Privacy-Preserving H.264 Video Encryption Scheme

SuGil Choi, Jong-Wook Han, and Hyunsook Cho

As a growing number of individuals are exposed to surveillance cameras, the need to prevent captured videos from being used inappropriately has increased. Privacy-related information can be protected through video encryption during transmission or storage, and several algorithms have been proposed for such purposes. However, the simple way of evaluating the security by counting the number of brute-force trials is not proper for measuring the security of video encryption algorithms, considering that attackers can devise specially crafted attacks for specific purposes by exploiting the characteristics of the target video codec. In this paper, we introduce a new attack for recovering contour information from encrypted H.264 video. The attack can thus be used to extract face outlines for the purpose of personal identification. We analyze the security of previous video encryption schemes against the proposed attack and show that the security of these schemes is lower than expected in terms of privacy protection. To enhance security, an advanced block shuffling method is proposed, an analysis of which shows that it is more secure than the previous method and can be an improvement against the proposed attack.

Keywords: H.264 video encryption, privacy protection, video surveillance, block shuffling.

I. Introduction

Security cameras have become ubiquitous in many countries. Whereas they previously appeared only in banks or other high-security areas, they are now entering public spaces such as malls, streets, and public transportation. Surveillance cameras have several benefits. One obvious benefit is that they assist the police in catching criminals during or after the act, and thus help in reducing crime. However, the biggest objection of security cameras concerns individual privacy. Many people feel that they should be able to travel or move around freely without being photographed or recorded because their personal information could be abused by corrupt authorities or hackers intercepting the video data during network transmission [1], [2].

While some misuse can be prevented by following video surveillance guidelines aimed at minimizing the impact on privacy and reducing potential law enforcement abuse, these measures alone are far from perfect. What is more greatly needed is a technology that protects video data against illegal access. This issue can be addressed by video encryption, and in the past decade, several algorithms, including those described in [3]-[14], have been reported on. These algorithms are called selective encryption because only a subset of the data is encrypted instead of the entire bit stream [12]. Thus, the amount of data to encrypt can be reduced. Through the encryption, a whole frame or region of interest (ROI) [15], containing privacy-sensitive information, can be secured. Since the difference between ROI and full frame encryption is the amount of data to be encrypted, we do not discuss them separately.

To provide sufficient privacy protection, the security of encryption algorithms must be evaluated. In security analysis, the recovery of partial information that is perceptually intelligible should be considered a security breach. In the case of privacy protection in video surveillance, the meaningful

Manuscript received Oct. 29, 2010; revised May 13, 2011; accepted June 29, 2011.

This work was supported by the IT R&D program of MKE, Rep. of Korea (Development of Video Surveillance Security Technologies for Preserving Personal Security).

SuGil Choi (phone: +82 42 860 1367, email: sooguri@etri.re.kr) Jong-Wook Han (email: hanjw@etri.re.kr), and Hyunsook Cho (email: hscho@etri.re.kr) are with the Software Research Labortory, ETRI, Daejeon, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.11.0110.0644

partial information for recognizing an individual is the face outline, which is generally considered contour information. In this paper, we propose an attack called a sign-only attack, which is specially designed for extracting contour information from encrypted video. More specifically, we focus exclusively on H.264, which is the current state-of-the-art video technology. If a person can be identified from the recovered face outline through the attack, the encryption algorithm is weak in the sense of privacy protection. As the name of the attack implies, the basic idea comes from the observation that discrete cosine transform (DCT) coefficient signs contain edge information. Unlike the proposed scheme in [16], where its naïve application to DCT coefficients in an H.264 encoded stream does not produce any contour information, we developed a customized algorithm for an H.264 stream that is capable of showing a contour image.

We evaluate the security of the existing H.264 video encryption schemes against the sign-only attack and show that the cost of a brute-force attack can be significantly reduced. After identifying the previous encryption schemes which are capable of defending against the sign-only attack, an appropriate solution to enhance its robustness against the attack is proposed. The new method, coupled with another encryption scheme, can make the visual degradation of recovered contour images more serious, and thus it enhances individual privacy in video surveillance systems based on H.264 technology.

This paper is organized as follows. Previous H.264 video encryption schemes are briefly reviewed in section II. The sign-only attack along with a security evaluation of existing encryption schemes is proposed in section III. Section IV proposes a method to improve privacy protection in H.264 video encryption. Finally, conclusions are drawn in section V.

II. Background

This section covers a brief overview of the H.264 selective encryption algorithms proposed so far. Firstly, variable-length coding (VLC) in H.264 standard [17] is introduced as background knowledge for understanding encryption algorithms.

1. H.264 Video Standard

H.264 encoder processes an input frame in units of a macroblock (corresponding to 16×16 pixels). All luma and chroma samples of a macroblock are either spatially or temporally predicted, and the resulting prediction residual is encoded using transform coding. For transform coding purposes, the residual macroblock is subdivided into smaller 4×4 blocks. Each block is transformed using an integer transform, and the transform coefficients are quantized and encoded using entropy coding methods. The standard specifies two types of entropy coding: context-based adaptive binary arithmetic coding and VLC.

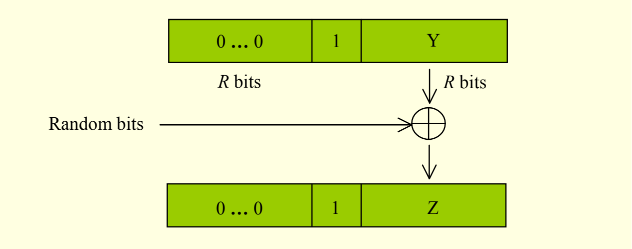


Fig. 1. Exp-Golomb encryption.

