Genetic Coding of Digital Watermarks to Enhance IoT Security

Dmitry Zaichenko, Irina Sineva
Moscow Technical University of Communications and Informatics “MTUCI”
Moscow, Russia
dmitryzaichenko@yandex.ru, iss@mtuci.ru

Abstract—The future of smart cities is already real, residents of megacities can easily exchange information anywhere, at any time. Through a network integrated with real-time monitoring systems, data is collected, processed and analyzed. The modern understanding of IoT focuses on three components—the devices themselves, the sufficiency of the channel capacity for their interaction and the information security technologies that should accompany this interaction. The proposed method based on genetic coding that hides messages between IoT devices is capable of detecting both internal and external attacks in the intellectual infrastructure of the Internet of things.

I. INTRODUCTION

The number of peripherals is growing exponentially around the world, providing quality services to the user. The Internet of things makes it easy to share information between users, devices and applications that are located in different locations around the world. This is a good start-to-start new research on the possibility of using portable devices to solve real optimization and machine learning problems.

It is expected that an increasing number of messages will generate many data, thus increasing the number of attacks for malicious users due to the openness, distributed nature and lack of control over the entire IoT environment.

II. THE PROBLEM OF DATA SECURITY

Users are increasingly employing smart devices for confidential operations, such as buying goods in online stores, banking operations, etc. They all store important personal data (e-mail, photos, pin codes, etc.) and corporate information (financial and economic activities of the organization, customer information, etc.). All of this has changed the way people, organizations and Government interact, and makes us live in a cyber society [1]. However, this brings a new dimension of complexity from the point of view of many potential problems, particularly in the areas of data security, including aspects of confidentiality and data integrity at the stages of data transmission in IoT in general.

The issue of security along with the device identification system is the most acute in the IoT. Numerous Recommendations of the International Telecommunication Union (above all X series: Data networks, open system communications and security and Y series: Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities) are devoted to this issue, including the real-time crypto and stego protection.

The number of malicious attack vectors is large because of the openness of the Internet of things, the distributed nature and the lack of control over the system. As more and more smart devices, connect to the IoT, the number of entry points for attacks increases. With the installation of third-party applications on these smart devices, the likelihood of malware also increases, which can ultimately make the system unsafe.

To build IoT as an efficient service platform, end users need to trust the system. As the number of information and communication technologies increases, the need for information security and data security increases. Ensuring the security of cyberspace and infrastructure in General is critical to economic growth, prosperity and security.

Cryptography and steganography are two of the most commonly used methods of information security in communication. Cryptography protects against threats by preventing attackers from getting anything useful from the device and an insecure communication channel. However, cryptography has a number of limitations, one of which is the presence of encrypted information in itself can be suspicious. On the other hand, steganography protects the information, hiding it inside the carrier, does not cause suspicion. It provides more privacy and security than cryptography because it hides the existence of the message, not just the protection of the content [2].

Text, images, audio, and video are the primary media that can be used to hide and transfer information in the IoT environment using steganography. Image files are the most common object on the Internet and can be used as a medium.

IoT applications are very popular in smart infrastructures such as smart home or smart city. Some peripherals of the Internet of things in the home scenario, for example, are the smart room temperature controller, smart door, smart TV, etc. (Fig. 1).

Fig. 1. Home IoT infrastructure

These devices communicate their status through the home router with the home server, where operations are stored,
processed, and performed. One of the smart home operation scenarios: when a person enters the smart home through the smart door, the smart TV and the smart room temperature controller turn on automatically. When a person leaves the room through a smart door, all edge devices are turned off.

In these infrastructures, attackers can be internal (a neighbor who has temporary access to the home router) and external (an Internet service provider who has access to the security gateway). An internal attacker can intercept and modify messages between IoT devices and between devices and the home server. An external attacker could alter the content of messages between the user and the home server. The proposed method, which hides messages between IoT devices, IoT devices and the home server, as well as the user and the home server, can prevent both internal and external attacks in the intellectual infrastructure of the Internet of things.

With regard to information technology, the increase in the exchange or transmission of information and its transmission through actually connected systems or communication networks increases the need for security exponentially. In today's world integrity, confidentiality, and authentication service are considered as the most important security principle.

III. GENETIC ALGORITHMS IN THE INTERNET OF THINGS APPLICATIONS

Today, genetic algorithms are a thriving field. Their application is quite diverse: modeling of innovation processes, stock market forecasting and investment portfolio planning, aerospace engineering, microchip design, biochemistry and molecular biology, schedules at airports and Assembly lines, and many others.

Genetic algorithms have recently been used to improve the performance of information hiding systems. Compared to other approaches, the most significant advantage of evolutionary computing is the increased flexibility and adaptability to the task at hand, combined with the high performance and nature of global search.

