IJTNR International Journal of Technology and Research

ENTROPY BASED DATA HIDING ON DOCUMENT IMAGES APPLIED ON DRDM APPROACH

Sadia Aslam
Departement of Software Engineering,
Faimah Jinnah University,
Islamabad, Pakistan
s4sadia_aslam@yahoo.com

Khurram Saleem Alimgeer
Center for advance studies in Telecommunication(CAST),
COMSATS Institute of Information Technology,
Islamabad, Pakistan
khurram_saleem@comsats.ede.pk

***Abstract*—This paper presents a data hiding application for document images providing security. Document images should be in bitmap format as well as grayscale in nature otherwise the developed system will automatically change it to grayscale. The system accepts any grayscale document image as an input and applies the developed technique on it. Developed technique is based on Entropy Based that selects the suitable areas in an image for that data insertion purpose. Output produced by the system is the processed document that is data embedded. Processed document should not look different from the original image means its visual quality should be high. The system is successfully tested for the test cases generated to check the effectiveness, quality and validation of the system. The system is developed using Matlab 7.7.0 (R2008b) for the development of front end as well as the back end.**

***Index Terms*—Component, formatting, style, styling, insert.** *(key words)*

I. INTRODUCTION

Authentication is the process of determination whether someone or something is, in fact who or what it is declared to be. It is an attempt to verify the digital identity of user. Document authentication provides good evidence of the substance of the electronic transaction to falsify or alter without detection of the signed object or the signature.

A technique of the watermarking inserts a number of data into an image which is hidden in order to perceive any hateful image fluctuation. Proposed technique uses Entropy measure and DRDM (Distortion Reciprocal Distortion Measure) approach for document authentication in grayscale document images and removes the limitations of existing technique.

Purpose of document:

Aim of this research is the comparison of existing techniques for the authentication of binary document images and development of new one. The features as well the limitations of four techniques are discussed in detail and a new techniques combining the benefits of existing ones and removing their limitations is designed and developed.

Background knowledge

Nature of Digital Documents

For the images in which the pixels take value from only a few potential, hiding data without causing noticeable artifact becomes harder. In particular, flipping black or white pixels that are not on the margin is likely to establish noticeable artifact in binary images. For a binary image, the human perceptual factor can be taken into report by studying each pixel and its instant neighbor to found a score of how hidden a change on that pixel will cause. The score is between 0 and 1, with 0 representing no flipping. Flip the pixels with top score usually introduce less artifact than flipping a lesser one.

The nature of most text images is binary having a forefront and backdrop color. The forefront can be the characters of dissimilar font sizes in text documents, scripts written by hand and statistics included in a bank cheques, or shape and signs in the drawing of engineering and architectural. Some documents have many gray levels and colors, but the amount gray levels and colors is typically few and each local region usually has a unvarying gray level or color, as different to the dissimilar levels and colors that are establish at single pixels of a continuous-tone image. Some documents having dual nature also contain grayscale images characterized as half-tone images, e.g. the photos in a newspaper. In such images, n x n binary sample are used to fairly accurate gray level values of a gray scale image, where n typically ranges from 2 to 4. Spatial incorporation of the well binary pattern within local region and perceives them as dissimilar intensities performed by the human visual system.

Data Hiding

Hiding the data show a class of process used to insert data, such as exclusive rights information into a variety of forms of medium such as an image, audio, or text with a least quantity of perceivable deprivation to the host signal. In recent digital watermarking, an idea of data hiding means a method to insert a series of bits in an image with small image weakening. A watermarking method makes use of a data-hiding method to embed some information in the original image, in classify to construct a claim about the image in the upcoming.

Ownership Protection

Ownership claim or patent security can be done by embed the information about the cause and the exclusive rights holder of the data as the watermark in the data. For this purpose it is important with a strong watermark that is hard to rub out or misrepresent so it can't be predictable. A watermark representing rights is set in the resource include CD

foundation. The watermark, which is already known only to the exclusive rights owner, is predictable to carry on ordinary dealing out and intended assault so that the holder can explain the existence in case of clash to express his/her rights of this watermark.

Authentication

A set of less important data is implanted in the resource like compact disk earlier, and used to decide if the original medium is tamper or not changed latterly. Strong watermarks are normally used for exclusive rights claim. They must not be simply impassive and have to oppose ordinary image utilization actions like to scale, to crop, loss comprehension, etc [1]. The strength against making the watermark unnoticeable is not a fear as such motivation from the point of view of the attacker is not there. On the other hand, falsify a suitable certification watermark in an illegal or tempered medium resource must be not permitted.