Since the following encryption algorithms make use of characteristics of the VLC, it is explained in more detail. When elements are coded using variable-length codes, residual coefficients are coded using a context-adaptive VLC (CAVLC) scheme, and other variable-length coded units are coded using Exp-Golomb codes. The Exp-Golomb code in binary form is composed of R 0's, one '1', and R bits of information Y, as shown in Fig. 1. More information on the Exp-Golomb code can be found in 9.1 of [17].

In the CAVLC, the number of non-zero quantized coefficients and the actual size and position of the coefficients are coded separately. Detailed information on CAVLC can be found in 9.2 of [17]. CAVLC encoding proceeds as follows.

i) Encode coeff_token (the total number of non-zero coefficients and the number of trailing +/–1 values). Trailing 1s (T1s) indicate the number of coefficients with an absolute value equal to 1 at the end of the scan.

ii) Encode the sign of each T1. A single bit encodes the sign (0 = +, 1 = –).

iii) Encode the levels (sign and magnitude) of the remaining non-zero coefficients. The level of non-zero coefficients tends to be higher at the start of the reordered array (near the DC coefficient) and lower towards the higher frequencies. CAVLC takes advantage of this by adapting the choice of a VLC look-up table for the level parameter depending on recently-coded level magnitudes. There are 7 VLC tables to choose from, VLC0 to VLC6.

iv) Encode positions of each non-zero coefficient by specifying the positions of 0's before the last non-zero coefficient. Total_zeros and run_before are the elements for conveying this information.

2. H.264 Selective Encryption Algorithms

According to the encryption process, algorithms are classified into two types: XORing selected bits with random bits and block shuffling.

A. XORing Selected Bits with Random Bits

As a reconstructed block using an incorrect intra-prediction mode or motion vector seriously degrades the visual quality, an intra-prediction mode and motion vector difference are proper candidates for encryption [4], [5].

Both intra-prediction mode and motion vector difference are encoded using the Exp-Golomb code. To maintain format compliance, only the *R*-bit suffix is XORed with random bits. The resulting suffix is another valid codeword, and the length is kept the same.

'T1s' and 'levels' are usually the selected components for encryption in CAVLC [4], [8]. As a single bit encodes each T1, the encryption of T1 is straightforward. Levels are encoded using an adaptive VLC table, and the codeword is similar to that of an Exp-Golomb code as depicted in Fig. 1. The last bit of the level suffix is used for signaling the +/– sign information, which is thus called a sign bit. To maintain format compliance, only the suffix bits are XORed with random bits. In some cases, only the sign bits are encrypted for greater efficiency.

B. Block Shuffling

Shuffling or permutation is a common cryptographic primitive operation. The spatial characteristic of visual data makes permutation a natural way to scramble the semantic meaning of multimedia signals [7]. While block shuffling is generally used for video encryption, the approaches for previous coding standards such as MPEG1 and MPEG4 cannot be used for H.264 because the neighboring blocks and codewords are context-sensitive [6]. The number of non-zero coefficients in neighboring blocks and the adaptive factors of each codeword in CAVLC blocks are correlated as shown in Fig. 2.

There are four possible look-up tables to use for encoding the coeff_token, and the choice of table depends on the number of non-zero coefficients in the upper and left-handed previously-coded blocks, *n*B and *n*A [17], [18]. If blocks A and B are available, *n*C=(*n*A+*n*B)/2, and *n*C selects the look-up table. Therefore, if permutation is carried out without considering this relationship, the resulting encrypted video may not be format-compliant. As a solution, the authors in [6] proposed an algorithm in which the residual blocks are categorized into different groups according to the number of non-zero coefficients and the *n*C value, and each group uses a different shuffling table.

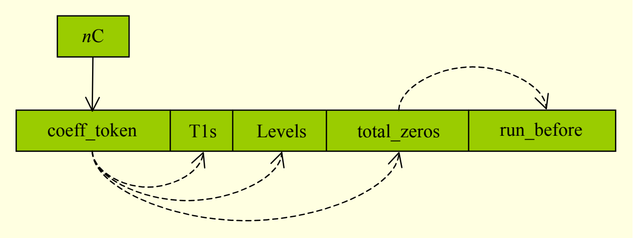


Fig. 2. Correlation between elements in CAVLC encoding.

III. Sign-Only Attack

In cryptography, a brute-force attack or exhaustive key search is a strategy that can in theory be used against any encrypted data. From Table 1, an attacker needs to recover 9,848 bits to accurately get an original I-frame, in case that News video sequence is encrypted by XORing T1s and sign bits of the levels. If a 256-bit key is used, the security against exact recovery by exhaustive search is determined by the key length rather than the number of encrypted bits. Therefore, it might be reasonable to say that a CIF-sized or larger encrypted video can be kept secure because 256-bit key length is generally recognized to provide long-term security [19].

However, according to Kerckhoff's principle, an encryption system should be secure even if the attacker knows all details about the system, with the exception of the secret key [19]. Since the brute-force attack treats the encryption algorithm as a black box, the robustness against the attack is not sufficient. An attacker would look for the weakest part in the algorithm or opt to recover approximate data rather than exact recovery by exploiting the internal structure of the encryption algorithm or the specific knowledge of target data. In this regard, there might be various ways of attacking video encryption algorithms. Here, we propose one such attack, called a sign-only attack, which is designed for extracting contour information from an encrypted H.264 video. Since the recovery of an approximate face outline through a sign-only attack can lead to an invasion of privacy, privacy-preserving video encryption algorithms should be resilient against the proposed attack. Thus, the sign-only attack can play an important role for measuring the strength of video encryption

Table 1. Number of bits for brute-force trials per frame (averaged over 90 frames of intra-coded CIF-sized video).

