Perceptual Hashing for Hardcopy Document Authentication Using Morphological Segmentation

Diego Massola Shimizu and Hae Yong Kim

Universidade de São Paulo (USP), Brazil
{dshimizu@lsi, hae@lps}.usp.br

1. Introduction

Semi-fragile authentication of hardcopy documents is a technique designed to detect any visually significant alteration in a document, while ignoring incidental alterations, like distortions resulting from print-scan operations, photocopies, rotations, scalings, translations and minor stains on the paper. It is meant to substitute the use of notarial authenticated photocopies. However, to our knowledge, there is still no functional authentication system for printed documents, only for documents in digital form [1].

A semi-fragile authentication system is composed of three sub-components: perceptual hashing, cryptography and data hiding. This work is concerned with the first sub-component. The perceptual image hashing $h(A)$ of an image A is a value that identifies A. It is also called robust visual hashing or media hashing [2], [3], [4]. Moreover, given two images A and B, the distance $D[h(A), h(B)]$ between the hashings must be somehow proportional to the perceptual visual difference of the images A and B.

To our knowledge, no perceptual hashing has been proposed for document authentication and perceptual hashings for continuous-tone images cannot be directly applied to authenticate documents. Behera et al. [5] have proposed a perceptual hashing that uses low resolution image of the document as the input data for document retrieval. However, it cannot be used for document authentication, because authentication must detect even the alteration of a single character and must use high-resolution images. This on-going work intend to propose a perceptual hashing for the document authentication.

2. Segmentation and Classification

Our method applies a bottom-up approach. First, a median filter reduces noise from the document image. Then, a morphological segmentation algorithm is used to separate the document in blocks and each block is classified as "text" or "halftone/image".

The segmentation algorithm is very simple and similar to the Run Length Smoothing Algorithm (RLSA) [5], [6]. However, we implemented it using morphological operations. Two closing filters are used to link together black pixels if the distance between them is less than a threshold. This process is performed using vertical and horizontal structuring elements, generating two images. The AND operation between them yields the segmentation (Figure 1).

After the block segmentation, our program classifies the blocks in two categories: texts and images. The classification is based on the blocks height, number of black pixels and number of black-white transitions. Text blocks usually have low height and low number of transitions, while images have a higher value for those parameters [6], [7].

3. Feature Extraction

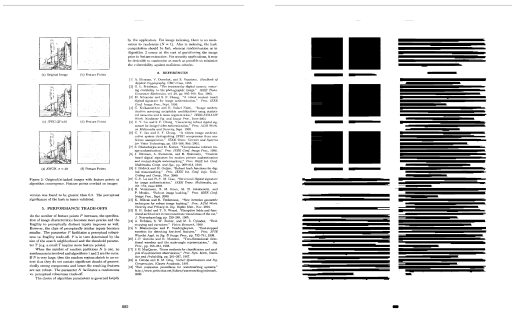
After the block segmentation and classification, the most adequate algorithm is used to extract the authentication index (or the string of features) from each block.

An OCR program recognizes the string of characters that makes up the text blocks and this string itself is used as the authentication index. The underlying idea is that the document is authentic only if its composing characters were not modified. It is possible to include also the sizes or the formats (italic, bold, etc.) of characters in the features.

The features of an image/halftone block are computed using an idea similar to [8]. However, we use a multi-scale quad-tree-based approach, instead of the original raster-order approach. We compute the mean grayscale m0 of the image. Then, the image is divided into four subregions and their mean grayscales m_1, m_2, m_3 and m_4 are calculated. The four extracted feature are the results of the comparisons between the mean grayscale m_0 of the parent image and of its offsprings m_1, m_2, m_3 and m_4. The result of each comparison can be less, more or undefined. This process is applied recursively, until obtaining enough number of information to authenticate the image.

The information extracted from the blocks can be compressed with a one-way hashing algorithm, like SHA-1. Then, it is ciphered using a public-key cipher, like RSA.

The document signature can be stored independently of the image, or can be inserted into the im-



(a) (b)

Figure 1. Segmentation using the morphological version of the RLSA algorithm. (a) Original document. (b) Boxes representing the document segmentation.

age using a barcode, or some data-hiding technique like [9]. Then, a document can be authenticated by extracting its signature from its content and comparing with the stored signature. If the signatures match, the document is successfully authenticated.

4. Partial Results

The segmentation was already implemented, using the morphological version of RLSA. Figure 1 shows a document image and its segmentation results. The block classification step is being finalized. After that, the remaining steps will be implemented.

References

[1] S. V. D. Pamboukian and H. Y. Kim, *New Public-Key Authentication Watermarking for JBIG2 Resistant to Parity Attacks*, Int. Work. Digital Watermarking (Siena), LNCS, vol. 3710, 2005, pp. 286–298.

[2] M. Schneider and S.-F. Chang, *A Robust Content Based Digital Signature for Image Authentication*, IEEE Int. Conf. Image Processing, Vol. 3, 1996, pp. 227–230.

[3] C.-S. Lu and C.-Y. Hsu, *Geometric Distortion-Resilient Based Digital Signature for Image Authentication*, Multimedia Systems **11** (2005), no. 2, 159–173.

[4] V. Monga and B. L. Evans, *Perceptual Image Hashing Via Feature Points: Performance Evaluation and Tradeoffs*, IEEE T. Image Processing **15** (November 2006), no. 11, 3452–3465.

[5] A. Behera, D. Lalanne, and R. Ingold, *Visual Signature Based Identification of Low-resolution Document Images*, Proc. ACM Symp. Document Engineering (Wisconsin, USA), 2004, pp. 178–187.

[6] K. Y. Wong, R. G. Casey, and F. M. Wahl, *Document Analysis System*, IBM J. Res. Develop. **26** (November 1982), no. 6.

[7] S. N. Srihari, *Document Image Understanding*, Proc. ACM Fall Joint Computer Conference (Dallas, USA), 1986, pp. 87–96.

[8] J. Oostven, T. Kalker, and J. Haitsma, *Visual Hashing of Digital Video: Applications and Techniques*, Proc. SPIE, Vol. 4472, 2001, pp. 121–131.

[9] H. Y. Kim and J. Mayer, *Data Hiding for Binary Documents Robust to Print-Scan, Photocopy and Geometric Distortions*, Sibgrapi - Brazilian Symp. on Comp. Graph. and Image Proc. (Belo Horizonte), 2007.