Chaotic Encryption Driven Watermarking of Human Video Objects Based on Hu Moments

Paraskevi K. Tzouveli, Klimis S. Ntalianis, Stefanos D. Kollias
National Technical University of Athens
Electrical and Computer Engineering Department
Iroon Polytexneiou 9, 15780, Athens, Greece
email: (tpar,kntal)@image.ntua.gr

Abstract. A novel human video object watermarking scheme is proposed in this paper, providing copyright protection of semantic content. The proposed method focuses on the existence of face and body regions within an initial image, however other cases can also be addressed. Initially detection of human video objects is achieved using two adaptive 2-D Gaussian models, one for skin color distribution modelling and the other for body localization modelling. A watermark is then designed using invariant Hu moments of each human video object, and the watermark insertion procedure is driven by an iterative encryption module based on chaotic functions. Performance of the proposed object based secure watermarking system is tested under various signal distortions and known cryptanalytic attacks.

1 Introduction

The copyright protection of digital images and video is an urgent issue of ownership identification. Many watermarking schemes [4]-[7] have been proposed for ownership protection, some of which providing significant resistance to image processing attacks. In this paper, a novel chaotic encryption driven watermarking scheme is proposed, based on Hu moments and applied to human video objects. The proposed system provides copyright protection of semantic content. In particular, the embedding method consists of two sub-modules: the automatic human video object detection sub-module and the watermark insertion sub-module, which is driven by chaotic encryption and modifies the Hu moments of the initial video object to provide the watermarked human video object.

Initially, human video object detection is performed for each candidate image, based on skin color distribution [1],[2]. Afterwards, an iterative cipher mechanism based on the logistic function is used in order to encrypt the pixel values of the

Please use the following format when citing this chapter:
Tzouveli, Paraskevi, Ntalianis, Klimis, Kollias, Stefanos, 2006, in IFIP International Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, eds. Maglogiannis, I., Karpouzis, K., Bramer, M., (Boston: Springer), pp. 673–680

human video object taking into consideration, in each iteration, the values of the previously encrypted pixels. As result, a chaotic "noise" is produced which is incorporated in producing the watermarked video object. This is achieved by modifying the Hu moments of the initial video object according to an additive scheme. In order to assure the robustness of the proposed watermarking method, the modification of moments is confined to a limited predefined interval. In the detection scheme a neural network classifier [3] is initially used in order to extract possible watermarked human video objects from each candidate image. Then, the watermark detection procedure is based on the comparison of Hu moments. Experimental results on real sequences further indicate the advantages of the proposed scheme.

2 Detection and Extraction of Human Video Object

The human face detection module that is used in the proposed method is based on the distribution of chrominance values corresponding to a human face [1], which occupy a very small region of the color space. The blocks of the image that are located at this small region can be considered as face block of the searching face class $\Omega_{f.}$ Using a Gaussian probability density function (pdf) [2], the histogram of chrominance values corresponding to the face class can be initially modelled as $P(\mathbf{x} \mid \Omega_f) = \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_f)^T \cdot \Sigma_f^{-1} \cdot (\mathbf{x}-\boldsymbol{\mu}_f)) \big/ 2\pi \cdot |\Sigma|^{1/2}$ where x=$[u \quad v]^T$ 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 class and $\boldsymbol{\Sigma}$ is the 2x2 variance matrix of the probability density function: $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_u^2 & \sigma_{u,v} \\ \sigma_{u,v} & \sigma_v^2 \end{bmatrix}$. 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 μ_f and Σ are estimated based on a set of several face images and using the maximum likelihood algorithm. Each B_i block of the image is considered belong to the face class, if the respective $P(\mathbf{x}(B_i)|\Omega_f)$ is high. The aspect ratio for face areas $R = H_f / W_f$ (where H_f is the height, while W_f is width of the head) was experimentally found to lie within the interval [1.4 1.6]. Using R and P, a binary mask, say M_f, is build containing the face area.

Detection of the body area can be achieved using geometric attributes that relate face and body areas. After the calculation of the geometric attributes (center $c_f = [c_x \quad c_y]^T$,width w_f and height h_f) of the face region, the human body can be localized by incorporating a probabilistic model, the parameters of which are estimated according to c_f, w_f and h_f. The probability of each block B_i to belong to a human body class, say Ω_b can be computed by:

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

where $\mu_x = c_x$, $\mu_y = c_y + h_f$, $\sigma_x = w_f$, $\sigma_y = h_{f/2}$, are the parameters of the human body location proposed model and $r(B_i) = [r_x(B_i) \quad r_y(B_i)]^T$ is the distance between the ith block and the origin. Similarly to human face detection, a block B_i belongs to the body class Ω_b, if the respective probability, $P(r(B_i)\,|\,\Omega_b)$ is high.

