SECURITY EVALUATION FOR COMMUNICATION-FRIENDLY ENCRYPTION OF MULTIMEDIA

Yinian Mao and Min Wu

ECE Department, University of Maryland, College Park

ABSTRACT

This paper addresses the access control issues unique to multimedia, by using a joint signal processing and cryptographic approach to multimedia encryption. Based on three atomic encryption primitives, we present a systematic study on how to strategically integrate different atomic operations to build a video encryption system. We also propose a set of multimedia-specific security metrics to quantify the security against approximation attacks and to complement the existing notion of generic data security. The resulting system can provide superior performance to both generic encryption and its simple adaptation to video in terms of a joint consideration of security, bitrate overhead, and communication friendliness.

1. INTRODUCTION

The development in digital multimedia and communication technologies has paved ways for people around the world to acquire, utilize, and share multimedia content. For the wide availability of multimedia information and successful commercialization of many related services, assuring that the multimedia information is used only by authorized users for authorized purposes has become essential. This paper discusses the confidentiality protection and access control for multimedia information, with an emphasis on system integration and security evaluation.

Content confidentiality and access control is generally addressed by encryption. In principle, digital multimedia can be encoded into a bitstream and encrypted in the same way as generic data [1]. By doing so, however, the encryption will wipe out the inherent structures and the syntax of multimedia data. Many functionalities provided by state-of-the-art signal processing for multimedia will disappear after encryption. A few of such examples are scalable coding, unequal error protection, and compressed domain search and indexing [2, 3]. To address this issue, a number of schemes have been proposed to take signal processing into consideration when encrypting multimedia [3, 4]. Among these schemes, transform/codeword domain shuffling and codeword domain index encryption are a few common choices.

Given the prevalence of multimedia content and the issues related to multimedia encryption as mentioned above, there is a need to study the encryption of multimedia in a systematic way. Our goal is to design encryption systems for multimedia that are friendly to communication and signal processing techniques, reduce the cost of such systems, and achieve an appropriate level of security. In [5], we proposed two atomic encryption operations for multimedia and provided analytical results on the bitrate overhead of the encrypted data. In this paper, we first introduce a notion of multimedia-specific security and two quantitative security metrics in Section 2. We then show in Section 3 how to integrate different encryption operations to build a video encryption system and discuss the tradeoff among security, compressibility, and compatibility to intermediate processing during transmission. Finally, conclusions are drawn in Section 4.

This work was supported in part by the Army Research Office under Award No. DAAD19-01-1-0494 and the National Science Foundation under Award No. CCR-0133704. The authors can be contacted at {ymao, minwu}@eng.umd.edu

2. SECURITY METRICS

In this section, we discuss why the security of multimedia encryption needs extra attention, and how to evaluate the security of multimedia encryption beyond the bit-by-bit representation.

2.1. Exact Knowledge Versus Approximation

In generic data encryption, a number of metrics can serve to measure the security against attackers' recovery of the exact plaintext from the ciphertext. A simple method is to count the number of brute-force trials an attacker needs to perform (on average) for the exact recovery. Since an attacker can either guess the clear-text or the decryption key, the number of brute-force trials is proportional to $\min\{|\text{clear-text space}|, |\text{key space}|\}$, where $|\cdot|$ denotes cardinality. An equivalent way to measure the security is to use binary bit as unit. The number of bits is obtained by taking the log of the size of clear-text space or key space. A more sophisticated notion of practically provable security was introduced in [6]. This notion quantifies the security strength of a system in terms of the amount of resources needed to break the system.

Due to the spatial and temporal correlation of multimedia, the encrypted content may be approximately recovered based on the syntax, context, and the statistical information known as *a priori*. This is possible even when the encrypted part is provably secure according to the generic security notion. For example, in MPEG video encryption, when motion vector fields are encrypted and cannot be accurately recovered, a default value 0 can be assigned to all motion vector fields [3]. This approach results in a fairly good approximation for slow-motion frames. Additionally, the statistical information, neighborhood patterns, and smoothness criterion can help estimate an unknown area in an image and automatically reorder shuffled image blocks [8]. Although the estimated signal may not be exact, it can still be perceptually meaningful and reveal a substantial amount of information when visualized or rendered. It is therefore important to introduce a notion of multimedia-specific security. We use visual data to illustrate the concept of visual security in this paper. Under the notion of visual security, the effectiveness of encryption should be measured against visual approximation attacks. For this purpose, we propose two visual security metrics below.

