264/AVC, also known as MPEG-4 Part 10, is the video coding standard of ITU-T and ISO/IEC. In H.264/AVC, each frame is divided in Macro-Blocks (MBs) of 16x16 pixels. These macro-blocks are encoded separately; the encoding method is an Entire Transform followed by quantization of the MB, a prediction between MBs in *intra* (I frame) or *inter* (P and B frames), and an entropy coding using either run length coding (CAVLC) or arithmetic coding (CABAC). In *intra* frame, the current MB is predicted spatially from neighboring MBs which were previously encoded and reconstructed. In *inter* frame, the current MB is predicted spatially and temporally from previous frames. The purpose of the reconstruction in the encoder is to ensure that both the encoder and the decoder use identical reference frames to create the predictions.

[*] loic.dubois@lirmm.fr
[†] william.puech@lirmm.fr
[‡] jacques.blanc-talon@dga.defense.gouv.fr

In the literature, several methods for video SE have been proposed [1]. SE, also known as partial encryption, is a encryption strategy which aims at saving computation time or enabling new system functionalities. In SE, a small part of the compressed bitstream is encrypted while still providing adequate data security with respect to total encryption which would encrypt the whole bitstream. Moreover, SE fulfills the main tasks of video encryption, namely visual confidentiality and data protection. These tasks are performed by applying a SE in certain segments of the bitstream with respect to total encryption which encrypts the whole bitstream. Another challenge in SE is that both encrypted and non-encrypted informations should be appropriately identified and displayed in order that the SE bitstream will remain compliant with the H.264/AVC video standard. In the field of video, different SE techniques have been developed which include permutation based on the Data Encryption Standard (DES) and the Advanced Encryption Standard (AES). Based on the location of the encryption stage in the video codec, these SE techniques can be divided into five broad categories. These are based on the spatial position, the video codec structure, the matrix of transformed coefficients, the entropy encoding or the bitstream. Encryption in the entropy encoding module is often efficient and has been adopted by several authors. Data security in *intra mode* is improved in [2], where each frame receives a specific and synchronized encryption key. Moreover, each type of MB is encrypted differently with chaotic sequences in order to improve protection against plain-text attacks. Perceptual encryption was also presented in [3], where encryption is done with an alternative transform of the DCT coefficients with a singular key.

The AES algorithm has also been used in SE-CAVLC [4] by encrypting only a part of the quantized coefficients in various VLC tables. SE-CAVLC is performed by using the AES algorithm in Cipher Feedback (CFB) mode on a subset of codewords/bin-strings. The data information is selectively encrypted for each MB, and header information is never encrypted because it is used for prediction of the next MBs. In the entropy coder, the SE is performed in the multiple VLC tables used in CAVLC. Only *signs of trailing ones* and *remaining non-zeros levels* are encrypted in order to keep the bitstream compliant. The encrypted spaces are VLC codes spaces which means that the same code lengths as a standard compression may be kept.

2.2. SSIM-Based Similarity Measures

SSIM [5] is an improved version of the universal image quality index. It is based on a top-down assumption that the HVS is highly adapted for extracting structural information from the scene, and therefore a measure of structural similarity should be a good approximation of the perceived image quality [5]. SSIM is generally useful in high quality image cases, and for detecting local degradation in an image due to the inter-dependence of the closest pixels. SSIM is also a future candidate for rate-distortion optimizations in the design of image compression algorithms. SSIM uses cross-variance coupled with the average and the variance. For video quality, SSIM is usually applied on each frame and the conclusions are based on the mean and variance of SSIM of the whole video. SSIM and the Scale-Invariant Feature Transform (SIFT) have been combined (SSIM_SIFT) [6] in order to evaluate systems which do not preserve the positions and/or shapes of objects. In [7], the authors present a specific SSIM-based metric called Structural Texture SIMilarity (STSIM). This metric is used to improve measurement of the quality of textures in images using both intra- and inter-subband correlations and with incorporation of the color composition. Note that few previous studies have presented methods that take temporal aspect into account.

