Entropy Based Data Hiding in Binary Document Images

Aihab Khan, Memoona Khanam, Saba Bashir, Malik Sikander Hayat Khiyal, Asima Iqbal, and Farhan Hassan Khan

Abstract—**This research paper has presented a data hiding technique for binary document images. Entropy measure method is used to minimize the perceptual distortion due to embedding. The watermark extraction is a blind system because neither the original image nor the watermark is required for extraction. The document image is similar to any other image. The proposed method discovers the specific regions where minimum distortion delay exists due to embedding. For embedding, the blocks that exist in the area of small font sizes are selected. Experimental results show that marked documents have excellent visual quality and less computational complexity.**

Index Terms—**Binary Documents, Controlled Dilation, Data Hiding, Document images, Entropy, Watermarking.**

I. Introduction

Digital media is getting more popular today. Many applications are based on binary document images like video and audio multi-level images. Most of existing image data hiding techniques has been used for grayscale or color images, but very few authors [1-3] have addressed the data hiding for binary images. Digital watermarking techniques are used for following purposes.

- Ownership protection
- Copy control
- Annotation and authentication of digital media.

AWTC (Authentication Watermarking by Template raking with symmetrical Central pixels) is a good technique with higher tamper resistance, robustness and good visual quality but with the limitation that no distortion measure is applied.

This research paper has presented a data hiding algorithm. It is based on entropy measure and it authenticates the digital documents for binary document image. Entropy measure is applied on the whole image. It selects the blocks that have minimum perceptual distortion for embedding. By applying this technique, all blocks in an image need not to be computed. It reduces the computational complexity as well as the response time and results in a good visual quality.

The organization of paper is such that Section 2 discusses the related work. Section 3 and 4 include an overview of AWTC and entropy measure. In section 5 the framework of proposed technique is presented. Section 6 and 7 show experimental results and conclusion respectively.

Manuscript received December 29, 2010; revised July 11, 2011.

Aihab Khan is with the Department of Software Engineering, Fatima Jinnah Women University, Rawalpindi, Pakistan. (aihabkhan@yahoo.com)

Memoona Khanam is with the Department of Electrical Engineering, Federal Urdu University of Arts, Science & Technology. Islamabad, Pakistan, (dr.mahayat@gmail.com).

Saba Bashir is with the Department of Computer Engineering, National University of Science and Technology. Islamabad, Pakistan. 12,(saba.bashir3000@gmail.com).

Malik Sikander Hayat Khiyal, (m.sikandarhayat@yahoo.com).

Asima Iqbal, (asima781@yahoo.com).

Farhan Hassan Khan,(mrfarhankhan@yahoo.com)*.

II. Related Work

Wu et al. [4] proposed an algorithm that is used for data hiding in digital binary images. The technique calculates the flipping scores using a set of rules. Uneven embedding capacity problem has been resolved using shuffling. Lu et al. [6] proposed a Distance-Reciprocal Distortion Measure (DRDM) for binary document images. The proposed technique has improved correlation with human visual perception as compared to PSNR (peak signal-to noise ratio), when binary document images are used [5].

H. Lu et al. [6] proposed a secure AWT (Authentication Watermarking Tecnique) scheme for digital binary images. Visual score based on pixel neighborhood is used to choose the pixels for embedding data in this scheme.

Kurup, et al. [7] proposed entropy based data hiding algorithm for binary document images. The proposed technique identifies the group of characters and other regions in the image. It identifies only those regions where the data can be hidden with minimum perception distortion.

K. Aihab, H.K.Sikander, Y.Rakhshanda [8] proposed a secure data hiding algorithm for binary document images. The technique is based on DRDM. It selects the pixels efficiently that are used to flip in embedding. Only the low visibility pixels are flipped after computing their distortion measure. It results in processed documents with excellent visual quality. By this experiment, even a single pixel flipping can be detected and document is marked to unauthentic.

III. Preliminaries

This section provides some information about entropy measure and AWTC.

Entropy Measure

1) Entropy Equation

$$H_k(a) = \sum pdf_k(i) \log (1/pdf_k(i)) \quad (1)$$

The above equation is used to calculate the entropy measure. The notations used in above equations are described such that pdf is probability density function that is used to determine the white and black pixels in an image, whereas k in subscript shows the block number. $H_k(a)$ is the entropy and k in subscript is the number of block under consideration. 'a' represents the regions covered by that image block. 'a' can has maximum two values, that are 0 and 1. 0 is used to represents the black pixels whereas 1 is for white pixels.

2) Entropy Algorithm

The algorithm given by [7] is as follows:

Consider a square block of size m*m.