2.2. Visual Security Metrics

Studies on human visual system suggest that two important types of information are extracted by an observer of a given image [7]. The first type is the color space or luminance information. The

0-7803-8554-3/04/$20.00 ©2004 IEEE.

second type is the edge and contour information, which describes the shape of the objects. Based on this observation, we introduce a color similarity score and an edge similarity score to quantitatively measure the distance between two images. For gray scale images, their color similarity becomes luminance similarity, which will be detailed below.

Luminance Similarity Score (LSS) To capture the luminance similarity, we introduce a block-based luminance similarity score. Two images of the same size are first divided into blocks. If the two images are not of the same size, they are resized and aligned by preprocessing modules. Then the average luminance values of the i-th block from both images, y_{1i} and y_{2i}, are calculated. We define the luminance similarity score as

$$LSS = \frac{1}{N}\sum_{i=1}^{N} f(y_{1i}, y_{2i}). \quad (1)$$

Here, the function $f(x_1, x_2)$ for each pair of average luminance values is defined as

$$f(x_1, x_2) = \begin{cases} 1 & \text{if } |x_1 - x_2| < \frac{\beta}{2}, \\ -\alpha \text{ round}(\frac{|x_1 - x_2|}{\beta}) & \text{otherwise,} \end{cases}$$

where the parameters α and β control the sensitivity of the score. Along with block-based aggregation, the α factor within the range from 0 to 1 and the quantization parameter β provide resistance to minor perturbation and noise. In our experiments, α and β are set to 0.1 and 3, respectively. A negative LSS value indicates substantial dissimilarity in luminance between two images.

Edge Similarity Score (ESS) The edge similarity score measures the degree of resemblance of the edge and contour information between two images. After the original images are partitioned into blocks, edge detection is performed for each block. The dominant edge direction in each block is extracted and quantized into one of the eight representative directions. The representative edge directions are equally spaced by 22.5 degrees in a polar coordinate system. We use indices 1 to 8 to represent these eight directions, and use index 0 to represent a non-edge block. Let e_{1i} and e_{2i} be the edge direction indices for the i-th block in two images, respectively, the edge similarity score (ESS) for a total of N image blocks is computed as:

$$ESS = \frac{\sum_{i=1}^{N} w(e_{1i}, e_{2i})}{\sum_{i=1}^{N} c(e_{1i}, e_{2i})}. \quad (2)$$

Here, $w(e_1, e_2)$ is a weighting function defined as

$$w(e_1, e_2) = \begin{cases} 0 & \text{if } e_1 = 0 \text{ or } e_2 = 0, \\ |cos(\phi(e_1) - \phi(e_2))| & \text{otherwise,} \end{cases}$$

where $\phi(e)$ is the representative edge angle for an index e, and $c(e_1, e_2)$ an indicator function that takes value 0 when both e_1 and e_2 are non-edge blocks ($e_1 = e_2 = 0$) and value 1 otherwise. The score ranges from 0 to 1, where 0 indicates that the edge information of the two images is highly distinct and 1 indicates a match between the edges in the two images. A special case arises when both images are very smooth, leading the denominator in (2) to 0. Although this is a match case, we assign an ESS score of 0.5 to it, because there is not much edge information extracted from either image. In our experiments, we partition the input images into non-overlapping 8x8 blocks and use the Sobel operator for edge detection.

3. SYSTEM STUDY ON VIDEO ENCRYPTION

3.1. Encryption Primitives

Atomic encryption primitives are basic building blocks for encrypting multimedia. We include three encryption primitives in our test system, i.e., generalized index mapping, coded block shuffling, and intra bitplane shuffling. All these encryption operations preserve the syntax prescribed by multimedia coding standards. As a result, many of the communication and signal processing techniques designed for unencrypted multimedia can also be applied to the encrypted data.