3. PROPOSED METHOD

3.1. Selective Encryption Method

Our proposed method, the so-called Smart Selective Encryption (SSE), aims to achieve a smart reduction in the encrypted AC coefficients with respect to the psychovisual measure of the encrypted video. We use a SE-CAVLC algorithm while encrypting just part of the non-zero coefficients, and we check whether the visual confidentiality remains efficient. An overview of the compression and encryption scheme is presented in Fig. 1. In the H.264/AVC codec, the prediction error is used to reduce the bitstream size of video sequences. This prediction error is the difference between the current MB and a previous neighbor MB. A scan of each previously neighboring encoded MB is achieved in order to find the MB yielding the smallest prediction error. For the intra frames, the prediction is done on the current frame. Moreover, this prediction is used in the temporal domain in order to encode inter frames. During the decoding step, because of the SE, a MB which has been decoded from an encrypted MB should be heavily distorted. We use this specificity to spread the encryption through each inter frame of a video sequence. The intra frames are selectively encrypted while the inter frames are not directly encrypted.

In our SE scheme, the three color channels, luminance and two chrominances of each frame are affected by the SE. For each GoP, we propose two solutions to apply the SE. In the first solution, only the first frame (intra frame) of the GoP_i frame is encrypted. This solution is sufficient to ensure the confidentiality of the entire GoP_i in the case of non-HD videos. In the second solution, we propose to encrypt the whole GoP_i. Next, non-zero AC levels of lower frequencies are encrypted in priority according to the selected number of non-encrypted coefficients. In the H.264/AVC compression scheme, high frequencies AC levels are encoded first due to the inverse zigzag scan in the entropic encoder. We integrated

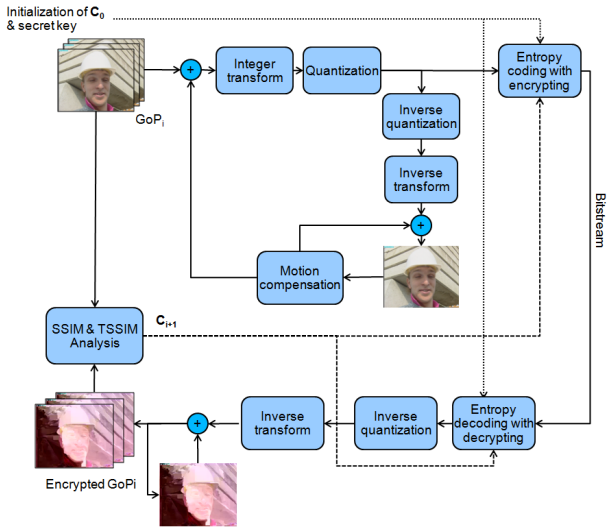

Fig. 1. Overview of the encryption and decryption method.

psychovisual metrics between GoP_i and GoP_{i+1}, these metrics measure the quality on the luminance component. These metrics control the amount of encrypted coefficients depending on the metric results of the previous GoP_i. The metrics used for our experimentations are SSIM and T-SSIM, which are combined and one maximum threshold and one minimum threshold for each of them are set. These psychovisual metrics control the visual confidentiality of each GoP_i frame and if one of the frame result metrics is above one of the maximum thresholds, then the next GoP will be encrypted with more encrypted coefficients ($C_{i+1} \leftarrow C_i + 1$), or if it is under the two minimum thresholds, we decrease the number of encrypted coefficients ($C_{i+1} \leftarrow C_i - 1$). With this control scheme, the encryption depends on the psychovisual measure and the confidentiality this changes in accordance with the video sequence and its different contents.

3.2. Psychovisual Metrics

The flickering phenomenon is a problem that disturbs the viewing of a video during the time. Research in this domain was previously carried out to reduce video flickering due to the compression algorithm. Generally, the flickering is considered and processed like noise. In [9] a block-based method for objective evaluation is presented, where each block in a given frame is classified with respect to its co-located block. This flickering phenomenon also appears in encrypted video sequences and increases the visual confidentiality. Indeed, the greater is the flickering, more annoying is the video sequence. This temporal visual effect is crucial to preserve the visual confidentiality of the contents. In order to measure and quantify the flickering phenomenon, we have developed a psychovisual metric based on the SSIM applied on successive video frames. Our proposed metric, called Temporal SSIM (T-SSIM), allows us to measure variations between the difference of two original compressed frames and two encrypted frames:

$$TSSIM_{(I_o,I_e)}(i) = SSIM(|I_o(i)-I_o(i-1)|, |I_e(i)-I_e(i-1)|), \quad (1)$$

where $I_o()$ is an original compressed frame and $I_e()$ an encrypted one. In fact, T-SSIM is the SSIM of the absolute differences of two successive original frames and the same successive frames of the encrypted video sequence. With T-SSIM, we can analyze flickering between two successive frames of an encrypted video sequence. The range of values for T-SSIM is nearly the same that for SSIM, *i.e.* above 0.8 the flickering is not really marked, and below 0.6 we consider that it is sufficient for the visual confidentiality of the video.