Weak or certification watermarks are simply degraded by doing any image-processing method. However, watermarks for inspection the image reliability and validity.

Fingerprinting or Labeling

Labeling aim at identifies each one with permission scattered duplicate of the data with a unusual watermark, a fingerprint, and thereby enable tracing of prohibited repetition and allocation. The watermark in this claim is used to copy the creator or recipient of a specific duplicate of the resource like CD. For example, dissimilar watermarks are implanted in dissimilar copies of resources like CD previous to distribute to a variety of recipient. The toughness next to obliterate and the capability to express a number of bits which are non-trivial are necessary.

Access control and Copy control

Duplicate avoidance and organize is about scheming what can be prepared with a scattered duplicate of the data. This is hard to get in unwrap atmosphere, and will necessitate manage over both the information with the implanted watermark and the tool used to examine the data and do something in accordance to the directions in the watermark.

Annotation

When the intention is for annotation or confirmation, digital watermarking is often called data hiding and the two terms of digital watermarking and data hiding are used by many authors interchangeably. Most of the techniques related to data hiding of the images in the literature are projected by binary or color images, while regarding grayscale images data hiding is only addressed by a few authors.

The implanted watermark in this request is predictable to express as many bits as probable with no use of the original unmarked copy in finding. While the strength against intended attack is not necessary, a sure quantity of strength next to ordinary dealing out like lossy compression may be preferred.

Motivation

A diversity of techniques has been proposed related to data hiding and digital watermarking. Nevertheless, the majority of the methods developed today is for binary and color images, where the color value of a selected group of pixels is changed by a small amount without causing visually perceptible artifacts. These techniques cannot be directly applied to grayscale document images where the pixels have different gray level values.

Therefore a dissimilar category of embedding method must be developed. This would have significant purpose in a broad diversity of text images that characterize as dual forefront and backdrop; e.g. depository checks, economic instrument, lawful papers, engineering maps, architectural drawings, road maps, driver licenses, birth certificates, digital books, etc. In anticipation of freshly, there has been little work on watermarking and data hiding techniques for grayscale document images. Over the previous few years, a rising but imperfect number of papers have been in print proposing new techniques and ideas for document image watermarking and data hiding.

The aim of this project is to implement a technique that can be applied to document images in order to provide authenticity to them without causing any visually noticeable different.

Research statement

The proposed project will cover “Entropy Based Data Hiding on document images using DRDM approach” using Matlab 7.7.0 (R2008b).

Objectives

1. To analyze existing authentication technique of digital documents in order to identify deficiencies in them.
2. To remove deficiency in existing technique and implemented updated version.
3. To propose a new technique that combines the advantages of existing techniques and removes their limitations.

Scope of the document

Paper is worth doing because the implemented technique will be able to provide authenticity to digital documents. The document covers the analysis of three different techniques, their limitations and the development of new one.

Outline

Section II gives the literature review of the topic, work that has been done so far on this topic and what is the state of the art today. In Section III, several data hiding techniques are presented along with their strengths and weakness. Comparative analysis of these techniques is also given at end of this section. Section IV provides the proposed model. Section V concludes the document with the findings and results and future directions.

II. Literature

In the literature, [10] there are many Authentication watermarking techniques by template rankings for continuous tone images. There are many data hiding methods for binary and halftone images. Nevertheless, only recently some secure Authentication watermarking by template rankings for dual images have been projected. We mean by protected Authentication watermarking by template rankings a method that has two properties: (1) it must identify *any* visually important image modification (both unintentional and spiteful); (2) its protection have to not stretch out on the confidentiality of the algorithm but only on the privacy of the

key. Hence, a secure Authentication watermarking by template rankings frequently relies upon cryptography. A cryptography-based Authentication watermarking by template rankings for dispersed-dot halftone images named AWST (Authentication Watermarking by Self Toggling) has been recently proposed [9]. It can be used with secret- or public-key ciphers. However, when this technique is applied to binary document images, visible salt-and-pepper noise appears. Another cryptography-based AWT for binary document images named AWTR (Authentication Watermarking by Template Ranking) has also been proposed. Document images watermarked by AWTR present excellent visual quality. The secret-key AWTR is secure. However, the public-key AWTR cannot securely authenticate "small" images due to the watermark adulterating technique called "parity attack."