Video	T1s + suffix bits of levels (a)	T1s + sign bits of levels (b)	Ratio (b/a)
News	11,587	9,848	85.0%
Stefan	30,603	25,293	82.5%
Foreman	12,927	11,509	89.0%
Paris	28,338	22,852	80.5%

algorithm in terms of privacy protection.

The attack development proceeds in steps, starting with the basic idea of a sign-only image, through the customization of the idea for use in intra-coded H.264 video, and finally reducing the required number of sign bits for creating a sign-only video. In short, it is the goal of a sign-only attack to create a sign-only image revealing intelligible contour information by using minimal number of sign bits. We also evaluate the security of the previous H.264 video encryption schemes against the proposed attack and show that the cost of a brute-force attack can be significantly reduced.

1. Basic Idea

DCT coefficient signs contain important edge information in an image, as shown in [16]. A DCT sign-only image is constructed by setting the amplitude of non-zero DCT coefficients to 1. In other words, all the positive coefficients are mapped to 1, and all the negative coefficients are set to –1. The effect of sign-only synthesis is similar to that of a high-frequency filter except that the amplitude of the coefficients does not affect the resulting image. Since the bits for coding the amplitude of the coefficients can be ignored without trying to recover original bits, the number of trials for exhaustive search can be reduced as illustrated in Table 1, where the average reduction per frame is 15.75%. The lowered complexity can be further improved, which will be described in the following subsections.

2. Intra-Coding and DCT Sign-Only Video

In contrast to previous video coding standards, H.264 introduces an intra-prediction scheme for encoding I-frames. However, this new mechanism hinders the creation of a DCT sign-only video. The main cause of the problem is that the coefficients of an intra-coded block in H.264 are the result of transforming only the residual data, whereas the coefficients of each block in [16] are obtained by applying a transformation to the original data. Figure 3(a) shows a normally decoded video, while Fig. 3(b) demonstrates the results of decoding only residual data. To obtain clear contour information, we use only the luma color component, and thus the resulting sign-only videos are in grayscale. When looking carefully at the image in Fig. 3(b), one can notice that a rough contour of the scene appears, although it is noisy. This observation hints at the possibility that the creation of a DCT sign-only video using only the coefficients from residual data might be feasible, and we provide a valid reason for this supposition as follows.

We exploit the fact that most regions of natural images except edges are covered by areas with smoothly changing pixel values. Since residual data is obtained by subtracting the predicted pixels from the original pixels, and because the predicted pixels tend to represent the smooth area that occurs across neighboring blocks, it is likely that smooth area is removed whereas edges remain in the residual data. Thus, during the process of creating a sign-only video, it is not necessary to recover the predicted pixels. This brings about another advantage. It is feasible to create a sign-only video for only a selected region within a frame since decoded pixels not within the target region are unnecessary. In this way, we can efficiently create a sign-only video consisting of only facial regions. Otherwise, all blocks to the top and left of the selected region should also be processed to obtain a sign-only image of the region.

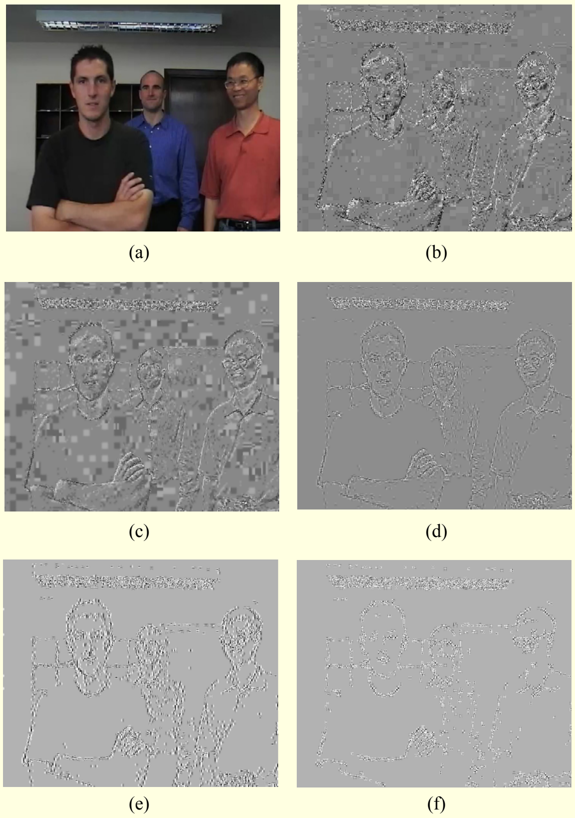


Fig. 3. DCT sign-only video: (a) original motinas_multi_face_frontal video in [20], (b) results of decoding only residual data, (c) sign-only video setting amplitude to 400, and (d)-(f) high-frequency filtered sign-only video. Number of cutoff coefficients are (d) 1, (e) 2, and (f) 3.

However, it is still not possible to obtain any meaningful contour information by applying the scheme in [16] to the coefficients of residual data, that is, the setting of the amplitude of non-zero DCT coefficients to 1. There are two reasons for this. First, according to the H.264 specifications, all coefficients are scaled by 64 before inverse transform. Second, the contour image reconstructed from the residual data is represented by close contrast values. Therefore, the amplitude of non-zero coefficients should be set to a larger value than 1 to deal with

Table 2. Number of sign bits for creating sign-only video with different numbers of cutoff coefficients.

Video	Original	1	2	3
News	9,848	7,926	6,686	5,380
Stefan	25,293	21,967	18,971	16,380
Foreman	11,509	10,057	8,151	6,370
Paris	22,852	19,428	17,003	14,747
motinas_multi_face_frontal	243,345	164,207	116,965	67,909

Table 3. Summary of video encryption algorithms with respect to their defense capability against sign-only attack.

