Noname manuscript No.
(will be inserted by the editor)

Smart Selective Encryption of H.264/AVC Videos using Confidentiality Metrics

L. Dubois · W. Puech · J. Blanc-Talon

the date of receipt and acceptance should be inserted later

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. Two visual quality measures, the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM), are 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.

Keywords 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

L. Dubois, W. Puech
LIRMM Laboratory, UMR 5506 CNRS, University of Montpellier II, 161, rue Ada, 34095 Montpellier Cedex 05, France
E-mail: {loic.dubois, william.puech}@lirmm.fr

J. Blanc-Talon
DGA,
7, rue des Mathurins, 92221 Bagneux Cedex, France
E-mail: jacques.blanc-talon@dga.defense.gouv.fr

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 and while saving computation time. In order to compress video data, redundancy in the images is reduced by the prediction error of the compression algorithms, and SE algorithms use this information vector to spread its encryption through the video frames.

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. In H.264 codec, a SE based on SE-CAVLC [1] is spread in the *inter* frames while only encrypting the *intra* frames. Moreover, we use similarity measures like Strutural Similarity (SSIM) and Peak Signal-to-Noise Ratio (PSNR) to analyze the perceptual effect of this SE, and create a psychovisual measure, *i.e.* TSSIM which analyzes the flicker phenomenon between two neighbor frames of the encrypted video. Furthermore, we present a new Reduced Selective Encryption (RSE) approach which controls the amount of encrypted coefficients with respect to a good visual protection of each frame, GoP by GoP, in terms of SSIM, PSNR and TSSIM.

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. The two main objectives are to highlight the propagation of an encryption effect through the predicted frames and the perceptual effect of this phenomenon. Furthermore, we analyze the proposed RSE method, which reduces the amount of encrypted coefficients in the entropic coder with respect to the perceptual confidentiality of the video. 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

In Section 2.1 we briefly present the H.264/AVC encoder. Next, a state of art on the SE of H.264/AVC encoder is described in Section 2.2. In Section 2.3, the psychovisual metrics that we used for our RSE method are presented.

2.1 H.264/AVC

H.264/AVC [2], 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), as presented in Fig. 1. 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.

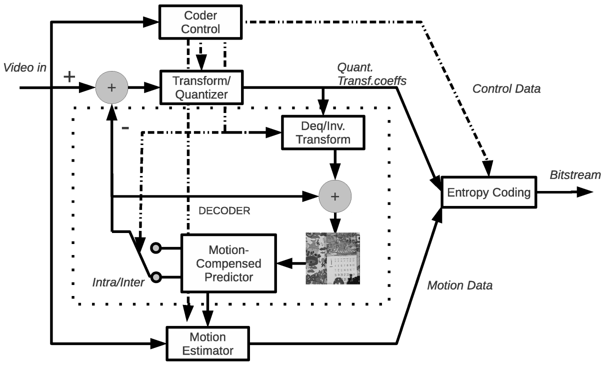


Fig. 1 Overview of the H.264/AVC encoder.

2.2 Selective Encryption

In the literature, several methods for video SE have been proposed [3]. 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 [4] 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 [5] 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) [6]. 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. The use of the Huffman entropy coder as encryption cipher was studied in [7]. Despite providing a compliant video bitstream, the scheme is hampered by a bitrate increase which makes it less appropriate and limits real-time applications. Moreover, a SE of MPEG-4 video standard was studied in [8], wherein DES was used to encrypt fixed length and variable length codes. In this approach, the encrypted bitstream is fully compliant with the MPEG-4 bitstream format but the bitstream size is increased. This particularity is a current observation in video SE. Furthermore, data security in *intra mode* is improved in [9], 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 [10], where encryption is done with an alternative transform of the DCT coefficients with a singular key. Chaotic-maps have also been used to randomize the macro-bloc positions in order to improve protection against plain-text attacks [11].

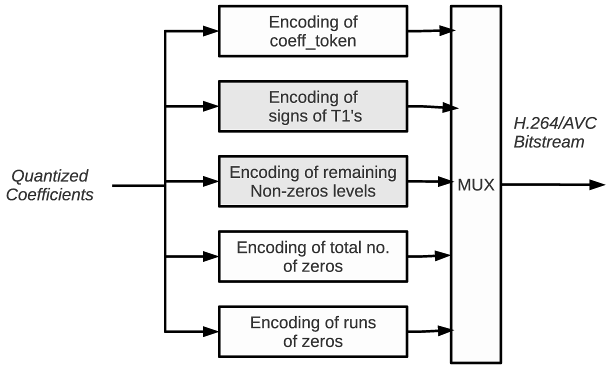

Quantized Coefficients
Encoding of coeff_token
Encoding of signs of T1's
Encoding of remaining Non-zeros levels
Encoding of total no. of zeros
Encoding of runs of zeros
MUX
H.264/AVC Bitstream

Fig. 2 Syntax elements used in the CAVLC entropy encoder of H.264/AVC. The encrypted coefficients in SE-CAVLC [1] are shown in grey.

The AES algorithm has also been used in SE-CAVLC [1] by encrypting only a part of the quantized coefficients in various VLC tables. SE-CAVLC [1] 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. In CAVLC, five syntax elements are used to code levels and runs: *coeff token*, *signs of trailing ones*, *remaining non-zero levels*, *total number of zeros* and *runs of zeros*, as shown in Fig. 2. 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. In [1], the entire parts of the experimentation were carried out on QCIF video sequences.

2.3 Similarity Measures

This section provides a brief description of some objective quality measures: PSNR and SSIM-based measures.

2.3.1 Peak-Signal-to-Noise-Ratio (PSNR)

PSNR [12] is widely used to measure the difference between two images based on pixel differences. For a N · M pixel image with pixel luminance values ranging from zero (black) to L_{max} (white) depending of the image dynamic, PSNR is defined as:

$$PSNR = 10 \cdot log_{10}(\frac{L_{max}^2}{MSE}), \tag{1}$$

where RMSE is the root mean square error defined as:

$$MSE = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} (I_o(i,j) - I_r(i,j))^2}{N \cdot M}, \tag{2}$$

where I_o and I_r are respectively the original image and the recorded one, and N and M represent the image resolution.

2.3.2 SSIM and specific methods

SSIM [13] is an improved version of the universal image quality index. It is based on a top-down assumption that the Human Visual System (HSV) 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 [14]. 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. Its mathematical computation can also be easily adapted for other types of signal. SSIM uses cross-variance coupled with the average and the variance, while PSNR uses only the mean square error (MSE). Mathematically, SSIM is defined as:

$$SSIM(x,y) = \frac{(2\mu_{I_o}\mu_{I_r} + c_1)(2cov_{I_o I_r} + c_2)}{(\mu_{I_o}^2 + \mu_{I_r}^2 + c_1)(\sigma_{I_o}^2 + \sigma_{I_r}^2 + c_2)}, \tag{3}$$

with μ the mean, σ the standard deviation, cov the covariance, and c_1 and c_2 are two constants to stabilize the division. This formula is principally applied on luma, where the correlation with the HVS is highest.

