An Automatic Human Video Objects Encryption Scheme Built on Stream and Block Ciphers and Based on Chaos

KLIMIS S. NTALIANIS
Electrical and Computer Engineering Department
National Technical University of Athens
9, Iroon Polytechniou str., Zografou 15773, Athens
GREECE
http://www.image.ntua.gr

Abstract: - The increasing popularity of multimedia applications creates an increasing need for secure storage and transmission techniques. Especially wireless communications (satellite, mobile, RF), which can be easily intercepted, should be protected by unauthorized persons. Towards this direction several encryption schemes have been proposed in literature, however most of them do not consider regions of interest (video objects - VOs). These regions may need better protection, may be the only regions that need protection, or should be accessed separately according to different privileges. For these reasons in this paper a human video objects encryption system is proposed based on chaos. Initially videoconference/videophone sequences are analyzed and human video objects are automatically extracted using a face and body detection method. Next the pixels of each human video object are properly arranged and the chaotic cipher module is activated to encrypt them. Finally the background area is also encrypted using a different key. The system presents robustness against known cryptanalytic attacks and enables efficient rate control as semantic information is separated from the background.

Key-Words: - chaos, face and body detection, human video object encryption, efficient rate control, stream cipher, block cipher.

1. INTRODUCTION

Multimedia traffic over communications networks is rapidly increasing and it is expected to continue increasing the following decades. Often content of high confidence and crucial importance is transmitted during special applications such as exchange of medical files, military communications, confidential videoconferencing, pay-TV, video surveillance and e-banking. This particularly sensitive content induces a growing need for security. When research community refers to security it covers three areas: content integrity, access protection and intellectual property.

The proposed system can be classified into the second category, where several attempts have been made in literature towards multimedia access protection. In early image scrambling schemes, methods such as line reversal, line dispersal and line segment swapping are proposed. In [1] a commercial system is implemented based on line permutation. These simple techniques are not robust to "correlation attacks", where correlation properties of typical images are employed for unauthorized decryption [2]. In [3] reordering of pixels is performed by changing the scan order according to a space-filling curve. Another possibility is to scramble the image in a transformed domain [4].

On the other hand, lately, some chaos-based cryptographic systems have also been proposed, as chaos presents many desired cryptographic qualities. In particular in [5], a chaotic key-based algorithm (CKBA) is proposed, functioning as value substitution cipher. An encryption algorithm that uses the iterations of the chaotic tent map is proposed in [6]. In [7], each character of the messages is encrypted as the number of iterations (at least 250) performed by the logistic equation. However multiple iterations of chaotic systems usually lead to slow ciphers. Another algorithm based on synchronized chaotic systems is proposed in [8], where each byte of a message is encrypted using a different chaotic map.

However, most of the proposed schemes do not consider video objects or any other kind of semantic information, just bits and pixels. Consequently, issues such as object – based secure distribution of multimedia content, layered access to regions of interest, and efficient

transmission / rate control over low bandwidth networks should be considered.

To face the previous issues, in this paper an automatic human video objects cryptographic scheme is proposed. The proposed system includes a human video objects detection module and an iterative cipher mechanism based on a triplet of chaotic maps. Initially human video objects are extracted according to a face and body detection module, which is based on Gaussian probability density functions. Afterwards the pixels of each human video object are properly arranged in sequential order, by stacking all rows of the video object. Next the stacked pixels are iteratively encrypted using a 256-bit key, which is generated by a chaotic pseudo-random bit generator. Afterwards the same procedure is followed for the background area, using this time a different 256-bit key. Finally the encrypted pixels are rearranged to produce the final image with encrypted video objects. The system is suitable for layered object-based access to videoconference/videophone sequences, while it can enable efficient object-based rate control.

The paper is organized as follows: in Section 2 the human video objects detection module is described. In Section 3 the proposed chaotic encryption scheme is presented while experimental results on real life videoconference sequences are shown is Section 4. Finally Section 5 concludes this paper.