Finally the face and body masks are fused and human video objects are extracted. In Figure 1, the phases of the proposed method are illustrated. Firstly, the human video object region is detected within the initial image (Figure 1a) and an object mask is produced (Figure 1b). This mask is used for the extraction of human video object (Figure 1c) is extracted using the object mask. The human face and body detection modules provide an initial estimation of the human video object, forming the training set, say Df.

Then, in order to have a more reliable training set, a region of uncertainty is initially created around the selected foreground mask (face (Mf) and body (Mb)). Particularly, for each connected component (representing face or body region), the confidence interval of the Gaussian pdf model is selected to be 80%, meaning that only blocks falling into this interval are considered as candidate training blocks. Finally, as several blocks (8x8 regions of pixels) fall into this confidence interval, the PCA method is incorporated and a small set of training blocks is eventually selected for the training phase of the neural network.

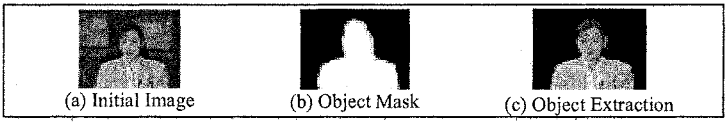
(a) Initial Image (b) Object Mask (c) Object Extraction

Figure 1. Human Video Object Extraction Method

3 Moments invariants

In the literature, moment invariants have been used for object recognition in an image [2-7] regardless of their particular position, orientation, viewing angle, and gray-level variations. Central moment $\mathrm{m}_{\mathrm{pq}}^{(\mathrm{f})}$ of order *(p+q)* of the image *f(x,y)* are non-negative integers, and can be computed by $\mathrm{m}_{\mathrm{pq}} = \sum_{\mathrm{x}}\sum_{\mathrm{y}} (\mathrm{x}-\bar{\mathrm{x}})^{\mathrm{p}} (\mathrm{y}-\bar{\mathrm{y}})^{\mathrm{p}} \mathrm{f}(\mathrm{x},\mathrm{y})$ where the coordinates $\bar{\mathrm{x}} = \mathrm{m}_{1,0}/\mathrm{m}_{0,0}, \bar{\mathrm{y}} = \mathrm{m}_{0,1}/\mathrm{m}_{0,0}$ denote the centroids of *f(x,y)*. The central moments of the image are invariant to translation as they are origin-independent. Scaling invariance can be achieved by normalizing the moments of the scaled image by the scaled energy $\eta_{\mathrm{pq}} = \mathrm{m}_{\mathrm{pq}}/\mathrm{m}_{00}{}^{\gamma}$ of the original (γ is the normalization factor $\gamma = (\mathrm{p}+\mathrm{q}/2)+1$). Hu [8] first introduced the mathematical foundation for two-dimensional moment invariants, based on methods of algebraic invariants, and demonstrated their application to shape recognition. Using nonlinear combinations of geometric moments, a set of seven invariant values, computed from central moments through order three, and independent of object translation, scale and orientation, can be calculated using the following equations:

$$\phi_1 = \mathrm{n}_{20} + \mathrm{n}_{02}, \quad \phi_2 = (\mathrm{n}_{20} - \mathrm{n}_{20})^2 + 4\mathrm{n}_{11}^2$$

$$\phi_3 = (\mathrm{n}_{30} - 3\mathrm{n}_{12})^2 + (\mathrm{n}_{03} - 3\mathrm{n}_{21})^2, \quad \phi_4 = (\mathrm{n}_{30} - \mathrm{n}_{12})^2 + (\mathrm{n}_{03} + \mathrm{n}_{21})^2$$

$$\phi_5 = (3n_{30} - 3n_{12})(n_{30} + n_{12}) \cdot \left[(n_{30} + n_{12})^2 - 3(n_{21} + n_{03})^2\right]$$
$$+ (3n_{21} - n_{03})(n_{21} + n_{03}) \cdot \left[3(n_{30} + n_{12})^2 - (n_{21} + n_{03})^2\right]$$
$$\phi_6 = (n_{20} - n_{02}) \cdot \left[(n_{30} + n_{12})^2 - (n_{21} + n_{03})^2\right] + 4n_{11}(n_{30} + n_{12})(n_{21} + n_{03})$$
$$\phi_7 = (3n_{21} - n_{03})(n_{30} + n_{12}) \cdot \left[(n_{30} + n_{12})^2 - 3(n_{21} + n_{03})^2\right]$$
$$+ (3n_{12} - n_{03})(n_{21} + n_{03}) \cdot \left[3(n_{30} + n_{12})^2 - (n_{21} + n_{30})^2\right]$$