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.

Multiscale SSIM (MS-SSIM) [15] is an extension of SSIM. It incorporates variations in the viewing condition in the calculation, which provides more adaptability for still images. The MS-SSIM parameters can be calibrated as a function of the relative importance of the different image scales.

A complex wavelet domain image similarity measure (CWSSIM) [16] has been developed to be insensitive to luminance changes, contrast changes and spatial translation. This metric uses magnitude and/or phase changes of local wavelet transforms. This SSIM-based method is robust for small geometric distortions. However, it cannot underline large displacements and non-geometric deformations.

SSIM and the Scale-Invariant Feature Transform (SIFT) have been combined (SSIM_SIFT) [17] in order to evaluate systems which do not preserve the positions and/or shapes of objects.

In [18], 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. In [19] temporal aspect of the video is analyzed to determine more precisely the quality. Different quality measures are used with standard deviation and mean, however flickering is not integrated in this method.

3 Proposed Method

An overview of our proposed method is presented in Section 3.1. In Section 3.2, we explain our encryption scheme where the prediction error of H.264/AVC is used as encryption vector. Section 3.3 introduces the flickering problem and our proposed psychovisual metric which analyzes this phenomenon. Moreover, we propose an adjustment of the encrypted coefficients depending on the psychovisual metrics in Section 3.4. In conclusion, in Section 3.5, we present different strategies to decode the encrypted video.

3.1 Overview

The goal of SE algorithms is to decrease the quantity of encrypted data in order to reduce the encryption process time, particularly during decoding. In H.264/AVC, the encryption of non-zero AC coefficients is generally sufficient to visually protect the video. Our proposed method 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. The strategy of our proposed scheme (presented in Fig. 3), is applied on each Group of Pictures (GoP) of a video sequence. The first GoP is selectively encrypted with an initial number of encrypted coefficients C_1, which guarantees a minimum degree of visual confidentiality. C_1 is then the number of encrypted coefficients by MB of the first frame (the Intra frame) of the GoP. After encryption, which is done during the entropy encoding, the full GoP is decoded and each frame is analyzed by a set of psychovisual metrics to check if the following GoP should be more encrypted or not with respect to the used metrics. This analysis allows us to set the value of C_2, which is the number of encrypted coefficients by MB for the second GoP. So for a GoP_i, we analyze the encryption of the GoP_{i-1} in order to set the value of C_i.

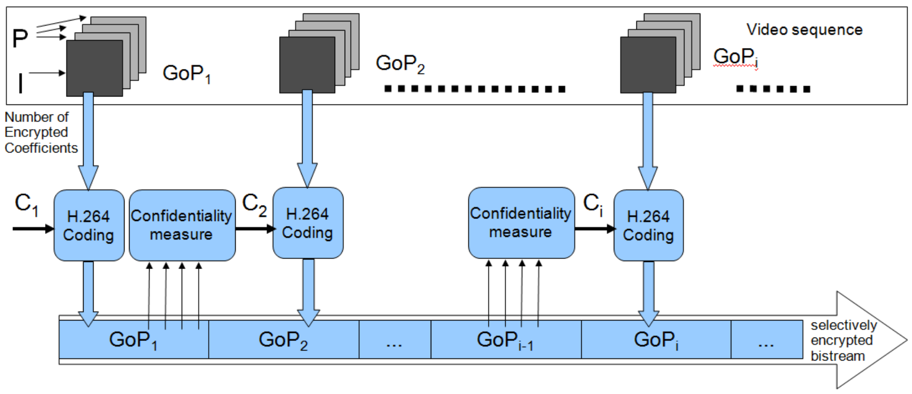


Fig. 3 Overview of the proposed Reduced Selective Encryption scheme.

3.2 Encryption through prediction errors

In this section, we present how the prediction errors of the H.264/AVC encoder are used to reduce the encryption ratio, which is the ratio between the number of encrypted bits and the video datasize (in bits): $Encryption\ ratio = \frac{Size\ of\ the\ encrypted\ bits}{Video\ datasize}$.

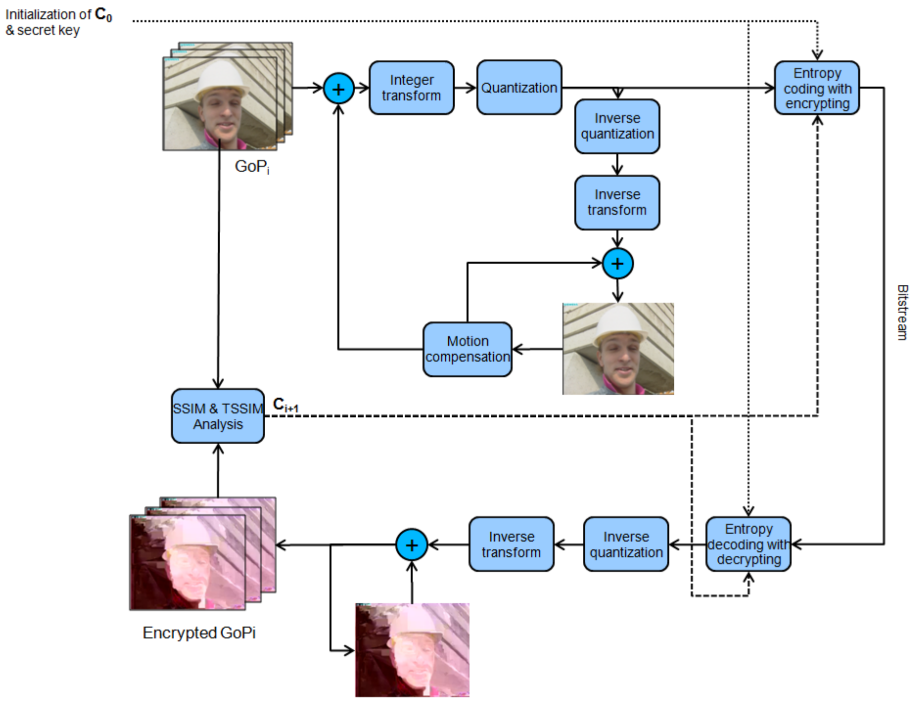


Fig. 4 Overview of the encryption and decryption method.

An overview of the compression and encryption scheme is presented in Fig. 4. 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 selective encryption, 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.

3.3 Psychovisual Metric

In this section, we present the flickering phenomenon, and we propose a new metric in order to measure this phenomenon in the case of encrypted videos.

3.3.1 Flickering phenomenon

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 [20–23]. In [20], a post-processing scheme is presented to reduce the discontinuity. This method is motion compensated-based. Sometimes the flickering is considered and processed like noise, as presented in [21, 22]. In [23] 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. For example, Fig. 5 presents four successive frames of an encrypted video sequence, and Fig. 6 illustrates the grey level variation of the luminance of one pixel of this video sequence with respect to the frame number. Note that there are high variations (more than 100 grey levels) between each frame. This phenomenon is highly annoying for a watcher, especially when the video is read at more than 25 frames per second.