In fact, evolutionary computation should be understood as a general adaptable concept for solving problems well suited to solving complex optimization problems [3], [4], [5], [6], [7]. Most modern implementations of evolutionary algorithms come from three strongly related but independently developed approaches: genetic algorithms, evolutionary programming, and evolutionary strategy.

The classic error-correcting coding algorithm is based on the addition redundancy. However adding extra redundancy to the message being transmitted may not be an acceptable price. Therefore, there is a problem of reducing the effects of possible distortion of characters in the transmission without making additional redundancy in the message. The proposed approach does not introduce additional redundancy in the transmitted message and, accordingly, does not lead to the errors' correction. Nevertheless, it allows you to decode the received message (possibly distorted) into the closest one as applied to the whole ensemble of messages. In this case, the gain reaches \( \sigma \) or more. The initial prototype algorithm is described in [8], the study of its properties and optimization are devoted to the works [9], [10], [11], [12]. Genetic coding algorithms were developed to improve noise immunity without introducing additional redundancy [13], [14], [15], [16], [17].

The uniqueness of the discussed class of algorithms is that the revealed property is not local, but global on the whole space of the message source. Strictly speaking, theoretically the properties of minimizing the use of the genetic algorithm have not been proven. Testing on small random arrays shows that the genetic algorithm really finds the minimum. For large arrays of points, the genetic algorithm works fine with the existing performance of a conventional computer, but full analysis is no longer acceptable. For example, for an array of 4096 points, more than \( 3.6 \times 10^{24} \) code combinations must be analyzed to find the minimum by exhaustive search.

Therefore, the main advantage of genetic algorithms is their application in hard-to-solve big data problems for which there are no specific methods. Where existing techniques have been successfully applied, it is possible to increase productivity by combining them with genetic algorithms.

IV. DESCRIPTION OF THE GENETIC CODING ALGORITHM

In contrast to the classical genetic algorithm-prototype, this paper uses a modified algorithm that allows you to take into account the difference in the probability of an error in different symbols of the binary code [18]. Let us move on to the description of the genetic coding procedure for the initial messages.

Our task will be to assign binary code combinations to points in such a way that the distortion of one character translates each point into one not very far from it. Thus, we will try to reduce the ensemble average error of the signal representation constructively. The points of the plane (pixel parameters) are fed to the algorithm input.

The first step is to select the starting point of the algorithm. The following two approaches are most appropriate: selection of the point of thickening of the array and the choice of the "center of gravity" of the entire array. In the second step, we look for the points of the first category, which will be further encoded by combinations with one unit. To find the least distant point, consider the possible distortion in each character, the probability of distortion of only the symbol \( i \) denote \( P_i \). If the code word length is distorted symbol \( i \), the probability of this event is

\[
P_i = (1 - p_i)(1 - p_j)...(1 - p_s)p_s(1 - p_j)...(1 - p_s) = \prod_{i} (1 - p_i) \frac{p_i}{1 - p_s} (1)
\]

In the third step, to find the least distant point, consider the possible distortion of the character pair. In the third step, we find the points of the second category, which together are the least distant from the points of the first category. Here it is necessary to take into account the possible distortion of a pair of characters (1). In the fourth step, threes of points are considered in the same way. The same scheme is used to search and encode points of the fourth and subsequent categories.

We describe step-by-step operation of point search taking into account the selected genetic operators (crossing over, mutation, population generation operators) and their parameters:
1) After starting the algorithm, a population of $2R$ individuals is created. Each individual is a number in the original array.

2) Individuals are estimated at the level of adaptability. The most adapted individuals are included in the new generation.

3) Selection is carried out by the method of pair tournament selection; as a result, a group of parents (individuals) who will give offspring is selected.

4) Crossbreeding operation is performed for selected individuals (parents). The creation of descendants is done using uniform crossover. In it, crossing is performed on the basis of a binary vector, the length of which is equal to the length of the genome at the loci. Each position of the vector determines the mutual replacement of the locus in the parent species.

5) The missing part of individuals ($R/2$) to the full volume of the new population ($2R$) is formed by mutation of already obtained individuals; part of individuals "mutates" - a random number of genes (bits) in each individual is inverted.

6) Steps 2 to 6 are repeated if the exit criterion is not reached. The criterion for stopping the algorithm is the invariance of the best individuals for a given number of generations or the achievement of a critical size of the number of generations. The high proportion of mutation of individuals and the applied exit criterion allows to reduce the probability of detection of the local optimum of the system. At the same time, the use of NSGA-II and tournament in breeding allows to speed up the process of global optimum allocation.

At the end of the algorithm, among the resulting $R$ individuals are selected unique, which are the optimal set of solutions to the problem.

The structure of this algorithm is presented in Fig. 2.

V. RESULTS OF THE GENETIC ALGORITHM

Let us consider how the coding algorithm results vary from different types of errors, modeled as random fields with different distributions.