2. HUMAN VIDEO OBJECTS EXTRACTION

In the current section a human video objects extraction technique is described for videoconference/videophone sequences, which incorporates a neural network classifier. The technique consists of two sub-modules: the human face and human body detection sub-modules.

2.1. Human Face Detection

Human face detection is an important task in numerous applications. For efficient detection, in the proposed approach, the two-chrominance components of a color image are used, as the distribution of the chrominance values corresponding to a human face, occupies a very small region of the color space [9]. Consequently, only blocks whose respective chrominance values are located at this small region can be considered as candidate face blocks.

Towards this direction, in the proposed approach, the histogram of chrominance values corresponding to the face class, say Ω_f, is initially modeled by a Gaussian probability density function (pdf) as:

$$P(\mathbf{x}|\Omega_f)=\frac{\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_f)^T\cdot\boldsymbol{\Sigma}_f^{-1}\cdot(\mathbf{x}-\boldsymbol{\mu}_f))}{2\pi\cdot|\boldsymbol{\Sigma}|^{1/2}} \quad (1)$$

where $\mathbf{x}=[u\ v]^T$ is a 2x1 vector containing the mean chrominance components u and v of an examined block, $\boldsymbol{\mu}_f$ is the 2x1 mean vector of a face area and $\boldsymbol{\Sigma}$ is the 2x2 variance matrix of the probability density function:

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_u^2 & \sigma_{u,v}\\ \sigma_{u,v} & \sigma_v^2\end{bmatrix} \quad (2)$$

where σ_u^2 is the variance of the chrominance component u, σ_v^2 is the variance of the chrominance component v and $\sigma_{u,v}$ corresponds to the covariance between u and v. Parameters $\boldsymbol{\mu}_f$ and $\boldsymbol{\Sigma}$ are estimated based on a set of several face images and using the maximum likelihood algorithm [10].

Afterwards each block B_i of the image belongs to the face area, if the respective probability of its chrominance values, $P(\mathbf{x}(B_i) \mid \Omega_f)$ is high, where the two chrominance components are extracted from block B_i, i.e., $u(B_i)$ and $v(B_i)$ and thus vector $\mathbf{x}(B_i)=[u(B_i)\ v(B_i)]^T$. In our case, in order to get more reliable results, we use only a sub-region of the Gaussian pdf. In particular a confidence interval of 80% is selected from the Gaussian model, so that only blocks inside this region are considered as face blocks. Therefore, for each test image of size N_1xN_2 a binary mask M is formed, with size $N_1/8\times N_2/8$ pixels (as block resolution is initially assumed); an 8x8 block with value equal to one indicates a possible face block, while a zero value indicates a non-face block.

However, as the aforementioned procedure takes into consideration only color information, the final binary mask M may also contain non-face blocks, which present similar chrominance properties. To confront this problem, shape information of human faces is also considered. In our case the method described in [9] is adopted, where rectangles with certain aspect ratios are used for shape approximation of face areas. In particular an aspect ratio for face areas is defined as:

$$R=H_f/W_f \quad (3)$$

where H_f is the height of the head, while W_f corresponds to the face width. After several experiments R was found to lie within the interval [1.4 1.6]. Finally a binary mask, say M_f, of size $N_1/8\times N_2/8$ is formed, in which pixels with value equal to one correspond to the initial face area, while zero values are not considered to belong to the initial face region.