Video encryption algorithm	Defense capability	Description
Intraprediction mode	Weak	Predicted pixels can be set to a constant regardless of the mode.
Motion vector	Weak	This is not used in I-frame.
T1 + level suffix	Medium	The encryption of bits other than sign bits are meaningless.
Block shuffling	Strong	Effective

the scaling-up and increase the image contrast. Figure 3(c) shows the results obtained when setting the amplitude to 400. After the inverse DCT, the pixel values are adjusted to the best range to improve visibility.

3. Reducing the Required Number of Sign Bits

When a sign-only video is created by setting the predicted pixels to a constant, the application of a high-frequency filter can reduce the required number of sign bits for creating a sign-only video while maintaining almost the same level of identifiability for individuals in the video.

Figure 3(c) shows a sign-only video prior to applying a high-frequency filter, while the images in Figs. 3(d) through 3(f) are the results of high-frequency filtering with different numbers of cutoff coefficients. The individuals in the image are still identifiable after removing the two lowest frequency coefficients as shown in Fig. 3(e), and the required number of sign bits is reduced by more than 50% as depicted in the last row of Table 2. Removing the two lowest frequency coefficients generally produces clear contour image while using less sign bits.

The above mentioned approach for creating a sign-only video from an H.264 encoded video is extensible to all video coding techniques based on DCT, and the general algorithm can be described as follows:

Step 1. Set the amplitude of coefficients to a constant.

Step 2. Throw away some low-frequency coefficients.

Step 3. Set the predicted values to a constant value (for example, intra-predicted pixels are set to 128).

Step 4. Convert a color image to grayscale.

Step 5. Perform image processing to maximize the visibility of the contour image.

4. Security of Previous Encryption Schemes against the Sign-Only Attack

In our analysis, four representative H.264 video encryption algorithms, shown in Table 3, are considered. We evaluate the strength of the encryption algorithms by counting the number of brute-force trials to recover the approximate contour image through the sign-only attack. Two schemes are shown to be ineffective defense measures against the attack, while the strength of another algorithm is lower than expected. Thus, when the purpose of evaluation is measuring the level of privacy protection, the security analysis of encryption algorithms against the exact recovery of original video from encrypted one is not appropriate.

In terms of visual degradation, encryption of intra-prediction modes is the most efficient tool. By efficiency, we mean that an encrypted video becomes unintelligible while encrypting fewer numbers of bits compared to other encryption schemes. However, a sign-only video can be created without recovering the original prediction mode because the predicted values can be set to a constant regardless of the original mode. Therefore, intra-prediction mode encryption is of no use.

In an H.264 video stream, a group of pictures is usually composed of one I-frame and several P-frames. P-frames provide more compression than an I-frame, and thus the residual data in P-frames contains much less information. Since the identification of individuals in one frame leads to a total compromise of privacy, attackers can focus on I-frames without wasting time working on P-frames. Motion vector encryption is effective only in P-frames, and thus the method cannot be used as a proper defense mechanism against the proposed attack.

Since only sign bits are used during the process of creating a sign-only video, the encryption of level suffixes other than sign bits is meaningless. Furthermore, some of the T1s and level suffixes can be discarded by high-frequency filtering, and therefore the level of contour distortion corresponds to only the remaining encrypted bits. For example, for a News video, the number of bits for T1s and level suffixes per frame is 11,587 as presented in Table 1. However, after discarding the two lowest frequency coefficients and all bits in the level suffixes except

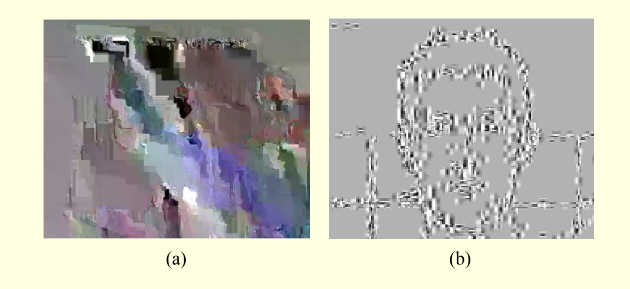


Fig. 4. Sign-only video created from encrypted motinas_multi_face_frontal video in [20]: (a) encryption of intra-prediction mode, T1, and level suffix and (b) face region in sign-only video created from (a).

sign bits, the remaining number of bits is only 6,686 as depicted in Table 2. Out of the encrypted bits, only 57.7% of them can contribute toward making the contour appear noisy. In addition, the encryption of T1s and levels in the chroma blocks is meaningless because the chroma value of each pixel is set to a constant to work in the grayscale.

There still exists some useful information as shown in Fig. 4(b), although the encrypted video contains almost no identifiable information. This is because some important information in an encoded block is not encrypted by the above analyzed encryption schemes. First, the levels encoded using a VLC0 table are not encrypted so as to maintain format compliance, and thus the contour information in those levels are kept as is. Second, although coefficient positions are important for reconstructing the block, the coefficients inside a block remain in place because total_zeros and run_before are not encrypted. Block shuffling might address these two problems because the unencrypted information can only be sensible in the correct context. The security of block shuffling against brute-force trials is explained in section IV.2.A.

IV. Advanced Block Shuffling Scheme

In the previous section, it was shown that two encryption algorithms can be used as defense mechanisms against the sign-only attack intended to reveal contour information of an individual, that is, face contour. However, there is no guarantee that the two encryption schemes are secure enough to provide sufficient privacy protection. Therefore, it is better to improve the security of the encryption schemes as much as possible, and we investigate a method to improve these schemes.