4. EXPERIMENTAL RESULTS

For our experimental results[1], we have used eight benchmark video sequences, four with CIF resolution (352×288 pixels); two with a 640×352 pixel resolution; two others in HD: *movie* in 1920×800 pixel resolution and *ar-drone* in 1280×720 pixel resolution. These video sequences show different combinations of motions, colors, contrasts and objects. The four CIF video sequences and the two 640×352 pixel video sequences are encoded with a GoP of 4 images, the two HD videos are encoded with a GoP of 2 images, and with a GoP of 4 images.

4.1. Selective encryption of only the I frame of a GoP

	SIMM		T-SSIM		ER-SSE	ER-SE
	Mean	Std	Mean	Std	(%)	(%)
Cit	0.46	0.17	0.40	0.24	**16.78**	19.31
For	0.48	0.16	0.53	0.28	**12.94**	15.83
Hal	0.51	0.05	0.72	0.38	**12.37**	17.53
Mob	0.33	0.08	0.39	0.20	**18.56**	25.51
BBB	0.53	0.17	0.71	0.33	**13.92**	20.15
Ven	0.57	0.08	0.57	0.27	**13.07**	13.76

Table 1. Results of SSE where just the I frame of a 4-frame GoP is encrypted. The results are presented in terms of mean and standard deviation (Std) for the SSIM and the T-SSIM and in percent for the encryption ratio for a SSE (ER-SSE) and a full SE (ER-SE) [4].

In the major parts of the results, we conclude that the encryption scheme tends to guide the confidentiality metrics between the thresholds of confidentiality, as presented in Tab. 1. However, the amount of encrypted coefficients varies depending on the video with an encryption ratio which varies between 12.37% and 18.56% for our set of video sequences.

[1] Visual results on these video sequences are available at: http://www2.lirmm.fr/∼dubois/PagesVideos.html

This result highlights that our scheme depends on the video content. For instance, textures are often more affected by the encryption than uniform regions of the videos due to the quantity of the AC coefficients of the entire transform, particularly for *mobile* video sequences. The encryption of only the first frame is not efficient, because between the P frames of a GoP, the T-SSIM is too high, and this method works only for short GoP cases.

With a subjective analysis of the confidentiality of HD video sequences which are encoded with a 4-frame GoP, we immediately notice that the encryption of the only I frame is not sufficient. In the fourth frame of a GoP, and sometimes in the third frame, some ROI are clearly visible due to the low prediction and the reconstruction of the image through the GoP. However, in the 2-frame GoP case, the results are still good, as presented in Fig. 2 which show that the encryption level depends on the content.

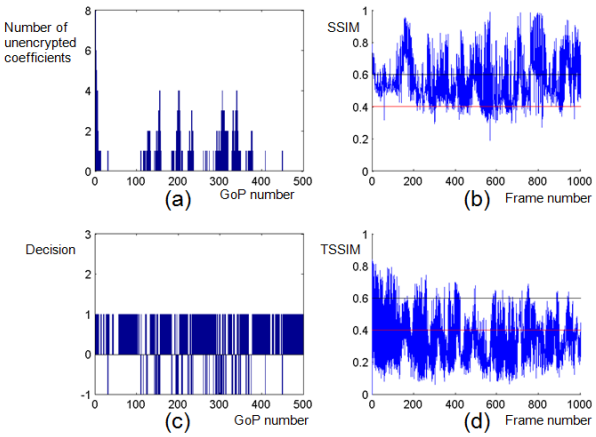

Fig. 2. SSE of only the I frame of a 2-frame GoP of the *ar-drone* HD video: a) The number of unencrypted coefficients, b) The SSIM of the encrypted video, c) The decision with respect to the GoP number: value -1 is the order for decreasing the encryption since both SSIM and T-SSIM are under their corresponding bottom thresholds. Value 0 is no change in the encryption. Value 1 is the increasing in the encryption due to a too high SSIM, value 2 is due to a too high T-SSIM and value 3 both of them. d) The T-SSIM of the encrypted video.