4 Logistic Map and Cryptography Scheme

Chaos theory is a set of ideas attempting to reveal structure in nonlinear dynamic systems [9]. Chaotic functions are very simple non-linear dynamical equations and can descript complex, chaotic behavior of a system [10]. Systems that present chaotic behavior are extremely sensitive to initial condition. Initializing a chaotic function by a key and after a number of iterations, this is able to generate random present different numerous [9], property that is very important for applications of cryptography. A great deal of chaotic behavior can be described by one, simple recursive function, the logistic map which is used in the proposed system. The logistic map function is expressed as: $x_{n+1} = r \cdot x_n(1 - x_n)$ where x takes values in the interval [0,1]. The remarkable features of the logistic map are the simplicity of its form (quadratic difference equation) and the complexity of its dynamics [9].

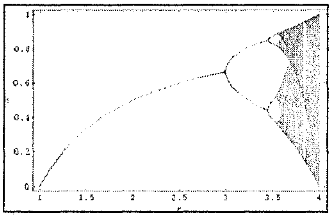

Figure 2. Diagram of the Logistic Map

In Figure 2, a plot of the logistic map is shown versus r with values obtained after a number of iterations. When $r \geq 3.57$, periodicity gives way to complete chaos. Finally for r=3.9 to 4, the chaos values are generated in the complete range of 0 to 1. The secrecy of encrypted information is depended only on the encryption key. In the proposed system, the key size is 256-bit splitted into 32 subkeys of 8 bits each (session keys). The unprotected data which consist of the human video object that is extracted from the original image is called plaintext (denoted by *p*). Applying a key-depended encryption algorithm to the plaintext, the ciphertext (denoted by *c*) is produced.

5 The Embedding Module

An overview of the proposed system's embedding module is depicted in Figure 3 which contains two sub-modules: the automatic human video object detection sub-module and the watermark insertion sub-module. Initially the human video object is extracted by the video object detection module, as described in section 2.

Afterwards, the pixels of the human video object are scanned from top-left to the bottom-right providing the p_i pixels (plaintext pixels). The plaintext is an input to the mapping function which contains the logistic map. The robustness of the system is further reinforced by a feedback mechanism, which leads the cipher to acyclic behavior so that the encryption of each plain pixel depends on the key, the value of the previous cipher pixel and the output of the logistic map. In particular, the feedback mechanism includes four operations.

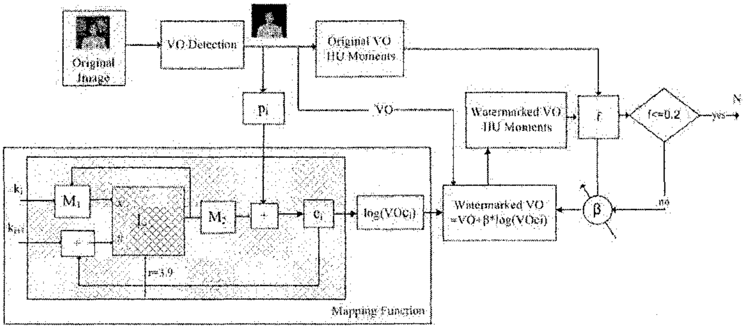

Figure 3. The Embedding Module

Firstly, the output value of the logistic map is input to box M1. The session key k_j is also input to M1 box. The box M_1 represents a mapping function from the input interval to the domain of the logistic map (real numbers in the interval [0, 1]) and fixes the initial value of x. The second operation interjects in the computation of the number of iterations the logistic map performs. Specifically, the cipher pixel is added to the k_{i+1} session key. The result of this addition provides the value of parameter x in order to serves control the number of iterations the logistic map performs. The box M2 serves with the purpose of normalizing the output of the logistic map. Normalization is performed by box M_2 which represents a fuzzy membership function mapping interval [0, 1] into the interval [0, 255]. Finally, the fourth operation is a summation of the plaintext and the normalized logistic map output which actually encrypts each plaintext p_i, producing the ciphertext c_i. Afterwards, the cipher video object (VOci) can be recomposed from the ciphertext. Eventually, the output of the mapping function is the modified video object *log(VOci)*. The watermarked video object $\tilde{O}$ is achieved adding the value *log(VOci)* multiplied by a weighed factor β to the video object which has been extracted from the original image (Figure 4).