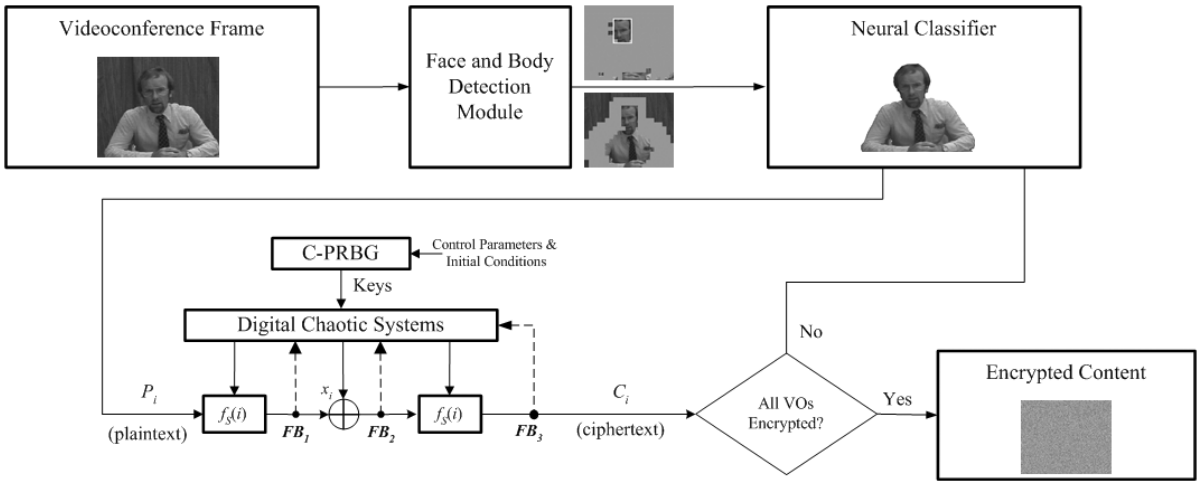

Figure 1: The proposed human video objects chaotic encryption scheme.

2.2. Human Body Detection

In this subsection, human body detection is performed by exploiting information derived from the human face detection task. In particular, initially the center, width and height of the face region, denoted as $\mathbf{c}_f=[c_x\ c_y]^T$, w_f and h_f respectively are calculated. Human body is then localized by incorporating a probabilistic model, the parameters of which are estimated according to $\mathbf{c}_f$, w_f and h_f.

In particular let us denote by $\mathbf{r}(B_i)=[r_x(B_i)\ r_y(B_i)]^T$ the distance between the *ith* block, B_i, and the origin, with $r_x(B_i)$ and $r_y(B_i)$ the respective x and y coordinates. In this paper the product of two independent 1-dimensional Gaussian pdfs is used to model the human body. Then for each block B_i of an image, a probability $P(\mathbf{r}(B_i) \mid \Omega_b)$ is assigned, expressing the degree of block B_i belonging to the human body class, say Ω_b.

$$P(\mathbf{r}(B_i)\mid\Omega_b)=\frac{\exp(-\frac{1}{2\sigma_x^2}(r_x(B_i)-\mu_x)^2)\exp(-\frac{1}{2\sigma_y}(r_y(B_i)-\mu_y)^2)}{(2\pi)\sigma_x\sigma_y} \quad (4)$$

where μ_x, μ_y, σ_x and σ_y are the parameters of the human body location model; these parameters are calculated based on the information derived from the face detection task, taking into account the relationship between human face and body. In our simulations, the parameters in (4) are estimated with respect to the face region as follows

$$\mu_x=c_x,\ \mu_y=c_y+h_f \quad (5a)$$
$$\sigma_x=w_f,\ \sigma_y=h_f/2 \quad (5b)$$

Similarly to human face detection, a block B_i belongs to body class Ω_b, if the respective probability, $P(\mathbf{r}(B_i) \mid \Omega_b)$ is high. Again a confidence interval of 80% is selected from the Gaussian model so that blocks belonging to a body region are reliably detected. Then, a binary mask, say M_b, of size $N_1/8\times N_2/8$, is formed and its pixels with value one correspond to the initial human body estimate, while pixels of value zero are not considered to belong to the initial body region.

