http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (8): 1608-1613, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.1608.1613

Robust Text Hashing for Content-Based Document Authentication

[1,2]Lina Tan and [3,1]Xingming Sun
[1]College of Information Science and Engineering, Hunan University,
Changsha, 410082, China
[2]Department of Information, Hunan University of Commerce, Changsha, 410205, China
[3]Jiangsu Engineering Center of Network Monitoring,
Nanjing University of Information Sciences and Technology, Nanjing, 210044, China

Abstract: Digital forgery and tampering of text documents create an urgent need for content-based document authentication. Observe that main geometric features of characters would approximately stay invariant under small perturbations. A robust document authentication scheme is proposed using a skeleton-based text image hash function. The algorithm is aimed at producing sufficiently randomized outputs which are unpredictable, thereby yielding required properties of image hashing. In the verification stage, the deskewing mechanism base on Hough transform is used to compensate for the distortions induced by rotation or hardcopy operations. Experimental results show that this technique withstands standard benchmark (e.g., Stirmark) attacks, including compression, geometric distortions of scaling and small-angle rotation and common image processing operations. Content-based modifications of text data are also accurately detected.

Key words: Robust text hashing, content-based authentication, primary skeleton, shuffling

INTRODUCTION

Observations (Anan *et al*., 2007; Varna *et al*., 2009) have been made that the crime rate due to the leakage or forgery of sensitive information is increasing every year. Hence, authentication technologies for text documents pose a broad prospect. However, to date, less work has been reported on this area, even if a comprehensive study of the authentication technologies has been undertaken (Al-Hamami and Al-Anni, 2005; Rabah, 2005; AL-Saraireh *et al*., 2006; Suri and Rani, 2006; Ibrahim, 2007; Wang *et al*., 2010). Contrarily to other multimedia signals, texts show a marked trait that the foregrounds contrast clearly with the background areas. Consequently we face the challenge for the little information redundancy which can be employed. Besides, printed documents undergo wear and tear by normal use and digitization that affect the accuracy of extracting character features. The information carried by text data is mostly retained even when the multimedia carrier has undergone moderate levels of filtering, geometric distortion or noise corruption.

Hashing algorithms can be used in multimedia protection applications, namely watermarking and authentication. Conventional hashing algorithms such as MD5 and SHA-1 are extremely sensitive to any slight change. However, digital images commonly undergo various manipulations, such as compression, enhancement, scaling and print-scan operations. Unlike cryptographic hash functions, robust image hashing is required to be robust against incidental image modifications. That is, the images being perceptually identical to the original image should be regarded as authentic. But for visually distinct images, the hash function should produce different hash values (Venkatesan *et al*., 2000; Mihcak and Venkatesan, 2001).

The underlying techniques for constructing robust image hashes can roughly be classified into methods based on image statistics and relations (Schneider and Chang, 1996; Venkatesan *et al*., 2000; Lin and Chang, 2001; Lu and Liao, 2003), preservation of coarse image representation (Fridrich and Goljan, 2000; Mihcak and Venkatesan, 2001; Kozat *et al*., 2004) and low-level image feature extraction (Dittman *et al*., 1999). Unfortunately, none of them can survive the binarization operation which cannot be directly applied to text documents, for most of scanned and computer-generated text images are binary.

The major challenge of the framework to generate a hash has been the feature extraction stage. Here we introduce a novel algorithm that utilizes structural features

Corresponding Author: Xingming Sun, College of Information Science and Engineering, Hunan University, No. 252, Lushan South Road, Yuelu District, Changsha, 410082, China
Tel: 86-025-58731575 Fax: 86-731-88821780

of characters that can capture the major content characteristics from a human perspective. The robustness against acceptable manipulations and the fragility against malicious attacks are guaranteed by the coarse components of host signals, i.e. the stroke and junction features. The security of this scheme just refers to the shuffling strategy with the difficulty of guessing the secret key.

PROPOSED METHOD

With the goal of offering good content representation and robustness to content-preserving modifications, two types of low-frequency features are used in the algorithm: stroke-based and junction-based. They are especially suitable for describing character since they have natural and inherent information. These features are then used to generate the hash values.