The *Generalized Index Mapping* [3, 5] is an encryption primitive that can be applied directly to symbols taking values from a finite set. Examples may include working with quantized coefficients, quantized prediction residues, and run-length coding symbols. The encryption/decryption process employs a bijective mapping between clear-text symbols and their binary indices. A core encryption primitive, such as AES or RSA, is applied on the symbol indices. In [5] and [9], we have developed analytical results on the bitrate overhead brought by index encryption, and have shown that by partitioning the input symbol range S into multiple subsets and restricting the encryption output to be in the same subset as the input symbol, the bit-rate overhead can be reduced at the expense of a reduced complexity for brute force attack.

Multimedia coding systems often partition input signals into segments and encode each segment into a self-contained unit. Shuffling the order of such units according to a cryptographic permutation table has the advantage of preserving the compressibility as well as the syntax of the coded bitstream [1, 3]. We refer to such operations as *coded block shuffling*. A major drawback for block shuffling is that an attacker can exploit the correlation across the blocks, such as the continuity of edges and similarity of colors and textures, and reassemble the shuffled blocks with a much smaller effort than that of a brute force search [8]. Therefore, block shuffling alone is often not a secure encryption operation. However, as a complementary building block, it can help achieve good visual/auditory scrambling effect for multimedia data.

Fine granularity scalability (FGS) is desired in multimedia communications to provide a near-continuous tradeoff between bitrate and quality. The *Intra Bitplane Shuffling* [5] is an encryption primitive compatible with bit-plane coding, such as the recently adopted MPEG-4 FGS. To encrypt the FGS layer video, random shuffling is applied in the transform coefficient domain on each bitplane of n bits. Using such an encryption approach, the scalable coded video can be protected without the loss of scalability in the encrypted bitstream, while maintaining a low bitrate overhead.

3.2. System Setup

Four video clips are used in our experiment, namely, *Football*, *Coastguard*, *Foreman*, and *Grandma*. Each video clip is 40 frames long and coded with MPEG-4 standard. The group of picture (GOP) size is set to 15 and all predictive frames are P frames. We identify three possible components in the base-layer video to which we apply the generalized index mapping. These components are: (1) the DC prediction residues of intra-blocks, (2) the motion vector (MV) residues of inter-blocks, and (3) a part of non-zero AC coefficients of intra-blocks. In addition, we also incorporate the random shuffling of macroblocks from both intra and predictively coded pictures.

We consider six encryption setting E1-E6 and three approximation attack settings A1-A3 in our experiments. Encryption settings E1-E3 are listed below, where the encryption of DC, AC,

and/or MV is based on the generalized index mapping with set partitioning. The DC and AC encryption ranges are chosen as [-63,64] and [-32,32], respectively [5].

(E1) encrypting intra block DC residues by index mapping;

(E2) encrypting inter block MV residues in the first two PVOP's in a GOP, and all intra block DC residues;

(E3) encrypting all the components listed in E2, plus the first two non-zero AC coefficients of intra blocks.

Setting E4-E6 correspond to E1-E3 plus macro-block shuffling in the compressed bit-stream, respectively. Corresponding to these encryption settings, the settings for approximation attacks (A1-A3) that emulate an adversary's action are:

(A1) set all intra block DC coefficients to 0;

(A2) set all intra block DC coefficients to 0 and set the encrypted motion vector values to 0;

(A3) including all the approximations in A2, plus set those encrypted AC coefficients to 0.

Fig. 1. Encryption results for *Coast-guard*. The encryption-approximation settings are: [top row, left to right] unencrypted, E1+A1 (PSNR 15.4 dB); [bottom row, left to right] E2+A2 (PSNR 16.2 dB), E4+A1 (PSNR 14.1 dB).

3.3. Encrypting Base-layer Video

First, we consider the complexity of exact recovery by brute-force trial. To exactly recover an encrypted I frame, an attacker needs to recover all the DC coefficients. For P frames, the values of motion vectors are also necessary. In the above configuration, each DC coefficient or motion vector component has 5 bits encrypted. From the discussion in Section 2, each I frame has 1980 equivalent DC bits encrypted and the encrypted motion vectors in a P frame is equivalent to 990 bits. Since a 128-bit AES key is used, the security against exact recovery by exhaustive search is determined by the key length (128 bits).