A. Field of the message source with a normal distribution

Fields of this type are examples of fields with strong localization, which should contribute to the efficient operation of the genetic coding algorithm. Fig. 3 shows an example of a genetic and random coding algorithm for a field consisting of 32 points.

The figures show that in case of random coding in the code space, the neighboring points are often located far enough in the source space. In the case of a genetic encoding algorithm, there are fewer such cases, i.e., close points in the code space are also close in the source space of the message.

In Fig. 4 shows a histogram of the distribution of the average random coding distances for 1000 implementations. This result is consistent with a normal distribution (mean = 150.8; standard deviation = 5.7), the significance level is 0.95.

The values on the histogram range from 130.8 to 169.6. The average distance for random coding is 150.8.

In Fig. 5 shows the histogram of the distribution of the average random coding distances in the last bit for 1000 implementations. This result is consistent with a normal distribution (mean = 30.2; standard deviation = 2.7), the significance level is 0.95.
The results of the genetic algorithm is always better than the randomly found minimum, they are summarized in Table I.

<table>
<thead>
<tr>
<th>TABLE I.</th>
<th>THE OPTIMIZATION OF GA PERFORMANCE MODIFICATIONS FOR GAUSSIAN FIELD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averaged distance obtained</td>
<td>Deviation from statistical mean</td>
</tr>
<tr>
<td>Sum of distances of all edges of the graph</td>
<td>106.0</td>
</tr>
<tr>
<td>Sum of edge distances in the last bit</td>
<td>18.0</td>
</tr>
</tbody>
</table>

VI. UNIFORM DISTRIBUTION OF MESSAGE SOURCES

The analysis of genetic coding algorithms for the space of the message source, which is described by the uniform distribution law (Fig. 6). $4.33 \cdot 10^{-15}$

A comparison of the modified genetic coding algorithm and random coding. For random coding, the average distance between the original and the distorted message is 216.3, and for the genetic coding algorithm for the uniform field: 158.8. (Fig. 7). This result is consistent with a normal distribution (mean = 209.1; standard deviation = 8.6), the significance level of 0.95.

![Fig. 6. Neighborhood graphs of code combinations obtained by the genetic prototype algorithm (left) and the genetic prototype algorithm with the highest probability in the last bit (right) for a uniform field](image)

The obtained histograms show that to reduce the total distance, the small distances of the original distance table are used almost in full and, if necessary, supplemented by large distances. This effect does not depend on the configuration of points or on the dimension of the array itself.

VII. APPLICATION IN PROBLEMS OF STEGANOGRAPHY

Recently, genetic algorithms have been used to improve the performance of information hiding systems. The power and invisibility of data can be increased by integrating genetic algorithms with steganography techniques. Solving the problem of information encoding using a genetic algorithm significantly reduces the error of message decoding in the metric of the initial space [19], [20].

Different types of media objects: text, image, audio or video files can be used to hide sensitive data. As already mentioned, the most popular type of steganography is embedding information in images. Here, a secret message is embedded in the carrier as noise, which is almost impossible to detect by the human eye. This type has a number of advantages: the ability to embed hidden information in container images, good resistance to external influences and attacks [21], [22].

For random encoding, the average distance between the original and the distorted message in the last bit is 41.7, and for the genetic encoding algorithm for the uniform field: 29.59. (Fig. 8). This result is consistent with a normal distribution (mean = 41.7; standard deviation = 4.1), the significance level of 0.95.

![Fig. 7. Histogram of average distances’ distribution for 1000 random coding implementations for uniform field](image)

![Fig. 8. Histogram of average distances’ distribution in the last bit for 1000 random coding implementations for uniform field](image)

The results of the genetic algorithm for the uniform field are summarized in Table II.

<table>
<thead>
<tr>
<th>TABLE II.</th>
<th>THE OPTIMIZATION OF GA PERFORMANCE MODIFICATIONS FOR UNIFORM FIELD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averaged distance obtained</td>
<td>Deviation from statistical mean</td>
</tr>
<tr>
<td>Sum of distances of all edges of the graph</td>
<td>158.8</td>
</tr>
<tr>
<td>Sum of edge distances in the last bit</td>
<td>29.59</td>
</tr>
</tbody>
</table>

Before discussing how information is hidden in an image file, the first thing to consider is how images are stored. An image file is a binary file that contains a binary representation of the color or light intensity of each image element (pixel).
Images are divided into three types: binary grayscale and RGB images. A binary image has one-bit value per pixel, representing 0 for black and 1 for white pixels. The number of bits in a color scheme, called bit depth, refers to the number of bits used for each pixel. The lowest bit depth in the current color schemes is 8, which means that 8 bits are used to describe the color of each pixel. Monochrome and grayscale images use 8 bits per pixel and can display 256 different colors or grayscale. Digital color images are typically stored in 24-bit files and use an RGB color model. RGB image is the most appropriate because it contains a lot of information that helps in hiding classified information with a small change in image resolution that does not affect the image quality and makes hiding the message more secure. For this reason, digital photos are used to hide messages on the Internet and other media.