Selected features: As a preliminary step, the gray image is converted to be binary. Characters on a binary text image can be viewed as connected components modeled as the combination of strokes. Figure 1 gives a specific example of the letter "K". Firstly character skeletons are abstracted by a thinning algorithm (Deng *et al*., 2000) as shown in Fig. 1b. To produce the primary skeleton illustrated in Fig. 1e, redundant segments such as spur skeleton points are removed using the junction type. The key of this technique is to position the junctions by detecting four types of junction: 1-fork (endpoint), 2-fork (corner point), 3-fork (trifurcate point) and 4-fork (cross point).

Junction detection: A model of Cross Number (CN) is formulated here for prospective junctions. Cross Number (CN) depends on the transitions of the eight neighbors in a 3×3 window around the considered pixel. The eight neighbors of the considered pixel p_c are p, p_2, ..., p_7 and p_8. Cross Number of p_c is given by:

$$CN(p_c) = \sum_{k=1}^{8} abs(p_{k+1} - p_k)/2 \quad (1)$$

where, $p_9 = p_1$. Supposing that the junction type of p_c in the skeleton image is m-fork, it can be deduced that:

$$m = \begin{cases} 1, & \text{if } CN(p_c) = 1 \\ 0, & \text{if } CN(p_c) = 2 \\ \forall, & \text{if } CN(p_c) > 2 \end{cases} \quad (2)$$

In the algorithm we care less about the case of $CN(p_c) = 2$, for which p_c is very likely to be an ordinary point excluded from junctions, namely 0-fork. For $CN(p_c) = 1$, p_c is evidently an endpoint. For $CN(p_c) > 2$, it needs further confirmation process for the junction type. All the junctions identified by $CN(p_c) > 2$ are labeled in Fig. 1c.

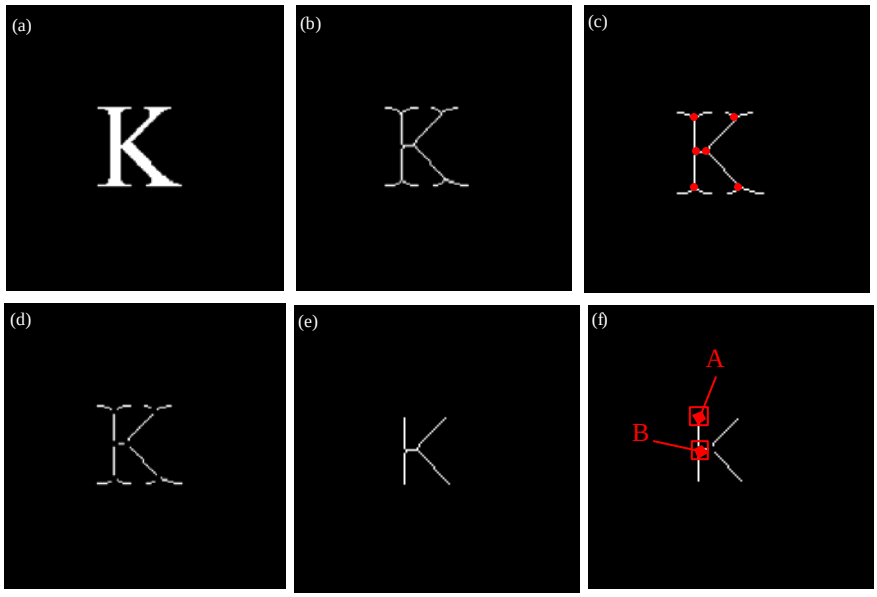

Fig. 1(a-f): Example of skeleton-based feature extraction. (a) Original image, (b) skeleton image, (c) junctions on the skeleton (marked by red points), (d) broken skeleton for junction type decision, (e) primary skeleton and (f) example of different junction types

Here all the direct neighbors of p_c are set to "0" to break the skeleton into separate pieces, as shown in Fig. 1d. After eliminating all the spur pixels to generate the primary skeleton as illustrated in Fig. 1e, the junction type is calculated by counting the number of connected objects in a small area surrounding p_c. To take Fig. 1f as an example, A is an endpoint and B is a 3-fork junction.