2.3. Training Set Construction and Final Segmentation

The human face and body detection modules provide an initial estimation of the foreground object forming the foreground training set, say D^f. Similarly, a background set, say D^b, should also be created. For this reason initially a region of uncertainty is created around the selected foreground masks (face (M_f) and body (M_b)). In particular for each connected component (representing face or body region), the confidence interval of the Gaussian pdf model increases further than 80%, leading to an expansion of the face and body areas. Under this consideration, the new blocks, which are classified to the face or body region, compose the region of uncertainty. Then the background mask D^b is comprised of the blocks that do not belong either to the face/body masks or to the uncertainty zone. As a result, the neural network training set consists of the blocks of sets D^f and D^b. Since there is a large number of similar training blocks, Principal Component Analysis (PCA) is incorporated to reduce their number and the remaining blocks are used for training the network. Finally the trained neural network classifies the image-pixels to separate the foreground video object (human) from the background [11].

3. THE PROPOSED CHAOTIC ENCRYPTION SCHEME

After extraction of human video objects, the proposed chaotic video objects cryptographic system is activated. An overview of the proposed system is given in Figure 1,

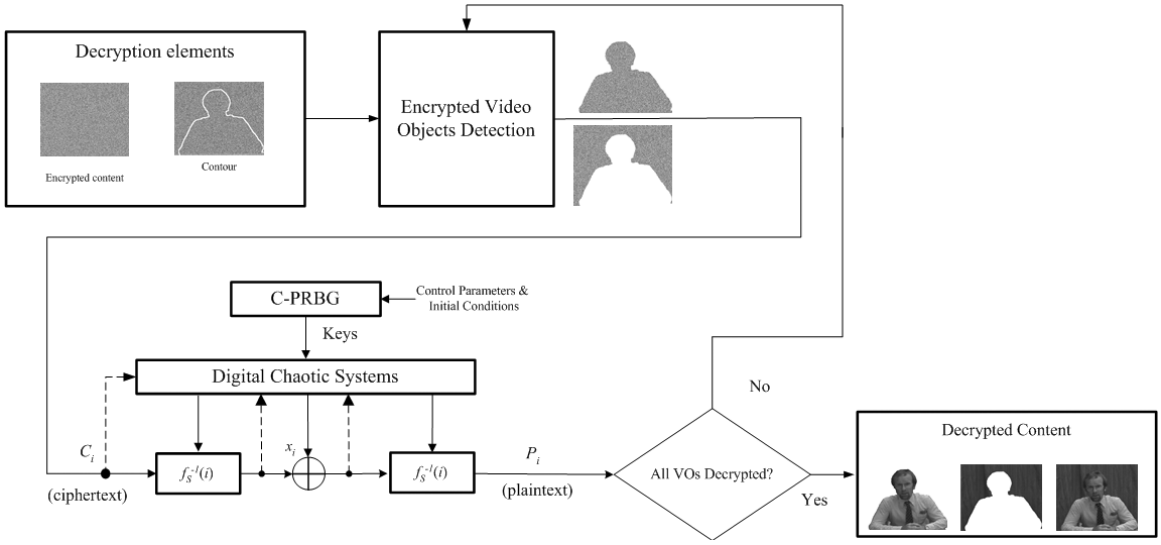

Figure 2: The decryption scheme

which consists of a chaotic pseudo-random bit generator and a chaos-based cipher module. More details are provided in the following subsections.

3.1. Keys Generation Based on C-PRBG

In most secure cryptographic schemes, the security of encrypted content mainly depends on the size of the key. In our system, for each video object a different key is used, which has size 256 bits, leading to a symmetric cipher. Each key is generated by a chaotic pseudo-random bit generator (C-PRBG). C-PRBGs based on a single chaotic system can be insecure, since the produced pseudorandom sequence may expose some information about the employed chaotic system [12]. For this reason in this paper we propose a PRBG based on a triplet of chaotic systems, which can provide higher security than other C-PRBGs as three chaotic systems are employed. The basic idea of the C-PRBG is to generate pseudo-random bits by mixing three different and asymptotically independent chaotic orbits.

