Cognitive based Data Classification in Image Steganography for Secured Data Transmission

Prof. Sangeetha K. N.[1] Dr. Usha B. A.[2]
[1]Assistant Professor [2]Associate Professor
[1]Department of Electronics & Communication Engineering [2]Department of Information Science and Engineering
[1]JSSATE, Bengaluru – 560060, India [2]BMSIT&M, Bengaluru – 560060, India

Abstract— Steganography is the art of hiding information within cover medium in such a way that it is undetectable. The main problem with steganography in the current system is that it does not differentiate between a highly sensitive document and non-sensitive document. Such a kind of steganography system fails when image manipulations are performed. A novel method to perform steganography based on the sensitivity of the data is developed to overcome the above stated problem. The sensitivity of the data is decided using the K-Nearest Neighbours (kNN) algorithm. As data classification is one of the most essential tools needed for data security such a sensitivity-based labelling is required before hiding data in a cover medium. One of the most important features of data classification is to find duplicate data to cut storage and backup costs. The decision to select the right algorithm is based on the security level assigned to data, size of data and the Peak Signal to Noise Ratio (PSNR) of the algorithms. With the kNN algorithm, an average accuracy of 90% has been achieved for data classification. In the data hiding aspect of the project, PSNR values ranging from 48dB to 70dB have been obtained. Data classification results in accurate results and the steganography that follows presents clear images without distortions.

Key words: Steganography, Discrete Wavelet Transforms (DWT), Discrete Cosine Transforms (DCT), Least Significant Bit (LSB), Data Hiding, And Data Classification

I. INTRODUCTION

Image Steganography is a technique of hiding information in digital images. In contrast to cryptography, it is not to keep others from knowing the hidden information but it is to keep others from thinking that the information even exists (Sharma, Mohd, & Sharma, 2013). Steganography became more important as more people joined the cyberspace revolution. Steganography is the art of concealing information in ways that prevents the detection of hidden messages. Steganography includes an array of secret communication methods that hides the message from being seen or discovered. In steganography, the message or encrypted message is embedded in a digital host before passing it through the network, thus the existence of the message is unknown. Besides hiding data for confidentiality, this approach of information hiding may be extended to copyright protection for digital media: audio, video and images (Amin, Salleh, Ibrahim, Katmin, & Shamsuddin, 2003).

A cognitive system is a one that performs the cognitive work of knowing, understanding, planning, deciding, problem solving, analysing, synthesizing, assessing, and judging as they are fully integrated with perceiving and acting (Usha, Srinath, Ravikumar, & Vismaya, 2015). In cognitive inspired image steganography, the algorithms employed are predicted for suitability. This is measured based on the signal to noise ratios for each of the various algorithms. This cognitive system uses data classification to categorize data on the basis of confidentiality.

The main objective of document categorization (Banchs, 2013) is to assign each document in a given data collection to a class or category, according to the nature of its textual content. In general, document categorization is used to directly address different practical tasks, such as spam filtering, press clipping and document clustering, just to mention a few; or, alternatively, it is used as a component of a larger system to tackle more complex tasks, such as, for example, opinion mining and plagiarism detection.

The present steganography applications does not provide document classification, this project aims to build an application which incorporate document cleaning, data classification in predefined categories so that suitable security level is assigned according to the sensitivity of data along with determining the best suitable algorithm for steganography to embed the data i.e., hide the data over an image using different steganography algorithms considering different parameters such as PSNR and MSE.

II. FUNDAMENTAL CONCEPTS

Data hiding techniques have constantly been an area for probable improvements like employing hybrid techniques combining different approaches. Cognitive systems are an upcoming branch of computer science which attempts to bring a human touch to decision making process. Traditional steganography fails to provide a personalized approach to the whole data hiding process since it provides a generic solution to the problem. Now that, steganography is collaborated with cognition, much importance is given to the semantic meaning of data in question. Depending on the security requirements particular to the application, the decision on the kind of steganography is made. This helps in reducing the overhead encountered in the computation and effectively the time complexity.

A. Text Mining

Text mining is an interrelated discipline that is based on information retrieval, information extraction, natural language processing, statistics, computational languages, and data mining (Ghosh, Roy, & Bandyopadhyay, 2012). Text Mining, sometimes also called as text analytics, may be defined as the process of extracting high quality information from text. It is achieved through continuous scanning of patterns and trends through ways called as statistical pattern learning (Ciarca, 2015). Structuring the input text involving parsing, addition or removal of some significant features, and subsequent insertion into a database is the first step of text mining. It is followed by analysing patterns within the structured data and finally evaluation and interpretation of the output. A combination of relevance, innovation and interestingness is what defines "High-Quality" in text mining. Text mining algorithms is either implemented using classification algorithms, association algorithms or clustering algorithms.

B. Document Classification

Document Categorization, in the field of computer science, is the task to either manually or algorithmically assign a document to one or more categories. The documents are classified either according to subjects or according to attributes like document type, magic bytes etc. Automatic Document Classification (Joshi & Nigam, 2011) is of three types: supervised classification where correct information on classification is provided by an external agent, unsupervised classification or document clustering where no external information is used for classification, semi supervised classification where parts of document are classified under the external agent's surveillance.