Next, we evaluate the perceptual security of different encryption configurations against approximation recovery attacks using the proposed perceptual similarity metrics. Table 1 lists the average ESS and LSS scores of three videos after approximation recovery. From the average LSS scores, we can see that the luminance information is well protected after DCs are encrypted and the score remains at a similarly low level as more video components are encrypted. However, from the average ESS scores we

Table 1. Perception based security measures for video encryption

	Football		*Grandma*		*Foreman*	
Settings	ESS	LSS	ESS	LSS	ESS	LSS
E1+A1	0.70	-0.78	0.64	-2.13	0.71	-1.42
E2+A2	0.53	-0.85	0.46	-2.13	0.43	-1.48
E3+A3	0.53	-0.86	0.30	-2.13	0.40	-1.48
E4+A1	0.12	-0.93	0.05	-2.13	0.07	-1.47
E5+A2	0.13	-0.92	0.05	-2.13	0.06	-1.45
E6+A3	0.12	-0.92	0.04	-2.13	0.05	-1.47

can also see that edge and contour information needs more protection than luminance information and block shuffling is an effective tool. Compared with the average PSNR (listed in the caption of Fig. 1), which do not change much with different encryption settings, the ESS and LSS metrics can better represent the amount of perceptually perceived information.

To examine the detailed ESS scores, we plot the frame-by-frame ESS score of *Coast-guard* under different encryption-attack settings in Fig. 2. The top curve is from the attacked video with DC encrypted only, which confirms that encrypting DC alone still leaves some contour information unprotected. The two middle curves are the results involving MV encryption for inter blocks and AC encryption for intra blocks, where the ESS scores are low at the beginning of a GOP and increase substantially toward the end of the GOP. This is because as it approaches the end of a GOP, motion compensation becomes less effective and the compensation residue provides a significant amount of edge information. The information leakage by encrypting DC and/or MV only has also been seen in Fig. 1. On the other hand, by incorporating the shuffling of macroblock coding units, the resulting ESS measurements are consistently around 0.1 or lower.

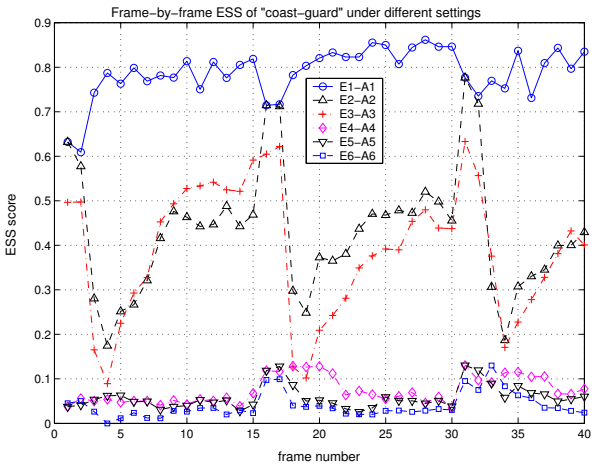

Fig. 2. Frame-by-frame ESS of *coastguard*

Through experiments, we have found that when an approximated image from attack has ESS lower than 0.5 and LSS lower than 0, the image is still perceptually scrambled to the effect that very little visually meaningful information can be revealed. Such encrypted images are sufficiently secure for most of applications.

The relative overhead of the encrypted videos are listed in Table 2. In general, fast motion videos will have a smaller relative overhead. Since shuffling coding units does not introduce bitrate overhead, the overhead of settings E4-E6 are identical to those of settings E1-E3, respectively. We found that the E5 setting, which

Table 2. Relative Compression Overhead of the Encrypted Videos

	Football	*Foreman*	*Coastguard*	*Grandma*
E1	1.29%	1.75%	3.15%	6.96%
E2	3.88%	6.41%	8.74%	11.11%
E3	6.47%	9.62%	11.54%	24.61%

is a combination of block shuffling and selective value encryption, provides a good tradeoff between security and overhead for many applications. More details can be found in reference [9].

3.4. Protecting FGS Enhancement Layer Video

We use 10 frames from the *Foreman* to demonstrate the protection of the enhancement layer while preserving the FGS characteristics. The proposed intra bitplane shuffling is applied within each 8x8 block. We also encrypt the sign bit of each coefficient using a stream cipher. Two encryption settings are used: (a) to shuffle the 1st FGS bitplane, and (b) to shuffle the first two bitplanes. To focus on the protection of the enhancement data, the encrypted FGS bitplanes are combined with a cleartext base layer video to show the visual effects of encryption.