Towards this direction let $F_1(x_1, p_1)$, $F_2(x_2, p_2)$ and $F_3(x_3, p_3)$ be three different one-dimensional chaotic maps: $x_1(i+1) = F_1(x_1(i), p_1)$, $x_2(i+1) = F_2(x_2(i), p_2)$, $x_3(i+1) = F_3(x_3(i), p_3)$, where p_1, p_2, p_3 are control parameters, $x_1(0)$, $x_2(0)$, $x_3(0)$ are initial conditions, and $\{x_1(i)\}$, $\{x_2(i)\}$, $\{x_3(i)\}$ denote the three chaotic orbits. Then a pseudo-random bit sequence can be defined as:

$$k(i) = \begin{cases} 1, & F_3(x_1(i), p_3) > F_3(x_2(i), p_3) \\ k(i-1), & F_3(x_1(i), p_3) = F_3(x_2(i), p_3) \\ 0, & F_3(x_1(i), p_3) < F_3(x_2(i), p_3) \end{cases} \quad (6)$$

According to this scheme the generation of each bit is controlled by the orbit of the third chaotic system, having as initial conditions the outputs of the two other chaotic systems.

3.2. The Encryption Module

After generating a pseudo-random key for each video object (human video object and background) the cipher module is activated (Figure 1). Initially for each video object the pixels are scanned from top-left to bottom-right providing a sequential arrangement of the plaintext pixels P_i. Next, taking into consideration the discussion at the introduction, multiple iterations of chaotic systems lead to slow ciphers, while a small number of iterations may raise security problems [13]. Thus, in order to make possible a single iteration of the chaotic systems while maintaining high security levels, the proposed scheme combines a simple chaotic stream cipher and two simple chaotic block ciphers (with time variant S-boxes) to implement a complex product cipher.

Considering Figure 1 the operation of the cipher module can be described as follows: assume that P_i and C_i represent the ith plaintext and ith ciphertext pixels respectively (both in n-bit formats). Then the encryption procedure is defined by

$$C_i = f_S(\{f_S(P_i, i) \oplus x_i\}, i) \quad (7)$$

where symbol $\oplus$ represents the XOR function, $f_S(\cdot, i)$ are time-variant $n \times n$ S-boxes (bijections defined on $\{0, 1, \ldots, 2^n - 1\}$) and x_i is produced from the states of three chaotic systems. Here, f_S are also pseudo-randomly controlled by the chaotic systems. The secret key provides the initial conditions and control parameters of the employed chaotic systems. The increased complexity of the proposed cipher against possible attacks is due to the

mixed feedback (internal and external): $f_S(P_i, i)$ at **FB1**, $f_S(P_i, i) \oplus x_i$ at **FB2** and ciphertext feedback C_i at **FB3**, which lead the cipher towards acyclic behavior.

The procedure is terminated after both the human video object and the background are encrypted. Then by combining the encrypted video objects (human video object and background) the final content is generated, which does not provide any clue about the contours of the video objects or the structure of the visual information.

3.2. The Decryption Module

A diagram of the proposed decryption module is given in Figure 2. The decryption module receives at its input the following elements: an image consisting of encrypted human video objects and background, the contours of the video objects (encrypted pairs of coordinates), the initial control parameters and initial conditions for the triplet of chaotic maps (C-PRBG module) and the initial cipher value C_0 (used at the first feedback).

Afterwards for each video object (detected by the corresponding contour) the pixels are scanned from top left to bottom right (as during encryption) and they are arranged in sequential order. Now the Digital Chaotic Systems produce the same specific values used during encryption, but now these values are incorporated for decryption. Decryption of the different video objects can be performed in parallel. The procedure is terminated after all pixels are decrypted, leading to regeneration of the plain-text video objects.