Content based classification is the one in which the weight or priority is given to a particular subject or word in the document to classify which category does it belong to. It could be the number of times a given word occurs in document.

The various techniques for document classification include: Expectation Maximization, Naïve Bayes Classifier, Latent Semantic Indexing (La Fleur & Renstrom, 2015), K-nearest neighbour classifier, decision trees (C3.5), Frequent Pattern Tree, Multiple Instance Learning etc. Document Classification methods find its application in fields like spam filtering, email routing, language identification, readability assessment, sentiment analysis.

C. Steganography

It is a practice of hiding a file, message, image, or video within other file, message, image or video. The advantage of steganography over cryptography is that the secret message does not catch attention as an object of investigation. Embedding data (Usha, Srinath, Narayan, & Tushara, 2014) in any medium requires two files: the first one being the cover file and the other being the secret message. Secret message may be either plain text, cipher text or an image. The file obtained after embedding is called the stego file. There are two most widely used image steganography techniques: Spatial Domain and Transform Domain.

In Spatial Domain method secret bits are embedded straight away in the cover file. Most widely used spatial domain method is Least Significant Bit (LSB) (Thangadurai, 2014). In this, the data to be hidden is embedded in the least significant bit of the cover image file. There are two types under it: LSB Matching & LSB Replacement. In LSB Matching if the message is same as the least significant bit of the cover image's pixel then it is remain unchanged otherwise it is incremented or decremented by one in a random fashion. In case of LSB replacement simply the least significant bit of cover pixel is replaced with the message to be protected.

In the case of transform domain technique, hiding of secret bits is carried out in the significant portions of the cover image file. The various methods under this technique are Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) (Bansal & Chhikara, 2014). In DCT, every colour component in the JPEG image format makes use of discrete cosine transform to make transformation from successive 8x8 pixel blocks of image into every 64 DCT coefficients. Under DWT, secret data is stored the least important coefficients of each 4x4 transformed blocks of pixels.

III. IMPLEMENTATION OF COGNITIVE INSPIRED IMAGE STEGANOGRAPHY

The most crucial part of a project is its implementation which is based on several important decisions such as selection of platform, the language used, etc. To run the project smoothly, several factors such as the real environment in which the system works, the speed that is required, other implementation specific details, etc. have to be taken care of.

The MATLAB platform has been selected to perform data classification as well as steganography. MATLAB being a mathematical scripting language is used to perform complex mathematical and graphical operations. It is also platform independent making it easy to compute, keeping the code secure and robust.

The following steps have been executed to get the results discussed in the next section.

The first step is to load the datasets into the system. This will help train the classifiers. A total of 600 samples are used for training the classifier.

The next step is to classify data file into different classes using kNN algorithm based on sensitivity of the document as per datasets. As a result of sensitivity based classification, a security level is assigned to the document. This assignment helps to determine the need for a robust algorithm for steganography.

After data file is classified, all the algorithm are executed for input data file and cover image. This gives a tabulated result of PSNR values which the user can see.

Suitable algorithm for steganography is selected on the basis of PSNRs, sensitivity of document, payload and robustness. User customization is provided to select a preferable algorithm.

Data hiding then takes place using selected algorithm and the stego-image is generated.

Stego-image is sent to receiver side and then data extraction takes place using same algorithm. Finally, data file is retrieved from stego-image having no noise at receiver side.

IV. EXPERIMENTAL RESULTS & ANALYSIS

This section lists the results of the project and the inferences made from the testing results. The evaluation metric have been listed and the results have been accordingly quantified.

A. Evaluation Metric

The main metric in the project is the PSNR values calculated based on the user security requirements, payload capacity, invisibility and robustness of the given file depending on which the most suitable image steganography algorithm is selected.

B. Performance Analysis

To check the accuracy of the application and to analyse the results obtained for a given input two cases are shown, one using default options and one with the user requirements.

The first set of results was obtained using default options suggested by the application. User enters the data file (category 6) then as the document requires less security its payload is less, invisibility is less and robustness is also low and for this combination the system is trained to select an algorithm which offers lesser security. But as per other case, if user want to change security level of document i.e. change of algorithm, then algorithm used for steganography is changed as per user request. This makes application more flexible and user friendly.

The Table 1 shows a comparison of different steganography parameters with respect to different algorithms. DWT, DCT and SS have high robustness, they are used to hide very sensitive data. Also, SS can handle high payload, so larger quantity of data is embedded. Though LSB and its variations are vulnerable to image manipulations, they have a least execution time.