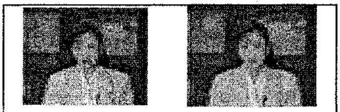

Figure 4. (a) Figure 5. (b)

Figure 4. Original image (a) and image with watermarked video object (b)

Now let us consider that $\Phi = \left[\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6, \phi_7\right]^T$ is the invariant Hu moments of the original human video object O. Let also Φ^* be the invariant Hu moments of the watermarked human video object $\tilde{\mathrm{O}}$. We can choose a function f, which can be any linear or non-linear combination of the invariant moments. In our case study, the function f is expressed as a sum value of the weighted average differences between Hu moments of the original human video object, ϕ, and the watermarked human video object, ϕ^*: $\mathrm{f}(\Phi^*, \Phi) = \sum_{i=1}^{7} \mathrm{w}_i \left(\frac{\phi_i^* - \phi_i}{\phi_i} \right)$ where weights w_i take the values $\mathrm{w}_1 = 1.5$, $\mathrm{w}_2 = 1.25$, $\mathrm{w}_3 = 1$, $\mathrm{w}_4, \mathrm{w}_5 = 0.75$ and $\mathrm{w}_6, \mathrm{w}_7 = 0.50$. These values have been set after several experimental tests, since the first and second Hu moments are the most robust among different moments [10], so the values of the weighted factors $\mathrm{w}_1, \mathrm{w}_2$ are higher than the other weighted factors. The output of function f is called factor N (Figure 3). The weighted factor β is controlled by feedback in order to ensure that $\mathrm{f}(\Phi^*, \Phi) \approx 20\%$.

6 The Detection Module

The detection module (Figure 5) includes two functions: the video object detector and the watermark detector. Firstly, when a candidate image is received, it passes through a neural network classifier [3] that detects video objects similar to the watermarked video objects.

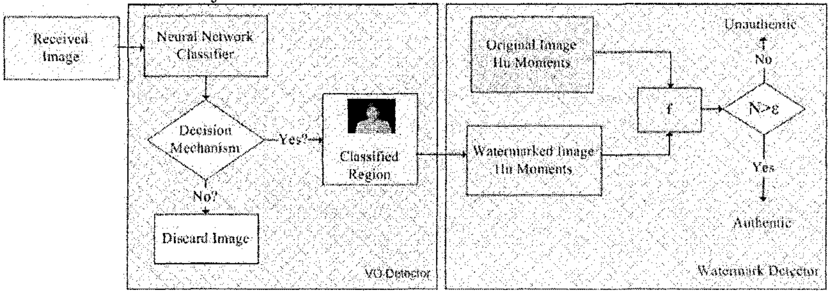

Figure 5. The Detection Module

Let us assume that the candidate image can be partitioned to a sequence of overlapping blocks of size 8x8 pixels, say B_i, and let b_i be a vector containing the lexicographically ordered values of the i^{th} block of the image or a transformed version of it, by extracting, for example, several block descriptors, such as color, motion or texture. The case of overlapping blocks is adopted since pixel resolution is required for accurate video object detection. Afterward, each pixel of the candidate image is classified to available classes ω_i, i=1,2,...p, to one of p available watermarked video objects. The output of p-classification problem is $\mathrm{y}(\mathbf{b}_i) = \left[\mathrm{p}^i_{\omega_1} \mathrm{p}^i_{\omega_2} \cdots \mathrm{p}^i_{\omega_p}\right]^T$ where $\mathrm{p}^i_{\omega_j}$ denotes the degree of coherence of b_i to class ω_j so each pixel is assigned to a class according to the highest degree of coherence. In order to efficiently perform the classification task, the neural network classifier is initially trained using blocks of each watermarked video object. Nevertheless, principal component analysis (PCA) is performed in order to reduce the number of similar training blocks in each watermarked video object. Finally, the classified pixels constitute the area of the respective video object.

In order to avoid a false detection, a decision mechanism is incorporated in order to calculate the ratio $R_i = \frac{N(\omega_i)}{F(vo_i)} \cdot 100$ (where $N(\omega_i)$, i=1,2,...,p, is the number of pixels of each class and $F(vo_i)$, i=1,2,...,p, is the number of pixels constituting each respective watermarked video object). If R_i is less than a threshold T then the candidate area is discarded, since it is considered too small for watermark detection. In other case, the watermark detection module is activated giving the classified region. Value of threshold T affects the false detection and false rejection rates and in the current work is selected by performing several experiments.