In Fig. 2, for the *ar-drone fly* HD video sequence, the different camera variations affect the video content and also the encryption quality. This is why we can notice fast variations in the number of encrypted coefficients. In Fig. 3, T-SSIM and SSIM are widely chaotic, especially T-SSIM, which is close to 1 at the end of the GoPs. In comparison with Fig 3, with SSE of only the I frame of a 4-frame GoP, the respective results of Fig. 2 are better in terms of SSIM and TSSIM. The results are similar for the *movie* HD video. This underlines the necessity of keeping a small GoP for HD video sequences. In high resolution cases, the MB size is small with

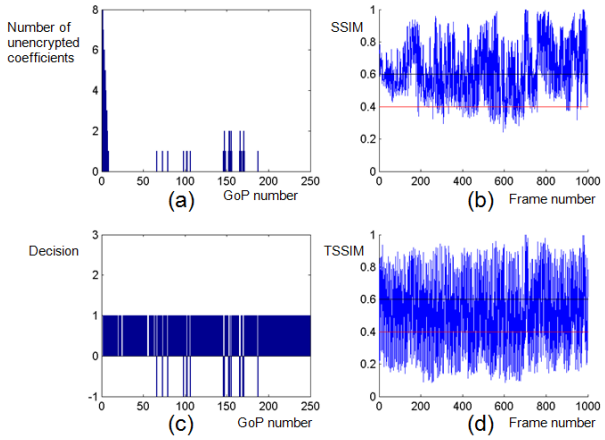

Fig. 3. SSE of only the I frame of a 4-frame GoP of the *ar-drone* HD video: a) The number of unencrypted coefficients, b) The SSIM of the encrypted video, c) The decision, d) The T-SSIM of the encrypted video.

respect to the size of the ROI of the video, also, and MBs do not have many AC coefficients after the transformation and quantization. This is why the visual protection for HD video sequences can be less efficient for the same number of encrypted coefficients with respect to other video sequences.

4.2. Selective encryption of all frames of a GoP

	SIMM		T-SSIM		ER-SSE	ER-SE
	Mean	Std	Mean	Std	(%)	(%)
Cit	0.39	0.09	0.35	0.19	**16.53**	19.52
For	0.38	0.12	0.46	0.26	**16.25**	17.34
Hal	0.40	0.06	0.72	0.39	**17.23**	18.01
Mob	0.32	0.08	0.37	0.19	**18.87**	26.51
BBB	0.32	0.14	0.67	0.35	**19.89**	20.29
Ven	0.54	0.08	0.55	0.27	**13.22**	13.93

Table 2. Results of the SSE where all of the 4-frames GoPs are encrypted. The results are presented in terms of mean and standard deviation (Std) for SSIM and T-SSIM in percent for the encryption ratio for a SSE (ER-SSE) and a full SE (ER-SE) [4].

Tab. 2 presents the results for six videos of the benchmark. Note that the encryption tends to be between the two desired thresholds with different encryption ratios, which vary between 13.12% and 19.89% for our set of video sequences. The results in terms of SSIM and TSSIM are better for the mean and the standard deviation which underline the fact that the videos are more annoying to watch and consequently the confidentiality is increased. We can notice that: even if the encryption ratio is similar, the encryption is more widespread through the GoP, which increases the visual confidentiality. In

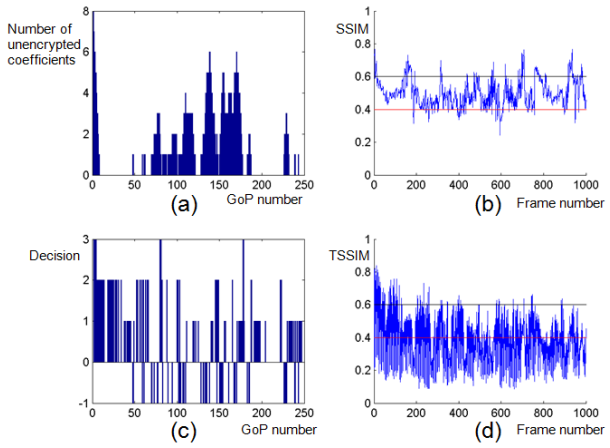

Fig. 4. SSE of a 4-frame GoP of the *ar-drone* HD video: a) The number of unencrypted coefficients, b) The SSIM of the encrypted video, c) The decision, d) The T-SSIM of the encrypted video.

terms of standard deviation, the high variations in the T-SSIM results underline that even if the *inter* frames are encrypted it is advised to keep a small size for the GoP. Moreover, the encryption is a pseudo-random process and we cannot precisely predict the visual confidentiality of a next GoP when the amount of encrypted coefficients is set. However, we can guide the encryption in a good way in order to have the desired confidentiality.