Fig. 3. Encryption results for *Foreman* FGS video. Top row, left to right: base layer plus 1 and 2 unencrypted FGS bitplanes; Bottom row, left to right: encryption results from settings (a) and (b).

We consider the complexity of exactly recovering only the second MSB bit-plane from the encrypted FGS video. Suppose in a bit-plane with QCIF size each shuffled block has 7 bit of "1"s, which is the average number in the 2nd MSB of the "Foremen" sequence. The number of brute-force trials an attacker has to perform for the exact recovery is proportional to $\binom{64}{7}^{396}$, which is approximately equivalent to 10^5 bits information encrypted. So the actual security is again determined by the key length, which is 128 bits.

Fig. 3 shows the unencrypted and the encrypted versions of the *Foreman* FGS video, and Table 3 lists the corresponding average PSNR, LSS and ESS. From these results we can see that, without encryption, the ESS, LSS, and PSNR increase with the addition of more bit-planes. With encryption, the edge and luminance similarity remains imperfect. This can be explained by viewing the encrypted FGS bitplanes as random noise added to the base-layer video. Since the ESS score is designed to be resilient to noise, the added noise does not affect the ESS score substantially. However, the LSS score in Table 3 captures the luminance degradation under encryption settings (a) and (b), as can be seen in Fig. 3. Overall, the results indicate that the video quality after encryption is almost the same as that of the base-layer video and much lower than the cleartext base-plus-enhancement video. Thus the premium quality version of the content can be encrypted in a FGS compatible way and discretionarily protected.

Table 3. Intra Bitplane Shuffling

	Base	+1BP	+2BP	(a)	(b)
PSNR	28.8	29.0	33.4	28.59	27.39
ESS	0.85	0.85	0.92	0.85	0.85
LSS	0.28	0.38	0.79	0.28	0.28

4. CONCLUSIONS

In this paper, we have addressed the importance and feasibility of incorporating signal processing into multimedia encryption. Regarding the security metrics, we pointed out the need of quantifying the security against approximation attacks that are unique to multimedia, and proposed a set of multimedia-specific security metrics to complement those for generic data. Using video as an example, we presented a systematic study on how to integrate different atomic operations together to build a video encryption system. Our experiment shows that by strategically integrating selective value encryption, intra-bitplane shuffling, and spatial permutation, the resulting scheme can achieve a good tradeoff among security, bitrate overhead, and compatibility to signal processing.

5. REFERENCES

[1] L. Qiao and K. Nahrstedt: "Comparison of MPEG Encryption Algorithms", *Inter. Journal on Computers & Graphics*, vol. 22, no. 3, 1998.

[2] Y. Wang, S. Wenger, J. Wen, and A. Katasggelos: "Error Resilient Video Coding Techniques", *IEEE Signal Processing Magazine*, vol.14, no.4, pp61-82, July, 2000.

[3] J. Wen, M. Severa, W. Zeng, M.H. Luttrell and W. Jin: "A Format-Compliant Configurable Encryption Framework for Access Control of Video", *IEEE Trans. on CSVT*, vol.12, no.6, pp545-557, June 2002.

[4] T-L. Wu and S.F. Wu: "Selective Encryption and Watermarking of MPEG Video", *Inter. Conf. on Image Science, Systems, and Technology(CISST'97)*, Las Vegas, NV, 1997.

[5] M. Wu and Y. Mao: "Communication-Friendly Encryption of Multimedia", *Proc. of IEEE MMSP'02*, Dec. 2002.

[6] M. Bellare: "Practice-Oriented Provable Security", *Proc. of First Inter. Workshop on Information Security(ISW'97)*.

[7] Z. Wang, L. Lu and A.C. Bovik: "Video Quality Assessment Based on Structural Distortion Measurement", *Signal Processing: Image Communications*, vol.19, no.1, Jan., 2004.

[8] A. Pal, K. Shanmugasundaram and N. Memon : "Automated reassembly of fragmented images", *Proc. of Inter. Conference on Multimedia and Expo*, Baltimore, Jul. 2003.

[9] Y. Mao and M. Wu: "A Joint Signal Processing and Cryptographic Approach to Multimedia Encryption", submitted for journal publication, June 2004.