The paper proposes a new AWT for binary document images, named AWTC (Authentication Watermarking by Template ranking with symmetrical Central pixel). It is completely immune to parity attacks and consequently can authenticate even "small" images using either secret- or public-key ciphers. Images marked by AWTC do not present visible salt-and-pepper noise. This technique can detect *any* image alteration, even a single pixel flipping. We did not apply any perceptual distortion measure to quantify the quality of watermarked images, because this analysis is beyond the scope of this paper. However, the pattern grade used in AWTC can be modified to reduce the misrepresentation in accordance to a particular perceptual copy. [10] Presented a new verification watermarking technique for binary images named Authentication Watermarking by Template Ranking with Symmetrical Central Pixel (AWTC). It can detect *any* modification of the watermarked image with the protection guaranteed by the cryptography supposition. Both covert- and public-key versions of Authentication Watermarking by Template Ranking with Symmetrical Central Pixel are totally protected against similarity attacks. Furthermore, just the pixels having low-visibility are flipped by the Authentication Watermarking by Template Ranking with Symmetrical Central Pixel watermark addition, resulting in watermarked images with admirable visual superiority.

In [11] Wu *et al.* suggest a data hiding algorithm for digital dual images. A set of laws is used to compute the score of flipping of pixels and shuffling is in use to managing the difficulty of not level implanted capability. Newly, Lu *et al.* [11] propose a DRDM (Distance-Reciprocal Distortion Measure) for binary document images. When functional to binary document images it has been shown that this measure has much better correlation with human visual observation than PSNR (peak signal-to-noise ratio). [1] Recommend a protected data hiding algorithm based on the DRDM measure for the purpose of authentication of digital documents in a binary image format. We unite a 2-D changing method with an odd-even inserting method and utilize the DRDM method to select the suitable pixels to turn over. Experiments show that the algorithm has good quality imperceptibility and can be used for alter validation.

[11] Digital text image dispensation is getting additional attention recently. Digital document images are fundamentally double images. In some binary document image applications, such as watermarking and data hiding, visual misrepresentation may be present, and it is necessary to determine such misrepresentation for performance evaluation or assessment [1]. There are two ways to compute visual distortion, as discussed in [2]. One is subjective measure, and the other is objective measure. Subjective measure is expensive, while it is significant, since a human is the eventual observer. On the other hand, objective measure is repeatable and easier to put into practice, while such a measure does not forever be in agreement with the subjective one. In this letter, we propose an objective distortion measure for binary document images that is based on human visual observation. Binary document images at this time submit to binary images that have pointed dissimilarity of black and white and there are clear margins between black and white areas in the images. The distance between pixels is found to play an essential role in human observation of misrepresentation or distortion in these images. Hence, the reciprocal of distance is used to measure visual distortion in digital binary document images. Subjective testing results show a good association between the projected objective measure and human visual perception.

Predetermined Partitioning of Images

This set of technique separation an image into preset blocks of size m x n, and calculates some pixel information or statistics from the blocks for inserting data. They can be applied to dual text images in common; e.g. papers with formatted content or manufacturing drawings.

State of the art today

There has been a increasing but incomplete number of papers proposing new techniques, methods and thoughts for document images watermark and data hiding. There are many techniques available for digital color and binary images in general.

Latest work done in this field is technique by Hae Yong Kim [10], Public Key Data Hiding Technique based on DRDM, Data hiding technique based on DRDM (Distance Reciprocal Distortion Measure) [12] and Entropy measure with AWTC approach. [9]

III. Data hiding Techniques

Technique # 1: Technique by Hae Yong Kim[10

Main Idea

An AWT for binary document images is proposed named AWTC. It is totally protected to equality attacks and as a result can verify even tiny images using covert or public key ciphers. Images marked by AWTC do not have perceptible noise such as salt and pepper noise. No perceptual misrepresentation measure is functional to calculate the excellence of watermarked images.

Features

- AWTC can detect *any* alteration of the watermarked image with the security assured by the cryptography theory.
- Only the low-visibility pixels are turned by the AWTC watermark embedding having good visual quality.
- Both covert- and public-key versions of AWTC are fully protected against equality attacks.