The algorithm used for encrypting levels can be modified to also encrypt levels encoded using a VLC0 table by adding or eliminating zeros in the prefix. Since the number of zeros in the prefix is usually small, it is highly likely that more zeros will be added to the prefix, leading to an increase in the amount of data in a video stream. This may not be acceptable considering that H.264 has been developed to deliver higher compression ratios. To the best of our knowledge, it is not feasible to improve the algorithm used for encrypting T1s and levels while satisfying H.264 format compliance and retaining the video stream size.

On the other hand, a block shuffling scheme can be further enhanced in such a way that it does not adversely impact the format compliance or compression friendliness. In this section, we present an advanced block shuffling scheme which is superior to the one in [6].

1. Algorithm Description

When shuffling is performed on a set of codewords corresponding to a coefficient block, an amount of visual degradation can be achieved. Through block-based shuffling, blocks are put into incorrect locations, but the information inside each block is kept as is. This consideration has lead to the development of an advanced block shuffling algorithm. In H.264, CAVLC is used to encode DCT coefficient blocks, and it consists of five components as depicted in Fig. 2. The components can be classified into two parts, namely P1 and P2, based on the information they code: the amplitude and position of non-zero coefficients. P1 is composed of coeff_token, T1s, and levels. Total_zeros and run_before are used for P2.

For each frame, P1s with the same num_coeff (the number of non-zero coefficients inside a block) and *n*C are put into the same group. Therefore, the number of shuffling groups for P1 is the number of distinct combinations of num_coeff and *n*C, and we represent each shuffling group as a pair (*n*C, num_ceoff) as shown in Fig. 5. P2s with the same num_coeff are put into the same group. The codeword derivation used to decode total_zeros is dependent on the num_coeff in the corresponding block, and total_zeros is input into the process of parsing run_before. Therefore, all P2s in the same group must be linked to the same num_coeff for format compliance. Each shuffling group is denoted as (num_coeff). Finally,

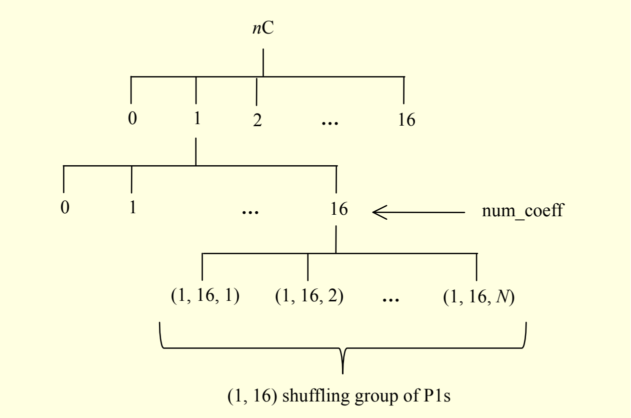
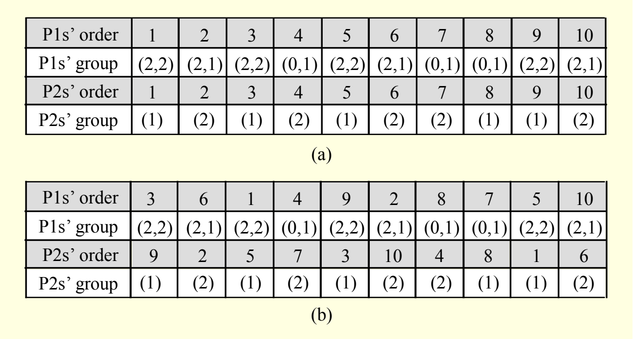


Fig. 5. Steps used for creating shuffling groups of P1s.



P1s' order	1	2	3	4	5	6	7	8	9	10
P1s' group	(2,2)	(2,1)	(2,2)	(0,1)	(2,2)	(2,1)	(0,1)	(0,1)	(2,2)	(2,1)
P2s' order	1	2	3	4	5	6	7	8	9	10
P2s' group	(1)	(2)	(1)	(2)	(1)	(2)	(2)	(1)	(1)	(2)

(a)

P1s' order	3	6	1	4	9	2	8	7	5	10
P1s' group	(2,2)	(2,1)	(2,2)	(0,1)	(2,2)	(2,1)	(0,1)	(0,1)	(2,2)	(2,1)
P2s' order	9	2	5	7	3	10	4	8	1	6
P2s' group	(1)	(2)	(1)	(2)	(1)	(2)	(2)	(1)	(1)	(2)

(b)

Fig. 6. Example of proposed block shuffling: (a) before shuffling and (b) after shuffling.

random shuffling is performed on each group, and as shown in Fig. 6, the order of P1s and P2s is scrambled within the corresponding group.

2. Security Analysis

In this part, we show the improved resistance to brute-force attacks compared to previous block shuffling [6] and evaluate the perceptual security against the sign-only attack.

A. Security against Exact Recovery by Exhaustive Search

The proposed shuffling scheme is more resistant to brute-force attacks than the one in [6] because the information within a block is not maintained as is. The information inside a block is categorized into two parts, the amplitude (P1) and the position (P2) of the coefficients, and the proposed shuffling method mixes the two parts independently within the corresponding shuffling group.

