Self-verifiable Paper Document and Automatic Content Authentication

Yibin TIAN*, Gang FANG**, Wei MING**

Abstract

This report describes a solution for self-verifiable paper documents. The content of the original document is extracted and encoded in high-capacity color barcodes, which are printed out on the same paper as the original image to form an authentic document. For content authentication, the document is scanned to obtain the barcodes and a 2nd image. The original image is reconstructed from barcode data, and compared to the 2nd image using multiple features. It utilizes regular office computers, multi-functional printers and papers, and supports multiple languages without using OCR. Thus it is inexpensive, robust, and more versatile compared to existing solutions. Evaluation on 353 documents in multiple Western languages scanned at 600 dpi achieved greater than 99 % alteration detection rate and less than 0.9 % false positive rate at the word/symbol level, in 25s per page.

1 Introduction

Paper document forgery is a significant issue for government agencies and businesses. Paper document authentication is the process of verifying the authenticity or originality of a printed document [1]. In content authentication, the content of a document is examined to ensure that it has not been altered. Alterations or changes may occur due to deliberate efforts of forgery or accidental events.

Manual verification of paper document content is time-consuming and prone to human errors. OCR has been utilized in various document analysis and management applications [2]. However, high accuracy of OCR is achievable only on printed text with good image quality [3]. In addition, OCR is language dependent. On the other hand, image-based content authentication is more versatile. It can verify documents containing handwritings and mixed-languages in a unified approach.

A number of image-based paper document content authentication methods have been proposed. Unfortunately, most of them can only detect whether alterations have occurred [4-6]. Yang et al. suggested a two-layer binary document that can detect where alterations have occurred, but there are strict requirements on the document [7]. Recently, Sankarasubramaniam et al. used Error Correction Code (ECC) to achieve pixel-level alteration localization, which is sensitive to photocopying noises, and produces too many false positives [8].

We have developed a simple solution for automatic creation and content authentication of a low-cost self-verifiable paper document that can achieve word/symbol-level alteration localization, which we believe is more meaningful for documents as the majority of the content is text. The image of an original document is decomposed to symbol templates and their corresponding locations. The resulting data is

* Delight Imaging, Inc.
** Konica Minolta Laboratory U.S.A.,Inc. (HLUS)

then compressed, encrypted, and encoded in high-capacity color barcodes. The original image and barcodes are printed on the same paper to form a self-verifiable authentic document. For content authentication, the paper document is scanned to obtain the barcodes and a 2nd image, and the original image is partially reconstructed from data decoded from the barcodes, which is then registered with and compared to the 2nd image using multiple image features. The authentication is carried out hierarchically from the entire image down to word and symbol levels. The proposed method is inexpensive, robust and fast. Some preliminary results have been presented at a conference [9].

The rest of the paper is organized as follows. In Sections 2 and 3, we describe the procedures for creating a self-verifiable paper document and verifying its content. In Section 4, we present preliminary evaluation results. Conclusion is given in Section 5.

2 Creation of a Self-verifiable Paper Document

A self-verifiable paper document has to store information of its original image. Watermark and barcode are two popular solutions for storing information on papers [10]. The data capacity of watermarks is usually low. High-capacity barcodes usually utilize spatial and color information. The highest data capacity of existing color barcodes is about 3 KB/in^2 [11, 12].

We tackle the problem from two fronts. First, we developed a lossy representation for the original document that can compress it to a smaller size than existing methods. Second, we designed a new color barcode that has higher data capacity than existing ones.

2. 1 Lossy representation of original images

The best-known compression method for document images is the international standard JBIG2 [13]. It separates a binary document image into text, image and other regions, and compresses them differently. We modified JBIG2's symbol dictionary approach to further improve the compression ratio. Text in the original image is decomposed into symbol templates and the bounding boxes in the following steps:

(1) Noise removal and quality enhancement: Noise is removed via edge-preserving filtering [14]. This step is not necessary if the original image is converted from electronic sources, such as Word or PDF files. Other image quality enhancement can be employed as well.

(2) Binarization, deskew and segmentation: Adaptive thresholding can be used for binarization [15, 16]. Deskew is carried out with Hough transform. Texture and connected component analysis, line detection, and morphological operations are all utilized to segment the image into text, table and graphics regions [16].

(3) Symbol classification: Text regions are further segmented into lines, words and symbols progressively using projections and white space clustering. Multiple image features are used for symbol classification, including geometric features (such as centroid, aspect ratio, density, etc.), zoning profile, side profiles, and topology statistics (such as number of holes, etc.)[17]. Symbols are normalized to a fixed size to compute zoning and side profiles. A naive tolerance based decision-tree classifier is used for classification [16].