Stroke extraction: Directional filtering is applied to extract stroke segments from the primary skeleton. Steerable filters have better orientation selectivity for feature estimation of oriented image structures, like edges under a certain angle. They provide linear combinations of basis filters, allowing each to be adaptively "steered" to any orientation and phase. On account of the characteristic of these filters we employ them to split the primary skeleton into stroke segments. A graphic description of the implementation is given in Fig. 2. More detailed representation of steering theorems is referred to Freeman and Adelson (1991). We aims at obtaining horizontal and vertical strokes of the primary skeleton by two orientations (i.e., 0° and 90°) using steerable filters. Figure 3 exemplifies the filtered results of the character K. The horizontal and vertical strokes are picked up by calculating the direction of each stroke segment present in the relevant filtered image. In Fig. 3a, only one vertical stroke is extracted, marked by a red ellipse. And none of horizontal strokes can be found in Fig. 3b.

Content-based feature codes: To develop discriminative features for characters we construct a variable-length sequence code of characters, as described in Fig. 4. Each bit of the sequence code has different meaning. J1, ..., J4, D1, D2 are fixed to appear but D3 and D4 are optional. J1, ..., J4 denote the number of 1, ..., 4-fork junctions present in the character skeleton, respectively. A special instance of feature codes for the connected components with fewer pixels, e.g., the punctuation ".", encode a 0 for short.

Before clarifying the definition of D1, ..., D4 we introduce the idea of Relative azimuth Coding (RAC). It is defined in terms of the direction of a designated point towards the reference point (the character centroid is used in the algorithm). The direction space is quantized to 8 bins as depicted in Fig. 5. The code "9" is located just in a small coverage (usually with the Euclidean distance ED<3 px) of the reference point.

D1 is associated with the horizontal strokes, among which the longest one H_l is firstly chosen. If there isn't any horizontal stroke, D1 is assigned 0. If H_l has only one 1-fork junction, D1 is assigned the RAC of the endpoint towards the centroid, otherwise calculated by the RAC of the stroke midpoint towards the centroid. The principle of D2 is similar to D1 and they only differ in that D2 deals with the vertical strokes.

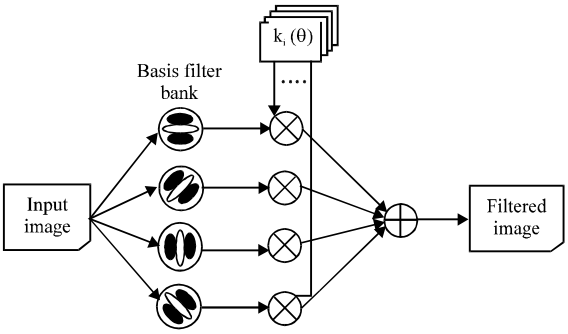

Fig. 2: Architecture for applying steerable filters. $k_i(\theta)$ gives the appropriate interpolation functions at each position and time

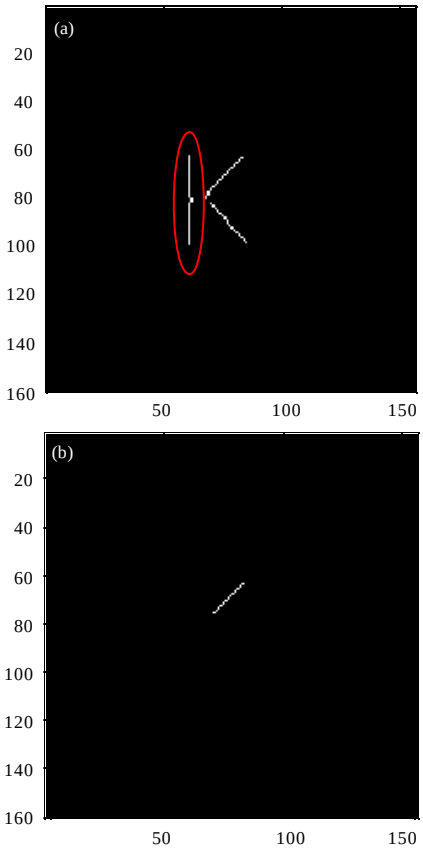

Fig. 3 (a-b): Directional feature images according to filtering angle of (a) 0° and (b) 90°