Graphical Model

AWTC Insertion

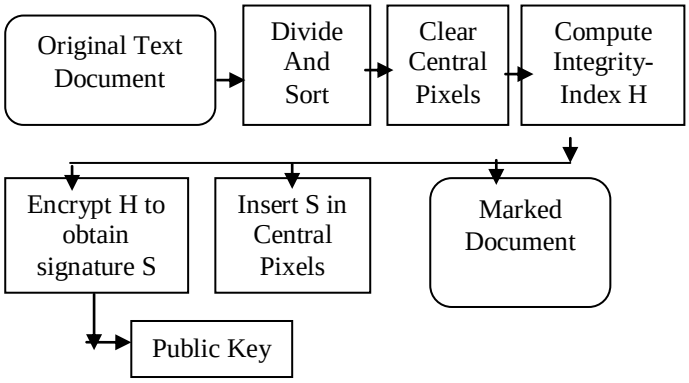

Fig. 1. Flow chart of AWTC Insertion

AWTC Verification

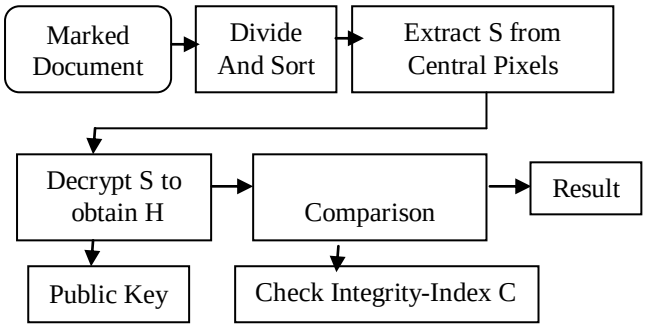

Fig. 2. Flow chart of AWTC verification.

Advantages

1. It can validate even tiny images by means of any covert or public key ciphers
2. This technique can detect any image alteration.
3. Both key versions of AWTC are totally protected in opposition to equality attacks.

Limitations

1. No perceptual misrepresentation measure is applied to calculate the excellence of watermarked images

Technique # 2: Public Key Data Hiding Technique based on DRDM

Main Idea

This technique is revised version of AWTC. Visual scores were being used in AWTC for selection of pixels to flip and hide data but in this technique distance reciprocal distortion measure (DRDM) of pixels is used in order to choose pixels for hiding data. Rest of the procedure is similar to AWTC.

Features

- The algorithm is based on DRDM that gives a well-organized method to choose the pixels to turn over in embedding.
- The misrepresentation due to turning over or flipping the pixel values is computed online and capable to obtain the outcome of a huge area of neighbor pixels into accounts.
- Data is inserted in central pixels of 3 x 3 blocks with low DRDM values.

Algorithm

3.2.4.1 Insertion

1. Divide the binary image into regular 3 x 3 blocks; unfinished pieces at image margins are not needed. Scan the image in sequence of raster.
2. Clear the central pixels of the blocks.
3. Using a hashing function, compute the hash function (say A) of the cleared image. Encrypt A to obtain an authenticate signature (say S).
4. Compute the DRDM for the central pixels of all the blocks (treating all the blocks as single image).
5. Insert S in the central pixels of n blocks with respect of DRDM, where n is the length of S.

Extraction and Verification

1. Divide image in blocks as before
2. Clear the central pixels and compute hash value (say B) using the hashing function.
3. Compute DRDM of central pixels and extract S from central pixels with respect to DRDM value as in insertion process. Decrypt S to obtain hash value A.
4. Compare A and B if they are same, image is verified otherwise it is unauthentic.

Advantages

1. DRDM gives a competent way to choose the pixels to turn over in inserting.
2. The misrepresentation due to turning over or flipping the pixel values is computed online and capable to obtain the outcome of a huge area of neighbor pixels into accounts.
3. Altering can be noticed in the removal.

Limitations

1. There is a problem of uneven embedding capacity for smaller values of m while calculating DRDM.

Technique # 3: Entropy Measure with AWTC approach

Main Idea

The technique provides efficient and fast data hiding in binary document images with less computations and time required. The technique will also overcome the drawbacks of existing data hiding techniques. The technique is developed using two existing concepts that are entropy measure and AWTC. Document verification feature after data hiding is also provided.

Features

Provide data hiding and document authentication of binary document images.