When the corresponding num_coeff is 4, the multiplicities of unique elements are listed in Table 4 and the number of P2 permutations is greater than 2^{200}. Since the number of permutations is much larger when num_coeff is 1, 2, or 3, the overall number of permutations is far more than 2^{200}. It would be easier for a brute-force attacker to guess the key assuming that the cryptographic key used for generating a secret shuffling table is less than 200 bits long. When the previous shuffling scheme is applied to CIF-sized video, the security against exact recovery by exhaustive search is also determined by the key length rather than the number of permutations [6]. Therefore, if the strength of the two shuffling schemes is to be judged solely on key length, they provide same level of security. However, brute-force attacks can be made much less effective by the proposed shuffling scheme. To recognize when an attacker has cracked the encryption, he/she has to build a shuffling table trying every possible key and evaluate some metric computed from the reconstructed video data. Since the number of shuffled elements in the proposed shuffling is twice the number of that in the previous scheme, the computational cost for testing every possible key will be increased as a result.

Table 4. Multiplicities of elements in P2 data (first frame of intracoded CIF-sized News video).

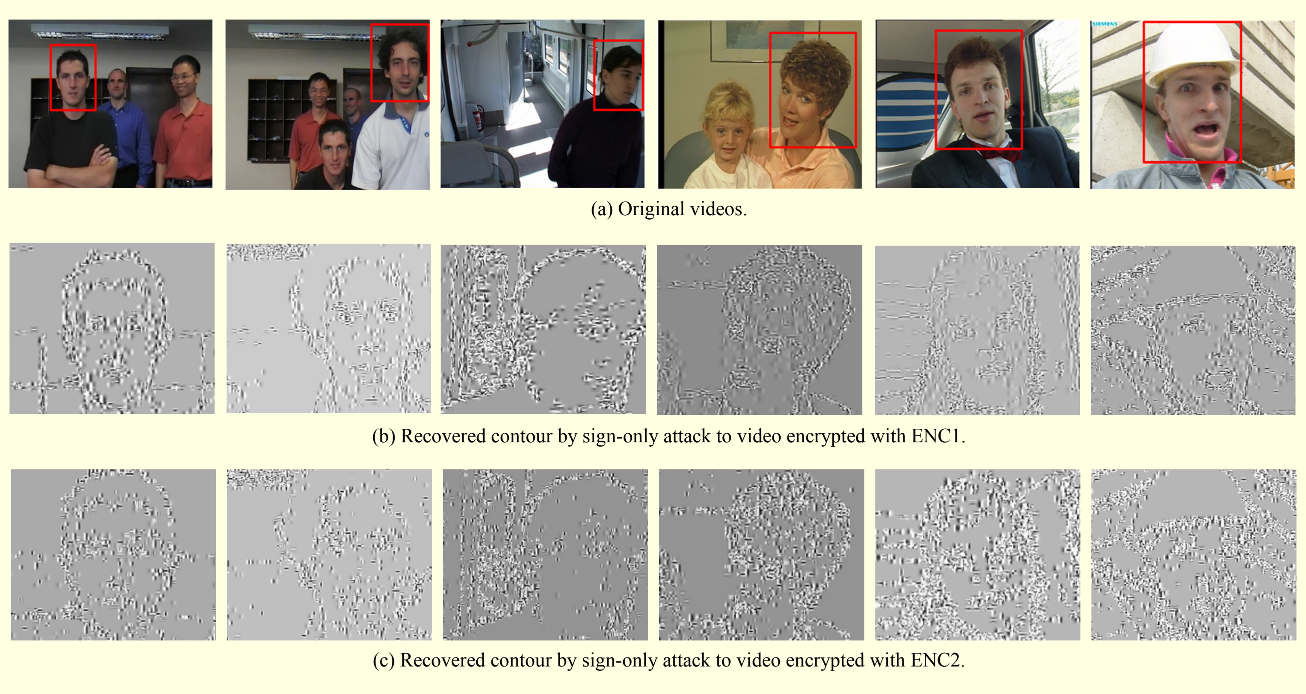
num_coeff	Multiplicities of elements						Sum
	1st	2nd	3rd	4th	5th	6th	
1	602	215	107	49	35	34	1,042
2	150	104	44	32	28	20	378
3	44	37	30	21	15	12	159
4	29	20	17	13	12	7	98

In the following implementation, the cost is at least doubled. The implementation consists mainly of two parts. First, random numbers are assigned to each element in the shuffling group, and the elements in the group are sorted according to an assigned random number. A shuffling table is then created based on the sorted sequence. In the second part, shuffling is performed according to the table. The computational cost for generating random numbers and swapping elements according to a shuffling table increases linearly with the number of elements. The run-time of comparison-based sorting algorithms is limited by a $O(n\log n)$ lower bound, so the increase in sorting complexity is greater than linear. Thus, it is proved that the proposed shuffling scheme is more robust to brute-force attacks than the one in [6].

B. Perceptual Security

We are going to show that visual degradation of recovered contour images by a sign-only attack gets more serious when advanced block shuffling is used as an encryption tool. Figure 7 shows visual examples for the effectiveness of the sign-only attack under different encryption settings. We encrypt test videos with and without advanced block shuffling, and then apply the sign-only attack to the encrypted video. Five video clips are used in our experiment motinas_multi_face_frontal (V1), sequence 14 'Faces' in [21] (V2), Mother-Daughter (V3), Carphone (V4), and Foreman (V5). We denote the encryption of an intra-prediction mode, T1, and level suffix by ENC1. The advanced block shuffling, coupled with the ENC1, is denoted by ENC2. Visual examination shows that encrypted video by ENC1 still leaks contour information after the sign-only attack as can be seen in Fig. 7(b). The ENC2 can scramble the content to a more unintelligible level as shown in Fig. 7(c).

We then use two methods, region shape descriptor (RSD) [22] and phase correlation [23], to quantitatively measure the visual difference between the contour image from the original



(a) Original videos.

(b) Recovered contour by sign-only attack to video encrypted with ENC1.

(c) Recovered contour by sign-only attack to video encrypted with ENC2.