	DWT	LSB	DCT	SS
Payload	LOW	HIGH	LOW	HIGH
Robustness	HIGH	LOW	HIGH	HIGH
PSNR	HIGH	HIGH	MED	MED
Invisibility	MED	HIGH	HIGH	HIGH

Table 1: Comparison of Steganography Parameters with respect to Algorithms Implemented

The Table 2 shows a comparison of PSNR values of all algorithms based on different cover image dimensions. The dimensions are taken in pixels. In Figure 1, PSNR of different algorithms for different Cover image dimension has been shown. It is observed that PSNR of spatial domain methods remains almost same for different cover image dimensions, but for transform domain methods PSNR varies for different cover image dimensions.

From the experimental analysis performed it may be infer that the DCT, DWT and SS algorithms have high robustness, so that they may be used to hide sensitivity data. Though LSB and its variations are vulnerable to image manipulations, they have a least execution time. Also, it is inferred that PSNR of spatial domain methods remain almost same for different cover image dimensions, but for transform domain methods PSNR will vary for different cover image dimensions. From our results it is also observed that execution time of Spatial Domain algorithms is very less as compared to Transform Domain embedding algorithms.

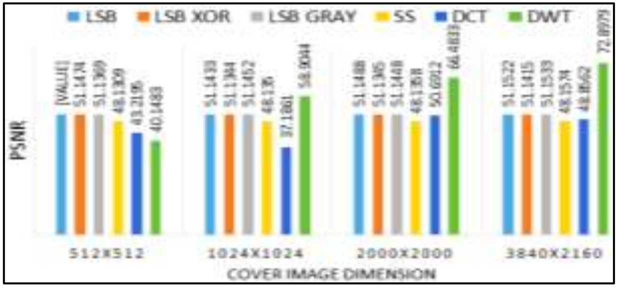

Fig. 1: Comparison of PSNR of Algorithms with Different Cover Sizes

	512* 512	1024* 1024	2000* 2000	3840* 2160
LSB	51.15	51.14	51.14	51.15
LSB XOR	51.14	51.13	51.13	51.14
LSB GRAY	51.13	51.14	51.14	51.15
SS	48.13	48.135	48.13	48.15
DCT	43.21	37.18	50.69	48.85
DWT	40.14	58.90	66.48	72.89

Table 2: PSNR of Algorithms WRT Cover Sizes

V. CONCLUSION

In a world where data thefts and attacks are increasing the need for data security is highly required. The main issue with steganography in the current system is that it fails when image manipulations are performed. This is overcome by performing a sensitivity-based labelling. The sensitivity of the data is decided using the kNN algorithm.

This work aims at suggesting the best suitable algorithm for a given data based on vectors like payload, invisibility and robustness. The selection of suitable algorithm is based on the sensitivity-based data classification. Robustness is assigned according to the data label. An algorithm is selected which provides the required robustness and can handle the payload of the entered data. Steganography is performed using the selected algorithm. kNN classification algorithm shows avg. accuracy of 92% and steganography results in PSNR ranging from 48dB to 70dB.

The above work can be extended to take care of multimedia files for data embedding. As only text classification is performed, it can also be extended to image classification. One of the most important feature that can be added is parallelism for data classification as well as steganography.

REFERENCES

[1] M. K. Sharma, N. Mohd and A. K. Sharma, “An Image Steganography Technique with High Hiding Capacity based on 24 Bit Color Image,” International Journal of Engineering Sciences and Research Technology, vol. 2, no. 11, pp. 3314-3319, 2013.

[2] M. M. Amin, M. Salleh, S. Ibrahim, M. R. Katmin and M. J. Shamsuddin, “Information Hiding Using Steganography,” in 4th National Conference on Telecommunication Technology, Johor, Malaysia, 2003.

[3] B. A. Usha, N. K. Srinath, C. N. Ravikumar and S. P. Vismaya, “Cognitive Prediction of the Most Appropriate Steganography Approach,” International Journal of Computer Applications, vol. 121, no. 8, pp. 58-63, 2015.

[4] R. E. Banchs, Text Mining with MATLAB, 1st ed., New York: Springer, 2013.

[5] S. Ghosh, S. Roy and S. K. Bandyopadhyay, “A tutorial Review on Text Mining Algorithms,” International Journal of Advanced Research in Computer and Communication Engineering, vol. 1, no. 4, pp. 223-233, 2012.

[6] C. A. Ciarca, “Statistical point pattern matching technique”. US Patent 9159164, 13 October 2015.

[7] S. Joshi and B. Nigam, “Categorizing the document Using MultiClass Classification in Data Mining,” in International Conference on Computational Intelligence and Communication Networks, Gwalior, 2011.

[8] Conceptual Indexing using Latent Semantic Indexing: A Case Study., 2015.

[9] B. A. Usha, N. K. Srinath, K. Narayan and C. K. Tushara, “Analysis of Data Embedding Technique in Image Steganography,” International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 6, pp. 54-59, 2014.

[10] K. Thangadurai, “An Analysis of LSB based Steganographic Techniques,” in International Conference on Compute Communication and Informatics, Coimbatore, 2014.

[11] D. Bansal and R. Chhikara, “A Study on Steganography Techniques,” International Journal of Engineering Research and Technology, vol. 3, no. 2, pp. 483-487, 2014.