(4) Template creation and down-sampling: One template is created for each symbol group using simple averaging. Template size is reduced via topology-preserving down-sampling [16], either to a fixed or variable size. In the latter case, templates that are more likely to be confused with others are down-sampled less.

2. 2 High-capacity color barcode

A 2D barcode divides an area into small cells, the majority of which are data cells. A small portion of them is for locating and decoding the barcode (such as locators and references, etc.). Colors have been used to improve data capacity by increasing the possible number of representations of each cell. To increase the data capacity of a color barcode, either the cell area is reduced or the number of representations increased. We developed a high-capacity color barcode using extremely small data cells (Fig. 1).

In one exemplary design, the barcode consists of 120 × 120 data cells in 1 in^2 (600 PPI), and each data cell is 5 × 5 pixels, of which 3 × 3 pixels in the center is colored (Cyan, Magenta, Yellow and Black), and 1 pixel on each side is left white to reduce color interferences between neighboring cells.

Another notable feature is that the color patches on the 4 sides act as locators, color references, and spatial references at the same time, all employed during the decoding process. Locators and color references are widely used in color barcodes, while spatial references are new in our design. On certain MFPs, there are different spatial drifts among the 4 printed colors (CMYK), which only become a

significant issue when the data cells are very small (Fig. 1c). We can compensate such spatial drifts of the different colors in barcode decoding using information derived from the spatial references. The design achieves 3.6 KB/in^2 data capacity limit for 4 colors (CMYK).

2. 3 Encoding of original document in color barcodes

The lossy representation of the original document is converted to bit stream, and further compressed and encrypted using standard lossless compression and encryption methods (such as LZMA and AES). The bit stream is then divided into packages such that 12-bit Reed-Solomon Error Correction Codes (ECCs) are generated for each package. The data is finally encoded into color barcodes in the CMYK space. If the data cannot fill one barcode, the remaining data cells are colored randomly. The color barcodes and the original image are combined and printed to form a self-verifiable paper document.

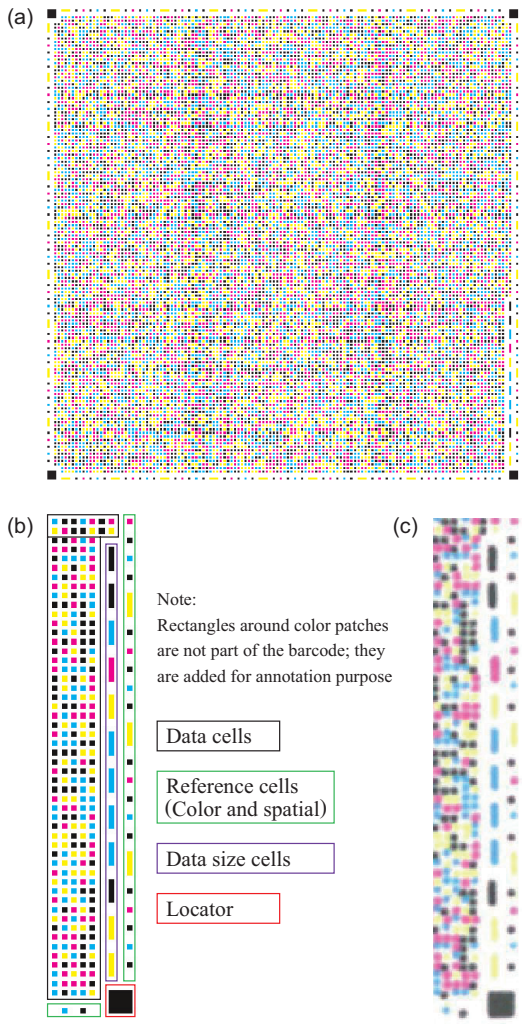

Fig. 1 Color barcode: (a) one complete barcode before printing; (b) right-bottom portion of (a) with annotations; (c) image of (b) after printing and scanning.

3 Automatic Authentication of Document Content

The self-verifiable paper document needs to be scanned to verify its content. The scanned image consists of two parts: the color barcodes and 2nd image. Two major steps are involved in verification: the data in the barcodes is first decoded; then the original image is reconstructed, registered with and compared to the 2nd image to find any alterations.

3. 1 Decoding color barcode

The scanned image is typically in RGB space. The image is rotated if the barcodes are not in the upright position, and individual barcode is extracted using projections. For each barcode, the decoding procedure is illustrated in Fig. 2.