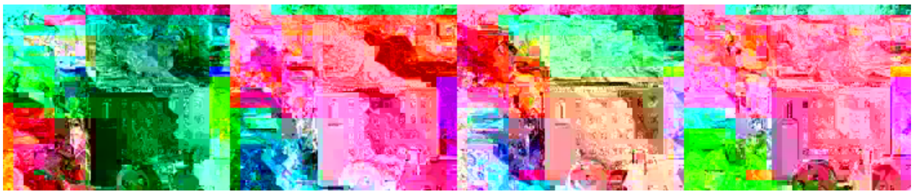


Fig. 5 Four successive frames of the encrypted *mobile* video sequence.

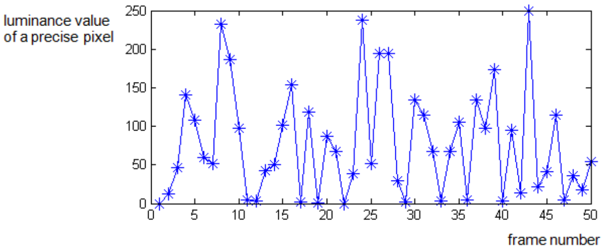


Fig. 6 Evolution of the luminance of one pixel of a selective encrypted *mobile* video sequence with respect to the frame number.

3.3.2 Temporal Structural SIMilarity

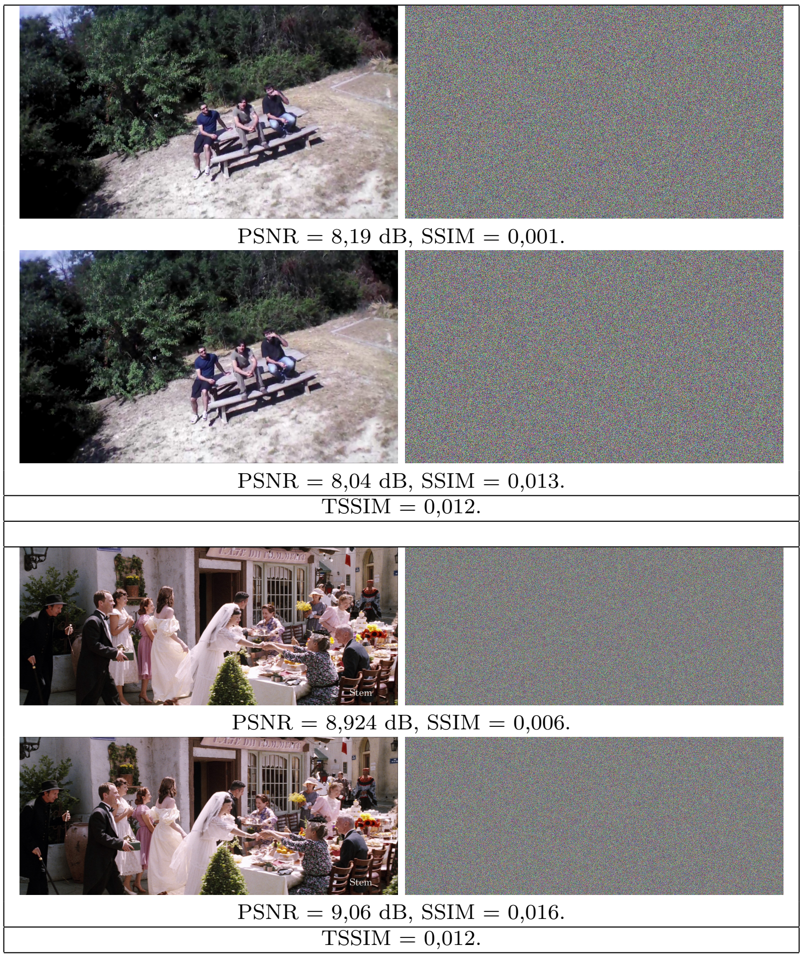
In order to measure and quantify the flickering phenomenon, we have developed a psychovisual metric based on the SSIM applied on successive video frames. We have chose the SSIM-based metric because SSIM is a good metric in terms of psychovisual sensitivity with respect to the Human Visual System (HVS), as presented in [14]. 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:

$$TSSIM_{(I_o,I_e)}(i) = SSIM(|I_o(i) - I_o(i-1)|, |I_e(i) - I_e(i-1)|), \qquad (4)$$

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 0.6 the flickering is not really marked, and below 0.4 we consider that it is sufficient for the visual confidentiality of the video. The last range between 0.4 and 0.6 is a transition range which needs to be analyzed case by case, so we consider that this last range is irrelevant for the visual confidentiality. Fig. 7 presents the quality measures of two successive images of two video sequences. Both of them are encrypted with a full encryption scheme from which we can extract the statistical results of PSNR, SSIM and TSSIM for perfect confidentiality: close to 7-8 dB for the PSNR and close 0 for SSIM and TSSIM.

3.4 Encrypted coefficient adjustment

In our SE scheme, the three color channels, luminance and two chrominances of each frame are affected by the selective encryption. 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 low resolution 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



PSNR = 8,19 dB, SSIM = 0,001.

PSNR = 8,04 dB, SSIM = 0,013.

TSSIM = 0,012.

PSNR = 8,924 dB, SSIM = 0,006.

PSNR = 9,06 dB, SSIM = 0,016.

TSSIM = 0,012.

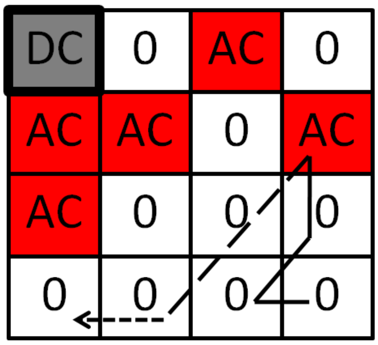
Fig. 7 Full encryption of two successive frames of *ar-drone* (on top right) and *movie* (on bottom right) with the corresponding original frames (on left). The PSNR and the SSIM between the encrypted frame and the original one is indicated below each case. Moreover, TSSIM between the successive frames is indicated below the second frames for each video sequence.

of non-encrypted coefficients, as presented in Fig. 8. In the H.264/AVC compression scheme, high frequencies AC levels are encoded first due to the inverse zigzag scan in the entropic encoder. The combination of these two points results in the scheme presented in Fig. 3. We integrated 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. This control is an algorithm which works like a selection trigger as presented in Algorithm 1. These psychovisual metrics control the visual confidentiality of each GoP_i frame and if one of the frame result metrics

Algorithm 1 Encrypted coefficient adjustment. Computation of C_i, the number of encrypted coefficients for the GoP_i, min and max are the upper and lower thresholds for the psychovisual metrics. $SSIM_i$ is the maximum SSIM for each frame of the GoP_i. $TSSIM_i$ is the T-SSIM of GoP_i or T-SSIM of the last frame of GoP_i relative to the first frame of GoP_i, depending on the encryption scheme.

```
if ($SSIM_i > SSIM_{max}$ $OR$ $TSSIM_i > TSSIM_{max}$) then
  $C_{i+1} \leftarrow C_i + 1$
else
  if ($TSSIM_i < TSSIM_{min}$ $AND$ $SSIM_i < SSIM_{min}$) then
    $C_{i+1} \leftarrow C_i - 1$
  else
    $C_{i+1} \leftarrow C_i$
  end if
end if
```

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 changes in accordance with the video sequence and its different contents.