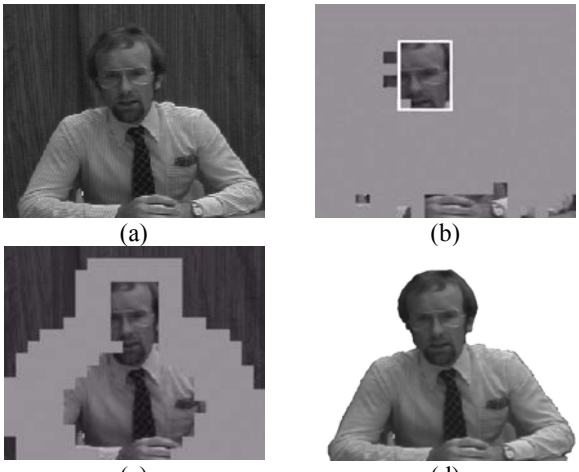

Figure 3: Human video object extraction, (a) One frame from 'Trevor' videoconference sequence, (b) Face detection, (c) Training set and region of uncertainty and (d) final classification (foreground video object).

4. EXPERIMENTAL RESULTS

For evaluation purposes the proposed human video objects cryptographic scheme is examined in terms of security and efficiency. In particular the proposed system is applied to the videoconference sequence 'Trevor', one frame of which is depicted in Figure 3(a). The frame size is 176 x 144 pixels.

Next in Figure 3(b) the performance of the human face detection module is illustrated. It should be mentioned that in this sequence, additional blocks are also selected as candidate face blocks, due to the fact that their chrominance characteristics are close to those of face regions, e.g., the hands. For clarity of presentation, when a block is classified to foreground, it is included as it is in the figures, while blocks classified to the background, are depicted with gray color. Finally the adopted shape-matching algorithm for face detection is incorporated. In our implementation, we set the range of aspect ratios of the bounding rectangles to the interval [1.2 1.7], also addressing cases of covered foreheads, exposed necks or when only a part of the face region has been initially estimated by the Gaussian pdf model.

As shown in Figure 3(b) a part of the human face is extracted, located inside the area depicted by a white rectangle, according to the shape-matching procedure. As observed, non-face regions are discarded, as they do not satisfy the shape constraints described in section 2.1.

Afterwards body detection is performed and the training set that contains the foreground and background areas is shown in Figure 3(c). In this figure, gray color is used to represent the region of uncertainty. As it can be observed, the region of uncertainty contains several blocks, located at the boundaries of the face and body areas and thus protects the neural network from ambiguous regions. Next foreground/background blocks are used to train the neural network classifier. Nevertheless, since there are several similar training blocks, Principal Component Analysis (PCA) is used to reduce their number, reducing the number of selected training blocks to 12. Finally, the trained network performs the final pixel-accurate segmentation and the human video object is presented in Figure 3(d).

Afterwards the encryption algorithm is activated, where in these experiments the three incorporated chaotic maps (both in the C-PRBG module and the cipher module) are piecewise linear chaotic maps (PWLCMs) of the form:

$$F(x,p)=\begin{cases} x/p, & x\in[0,p) \\ (x-p)/\left(\frac{1}{2}-p\right), & x\in[p,\frac{1}{2}] \\ F(1-x,p), & x\in(\frac{1}{2},1] \end{cases} \quad (8)$$

where $0 < p < ½$, and initial control parameters p_1=0.15, p_2=0.27 and p_3=0.43. Eventually the encryption module is also applied to the background and the final encrypted content is produced (Figure 4). As it can be observed, the final content looks completely random and does not provide clues relevant to the number or location of video objects. This fact is further clarified in Figures 5(a) and 5(b), where the histograms of the foreground (human video object) and the background are presented respectively. Both histograms approximate the histogram of a table with random values. This is a very important security merit, as the encrypted content approximates the statistics of randomly generated pixels, independently of the plaintext.

Afterwards the encrypted content can be transmitted to a recipient. In this case just the contour of the human video object should be transmitted, together with the encrypted image. In the case under consideration the total extra elements (contour, initial control parameters and initial conditions, C_0) correspond to a bit rate increase of about 1%.

