Tamper Detection in Text Document

Ali Kadhim Mousa*

Date of acceptance 9/3/2008

Abstract

Although text document images authentication is difficult due to the binary nature and clear separation between the background and foreground but it is getting higher demand for many applications. Most previous researches in this field depend on insertion watermark in the document, the drawback in these techniques lie in the fact that changing pixel values in a binary document could introduce irregularities that are very visually noticeable. In this paper, a new method is proposed for object-based text document authentication, in which I propose a different approach where a text document is signed by shifting individual words slightly left or right from their original positions to make the center of gravity for each line fall in with the middle point of intended line. Any modification, addition or deletion in a letter, word, or line in the document will be detected.

Introduction

Binary images are commonly used for archiving text document and logo images. It is often necessary to develop appropriate methods to verify their fidelity and integrity for security protection in various applications. So far, there were only a few studies on data hiding in binary images and very few researches on authentication of binary images. Wu, Tang, and Liu [1] embedded bits in image blocks by pattern matching; the method can be used both for data hiding and image authentication. The method proposed by Pan, Chen, and Tseng [2] changed pixel values in image blocks to hide secret data by mapping block contents into the secret data. And the method proposed by Tseng and Pan [3] modified the method by Pan, Chen, and Tseng [2] to control the image quality. In Koch and Zhao [4], a bit 1 or 0 was embedded in an image block by enforcing the ratio of the number of black pixels in the block to that of white ones to be larger or smaller than the value 1, respectively. A text document contains some objects such as characters, words, lines and paragraphs. A watermark embedded and extracted in the document for authentication purposes by shifting these objects are known as object-based text document authentication [5]. Authentication method for text document images was proposed in which the maximum data hiding capacity of the host document was utilized for watermark embedding. For the embedding process, maximum data hiding capacity of the document was calculated before actually the watermark is embedded. Then the watermark was generated from the owner secret key according to the calculated capacity and embedded within the whole document to ensure the authenticity and integrity of the document [6]. For the sake of convenience, we propose a new content-based scheme. The authentication process doesn't need the original document. We study the characteristics of the center of gravity (centroid) for binary images. Experimental results show that centroid can be used to represent the binary image and to decide the authenticity of the image effectively. The paper is organized to describe the definition of centroid, steps for calculate centroid, cropping the body of binary text from image document by

* Department of Computer Science, College of Science for Women, University of Baghdad

applying trimming process , image authentication and alteration localization through line authentication algorithm.

The proposed method

1 Center of Gravity

The center of gravity (CG) is the center of an object's weight distribution, where the force of gravity can be considered to act. It is the point in any object about which it is in perfect balance no matter how it is turned or rotated around that point. For a finite set of point masses, CG may be defined as the average of positions weighted by mass. That is, the {Sum of (weight X position) / (Sum of weights)}.

To find the CG of a two dimensional object consist of elements,

1. Calculate the weights of the basic elements.
2. Choose a starting point. This is called the datum. This point is arbitrarily placed at one end of the object
3. Measure the distances from the datum to the center of each element.
4. Multiply each distance by the respective weight. This gives the moment for each element.
5. Add the weights of all the elements.
6. Divide the total moment by the total weight. This is the distance from the datum to the center of gravity.

We can apply the characteristics of CG on binary text image. The digital text image is a two-dimensional binary image g(x,y) whose pixels' values (weights) are ones or zeros corresponding to light and dark points on the original image. Where each element is, for example, assigned the value 1 if the *i* th cell contain a portion of the character, and is assigned the value 0 otherwise

A text document image is represented by the following function :

$f(x, y) \varepsilon [0, 1]$ Where x=0, 1, 2,… , D and y=0, 1, 2,……, L

D and L represent the width and length of the document in pixels.

Use the formulas:

$X_{cg} = \sum xw / \sum w$ to find the CG along the x-axis …….. (1)

$Y_{cg} = \sum yw / \sum w$ to find the CG along the y-axis …….. (2)

The point at which they intersect is the Center of Gravity (centroid).

2 Image trimming

Text document image is an area that contains binary image. This area may consist of additional lines and columns that have no data (spaces), so these empty lines are eliminated by tracing from outside margins towards inside and stop at a first occurrence of on-pixel at each side of the four edges, see fig (1).