DC	0	AC	0
AC	AC	0	AC
AC	0	0	0
0	0	0	0

Fig. 8 Example of 4x4 MB, with the inverse zigzag scan during entropic compression. The DC level is encoded independently, and AC non-zero levels are colored in red.

3.5 Decoding step

During the decoding step, the process has to know the amount of encrypted coefficients per MB (value of C_i) in order to correctly decode the video frame. More precisely, the decoder has to know only if the C_i value has to be incremented or decremented compared to the previous C_{i-1} value. Indeed, the initial frame is the only frame which requires the initial number of encrypted coefficients (C_0). In order to solve this problem, two methods can be applied: firstly, the information can be embedded in the header data of each MB. Secondly, the number can be hidden in the DC coefficient of the first MB of each frame by using a watermarking method. This last method is more advisable. The DC coefficient can be watermarked because the DC coefficients are encoded independently of the AC coefficients. Moreover, they are not encrypted. This data-hiding has to be performed during the loop filter in order to be considered for the prediction error as presented in [24]. If not, drifts would occur in the decoded frames. Another decoding method, takes into account the coefficient distributions in order to know if the macro-blocks are encrypted or not, as presented in [25]. In fact, this last method involves calculating the values during the decoding.

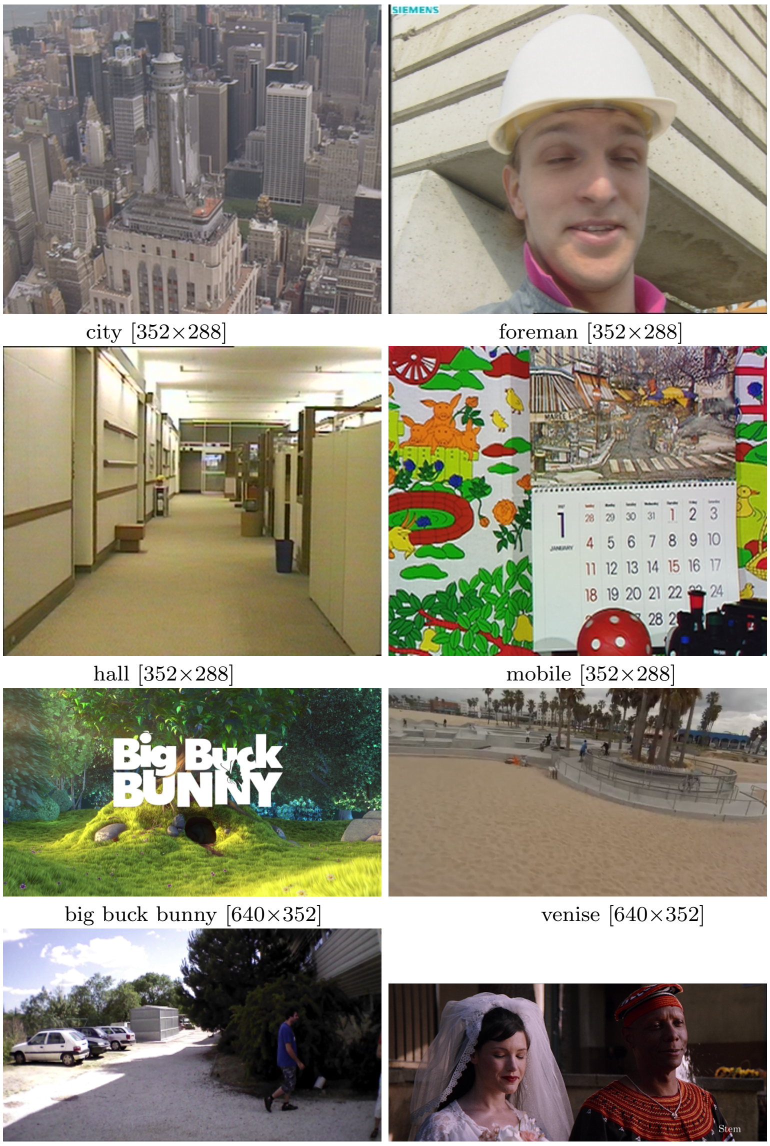
4 Experimental Results

For our experimental results[1], we have used eight benchmark video sequences, four with CIF resolution (352×288 pixels): *city*, *foreman*, *hall* and *mobile*; two with a 640×352 pixel resolution: *big buck bunny* and *venise-fly*; two others in HD: *movie* with a 1920×800 pixel resolution and *ardrone-fly* with a 1280×720 pixel resolution. Fig. 9 presents an image of each video. These video sequences show different combinations of motions, colors, contrasts and objects. The results are presented with the most representative samples. In terms of encryption, we consider that we preserve the confidentiality if the PSNR is less than 13 dB, the SSIM less than 0.6 and the TSSIM less than 0.6. These thresholds have been determined with a subjective experiments. Human users have watched several encrypted videos with different levels of encryption. Also, we have compared these results with respect to the results of quality measures to find the confidentiality thresholds. All of the videos were compressed with a QP of 24, which represents a good trade off between quality and compression with a PSNR of around 35 dB for a compressed video sequence. The GoP size, which optimizes the encryption ratio and confidentiality, also has to be analyzed. A first study was proposed in [26] and an interesting GoP size should be less than 16 frames for video sequences with 30 frames per second. If the GoP size is above than 16, then flickering due to the encryption algorithm is no longer visually annoying to ensure visual confidentiality. 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. The results are separated into: low resolution (6 videos) and high resolution (2 videos). The experimental results on CIF and HD resolutions are presented because they are being used increasingly on new video devices, like smartphones, touchpads and computers.

4.1 Selective encryption of only the I frame of a GoP

In this section, we present experimental results on the encryption of only the I frame for a GoP. This section is divided into two parts: Section 4.1.1 for low resolution cases and Section 4.1.2 for high resolutions cases.

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



city [352×288]

foreman [352×288]

hall [352×288]

mobile [352×288]

big buck bunny [640×352]

venise [640×352]

ardrone-fly [1280×720]

movie [1920×800]

Fig. 9 Benchmarks of videos with their corresponding names.