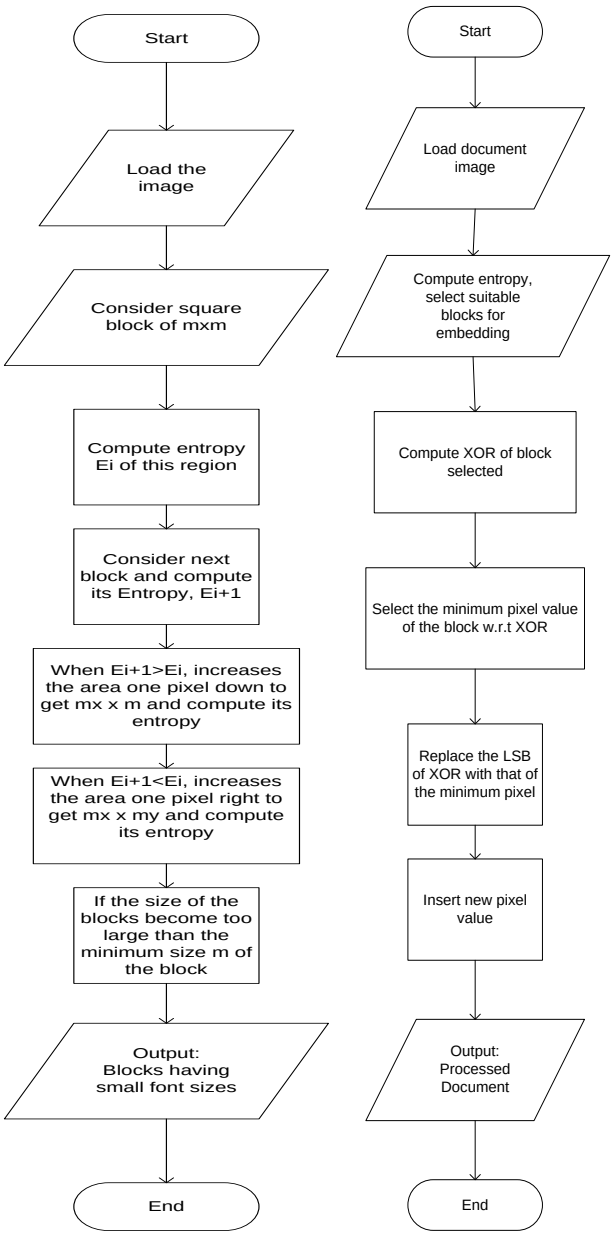

Fig. 3. Flow chart of a)Entropy measure b)Data insertion.

Based on the method that has less computational complexity as compare to other techniques.

- Based on Entropy Measure and AWTC concepts.
- Entropy Measure selects automatically select the blocks having less distortion thus requiring less calculations and time.
- AWTC is used for data hiding in the blocks selected by Entropy Measure

Advantages

1. Require less computational cost and time than other techniques.
2. Provide combined benefits of all the existing techniques.
3. Overcomes the limitations of existing techniques.