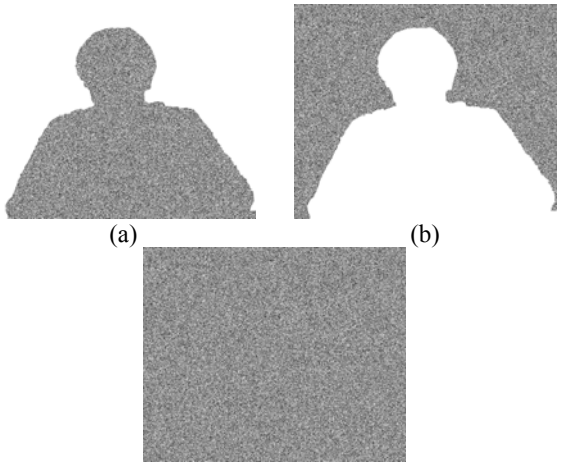

Figure 4: Final encrypted content. (a) Human video object encrypted with 2 key-1, (b) Background video object encrypted with key-2 and (c) Final content (foreground & background).

Now the security of the proposed scheme is further examined. Due to the feedback operations, the time-variant S-boxes and the multiple chaotic systems that control the S-boxes and the input to the XOR operation, it looks very difficult to use statistical methods to cryptanalyze the cipher. For this reason some tests are performed. Let us assume that an unauthorized user knows the contour of the human video object and tries to decrypt the content by brute force attack. If the exact key is detected then the human video object can be decrypted (Figure 6(a)). However if the key differs even by just one bit, the content will not be decrypted as it can be seen in Figure 6(b). Furthermore assuming that an unauthorized user knows the exact key but does not know the contour of the video object he/she cannot decrypt the visual content as shown in Figure 6(c). Finally in case that a user knows only the contour and the key of the background he/she can decrypt only the background video object (Figure 6(d)). According to the previous result, the proposed scheme can be used for layered access to visual content, where users can view only the content they are authorized to view.

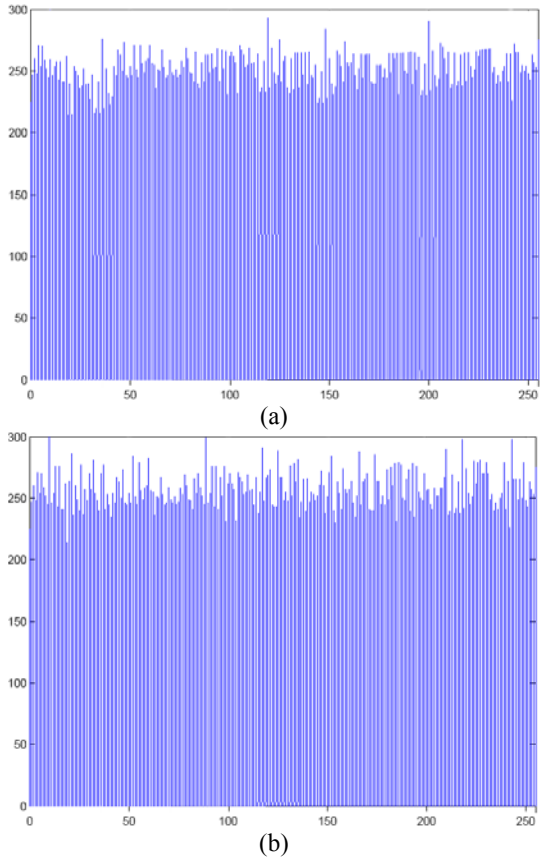

Figure 5: Histograms of encrypted video objects. (a) Histogram of the human video object and (b) Histogram of the background video object.

5. CONCLUSIONS

In this paper a chaotic encryption system is designed to confront the problem of human video objects security. The case of videoconference/videophone sequences was addressed where face and body areas can be estimated according to Gaussian pdfs. The final content does not provide any clue relative to the number and locations of video objects and enables layered access to the visual

material. Furthermore topology issues due to segmentation of the frame into different regions may lead to enhancement of the overall security (not from strict cryptographic viewpoint).