4.1.1 Results for low resolution video sequences

In the major parts of the results for low resolution videos, 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. This result highlights that

	SIMM		T-SSIM		AoUC	ER-SSE	ER-SE [1]
	Mean	Std	Mean	Std	Mean	(%)	(%)
Cit	0.46	0.17	0.40	0.24	9.05	**16.78**	19.31
For	0.48	0.16	0.53	0.28	9.53	**12.94**	15.83
Hal	0.51	0.05	0.72	0.38	12.1	**12.37**	17.53
Mob	0.33	0.08	0.39	0.20	42.3	**18.56**	25.51
BBB	0.53	0.17	0.71	0.33	28.6	**13.92**	20.15
Ven	0.57	0.08	0.57	0.27	0.81	**13.07**	13.76

Table 1 Results of reduced selective encryption of the six low definition videos 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, in mean for the amount of unencrypted coefficients (AoUC) per macro-block and in percent for the encryption ratio for a Smart SE (ER-SSE) and a full SE (ER-SE).

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. But the T-SSIM is just measured during the switching of GoPs, and, in these cases, the T-SSIM is always under the bottom threshold, as presented in Fig. 10 for the *foreman* video sequence. This particular example is a good outline for the other videos.

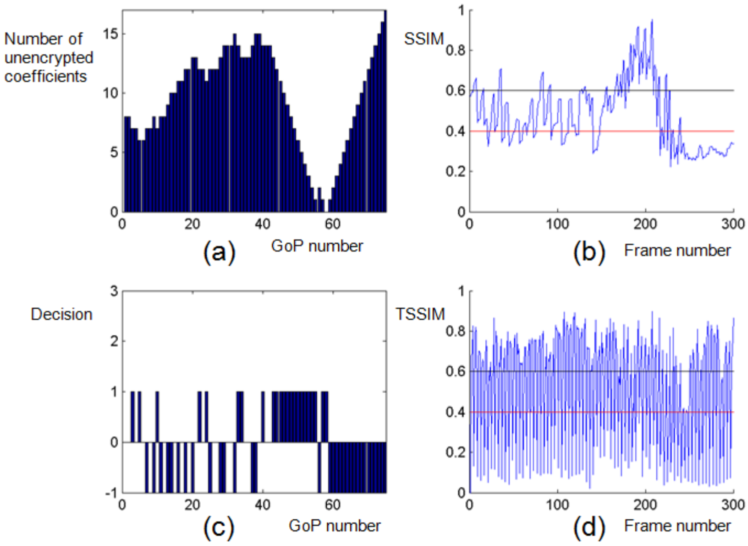


Fig. 10 Reduced selective encryption of only the I frame of a 4-frame GoP of the *foreman* video (300 frames ← 75 GoPs): a) The number of unencrypted coefficients with respect to the GoP number, b) The SSIM of the encrypted video with respect to the frame number, c) The decision with respect to the GoP number, d) The T-SSIM of the encrypted video with respect to the frame number.

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 as we will discuss in Section 4.1.2.

4.1.2 Results for high resolution video sequences

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. A typical example is presented in Fig. 11: in the ROI containing a car; we can clearly see that this ROI is not well protected in the last two frames of the 4-frame GoP whereas the protection remains efficient for a 2-frame GoP. However, in the 2-frame GoP case, the results are still good, as presented in Fig. 12 and Fig. 13, which show that the encryption level depends on the content.

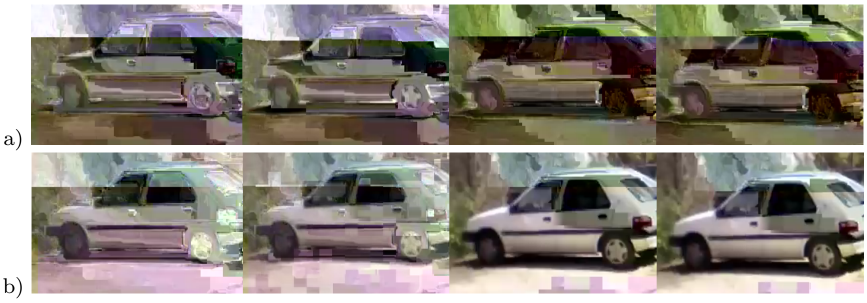


Fig. 11 A ROI of the *ar-drone* video sequence: a) A reduced selective encryption of only the I frame of a 2-frame GoP, b) A reduced selective a encryption of only the I frame of a 4-frame GoP.

In Fig. 12, for the *ar-drone fly* 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. 13, for the *movie* video sequence, switching of scenes during the movie markedly affects the encryption. We can notice in this video that although the encryption is at its maximum value, the SSIM remains too high due to the content of the end at the video in this particular scene.

In Fig. 14, T-SSIM and SSIM are correctly controlled, and the number of unencrypted coefficients is at a stable level at the end of the video sequence. In Fig. 15, 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 14 and Fig. 15, with reduced selective encryption of only the I frame of a 4-frame GoP, the respective results of Fig. 12 and Fig. 13 are better in terms of SSIM and TSSIM. This underlines the necessity of keeping a small GoP for HD video sequences. In high resolution cases, the MB size is small with respect to the size of the ROI of the video, also, and MBs do not have many AC coefficients after the transformation and quantization.

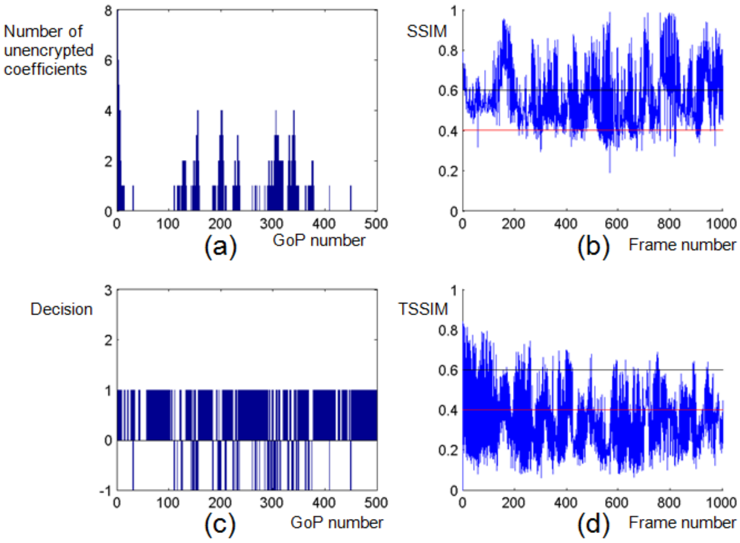


Fig. 12 Reduced selective encryption of only the I frame of a 2-frame GoP of the *ar-drone* video (1000 frames ← 500 GoPs): a) The number of unencrypted coefficients with respect to the GoP number, b) The SSIM of the encrypted video with respect to the frame number, c) The decision with respect to the 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 with respect to the frame number.

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

The experimental results of the encryption of all frames of a GoP are presented in section 4.2, which is divided into two parts: Section 4.2.1 for low resolution cases and Section 4.2.2 for high resolution cases.