A. Starting from top left corner, calculate the entropy E_i of first block.
B. Now consider next adjacent square block of same size and calculate entropy E_{i+1}.
C. Continue to calculate entropy of adjacent m*m blocks until $E_{i+1}>E_i$
D. When $E_{i+}1>E_i$, then

a) Increase the area of region by moving one pixel down to get a new region m_x*m where
m_x= m+<(x-direction)
b) Compute the entropy E_{xi} of this region.
c) Again increase the covered area in vertical direction and compute new entropy E_{xi}+1.
d) When $E_xi_{+1}<E_{xi}$, then 1.
e) Increase the area of region by moving one pixel towards the right to get a new region m_x*m_y where
f) m_y= m+. (y-direction).
g) Continue to increase m_y by < in a similar manner until E_{yi}+1.E_{yi}
h) The bounds of the region covered by the above process gives the size of character.

3) Review of AWTC

AWTC technique is used to detect the alternations in an image, even when a single pixel flipping exists. Salt and pepper noise is not visible in AWTC marked images. AWTC technique is resistant to parity attacks and as a result it can authenticate even small images using either secret- or public-key ciphers.

4) AWTC insertion algorithm

AWTC insertion algorithm proposed by [6] is as follows.

1. Consider a binary image Z. Divide this image into sequence of small images called v. The pieces of images should be non-overlapped. Sort v according to visual score of patterns.

2. Clear the first n central pixels of the sorted sequence v, where n is the size of adopted AS.

3. Compute the integrity-index H of this binary image Z by using a cryptographically secure hashing function. Encrypt the integrity-index H with the secretor private-key, obtaining AS S.

4. Insert n bits of S in n first central pixels of sorted sequence v, obtaining the watermarked image [6].

5) AWTC verification

The algorithm is as follows.

1. Consider a binary image Z. Divide this image into sequence of small images called v. The pieces of images should be non-overlapped. Apply the same procedure for sorting v as in insertion algorithm.

2. Select the sorted v and extract the AS S from n first central pixels. Decrypt S with secret- or public-key and obtain the integrity index H.

3. Clear first n central pixels of sorted v and compute the check integrity-index C of the now-cleared image Z, using hashing function.

4. If extracted integrity-index H and the check integrity-index C are same, then watermark is verified. Otherwise, the image is modified [6].

IV. Proposed Technique

The proposed technique is revised version of AWTC with introduction of entropy measures. Visual scores have been used in AWTC for selection of pixels to flip and hide data. In the proposed technique, entropy measure is used to choose pixels for hiding data.

Modified AWTC insertion algorithm

The proposed insertion algorithm with introduction of entropy is as follows.

Divide the binary image into regular 3×3 blocks and compute the entropy of each block.

The blocks that lie in small font characters are selected.

Clear the central pixels of those blocks.

Using a hashing function, compute the hash value (e.g H1) of the cleared image. Encrypt H1 to obtain an authentication signature (e.g S).

Insert S in the central pixels of n blocks where n is the length of S.

1) Graphical Model

Data insertion process requires document image as an input. Entropy measure and hash values are calculated to embed the signature in image. Data insertion is represented in Fig 1.

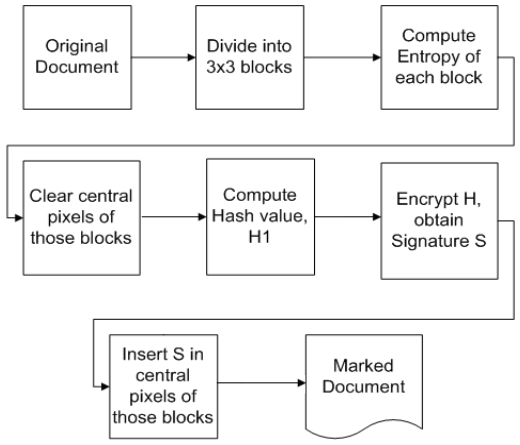

Fig 1: AWTC Insertion Algorithm

2) Pseudo code

The data insertion algorithm is described in Fig 2.

```
Algorithm: data_insertion
Input: Document Image, X
Key: K
i = index;
Output: Marked image, G

1)  For each 3x3 block, compute Entropy, E
2)  Select blocks lying in the region of small font sizes.
3)  for each [row,col]
        X(row+1,col+1) = 0
4)  H = hashindex()
5)  S = encrypt (H,K)
    X(row+1,col+1) = S(i)
```

Fig 2: Data Insertion Algorithm

3) Modified Extraction & Verification

1. Divide image into 3x3 blocks.
2. Compute entropy of each block and select the blocks with small font sizes.
3. Now remove the central pixels of those blocks and calculate the hash value (e.g H2) using the hashing function.
4. Extract S from central pixels with respect to entropy as described in insertion process. Decrypt S to obtain hash value H1.
5. Compare H1 and H2. If they are same, then image is authentic otherwise not.