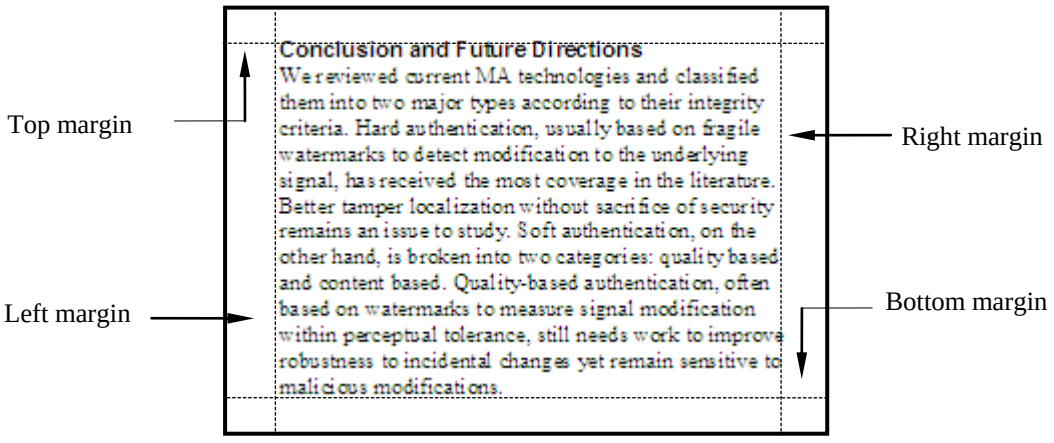

Fig 1: Margins of text body

3 Image authentication

After producing the digital image, a trimming procedure is applied, according to this process the empty areas surrounding the text image has to be removed, then we can apply equations (1) & (2) by using bottom margin and left margin as a datum respectively to get the center of gravity for whole image in both direction which consist of two numbers represent coordinates of centroid. These two numbers inserted in the header of file. To check the authenticity of the image the previous steps repeated by the recipient of the file and compare the calculated centroid with the centroid that has impeded in the header of file. If they are not identical, this mean the document is forged where each deletion or modification of word or letter in the document will change the centroid coordinates. The centroid is computing finally after applying Line Authentication Algorithm which discussed in section 2.4

4 Localization of modification

To localize the alteration may have occurred in the document we apply equation (1) on each line of document after finding the boundaries of its individual words; this can be accomplished with one of sophisticated edge detection techniques. The proposed method requires leaving double space between words of document for two reasons; first, to utilize these spaces in defining the threshold for word segmentation and boundary detection, second, to give a tolerance to decrease (not less than predefined threshold) or increase these spaces as will explained later. A threshold value has to be established to perform the separation. Segmentation of line's words is shown in fig. (2). We find the center of gravity along the x-axis for this line by calculate the weight and center of gravity of each word, then find their distances from the datum (left margin). Then we compute the mid length of intended line. If the distance from reference line to the mid length not identical with the Center of Gravity for the line, we shift the words forward or backward by changing the length of spaces between the words of the intended line to make the Center of Gravity along x-coordinate fall in with the mid point of line length. The procedures shown in Line Authentication Algorithm.

To check the authenticity of document's lines, the previous steps repeated by the recipient of the file and compare the calculated Center of Gravity along the x-axis with the mid length for each line. If they are not identical, this indicates that this line has altered. If one (or more) line has deleted or added then the centroid of whole document will change.

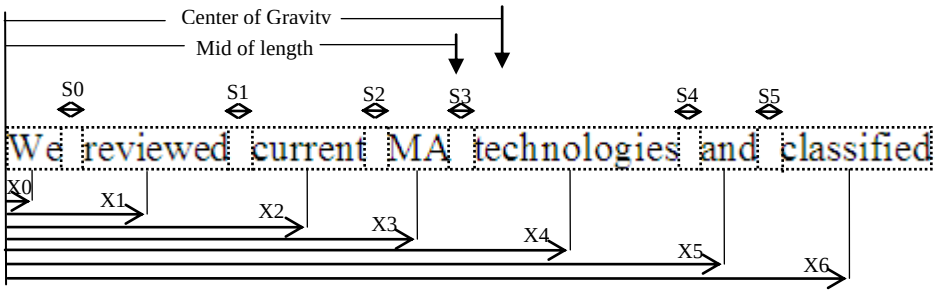

Fig 2: Margins of individual words in one

Line Authentication Algorithm

Variable used:

D Width of trimmed binary text image
L Length of trimmed binary text image
MD Distance from datum to mid point of line length
CGL Center of Gravity of line along x-axis
CGW(i) Center of Gravity of the word of order i
W(i) Total weight of the word of order i
X(i) Distance from datum to Center of Gravity of word(i)
S(j) Space between word(i) and word(i+1) in line
NS Number of spaces between words in line
THR Threshold represents minimum allowable space between words

Program body:

1 SET *THR*
2 SET *i* = 0, *j* = 0, *NS* = 0
3 EXECUTE segmentation of lines for binary text image
4 WHILE not end of lines DO
5 EXECUTE segmentation of words for intended line
6 WHILE not end of spaces DO
7 compute *S(j)*
8 *NS* = *j* + 1
9 *j* = *j* + 1
10 ENDWHILE
11 WHILE not end of words DO
12 compute *W(i)*
13 compute *CGW(i)*
14 compute *X(i)*
15 *i* = *i* + 1
16 ENDWHILE
17 compute *MD*
18 compute *CGL*
19 Compute ∑ *S(j)* *for j =0 to NS-1*
20 *Difference* = *CGL* – *MD*
21 IF *Difference* > 0 AND *Difference* > (∑*S(j)* – *NS* * *THR*) THEN
22 shift last word to next line : GOTO step 2
23 ELSE IF *Difference* < 0 AND ABS*(Difference)* > (D – *2* * *MD*) THEN
24 shift last word to next line : GOTO step 2
25 ENDIF
26 IF *Difference* = 0 THEN : GOTO step 43
27 WHILE *difference* > 0 DO
28 SET *j* = 0
29 WHILE *j* < *NS* AND *difference* > 0 DO
30 *S(j)* = *S(j)* – Point
31 *j* = *j*+ 1
32 compute new *difference*
33 ENDWHILE
34 ENDWHILE
35 WHILE *difference* < 0 DO
36 SET *j* = 0
37 WHILE *j* < *NS* AND *difference* < 0 DO
38 *S(j)* = *S(j)* + Point
39 *j* = *j*+ 1
40 compute new *difference*
41 ENDWHILE
42 ENDWHILE
43 Next Line
44 ENDWHILE

Conclusion

Digital Text document is easily reproduced and modified without any trace of manipulations. The big important for text documents in our live lead me to find a novel method for verify the genuineness of text documents. A physical property of materials; the Center of Gravity, was exploited and applied on text documents. Very encouragement results has fulfilled when suggested algorithm achieved.

References

1. Tang, M. Wu, E. and B. Liu, 2000. "*Data Hiding in Digital Binary Images*", presented at the IEEE International Conference on Multimedia and Exposition, New York.

2. Pan, H. K., Y. Y. Chen, Y. C. Tseng, 2000. "*A Secure Data Hiding Scheme for Two-Color Images*", IEEE ISCC.

3. Tseng Y. C. , H. K. Pan, 2001. "*Secure and Invisible Data Hiding in 2-color Images*", in Proc. IEEE INFOCOM, The Conference on Computer Communications.

4. Koch, E. , J. Zhao, 1995. "*Embedding Robust Labels into Images for Copyright Protection*". Proceedings of International Congress on Intellectual Property Rights for Specialized Information, Knowledge and New Techniques, Munich, Germany.

5. Huijuan Yang and Alex C. Kot, 2005. "*Data Hiding for Text Document*

Image Authentication by Connectivity Preserving", IEEE ICASSP, Philadelphia, March 2005

6. Imtiaz Awan, S.A.M. Gilani, and S.A. Shah, 2006. " *Utilization of Maximum Data Hiding Capacity in Object-based Text Document Authentication*" Proceedings of the international Conference on Intelligent Information Hiding and Multimedia Signal Processing,.

اكتشاف التلاعب في الوثائق النصية

علي كاظم موسى*

*مدرس/ قسم علم الحاسبات / كلية العلوم للبنات / جامعة بغداد

الخلاصة:

رغم صعوبة عملية التحقق من صحة الوثائق النصية بسبب طبيعتها كونها ثنائية اللون حيث وضوح الفرق بين النص الأسود وخلفيته البيضاء، إلا إنها أصبحت مطلوبة وبشكل كبير في تطبيقات عديدة. أغلب البحوث السابقة في هذا الحقل اعتمدت على تقانات طمر علامة مائية في الوثيقة. المأخذ على هذه التقانات يستند إلى حقيقة مفادها إن هذه العلامة تقوم بتغيير بسيط في قيمة النقاط المكونة للحروف وهذا قد ينتج عنه تشوه يمكن أن يرى بالعين المجردة. في هذه الورقة البحثية نقدم طريقة جديدة تعتمد على محتوى الوثيقة النصية للتحقق من صحتها. حيث جرى اقتراح منحى مختلف يقوم بتزحيف بعض المقاطع النصية المكونة للوثيقة إلى الأمام أو إلى الخلف على وفق ضوابط لتحقيق تطابق مركز ثقل الكلمات في كل سطر مع نقطة المنتصف لطول ذلك السطر. وبذلك يمكن الضمان إن أي تغيير ناتج عن تبديل أو حذف أو إضافة لحرف أو كلمة أو سطر لهذه الوثيقة سيجرى اكتشافه لاحقاً.