4.2.1 Results for low resolution video sequences

Tab. 2 presents the results for the 6 low resolution 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. So this level depends on the video content as we predicted and with respect to the previous results presented in Tab. 1 we can notice that: even if the encryption ratio is similar, the encryption is more widespread through the GoP, which increases the visual

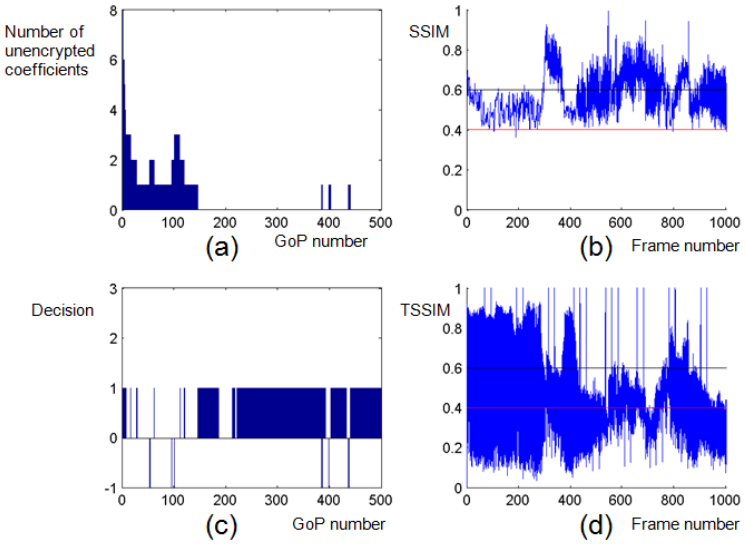


Fig. 13 Reduced selective encryption of only the I frame of a 2-frame GoP of the *movie* video (1000 frames ← 500 GoPs): a) The number of unencrypted coefficients with respect to the GoP number, b) The SSIM of the encrypted video with respect to the frame number, c) The decision with respect to the GoP number, d) The T-SSIM of the encrypted video with respect to the frame number.

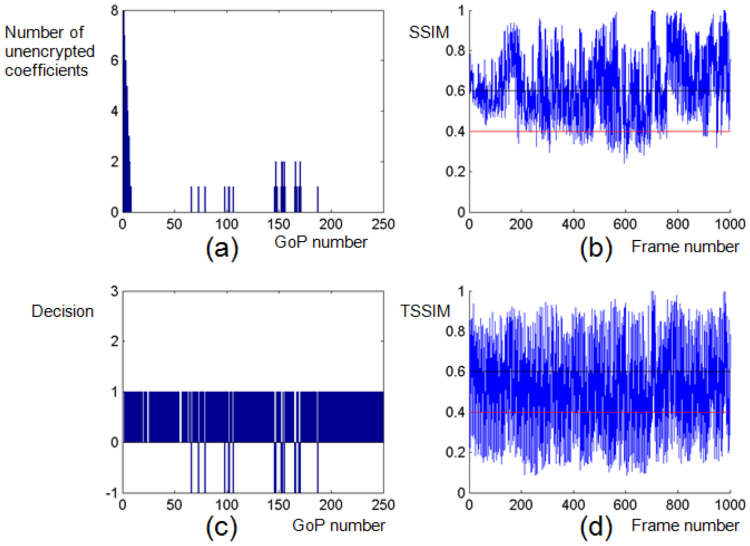


Fig. 14 Reduced selective encryption of only the I frame of a 4-frame GoP of the *ar-drone* video (1000 frames ← 250 GoPs): a) The number of unencrypted coefficients with respect to the GoP number, b) The SSIM of the encrypted video with respect to the frame number, c) The decision with respect to the GoP number, d) The T-SSIM of the encrypted video with respect to the frame number.

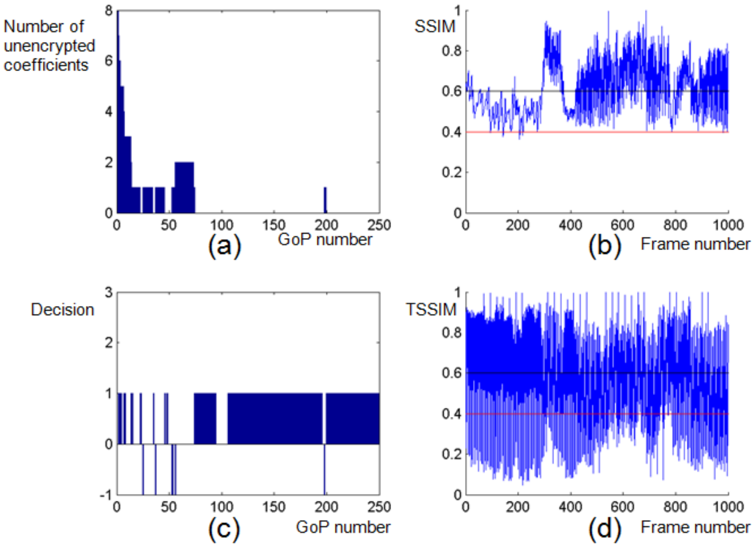


Fig. 15 Reduced selective encryption of only the I of a 4-frame GoP of the *movie* video(1000 frames ← 250 GoPs): a) The number of unencrypted coefficients with respect to the GoP number. b) The SSIM of the encrypted video with respect to the frame number, c) The decision with respect to the GoP number, d) The T-SSIM of the encrypted video with respect to the frame number.

	SIMM		T-SSIM		AoUC	ER-SSE	ER-SE [1]
	Mean	Std	Mean	Std	Mean	(%)	(%)
Cit	0.39	0.09	0.35	0.19	10.22	**16.53**	19.52
For	0.38	0.12	0.46	0.26	1.44	**16.25**	17.34
Hal	0.40	0.06	0.72	0.39	0.48	**17.23**	18.01
Mob	0.32	0.08	0.37	0.19	41.04	**18.87**	26.51
BBB	0.32	0.14	0.67	0.35	0.21	**19.89**	20.29
Ven	0.54	0.08	0.55	0.27	0.14	**13.22**	13.93

Table 2 Results of the reduced selective encryption of the 6 low definition videos, 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 mean for the amount of unencrypted coefficients (AoUC) per macro-block and in percent for the encryption ratio for a Smart SE (ER-SSE) and a full SE (ER-SE).

confidentiality. In 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, as presented in Fig. 16.