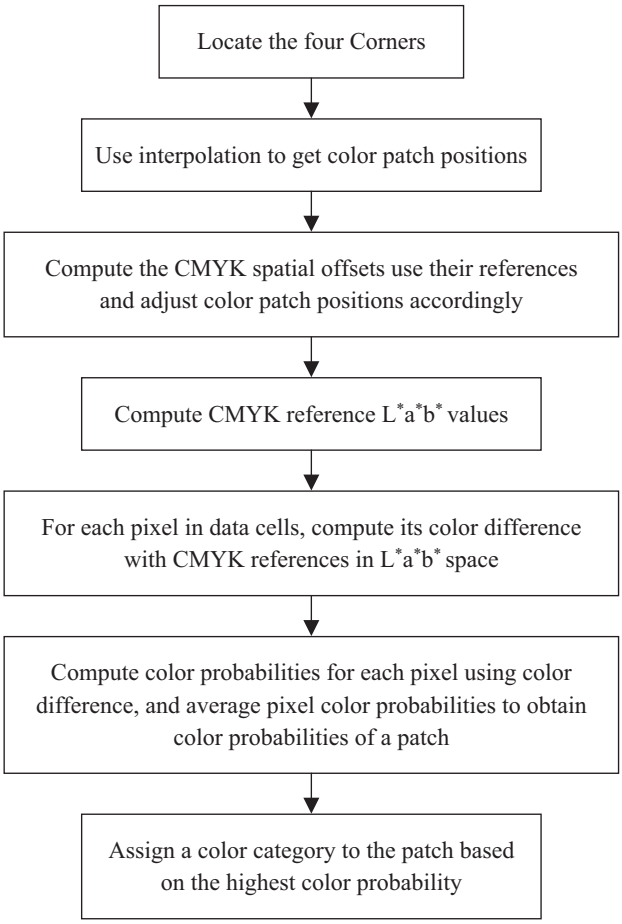

Fig. 2 Decoding procedure of one color barcode.

3. 2 Content authentication using image features

The original document image (text portion) can be easily reconstructed from decoded data that contains the bounding box of each symbol and the corresponding down-sampled template. The same pre-processing steps (steps 1 & 2) described in Section 2.1 are applied to the 2nd image. The two images are registered by

Fourier-Mellin transform [17]. The 2nd image is adjusted to match the original image using the RST (Rotation, Scaling and Translation) registration model. The original and adjusted 2nd images are divided into words, and verified word-by-word. Two words are marked the different if their Hausdorff distance and pixel differences exceed a set of upper limits, on the other hand, if their Hausdorff distance and pixel differences fall below a set of lower limits they are marked as the same. Otherwise, they are divided into symbols and further compared using image features.

The image features used in content authentication are similar to those used in the symbol classification described in Section 2.1. And symbol-level comparison is similar to the decision-tree classification process. The upper and lower limits are determined empirically from character samples. This hierarchical authentication process is tolerant of print-and-scan noise, and runs fast because in many cases the comparison ends at the early stage of the cascade. It should be noted that some steps may be dropped altogether depending on the authentication requirements of a specific application.

4 Evaluations and Results

The proposed solution has been implemented in C++ using the OpenMP® API and OpenCV® library, with a web-based graphic user interface. It has been evaluated on 216 character tables and 137 real documents using a Dell Latitude laptop (Dell® Latitude E6520 with Intel® Core i7-2720QM 2.20 GHz CPU, 6 GB RAM, and 32 and 64-bit Windows® 7 OS) and multiple Konica Minolta bizhub multi-function printers (MFPs).

4. 1 Character tables

Each character table contains the upper and lower cases of 26 English letters, 10 Arabic numerals and 12 special characters (a total of 72 symbols in 3 rows). They were repeated from font sizes 8 to 20. The table was printed on letter-size paper and scanned at 600 dpi in grayscale to obtain the original image. Altered images were created from the original image via electronic image manipulations (cut, paste and erase in Photoline) of each of the symbols to obtain 72 manipulated images. They were printed and scanned in the same conditions to obtain the 2nd images. The process was repeated for 3 different fonts (Arial Bold, Courier Italic and Times Standard). The total altered images are 216.

4. 2 Real documents

A diverse set of real documents was used. It covers at least the following scenarios: pure text (single- and multi-column), text and table mixed, text and image mixed, text and signature mixed. The majority of the documents (100) are in English, and the others (37) have multiple languages and their mix (English, French and German) (Fig. 3). The original and 2nd images were obtained in similar way as described above. The alterations include words, numbers, and signatures.

4. 3 Evaluation results

The true positive (TP) and false positive (FP) detection rates are shown in Table. 1. It should be noted that the counting units for true and false positives were different for character tables and real documents: symbols for the former and words for the latter. For example, in Fig. 3 (b), there are 4 word alterations.

Table 1 Evaluation results from 216 character tables and 137 real documents.