Fig. 4 presents SSE applied to *ar-drone* HD video. Note that the encryption is sufficient enough in terms of SSIM and T-SSIM. We can clearly see that with this SSE of a 4-frame GoP, the T-SSIM is less chaotic than the previous method, where only the *intra* frame was encrypted. The encryption is also efficient in terms of SSIM in the *movie* video but the flickering of the encrypted frames is not perfectly affected and this confirms the necessity of keeping small GoPs in a full SE case. This is probably caused by the lack of information in each MB of the video, if there are no-coefficients to scramble, the encryption cannot be properly done. Note that we obtain similar results for *movie* HD video. These results underlined that in HD cases the effectiveness of the encryption is closely correlated with the amount of texture regions in the video. If the video presents some scenes where the ROI are expanded, like zooms or closeups, the amount of AC coefficients is not wide enough for the encryption. With the proposed method, SE is tailored to the switching of scenes.

5. CONCLUSION

The SSE scheme we developed allows to control the encryption of video data depending on the desired confidentiality level. The use of psychovisual metrics is a good way to ensure the visual confidentiality, especially the proposed T-SSIM, which takes into account the temporal visual protection generated by the encryption, *i.e.* flickering. T-SSIM allows us to have better control of the temporal visual confidentiality over the time. In most of the videos, the visual protection is adequate while minimizing the amount of encryption with respect to the previous SE scheme. We have an mean encryption ratio of 16% with a mean SSIM of 0.4 for our experimentation. However, the control of the visual protection through encryption is efficient but not integral, because the encryption method is pseudo-random, also we cannot predict the exact deterioration of the visual content.

6. REFERENCES

[1] T. Stütz, A. Uhl, A Survey of H.264 AVC/SVC Encryption, IEEE Transactions on Circuits and Systems for Video Technology 22(3) (2011) 325–339.

[2] J. Jiang, Y. Liu, Z. Su, G. Zhang, S. Xing, An Improved Selective Encryption for H.264 Video based on Intra Prediction Mode Scrambling, Journal of Multimedia 5 (2010) 464–472.

[3] S.-K. Au Yeung, S. Zhu, B. Zeng, Perceptual Video Encryption using multiple 8x8 transforms in H.264 and MPEG-4, IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic (2011) 2436–2439.

[4] Z. Shahid, M. Chaumont, W. Puech, Fast Protection of H.264/AVC by Selective Encryption of CAVLC and CABAC for I & P frames, IEEE Transactions on Circuits and Systems for Video Technology 21 (5) (2011) 565–576.

[5] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, L. K. Cormack, Study of Subjective and Objective Quality Assessment of Video, IEEE Transactions on Image Processing 19(6) (2010) 1427–1441.

[6] M. Decombas, F. Dufaux, E. Renan, B. Pesquet-Popescu, F. Capman, A New Object Based Quality Metric Based On SIFT and SSIM, IEEE International Conference on Image Processing, Orlando, Florida, U.S.A. (2012) 1493–1496.

[7] J. Zujovic, T. Pappas, D. Neuhoff, Strutural Similarity Metrics for Texture Analysis and Retrieval, IEEE Conference on Image Processing, Cairo, Egypt (2009) 2225–2228.

[8] S. Kanumuri, O. G. Guleryuz, M. R. Civanlar, Temporal Flicker Reduction and Denoising in Video using Sparse Directional Transforms, Proceedings in SPIE 7073 (2008) 10.1117/12.796747.

[9] S. Chebbo, P. Durieux, B. Pesquet-Popescu, Objective Evaluation of Compressed Video's Temporal Flickering, IEEE International Conference on Image Processing Theory, Tools and Applications, Paris, France (2010) 177–180.