Fig. 17 presents a comparison in terms of SSIM with the previous method SE-CAVLC [1]. The encryption ratio is decreased by 1.09% (17.34% of the encryption ratio for the full SE of the *foreman* video sequence) and we maintain visual confidentiality between the two thresholds. The mean SSIM is increased by 0.02 for our method (0.36 as mean SSIM for a full SE of the *foreman* video sequence). Also,

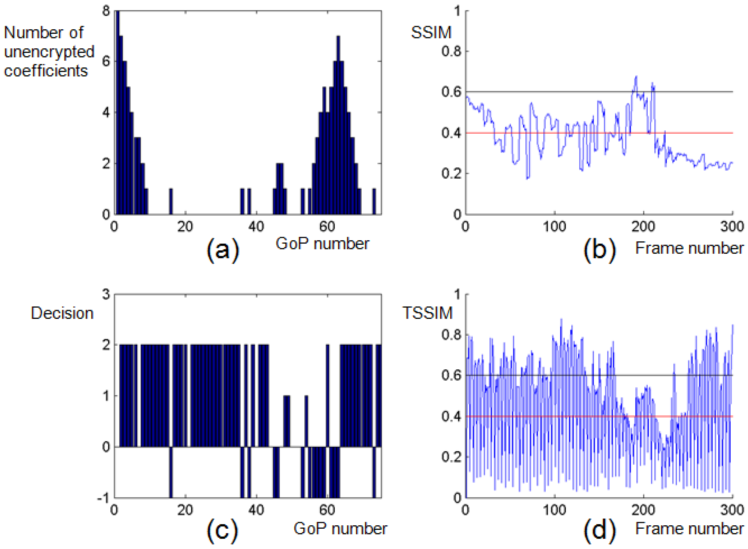


Fig. 16 Reduced selective encryption of a 4-frame GoP of the *foreman* video (300 frames ← 75 GoPs): a) The number of unencrypted coefficients with respect to the GoP number, b) The SSIM of the encrypted video with respect to the frame number, c) The decision with respect to the GoP number, d)The T-SSIM of the encrypted video with respect to the frame number.

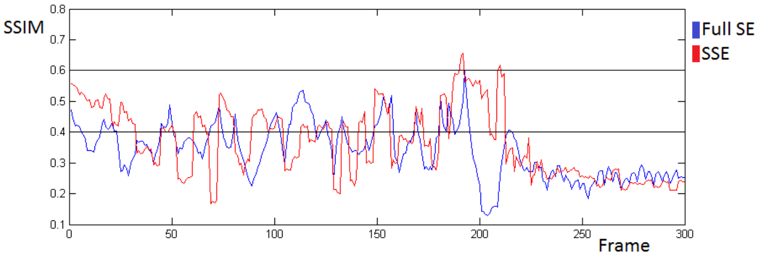


Fig. 17 SSIM of the encrypted *foreman* video sequence with respect to the frame number. In blue with full SE, and in red with Smart SE.

when the encryption is sufficient, we can lighten it when the previous method keeps a full SE.

4.2.2 Results for high resolution video sequences

Fig. 18 and Fig. 19 present reduced SE applied to HD videos. Note that the encryption is sufficient enough for the *ar-drone* video in terms of SSIM and T-SSIM. We can clearly see that with this Smart SE 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 even if the flickering is stronger 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.

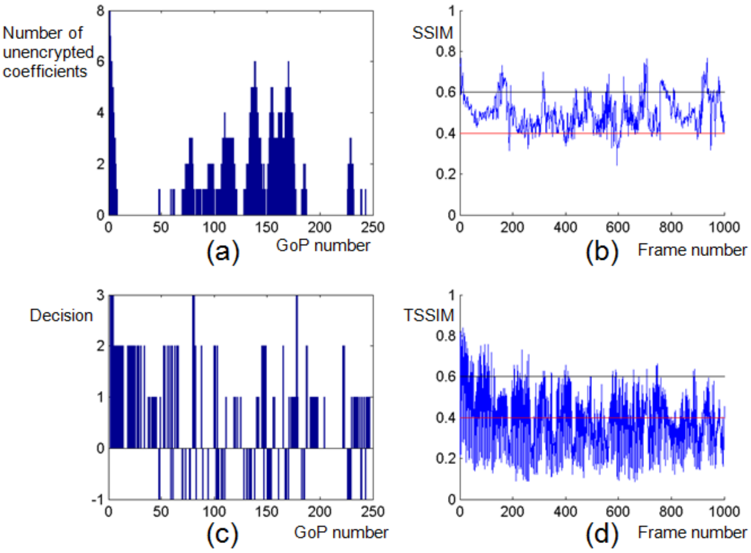


Fig. 18 Reduced selective encryption of a 4-frame GoP of the *ar-drone* video (1000 frames ← 250 GoPs): a) The number of unencrypted coefficients with respect to the GoP number, b) The SSIM of the encrypted video with respect to the frame number, c) The decision with respect to the GoP number, d) The T-SSIM of the encrypted video with respect to the frame number.

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. In this complete example, we can notice a control of the amount of encrypted coefficients with successively a decision due to the T-SSIM, the SSIM and both of them.

4.3 Global analysis and discussions

A first subjective analysis has given several confidentiality thresholds based on quality and flickering measures. These thresholds are used to control the proposed SSE-CAVLC to protect a video sequence with respect to SSIM and T-SSIM results. The proposed method presents several good results in terms of encryption ratio. With respect to [1], the global encryption ratio of our method is 5% lesser in mean. However, the visual confidentiality of the encrypted videos is in mean near 11 dB which is adequate. To conclude Section 4, Fig. 20 presents some visual results of the encrypted video sequences.