Having received a region that the neural network has classified as possible watermarked video object, the watermark detector module estimates the Hu moments of this region. In parallel, the seven values of the Hu moments of the respective original human video object are received as input. Afterwards, the value of function $f(\Phi^*, \Phi)$ can be computed and the value of N can be defined. Matching of the received human video object can be achieved by checking the validity of the equation $N \leq \varepsilon$ where ε is the margin of acceptable error between the two video objects. Then, the detection procedure returns either 1, if the candidate video object is the watermarked video object, or 0, if the candidate video object is not watermarked.

7 Experimental Results

This section presents results that prove the robustness of the proposed system. To illustrate the security of the algorithm, a hacker's approach to crack a watermarked image is considered as a case. Since the 256-bit sequence of key is used as an input to the logistic map, this mean that the keyspace has 2^{256} different values. Additionally the number of iterations supported by the logistic map module is between 0 and 767, as cipher pixels take values in the interval [0, 512] and the session keys take values in the interval [0, 255].

So, the only way to break the proposed system is by brute–force attack scanning the whole keyspace that is to try 2^{256} different keys! Furthermore as the cipher of the proposed system is based on feedback mechanisms, periodicities in the encrypted data do not appear. The proposed system is suitable for real time applications as it spends less time in watermark insertion as only human video objects are considered and not the whole image Experimental results [3] have proved that the neural network classifier can sufficiently extract the human video object of an image which has passed the watermarking embedding procedure and has been attacked.

The watermarked video object can be detected under different attacks, (Table I), except from cropping where the attacked video object region is not similar to the watermarked one also and their moments are very different. The proposed watermarking scheme serves as a 1-bit watermarking system as it answers the yes/no question of authenticity. The threshold is set so as the attacked video object can be different from the watermarked object only by 1%. Robustness of the proposed scheme to geometric manipulations is guaranteed using invariant Hu moments. Only the value of the moments of the original video object is received from the detection module.

8 Conclusion

Nowadays, protection of digital media through networks is a crucial issue. Most watermarking systems are not taking into consideration regions of semantic information comprising the content that should be protected. Our proposed system takes as input images that contain human video objects. The human video objects are extracted by a face and body region detection module. Finally, each human video object is watermarked by modifying its Hu moments. The procedure is driven by a chaotic encryption module. Experimental results illustrate the robustness of the proposed scheme to attacks as well as noise addition, and image distortions.

Table I: Results of detecting the watermarked akiyo video object after several attacks

Attack			$f(\Phi^*, \Phi) = \frac{1}{7}\sum_{i=1}^{7} w_i \left(\frac{\phi_i^* - \phi_i}{\phi_i} \right)$		Watermark Detection	
Filtering	Gaussian	Median	0.004023	0.004023	Pass	Pass
JPEG	Q=50%	Q=10%	0.001229	0.004691	Pass	Pass
Rotation	5°	-1°	0.044546	0.000012	Pass	Pass
Scaling	50%	110%	0.002085	0.003790	Pass	Pass
Cropping 10%			0.40962		Fail	
Flipping			0.002162		Pass	

References

1. H. Wang and S-F 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.
2. N. Tsapatsoulis, Y Avrithis, S. Kollias "Facial Image Indexing in Multimedia Databases" in *Pattern Analysis & Applications*, vol. 4, pp. 93-107, 2001.
3. 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" in *IEEE Transactions on Neural Networks*, Vol. 14(3), pp. 616-630, May 2003
4. F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Attacks on copyright watermarking systems," in *Proc. 2nd Int. Workshop Information Hiding*, pp. 218–238, 1998.
5. S. Pereira and T. Pun, "Robust template matching for affine resistant image watermarks" *IEEE Trans. Image Processing*, vol. 9, pp. 1123–1129, July 2000.
6. M. Kutter, S. K. Bhattacharjee and T. Ebrahimi, "Towards second generation watermarking schemes" in *Proc. IEEE Int. Conf. Image Processing* 1999, pp. 320–323.
7. C. Lin, M. Wu, J. Bloom, I. Cox, M. Miller, Y. Lui, "Rotation, scale, and translation resilient watermarking" *IEEE Trans. Image Processing*, vol. 10, pp. 767–782, May 2001.
8. M. K. Hu, "Visual pattern recognition by moment invariants," *IEEE Trans. Inform. Theory*, vol. 8, pp. 179–187, 1962.
9. G. Jakimoski and L. Kocarev, "Chaos and Cryptography: Block Encryption Ciphers Based on Chaotic Maps", *IEEE Trans.Circuits and Systems*, vol. 48, no. 2, Feb. 2001.
10. R. Devaney, "An Introduction to Chaotic Dynamical Systems", 2nd ed. *CA: Addison-Wesley*, 1989.