Fig. 4: Illustration of feature sequence code

Table 1: Feature detection on 10 pt TNR font

Character	Feature code	Character	Feature code	Character	Feature code
a	1110083	s	20000051	K	40200413
b	1110033	t	30102963	L	21007227
c	20000071	u	21000331	M	23000575
d	1110011	v	21000013	N	22000551
e	1110908	w	23000031	O	000000
f	40019661	x	40010031	P	1110055
g	1030001	y	30100051	Q	1010007
h	30100435	z	22003037	R	21202557
i	0/20000926	A	21209057	S	20000051
j	0/20000252	B	022024	T	30102963
k	40200337	C	20000071	U	21000331
l	20000926	D	020004	V	21000013
m	31100557	E	32101417	W	23000031
n	21000557	F	31101551	X	40010031
o	000000	G	20000818	Y	30100051
p	1110055	H	40209431	Z	22003037
q	1110077	I	20000926		
r	21000661	J	20000252		

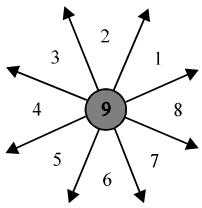

Fig. 5: RAC with the reference point located in the center

D3 and D4 are the RAC of the first and second furthest endpoints of the primary skeleton towards the centroid, respectively. Obviously, the existence of the endpoints brings D3 and D4 into being.

Robust document authentication: We briefly explain the involved steps to generate the digital signature:

- Split a document into lines of text using the horizontal profile (Culnane *et al*., 2006)
- Sort the connected components (characters) by their centroids according to text lines
- Shuffle the ordered connected components (characters) throughout the document with a secret key
- Generate a feature vector by computing the feature code for each connected component within the shuffled domain
- Generate a hash code from the feature vector using a cryptographic hash function

Our scheme has two objectives: robust against content-preserved attacks while fragile to fraudulent alterations and impersonations. For the segmentation of text lines we assume good compensation (document deskewing) of a possible document misalignment while print-scanning. In the verification stage, after the scanned image gets binarized, morphological filtering is applied to remove the noise. The classical Hough Transform (HT) is extended here to identify skew angle of the text lines. Each centroid of a connected component is mapped to all values of the parameters determining the location and orientation of the lines in the image. Then the orientations performed by steerable filters and RAC must be rectified by the skew angle. The detector would operate well for both digital and printed text documents.

Table 2: Test performance under various attacks

Attacks	False rejection rate (%)	False acceptance rate (%)
1° rotation	0.0	2.0
3° rotation	2.0	0.0
5° rotation	4.0	0.0
1% noise	0.0	2.0
3% noise	0.0	1.0
5% noise	3.0	0.0
JPEG_10%	0.0	2.0
JPEG_30%	0.0	2.0
JPEG_50%	0.0	2.0
Median filtering (2×2 window)	9.0	0.0
Print-scanning	0.0	2.0

EXPERIMENTAL RESULTS

The implementation of this method in a digital-only environment is straightforward. We tested the discriminative capability of the feature codes on a set of Times New Roman (TNR) font characters of size 10 pt. Table 1 contains test results for this. The feature codes are not case sensitive, i.e., allowing for the same codes belonging to a letter's upper and lower case. It can be observed that there are only two characters "l" (lower case of "L") and "I" (upper case of "i") whose feature codes are identical, resulted from the inherent structure of the two characters. An adversary in authentication applications would have much more incentive to counterfeit valid text content, so this may be resolved in conjunction with the semantic environment.

In terms of robustness and security we carried on exhaustive tests using documents distorted by both admissible and malicious modifications. 100 original and 100 tampered images underwent various distortions. Figure 6 shows text images after applying attacks. The quality factor Q mentioned here is inversely proportional to the compression level. In the print-scan tests we used a HP Laserjet P1505n for printing the documents at a default resolution of 600 dpi. The hardcopy documents were scanned on a Microtek scanner at 600 dpi.