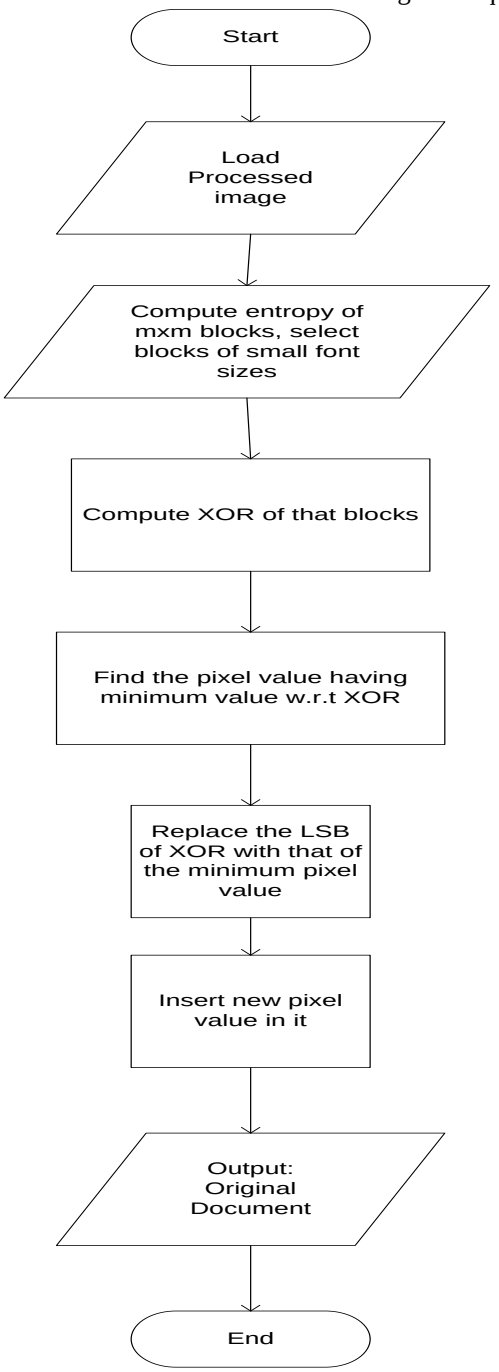

Fig. 4. Flow chart of Data Extraction

Limitations

1. The number of bits to be watermarked is less in number.

Comparative Analysis

The techniques are compared on the basis of certain properties; these properties distinguish the strong points and weak points of the techniques. The comparative significance of each property depends on the function performed by

different techniques. The features used in the analysis are as follows:

Capacity of Hiding

Capability of hiding is the range of information that can be hidden comparative to the size of the secret. A large capacity of hiding permits the use for a message of predetermined size, and thus decreases the bandwidth necessary to broadcast the stego-image.

Visual Quality/Perceptual Simplicity

The work of hiding the message in the cover needs some noise inflection or misrepresentation of the secret image. It is essential that the embedding occur without significant deprivation or failure of perceptual excellence of the secret. To preserve perceptual simplicity in an embedded
watermark for exclusive rights security is also of supreme important because the reliability of the unusual work must be retained. For purpose where the perceptual simplicity of implanted data is not serious, permit extra alteration in the stego-image can enhance capacity of hiding, strength, or both.

Robustness

The ability of embedded data to remain intact if the stego-image go through alteration, such as linear and non-linear straining (filtering), accumulation of casual noise, sharpening or blurring, lossy comprehension, scaling and rotations, cropping or decimation, and conversion from digital to analog from and then reconversion back to digital from (for instance in the case when a hard copy of a stego-image is printed and then a digital image is produced by consequently examine the hardcopy.) is referred as robustness.

Robustness is critical in exclusive rights security watermarks because pirates will attempt to pass through a filter and demolish a few watermarks inserted in images. Anti-watermarking software is previously available on the Internet and has been shown efficient to eliminate some watermarks. These techniques can also be used to demolish the message in a stego-image.

Tamper Resistance

Away from strength to damage, tamper-resistance (alteration-conflict) refers to the complexity for an invader to modify or falsify a message once it has been inserted in a stego-image; for an example a pirate substituted a patent mark with one argues officially permitted rights. Applications that demand high strength generally also insist a tough degree of tamper resistance. In a exclusive rights security application, getting good alteration conflict can be complex because a exclusive rights is efficient for many years and a watermark should stay challenging to tampering even when a pirate tries to change it using computing technology decades in the upcoming.

IV. PROPOSED MODEL

Proposed Model

Main Idea

An Entropy based data hiding approach for grayscale document images is proposed. The region is automatically selected by computing entropy measure. And distance reciprocal distortion measure (DRDM) of pixels is computed of the marked document. Calculating DRDM of the marked image to validate the document image and to verify that document has good visual quality.

Features

- Entropy Measure selects automatically select the blocks having less distortion thus requiring less calculations and time.
- Based on Entropy Measure and find out DRDM value instead of computing PSNR.
- Based on the method that has less computational complexity as compare to other techniques.
- The algorithm after insertion computes DRDM which gives the actual result of human visual perception and tells about the quality of the document
- Provide good data hiding of grayscale document images.

Advantages

1. Provide combined benefits of all the existing techniques.
2. Overcomes the limitations of existing techniques.
3. DRDM provides an efficient way of good visual quality of document image.
4. The distortion due to flipping is calculated online and able to take effect of a large area of neighbor pixels into accounts
5. Marked image has good quality and tampering can be detected in the extraction.
6. Entropy Measure selects automatically select the blocks having less distortion thus requiring less calculations and time.