Consider the algorithm of the least significant bit replacement method with the integration of a genetic algorithm [23], [24], [25].

Image steganography is one of the common methods used to hide information in a cover image. LSB is a very efficient algorithm used to embed information into a cover file.

Least significant bit (LSB) steganography is a popular and widely used method in the spatial domain. Conventional methods used in LSB-based steganography mainly focus on increasing the capacity of embedded information and invisibility, while the security issue still needs to be addressed because LSB embedding is vulnerable to several common data attacks, such as additive white Gaussian noise attack (AWGN), geometric attacks, and more.

The first action scans the original image, line by line, and encodes it in binary format. The size of the medium and the secret message is checked. In the second stage, a genetic algorithm encodes the secret carrier message. Then, the resulting code combination is hidden in the original image in part of the pixel in the least significant bits. The output is an encrypted stegofile with greater noise immunity without introducing additional redundancy into the transmitted message. The third stage is the analysis of the stegofile.

Image quality can be determined by statistical, spectral, brightness characteristics of the image. In most practical applications, quality is seen as a measure of the proximity of two images: the converted image and the original image. With this approach, it is possible to evaluate both the subjective degree of similarity of images, and to obtain objective estimates of the parameters of the image signals: the moments of the first and second order of the difference signal of the compared images, such conversion parameters as the signal-to-noise ratio, information compression ratios, and others.

A comparison of the classical method of replacing the least significant bit and the method with the addition of a preliminary coding step using a genetic algorithm is carried out by assessing the ratio of the peak signal-to-noise ratio (PSNR). Typical PSNR values in image processing range from 30 to 40 dB. A higher PSNR value corresponds to a better image quality in the background of noise. As can be seen in the tables III-IV PSNR values exceed 50 dB, which indicates high quality concealment of embedded information. Preliminary use of genetic coding reduces the chances of information disclosure, that is, it increases cryptostrength.

The experiment presented in Table III used a monochrome 8-bit image.

<table>
<thead>
<tr>
<th>Original image</th>
<th>Secret message</th>
<th>Stegofile</th>
<th>PSNR value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LSB</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSB with GA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>51.123 dB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>51.875 dB</td>
</tr>
</tbody>
</table>

As can be seen from the above example, stegofile has not only greater noise immunity in the communication channel, but also gives a gain in assessing the ratio of the peak signal-to-noise ratio, even on such complex objects.

Let us conduct experiment 2. The results are presented in Table IV.

<table>
<thead>
<tr>
<th>Original image</th>
<th>Secret message</th>
<th>Stegofile</th>
<th>PSNR value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LSB</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSB with GA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>51.145 dB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>51.163 dB</td>
</tr>
</tbody>
</table>

The use of a genetic algorithm with LSB steganography showed a better result than the use of the classical method of the least significant bit

The problem of noise reduction in images is one of the classical problems of image processing. Consider the difference image (Fig. 9), which is the difference between the original image and the stegofile.

This differential frame show the changes occurring with the image in the process of noise suppression. The desirable feature of the difference frame is their randomness: no visible correlation with the original image.

![Fig. 9. Difference image for experiment 1 (a), experiment 2 (b)](image-url)

From the pictures, you can see that there is a decorrelation of the image. Correlations between the counts do not remain in
the difference signal, which is a positive assessment of the genetic algorithm.

On the methodology, prospects of this direction, the study of the possibility of using genetic algorithms in steganography to create a method of persistent regular-singular (RS) model based on the mechanism of natural genetics and the theory of evolution is presented in [26].

The eye is a perfect invention of nature; it cannot compete with estimates such as PSNR and others. Therefore, some results considered in terms of objective assessments as the same, visually may be perceived differently. The functioning of automatic computer systems is fully subject to mathematical criteria, and only objective indicators evaluate the quality of their work. It is clear that the quality of images used in these systems should also be evaluated only by objective criteria.

VIII. SUMMARY

In this paper, in contrast to the classical genetic algorithm-prototype, an algorithm that allows you to take into account the variation of the error probability in different symbols for the binary code. The combination of genetic coding and LSB algorithm improves the noise immunity characteristics of the stego-system without introducing additional redundancy into it, which serves to ensure the safety of embedded watermark systems in IoT.

For stepsaway systems, IoT with LSB method is actually the distortion of information in a specific bit rate. The proposed modification of the genetic algorithm takes into account this fact and has greater noise immunity. The proposed model can use any existing method of steganography to embed the encoded message in the image.

Stego images are a relatively new object in the field of information concealment. Although much research has been done in this area, some issues remain to be explored.

REFERENCES