	Character tables		Real documents	
	Add Delete	***Replace***	***Add Delete***	***Replace***
TP	99.5	99.2	100	99.1
FP	0.62	0.53	0.69	0.85

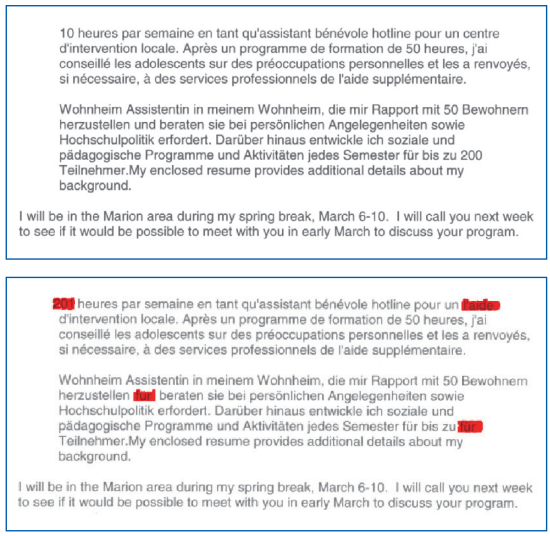

Fig. 3 Authentication example: (a) original image; (b) 2nd image with detected alterations in red.

The average size of compressed text data of the real documents (excluding graphics/images) is 11.6 KB. So on average each page of letter size document scanned at 600 dpi can be stored in less than 4 color barcodes (1 in^2 per barcode). The average verification time (including barcode decoding) is less than 25s.

5 Conclusions

A method for creating a self-verifiable paper document and automatically verifying its content has been proposed, implemented and evaluated. The proposed solution does not need OCR, and is capable of handling document with multiple Western languages. It also provides meaningful alteration localizations at word/character level and is tolerant to noises arising from the print-and-scan process.

6 References

[1] Doermann D., "The evolution of document authentication", ICFHR, 3 (2010).

[2] Schantz H. F., "The History of OCR, Optical Character Recognition". RTUA (1982).

[3] Rose H., "How good can it get? Analyzing and improving OCR accuracy in large scale historic newspaper digitization programs", D-Lib Magazine, vol. 15 (2009).

[4] Chen M., Wong E.K., Memon N. and Adams S., "Recent developments in document image watermarking and data hiding", Proc of SPIE 4518, 166 (2001).

[5] Hunag P.M., Wu D.C. and Tsai W.H., "A novel block-based authentication technique for binary images by block pixel rearrangements", ICME, 903-906 (2004).

[6] Jiang M., Wong E.K. and Memon N., "Robust document image authentication", ICME, 1131-1134 (2007).

[7] Yang H. and Kot A.C., "Two-layer binary image authentication with tampering localization", ICASSP, 309-312 (2006).

[8] Sankarasubramaniam Y., Narayanan B., Viswanathan K. and Kuchibhotla A., "Detecting modifications in paper documents: A coding approach", Proc of SPIE 7534, 0A1 (2010).

[9] Tian Y., Zhan X., Wu C. and Ming W., "Self-verifiable paper documents and automatic content verification", Proc of SPIE 9028, 0L1 (2014).

[10] Millter M.L., Cox I.J., Linnartz J.P. and Kalker T., "A review of watermarking: principles and practices", in Digital Signal Processing in Multimedia Systems, Parhi K.K. and Nishitani T. Eds., 461-485 (1999).

[11] Parikh D. and Jancke G., "Localization and segmentation of a 2D high capacity color barcode", Applications of Computer Vision, 1-6 (2008).

[12] Bulan O. and Sharma G., "High capacity color barcodes: per channel data encoding via orientation modulation in elliptical dot arrays", IEEE Trans. Image Processing, vol. 20, 1337-1350 (2011).

[13] Ono F., Rucklidge W., Arps R. and Constantinescu C., "JBIG2-the ultimate bi-level image coding standard," ICIP, 140-143 (2000).

[14] Perona P. and Malik J., "Scale-space and edge detection using anisotropic diffusion", IEEE Trans Pattern Anal Mach Intell, vol. 12, 629-639 (1990).

[15] Sauvola J. and Pietikainen M., "Adaptive document image binarization", Pattern Recognition, vol. 33, 225-236 (2000).

[16] Tian Y. et al., US patent application 13/607,667, and other patent applications that have not been published (2012).

[17] Trier O.D., Jain A.K. and Taxt T., "Feature extraction methods for character recognition-A survey", Pattern Recognition, vol. 29, 641-662 (1996).

[18] Chen Q., Defrise M. and Deconinck F., "Symmetric phase-only matched filtering of Fourier-Mellin transforms for image registration and recognition", IEEE Trans Pattern Anal Mach Intell, vol. 16, 1156-1168 (1994).

Source

This paper is reprinted by minor revision the "Proceedings of Imaging Conference Japan 2014". The Imaging Society of Japan grants the copyright of this paper.