We see from Table 2 that very low false rejection rate and false acceptance rate were obtained in our tests. It

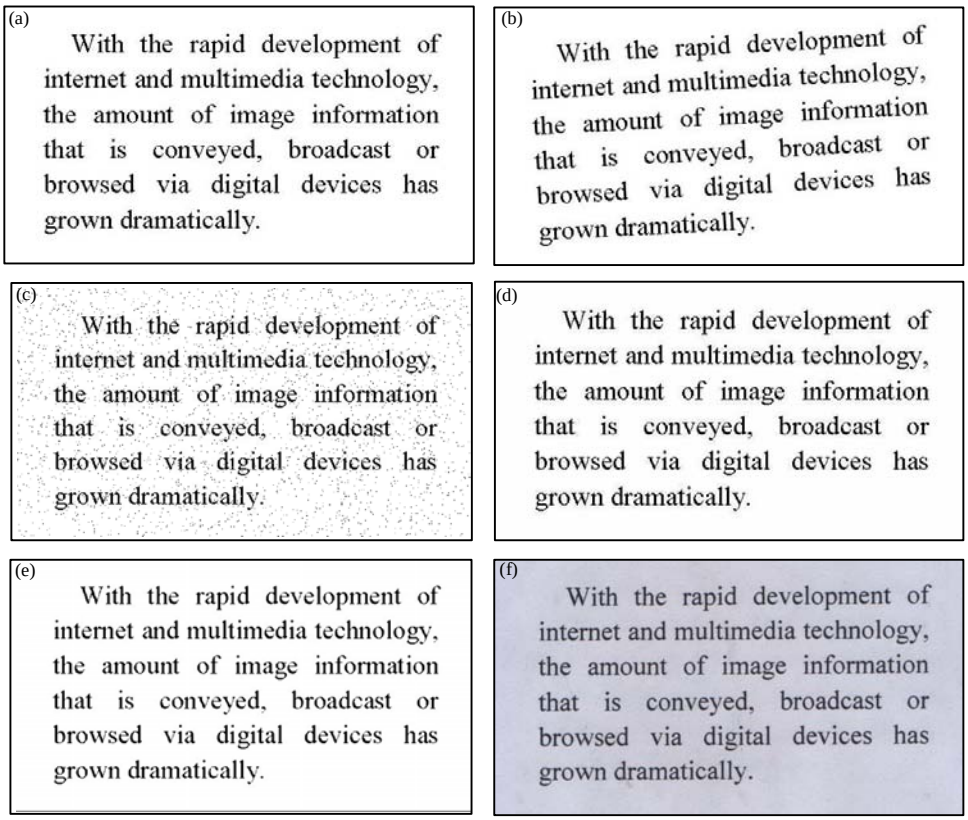

Fig. 6 (a-f): Various types of attacks. (a) Original text, (b) 3° rotation, (c) 3% salt and pepper noise, (d) JPEG lossy compression with Q = 10%, (e) median filtering using 2×2 window size and (f) print-scanning

can be concluded that this scheme can resist affine transforms, JPEG compression and low-level noise but not very robust against median filtering. The amount of the altered documents that are falsely accepted is decreased against the increasing degree of attacks.

CONCLUSION

We have presented a typical skeleton-based hashing algorithm that has features of good robustness for document authentication. The proposed scheme also has good discriminative capabilities and can identify malicious manipulations, such as a cut-and-paste type of editing that do not preserve the content of the image. Hashing can be used in place of watermarking with the benefit that nothing is added to images. Besides content authentication, it can provide a robust and secure representation of images for numerous applications, such as large databases and anti-piracy search.

But then, our technique requires ideal approaches to character features. An adaptive binarization algorithm is also beneficial to improve the verification performance. More effort is needed to address these problems.

ACKNOWLEDGMENTS

This study was supported by National Basic Research Program 973 of China (Grant No. 2010CB334706, 2010CB334706) and National Natural Science Foundation of China (Grant No. 60736016, 60973128, 60973113, 61073191, 61070195 and 61070196).