4) Graphical Model

Data extraction process requires marked document image as an input. Entropy measure and hash values are calculated to extract the signature from image. Data extraction process is shown in Fig 3.

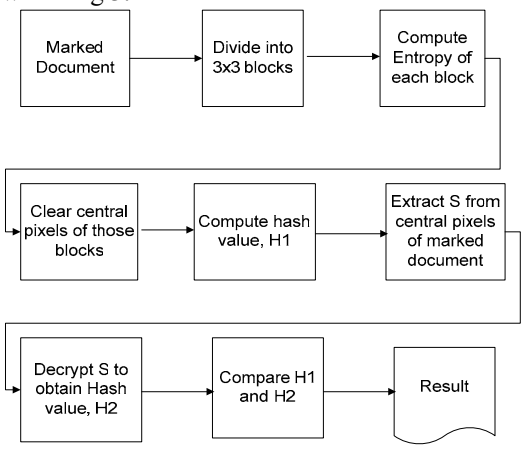

Fig 3: AWTC Extraction Algorithm

5) Pseudo Code

The data extraction algorithm is described in Fig 4.

```
Algorithm: data_extraction
Input: Marked Image, G
Key: K
Output: Original image, X

1) For each 3x3 block, compute entropy E,
      Select blocks lying in the region of small
      font sizes.
2) H1 = hashindex()
3) S = G(row+1,col+1)
4) H2 = decrypt (S,K)
5) Compare H1and H2
6) if(H1=H2)
   Output = original document
```

Fig 4: Data Extraction

V. Proposed Framework

Data flow diagram in Fig 5 represents the framework of proposed technique. A user can load image, hide data and save document. These tasks are represented in detail in the figure.

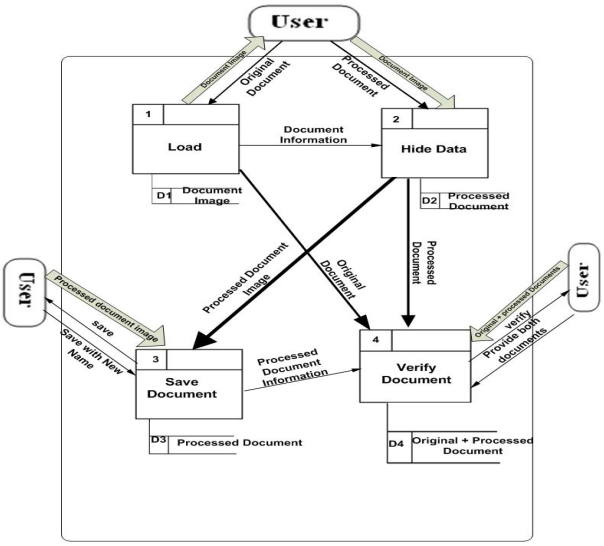

Fig 5: Proposed Framework

VI. Experimental Results

Fig 6 shows the original image under consideration. Following values are used for analysis.

- m = 3 i.e. the size of weight matrix for computing entropy.
- Key = 8.7e+037 which is converted into binary.
- Signature = 58 bits.
- Elapsed time for data hiding is 4.906000 seconds.

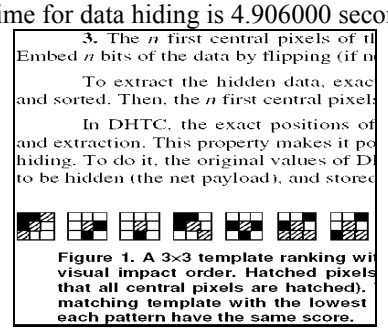
3. The *n* first central pixels of tl
Embed *n* bits of the data by flipping (if n
To extract the hidden data, exac
and sorted. Then, the *n* first central pixel
In DHTC, the exact positions of
and extraction. This property makes it po
hiding. To do it, the original values of DI
to be hidden (the net payload), and stored
Figure 1. A 3×3 template ranking wi visual impact order. Hatched pixels that all central pixels are hatched). matching template with the lowest each pattern have the same score.

Fig 6(a): Original Image

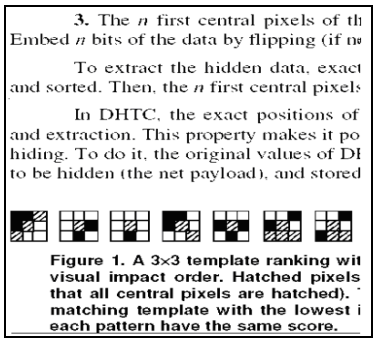
3. The *n* first central pixels of th
Embed *n* bits of the data by flipping (if n
To extract the hidden data, exact
and sorted. Then, the *n* first central pixels
In DHTC, the exact positions of
and extraction. This property makes it po
hiding. To do it, the original values of DI
to be hidden (the net payload), and stored
Figure 1. A 3×3 template ranking wit visual impact order. Hatched pixels that all central pixels are hatched). matching template with the lowest i each pattern have the same score.