Comparison with other techniques

In the previous technique of [9], proposed method reduces perceptual misrepresentation due to inserting and allows watermark removal without the necessity of any side information at the decoder end. They change the chosen block in such a way that the customized block does not appear perceptually different from the original block of the image but very less blocks are changed. Watermark only inserted in the blocks having font size which occurs a lot in the document. The no. of bits which were encoded is very less in number. The technique given in [13], there is a problem of uneven embedding capacity for smaller values of m while calculating DRDM. The embedding capacity is not proper but is uneven. Including the uniform blocks also creates the problem. The technique of [14], also have limitations because in this the bits which is watermarked are very less in number and there is no perceptual misrepresentation calculation is applied to compute the excellence of the image.

In comparison with all the techniques, proposed technique has number of watermarks inserted into the document without having change in visibility of the document. Similarly the blocks will be automatically selected depending upon the entropy of those blocks having small font sizes, so specific blocks will be selected for embedding watermarks. And also proposed technique measures Distance Reciprocal Distortion Measure for analyzing and quantifying the visual quality of the document image. It has been shown in [11] that DRDM is

a better distortion measure for document images than PSNR. From the experience, a d of value below 0.2 is generally considered of good quality.

V. Conclusion

A data hiding technique for grayscale images is developed for the purpose of security of document images. The algorithm is entropy based that provides an efficient way to choose the pixels to flip in embedding. Only the pixels of blocks having small font sizes are flipped after computing their entropy measure, resulting in processed documents with excellent visual quality. Hiding capacity is very good length of hidden data varies with size of document so this system can be used even for documents of very small sizes. System is performing efficiently.

References

[1] L. Haipang , Alex C. Kot, and Yun Q. Shi , "Distance-Reciprocal Distortion Measure for Binary Document Images," in *IEEE Signal Processing Letters*, Vol. 11,No. 2,February 2004.

[2] Y. Huijuan , Alex C. Kot, "Text Document Authentication By Integrating Inter Character And Word Spaces Watermarking" ", *The 2004 IEEE* International Conference on Multimedia and Expo. (ICME'2004), June 26-30, 2004.

[3] Y. Huijuan , Alex C. Kot, "Data hiding for Bi-level Documents Using Smoothing Techniques", *The 2004 IEEE* International Symposium on Circuits and Systems(ISCAS'2004), Vol. V, pp. 692-695, May 23-26, 2004.

[4] Young-Won Kim, Kyung-Ae Moon, and Il-Seok Oh, "A Text watermarking Algorithm based on Word Classification and Inter-word Space Statistics", *IEEE, Seventh International Conference on Document Analysis and Recognition Volume II*, August 03-06,2003, Edinburgh, Scotland.

[5] H. Ding , Hong Yan "Interword Distance Changes Represented by Sine Waves for Watermarking Text Images", *CirSysVideo*, Vol 11, No. 12, December 2001.

[6] K. Jonathan Su, Frank Hartung, and Bernd Girod ,"Digital Watermarking of Text, Images and Video Documents", Computer Graphics International 98, Hannover, Germany, June 1998.

[7] C. Nopporn, "Document Image Data Hiding Using Character Spacing Width Sequence Coding", in ICIP'99, pg II:250-254.

[8] S. Pavan, Sridhar Gangadharpalli, Sridhar V., "Multivariate Entropy Detector Based Hybrid Image Registration Algorithm", *IEEE International Conference on Acoustics, Speech, and Signal Processing*,March 18-23, 2005 (ICASSP).

[9] Swetha Kurup, Sridhar G., and Sridhar V. "Entropy Based Data Hiding for Document Images", Proceedings of world academy of science, engineering and technology volume 5 April 2005 ISSN 1307-6884

[10] *Hae Yong Kim , "*A new public-key authentication watermarking for binary document images resistant to parity attacks", *0*-7803-9134-9/05/$20.00 ©2005 IEEE

[11] H. Lu, J. Wang, A. C. Kot, and Y. Q. Shi, "An objective distortion measure for binary document images based on human visual perception," in *Proc. Int. Conf. Pattern Recognition*, vol. 4, Quebec, Canada, Aug. 2002, pp. 239–242.

[12] *Haiping Lu, Alex C. Kot and Jun Cheng, "*Secure Data Hiding In Binary Document Images for Authentication," School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798.

[13] Rakhshanda Yousaf, dissertation of department of software engineering "Data Hiding in Binary Document Images Based on DRDM for Authentication", BSE (FJWU) 2008.

[14] Asima Iqbal, Dissertation of department of software engineering "Data Hiding Technique for Binary Document Images Based on Entropy Measure with AWTC for Authentication", BSE (FJWU) 2009.