REFERENCES

Al-Hamami, A.H. and S.A. Al-Anni, 2005. A proposal for comprehensive solution to the problems of passport`s authentication. Inform. Technol. J., 4: 146-150.

Al-Saraireh, J., S. Yousef and M. Al-Nabhan, 2006. Analysis and enhancement of authentication algorithms in mobile networks. J. Applied Sci., 6: 872-877.

Anan, T., K. Kuraki and S. Nakagata, 2007. Watermarking technologies for security-enhanced printed documents. Fujitsu Sci. Tech. J., 43: 197-203.

Culnane, C., H. Treharne and A.T.S. Ho, 2006. A new multi-set modulation technique for increasing hiding capacity of binary watermark for print and scan processes. Proceedings of the Digital Watermarking 5th International Workshop. Nov. 8-10, Jeju Island, South Korea, Springer-Verlag, pp: 96-110.

Deng W., S.S. Iyengar and N.E. Brener, 2000. A fast parallel thinning algorithm for the binary image skeletonization. Int. J. High Perform. Comput. Appl., 14: 65-81.

Dittman J., A. Steinmetz and R. Steinmetz, 1999. Content based digital signature for motion picture authentication and content-fragile watermarking. Proceedings of the IEEE Int. Conf. Multimedia Comput. and Systems. June 7-11, Florence, Italy, IEEE, pp: 209-213.

Freeman, W.T. and E.H. Adelson, 1991. The design and use of steerable filters. IEEE Trans. Pattern Anal. Machine Intellig., 13: 891-906.

Fridrich, J. and M. Goljan, 2000. Robust hash functions for digital watermarking. Proceedings of the International Conference on Information Technology: Coding and Computing, Mar. 27-29, Las Vegas, NV, USA., pp: 178-183.

Ibrahim, Y.K., 2007. Password-based key authentication model-a new approach. Trends Applied Sci. Res., 2: 456-459.

Kozat, S.S., R. Venkatesan and M.K. Mihcak, 2004. Robust perceptual image hashing via matrix invariants. Proceedings of the International Conference on Image Processing (ICIP). Singapore, Oct. 24-27, IEEE, pp: 3443-3446.

Lin, C.Y. and S.F. Chang, 2001. A robust image authentication method distinguishing JPEG compression from malicious manipulation. IEEE Trans. Circuits Syst. Video Technol., 11: 153-168.

Lu, C.S. and H.Y.M. Liao, 2003. Structural digital signature for image authentication: An incidental distortion resistant scheme. IEEE Trans. Multimedia, 5: 161-173.

Mihcak, M.K. and R. Venkatesan, 2001. New iterative geometric techniques for robust image hashing. Proceedings of the ACM Workshop on Security and Privacy in Digital Rights Management. Nov. 5, Springer-Verlag, Philadelphia, PA, USA, pp: 13-21.

Rabah, K., 2005. Secure implementation of message digest, authentication and digital signature. Inform. Technol. J., 4: 204-221.

Schneider, M. and S.F. Chang, 1996. A robust content based digital signature for image authentication. Proceedings of the IEEE Conf. Image Process. Sep. 16-19, IEEE, Lausanne, Switzerland, pp: 227-230.

Suri, P.R. and S. Rani, 2006. Avoidance of intruder attack with changed bluetooth authentication procedure. Inform. Technol. J., 5: 1033-1037.

Varna, A.L., S. Rane and A. Vetro, 2009. Data hiding in hard-copy text documents robust to print, scan and photocopy operations. Proceedings of the ICASSP. Apr. 19-24, Taipei, Taiwan, IEEE, pp: 1397-1400.

Venkatesan, R., S.M. Koon, M.H. Jakubowski and P. Moulin, 2000. Robust image hashing. Proceedings of the IEEE Conf. Image Processing. Sep. 10-13, Vancouver, BC, Canada, IEEE, pp: 664-666.

Wang, X., L. Yang, X. Sun, J. Han, W. Liang and L. Huang, 2010. Survey of anonymity and authentication in P2P networks. Inform. Technol. J., 9: 1165-1171.