Fig. 7. Visual examples for effectiveness of the sign-only attack. The videos are (from left to right): V1, V1, V2, V3, V4, and V5.

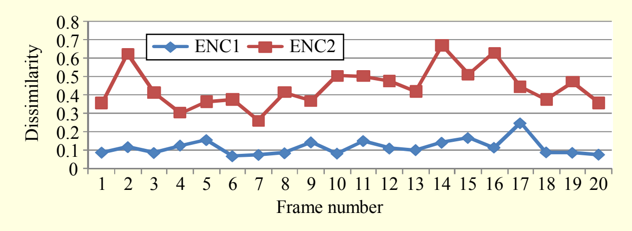


Fig. 8. Frame-by-frame RSD dissimilarity of recoverd contour image from V1 encrypted with ENC1 and ENC2.

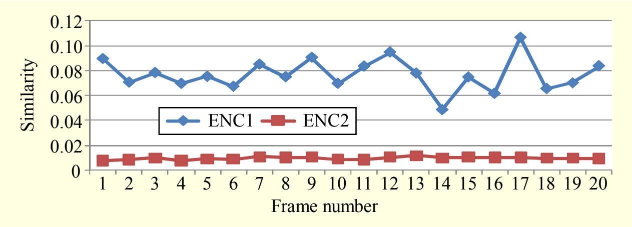


Fig. 9. Frame-by-frame phase correlation similarity of recoverd contour image from V1 encrypted with ENC1 and ENC2.

video data and the attacker's recovered copy from the encrypted video. The MPEG-7 standard [22] provides a rich set of standardized tools to describe multimedia content, and shape descriptors are one of the MPEG-7 Visual Descriptors.

In the case of '2D' shapes, there are two descriptors: region shape and contour shape. Since the RSD can describe complex objects consisting of multiple disconnected regions, RSD is better suited for describing contour image of complex scene. Phase correlation is a method to check the similarity of two images with equal size. Unlike many spatial-domain algorithms, it uses a frequency-domain approach, thus it is resilient to noise. The RSDs are extracted by using XM [24], and OpenCV [25] is used to perform phase correlation.

Once the advanced block shuffling is incorporated in the encryption, the contour images recovered by the sign-only attack become more dissimilar to the contour images from the original video as shown in Fig. 8 and Fig. 9. The RSD dissimilarity score ranges from 0 to 6.74, where 0 indicates a match between two images. The phase correlation similarity score ranges from 0 to 1, where 1 indicates a match. Perfect dissimilarity of two images, that is, RSD dissimilarity of 6.74 and phase correlation similarity of 0, might not happen in practice. From the experimentation of comparing different images, we found that the average RSD dissimilarity is 1.3631 and average phase correlation similarity is 0.0132. These two scores can be used to decide if two images are different or not.

Table 5 lists the average RSD dissimilarity and phase correlation similarity of videos after the sign-only attack. For the videos encrypted with ENC1, all the maximum values in phase correlation are found at (0, 0) position in an image. However, phase correlation detects translative movements in V1, V2, and V5 encrypted with ENC2. This means the recovered contour images from the videos encrypted with ENC1 are identical to the original contour images with some noise in it. When V2, V3, and V4 are encrypted with ENC2, the phase correlation similarity is similar to or lower than

Table 5. Visual difference between contour image from original video and attacker's recovered copy from encrypted video under two encryption settings (averaged over 50 frames).

Video	RSD dissimilarity		Phase corr. similarity	
	ENC1	ENC2	ENC1	ENC2
V1	0.2274	0.4141	0.151	0.0459
V2	0.1954	0.3694	0.0925	0.0146
V3	0.1887	0.5813	0.0864	0.0145
V4	0.1176	0.4347	0.0866	0.0098
V5	0.1318	0.3017	0.2244	0.0725

0.0132, thus it could be said that ENC2 made those videos look like another video when judged by phase correlation.

Although it is hard to set thresholds to determine if an encrypted video passes the similarity tests against the sign-only attack, the average scores in Table 5 shows that, by incorporating the advanced block shuffling, the resulting similarities are consistently lower. Therefore, it can be concluded that the advanced block shuffling scheme, coupled with previous encryption schemes, can make the contour image more chaotic.

However, if the video is of high resolution and the face area occupies a large portion of the frame, the recovered contour from the video encrypted with ENC2 may leak more information than visible in Fig. 7(c). To make the contour image created from encrypted video totally chaotic, the applied encryption must be able to add random noise in the blocks where there is no non-zero coefficient after high-frequency filtering. However, this is not considered to be a goal of the proposed shuffling scheme because it is not possible to do that without adversely impacting the compression friendliness which is the main objective of the H.264 standard. The tradeoff between the level of privacy protection and compression efficiency must be evaluated before choosing a proper encryption algorithm.

V. Conclusion

In this paper, we demonstrated the feasibility of obtaining important information, such as a face outline for personal identification, from an encrypted video. By creating an attack that can extract contour information from an encrypted video, we have provided an analysis method for the security of well-known selective encryption schemes for H.264, and have demonstrated that the security of these schemes is lower than expected in terms of privacy protection. This rather striking result reveals that, beyond an exact recovery, it is also important to ensure that partial perceptually intelligible information is not leaked from an encrypted video. We specially crafted our attack for an H.264 encoded video by taking into account the unique features of the H.264 codec. We have also pointed out the need for enhancing security against our attack and have proposed an advanced block shuffling algorithm for this purpose. Our experiments have shown that the new method, coupled with another encryption scheme, can render the face outline more unintelligible, and thus the privacy of individuals monitored through a video surveillance system can be enhanced. As future work, we plan on devising more sophisticated metrics for measuring the level of privacy protection gained by video encryption.