Fig 6(b): Processed Image

Figure 6 (c): Difference image

Fig 7(a): Original Image

Fig 7(b): Processed Image

Figure 7 (c): Difference image

It is analyzed from the above results that the image after embedding has good quality and flipping is hard to perceive. Also that the computational cost is low and it makes the algorithm efficient.

VII. Conclusion & Future Work

This paper has presented a new approach for data hiding in document images. The proposed method identifies the group of characters and other regions in the image where minimum perception distortion causes the data hidden. Further the technique is analyzed by applying on different images.

It is clear from the experimental results that the proposed technique has high visual quality and requires less computational cost. Verification is performed in the absence of original image.

Future research should aim at addressing data hiding in color and grayscale images using the technique presented in this research.

References

[1] M. Chen, E. K. Wong, N. Memon, and S. Adams, "Recent developments in document image watermarking and data hiding," in *Proc. SPIE Conf. 4518: Multimedia Systems and Applications IV*, Aug. 2001, pp. 166–176

[2] H. Lu, X. Shi, Y. Q. Shi, A. C. Kot, and L. Chen, "Watermark embedding in DC components of DCT for binary images," in Proc. Int. Workshop on Multimedia Signal Processing, US Virgin Islands, Dec. 2002.

[3] M. Wu, E. Tang, and B. Liu, "Data hiding in digital binary image," in Proc. *IEEE Int. Conf. on Multimedia & Expo*, New York, NY, 2000, pp. 393–396.

[4] M. Wu, E. Tang, and B. Liu, "Data hiding in digital binary image," in *Proc. IEEE Int. Conf. on Multimedia & Expo*, New York, NY, 2000, pp. 393–396.

[5] Y. Q. Shi and H. Sun, Image and Video Compression for Multimedia Engineering: Fundamental, Algorithm, and Standards. Boca Raton, FL: CRC Press LLC, 1999.

[6] 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.

[7] Swetha Kurup, Sridhar G., and Sridhar V, "Entropy Based Data Hiding for Document Images" Proc of WASET, vol. 5 April 2005 ISSN 1307-68

[8] Rakhshanda Yousuf, Aihab Khan, Malik Sikander Hayat Khiyal , "Data Hiding in Binary Document Images based on DRDM for Authentication",Thesis submitted to dept. of software engineering. June 2008.

Aihab Khan, works in Dept. of Computer Sciences Fatima Jinnah Women University Pakistan. His research interests are in the field of Data Mining, Data Warehousing as well as Information security.

Memoona Khanum, works in Dept. of Computer Sciences Fatima Jinnah Women University Pakistan. His research interests are in the field of Data Mining, Data Warehousing, Artificial Intelligence.

M. Sikandar Hayat Khiya,l born at Khushab, Pakistan. He is Chairman Dept. Computer Sciences and Software Engineering in Fatima Jinnah Women University Pakistan. He Served in Pakistan Atomic Energy Commission for 25 years and involved in different research and development program of the PAEC. He developed software of underground flow and advanced fluid dynamic techniques. He was also involved at teaching in Computer Training Centre, PAEC and International Islamic University. His area of interest is Numerical Analysis of Algorithm, Theory of Automata and Theory of Computation. He has more than hundred research publications published in National and International Journals and Conference proceedings. He has supervised three PhD and more than one hundred and thirty research projects at graduate and postgraduate level. He is member of SIAM, ACM, Informing Science Institute, IACSIT. He is associate editor of IJCTE and Co editor of the journals JATIT and International Journal of Reviews in Computing. He is reviewer of the journals, IJCSIT, JIISIT, IJCEE and CEE of Elsevier.

Saba Bashir, is doing PhD in computer software Engineering from National University of Science and technology. Ms from National University of Science and technology. BS from Fatima Jinnah Women University, Rawalpindi. Pakistan. Currently doing job as Assistant Professor in Federal Urdu University of Arts, Science and Technology, Islamabad.

Asima Iqbal, is a graduate from Dept. of Software Engineering, Fatima Jinnah Women University, Pakistan

Farhan Hassan Khan, is doing PhD in Computer Software Engineering from National University of Science and Technology. Ms from National University of Science and Technology. BS from University of Engineering and Technology, Taxila. Currently serving as Project Manager in Software firm