Experimental results illustrate the security of the proposed scheme showing that even in the case of a brute-force attack, if unauthorized users do not know the exact contour it is very difficult to decrypt the content. Further experiments and analysis should be carried out to illustrate the full capabilities, applicability and disadvantages of the proposed scheme.

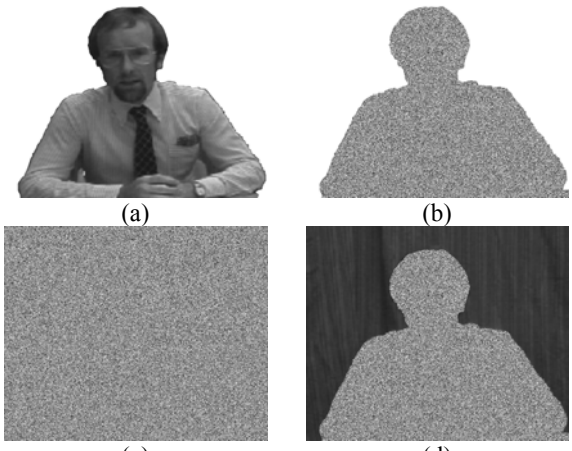

Figure 6: Security of the proposed system. (a) Decryption result when both the key and the contour of the human video object are known, (b) Decryption result when key is different by one bit and contour is known, (c) Decryption result when keys are correct but contour is unknown and (d) Decryption result when only the contour and the key of the background video object are known.

References:

[1] A. Kudelski, "Method for scrambling and unscrambling a video signal," *U. S. Patent* No. 5375168, Dec. 20, 1994.

[2] M. G. Kuhn. (1998), "Analysis of the Nagravision Video Scrambling Method," Online Available: http://www.cl.cam.ac.uk/~mgk25.

[3] Y. Matias and A. Shamir, "Video scrambling apparatus and method based on space filling curves," *U.S. Patent* No. 5058158, Oct. 15, 1991.

[4] W. Zeng and S. Lei, "Efficient frequency domain selective scrambling for digital video," *IEEE Trans. Multimedia*, Vol. 5, pp. 118–129, March 2003.

[5] Jui-Cheng Yen and Jiun-In Guo, "A new chaotic key-based design for image encryption and decryption," in *Proc. IEEE Int. Conf. Circuits and Systems*, 2000, Vol. 4, pp. 49–52.

[6] T. Habutsu, Y. Nishio, I. Sasase, and S. Mori, "A secret key cryptosystem by iterating a chaotic map," in *Proc. Advances in Cryptology*, Berlin, Germany: Springer-Verlag, 1991, pp. 127–140.

[7] M. S. Baptista, "Cryptography with chaos," *Phys. Lett. A*, Vol. 240, pp. 50–54, 1998.

[8] Bruce Schneier, "Applied Cryptography," Second Edition.

[9] H. Wang and Shih-Fu Chang, "A Highly Efficient System for Automatic Face Region Detection in MPEG Video Sequences," *IEEE Trans. CSVT*, vol. 7, No. 4, pp. 615-628, August 1997.

[10] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw Hill, 1984.

[11] N. D. Doulamis, A. D. Doulamis, K. S. Ntalianis, and S. D. Kollias, "An Efficient Fully-Unsupervised Video Object Segmentation Scheme Using an Adaptive Neural Network Classifier Architecture," in *IEEE Transactions on Neural Networks*, Vol. 14(3), pp. 616-630, May 2003.

[12] V. A. Protopopescu, R. T. Santoro, and J. S. Tollover, " Fast and secure encryption – decryption method based on chaotic dynamics," *US Patent* No. 5479513, 1995.

[13] S. Li, X. Zheng, X. Mou, and Y. Cai, "Chaotic encryption scheme for real-time digital video," in *Real-Time Imaging VI*, Proceedings of SPIE, Vol. 4666, pages 149–160, 2002.