References

[1] D.N. Serpanos and A. Papalambrou, "Security and Privacy in Distributed Smart Cameras," *Proc. IEEE*, vol. 96, no. 10, Oct. 2008, pp. 1678-1687.

[2] W.H. Widen, "Smart Cameras and the Right to Privacy," *Proc. IEEE,* vol. 96, no. 10, Oct. 2008, pp. 1688-1697.

[3] J. Wen et al., "A Format-Compliant Configurable Encryption Framework for Access Control of Video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, June 2002, pp. 545-557.

[4] S. Lian et al., "Secure Advanced Video Coding Based on Selective Encryption Algorithms," *IEEE Trans. Consum. Electron.*, vol. 52, no. 2, May 2006, pp. 621-629.

[5] S. Lian et al., "Efficient Video Encryption Scheme Based on Advanced Video Coding," *Multimedia Tools Appl.*, vol. 38, May 2008, pp. 75-89.

[6] S.B. Liu et al., "A Novel Format-Compliant Video Encryption Scheme for H.264/AVC Stream Based on Residual Block Scrambling," *Int. Conf. Intell. Comput.*, vol. 5226, 2008, pp. 1087-1094.

[7] C. Bergeron and C. Lamy-Bergot, "Compliant Selective Encryption for H.264/AVC Video Streams," *Int. Workshop Multimedia Process.*, Oct.-Nov. 2005, pp. 477-480.

[8] L.F. Wang et al., "Perceptual Video Encryption Scheme for Mobile Application Based on H.264," *J. China Universities Posts Telecommun.*, vol. 15, Sept. 2008, pp. 73-78.

[9] Y. Zou et al., "H.264 Video Encryption Scheme Adaptive to DRM," *IEEE Trans. Consum. Electron.*, vol. 52, no. 4, Nov. 2006, pp. 1289-1297.

[10] Y. Mao and M. Wu, "A Joint Signal Processing and Cryptographic Approach to Multimedia Encryption," *IEEE Trans. Image Process.*, vol. 15, no. 7, July 2006, pp. 2061-2075.

[11] M. Abomhara, O. Zakaria, and O.O. Khalifa, "An Overview of Video Encryption Techniques," *Int. J. Comput. Theory Eng.*, vol. 2, no. 1, Feb. 2010, pp. 103-110.

[12] A. Massoudi et al., "Overview on Selective Encryption of Image

and Video: Challenges and Perspectives," *EURASIP J. Info. Security*, , vol. 2008, Article ID 179290, 2008, pp. 1-18.

[13] D. Xie and C.-C. Jay Kuo, "Multimedia Encryption with Joint Randomized Entropy Coding and Rotation in Partitioned Bitstream," *EURASIP J. Info. Sec.*, vol. 2007, Article ID 35262, Jan. 2007, pp. 1-18.

[14] S. Li et al., "On the Design of Perceptual MPEG-video Encryption Algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 2, Feb. 2007, pp. 214-223.

[15] F. Dufaux and T. Ebrahimi, "Scrambling for Privacy Protection in Video Surveillance Systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, Oct. 2008, pp. 1168-1174.

[16] H. Kondo et al., "Binary Signal Compression Using DCT Signs," *Int. Conf. Autonomous Robots Agents*, Dec. 2004, pp. 240-243.

[17] "Advanced Video Coding for Generic Audiovisual Services," ITU-T Rec. H.264/ISO/IEC 14496-10, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVTG050, Nov. 2007.

[18] D.H. Yeo and H.C. Shin, "High Throughput Parallel Decoding Method for H.264/AVC CAVLC," *ETRI J.*, vol. 31, no. 5, Oct. 2009, pp. 510-517.

[19] C. Paar and J. Pelzl, *Understanding Cryptography: A Textbook for Students and Practitioners*, Springer, 2010.

[20] E. Maggio et al., "Particle PHD Filter for Multi-target Visual Tracking," *IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Apr. 2007, pp. 1101-1104.

[21] BOSS (On Board Wireless Secured Video Surveillance) project, Oct. 2006 to June 2009.

[22] B.S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, Wiley, July 2002.

[23] E. De Castro and C. Morandi, "Registration of Translated and Rotated Images Using Finite Fourier Transforms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, Sept. 1987, pp. 700-703.

[24] ISO/IEC 15938-6 Information Technology - Multimedia Content Description Interface Part 6: Reference Software, 2003.

[25] G.R. Bradski and V. Pisarevsky, "Intel's Computer Vision Library: Applications in Calibration, Stereo Segmentation, Tracking, Gesture, Vace and Object Recognition," *Int. Conf. Comput. Vision Pattern Recognition*, vol. 2, 2000, pp. 796-797.

SuGil Choi is a senior member of the engineering staff of the Knowledge-Based Information Security & Safety Research Division at ETRI, Daejeon, Rep. of Korea. He received his BS in industrial engineering from Korea University in 2000, and MS degree in information engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2004. His major interests are in multimedia security, trusted computing, and machine learning.

Jong-Wook Han is a team leader of the Convergence Service Security Research Team of the Knowledge-Based Information Security & Safety Research Division at ETRI, Daejeon, Rep. of Korea. He received the BS, MS, and PhD from Kwangwoon University in 1989, 1991, and 2001, respectively. His major interests are in multimedia security, home network security, and surveillance systems.

Hyunsook Cho is the managing director of the Knowledge-Based Information Security & Safety Research Division at ETRI. She received her PhD from Chungbuk National University, Rep. of Korea, in 2001. Her major interests are broadly in the area of applied cryptography, network security, multimedia security, and digital forensics.