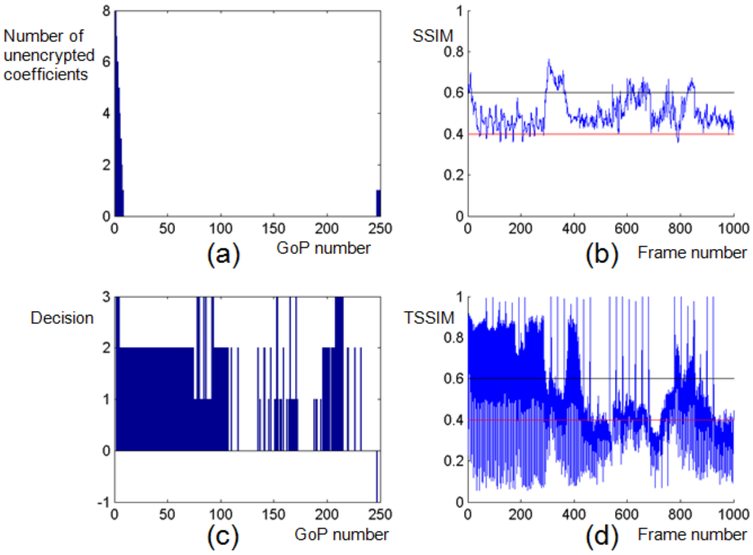
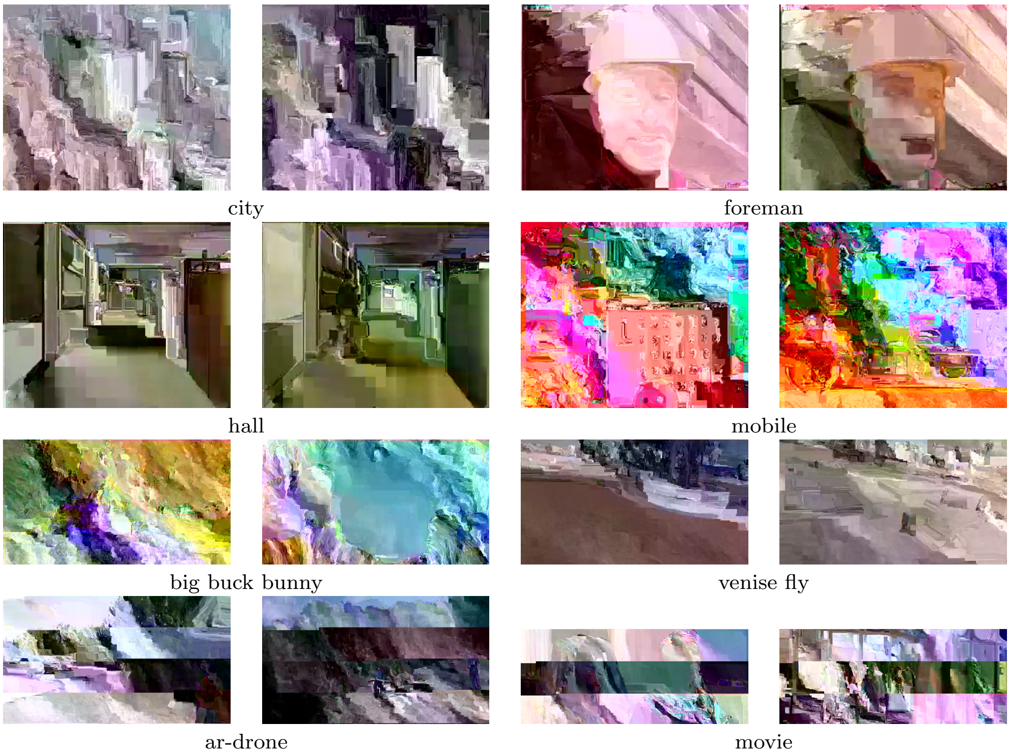


Fig. 19 Reduced selective encryption of a 4-frame GoP of the *movie* video (1000 frames ← 250 GoPs): a) The number of unencrypted coefficients with respect to the GoP number, b) The SSIM of the encrypted video with respect to the frame number, c) The decision with respect to the GoP number, d) The T-SSIM of the encrypted video with respect to the frame number.

5 Conclusion

The reduced selective encryption 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. This T-SSIM quality metric 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 a 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.

Further studies are required to enhance these results. Firstly, we need to develop other psychovisual metrics which are closely correlated with the HVS in terms of low quality and encrypted images. Currently, psychovisual metrics are developed in order to be efficient in high quality cases. Secondly, we need to improve the T-SSIM while taking into account the HVS integration times because the T-SSIM measures two successive frames regardless of the frame frequency. T-SSIM needs also to be extended in order to measure the flickering on several frames for a more precise local measure. Moreover, the encryption can be improved by including some ROIs from objective measures. Finally, this encryption scheme is efficient for most video resolutions, but it deserves to be improved for HD videos.



city

foreman

hall

mobile

big buck bunny

venise fly

ar-drone

movie

Fig. 20 Visual results of two successive encrypted frames of the different video sequences with their respecting names.

References

1. 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.
2. Joint Video Team, Draft ITU-T Recommendation and Final Draft International Standard of Joint VIdeo Specification (ITU-T Rec. H.264 / ISO/IEC 14496-10 AVC), Doc. JVT-G050 Tech. Rep.
3. 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.
4. T. Lookabaugh, D. Sicker, Selective Encryption for Consumer Applications, IEEE Communications Magazine 42 (5) (2004) 124–129.
5. H. Chen, X. Li, Partial Encryption of Compressed Images and Videos, IEEE Transactions on Signal Processing 48 (8) (2000) 2439–2445.
6. A. Uhl, A. Pommer, Image and Video Encryption - From digital Rights Management to Secured Personal Communication, Springer, 2005.
7. C. Wu, C. Kuo, Design of Integrated Multimedia Compression and Encryption Systems, IEEE Transactions on Multimedia 7 (2005) 828–839.
8. J. Wen, M. Severa, W. Zeng, M. Luttrell, W. Jin, A Format-Compliant Configurable Encryption Framework for Acess Control of Video, IEEE Transactions on Circuits and Systems for Video Technology 12 (6) (2002) 545–557.
9. 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.

10. 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.
11. S. Choi, J. Han, H. Cho, Privacy-Preserving H.264 Video Encryption Scheme, Information, Telecommunication & Electronics - ETRI Journal 33 (6) (2011) 935–944.
12. I. Avcibas, B. Sankur, K. Sayood, Statistical Evaluation of Image Quality Measures, Journal of Electronic Imaging 11 (2002) 206–223.
13. Z. Wang, A. C. Bovik, R. Hamid, Sheik, E. P. Simoncelli, Image Quality Assessment: From Error Visibility to Structural Similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.
14. 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.
15. Z. Wang, A. C. Bovik, E. P. Simoncelli, Multi-scale Structural Similarity for Image Quality Assessment, IEEE Asilomar Conference Signals, Systems and Computers (2003) 1398–1402.
16. Z. Wang, E. P. Simoncelli, Translation insensitive image similarity in complex wavelet domain, IEEE International Conference In Acoustics, Speech, and Signal Processing, Montery, California, U.S.A. (2005) 573–576.
17. 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.
18. J. Zujovic, T. Pappas, D. Neuhoff, Strutural Similarity Metrics for Texture Analysis and Retrieval, IEEE Conference on Image Processing, Cairo, Egypt (2009) 2225–2228.
19. S. Chikkerur, V. Sundaram, M. Reisslein, L. J. Karam, Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison, IEEE Transactions on Broadcasting vol. 57, no. 2, (2011) 165–182.
20. Y. Kuszpet, D. Kletsel, A. Levy, Post-processing for flicker reduction of H.264/AVC, Picture Coding Symposium.
21. 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.
22. B. C. Song, K. W. Chun, Noise Power Estimation For Effective De-noising in a Video Encoder, IEEE International Conference on Acoustics, Speech and Signal Processing, Monterey, California, U.S.A. (2005) 357–360.
23. 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.
24. Z. Shahid, M. Chaumont, W. Puech, Considering the reconstruction loop for data hiding of intra- and inter-frames of H.264/AVC, Signal, Image and Video Processing, Springer (2011) 1–19.
25. N. Islam, W. Puech, Noise removing in encrypted color image by statistical analysis, SPIE, Electronic Imaging, Media Watermarking, Security and Forensics 8303 (2012) 10.1117/12.910497.
26. L. Dubois, W. Puech, J. Blanc-Talon, Reduced Selective Encryption of Intra and Inter Frames of H.264/AVC using Psychovisual Metrics, IEEE International Conference on Image Processing, Orlando, Florida, U.S.A. (2012) 2641–2644Orlando, USA.