American Journal of Information Science and Computer Engineering
Vol. 1, No. 3, 2015, pp. 113-116
http://www.aiscience.org/journal/ajisce

Scanned Documents Forgery Detection Based on Source Scanner Identification

Ramzi M. Abed*

Faculty of IT, Islamic University of Gaza, Gaza, Palestine

Abstract

With the increasing number of digital image editing tools, it becomes an easy task to modify any digital image by any user with any level of experience in image editing. One important type of digital images is the scanned documents as they can be used as legal evidence. Therefore, some legal issues may arise when a tampered scanned document cannot be distinguished from an authentic one. In this work we are proposing a novel technique to detect scanned documents tampering, this proposed technique is based on the used scanner identification using features intrinsic to a data-generating sensor.

Keywords

Forgery, Tampering, Graylevel Co-occurrence Matrix, GLCM, Image Forensics

Received: July 4, 2015 / Accepted: July 18, 2015 / Published online: August 17, 2015

1. Introduction

Forgeries are a very old problem. In the past it was limited to art and literature but did not affect the general public. Nowadays, due to the advancement of digital image processing software and editing tools, an image can be easily manipulated and modified [1]. It is very difficult for humans to identify visually whether the image was modified or not. There is a rapid increase in digitally manipulated forgeries in mainstream media and on the Internet [2]. This trend indicates serious vulnerabilities and decreases the credibility of digital images. Therefore, developing techniques to verify the integrity and authenticity of the digital images is very important, especially considering that the images are presented as legal evidences, as news items, as a part of medical records, or as financial documents. In this sense, image forgery detection is one of the primary goal of image forensics [3, 4].

In today's digital world securing different forms of content is very important in terms of protecting copyright and verifying authenticity [5, 6, 7]. One example is watermarking of digital audio and images. We believe that a marking scheme analogous to digital watermarking but for documents is very important [8]. Scanned documents are direct accessory to many criminal and terrorist acts. Examples include forgery or alteration of scanned documents used for purposes of identity, security, or recording transactions. In this case, the ability to identify the device used to scan the material in question would provide a valuable aid for detecting any forged regions in this document especially when the forgery is applied by adding external content to the scanned documents [9].

There are various levels at which the image source/sensor identification problem can be addressed [10]. A number of robust methods have been proposed for source scanner identification. In [12], techniques for classification of images based on their sources: scanner, camera and computer generated images, are presented. Sensor pattern noise can be successfully used for source camera identification and forgery detection [11, 13]. Also, source scanner identification for photographs can be done using statistical features of sensor pattern noise [14]. All these methods for source scanner identification focused on scanned versions of photographs and not on scanned versions of printed text documents. Since the methods utilising sensor pattern noise

* Corresponding author
E-mail address: rabed@iugaza.edu.ps

for source identification mainly use Photo- Response Non-uniformity (PRNU) as the sensor's signature and the PRNU is almost absent in "saturated" regions of an image [13]. In [15] authors proposed a method to identify the scanner used to scan a text document depending on the use of texture analysis. In this paper we present a method for verifying the authenticity of scanned documents, that have been captured by flatbed desktop scanners, using texture features based on the work in [15].

2. The Proposed System

Figure 1 shows the block diagram of the proposed scanned document authenticity detection system. Given a scanned text document digital image, henceforward referred to as the tested scanned document, our proposed system is supposed to decide if this tested scanned document is a forged or an authentic one.

Starting with a tested scanned document, the system will begin by extracting all the letter "e"s in the document, as it is the most frequently occurring character in the English language. Next, the system will extract a set of features from each group of n_e characters ("e's"), then it will form a feature vector for them by dividing the tested scanned document into non-overlapping blocks of size $N \times N$ pixels. A different set of features are extracted from each of these blocks. Each of these feature vectors are then tested and classified separately using Support Vector Machine (SVM) classifier, which will decide whether the tested scanned document image an authentic or a tampered one.

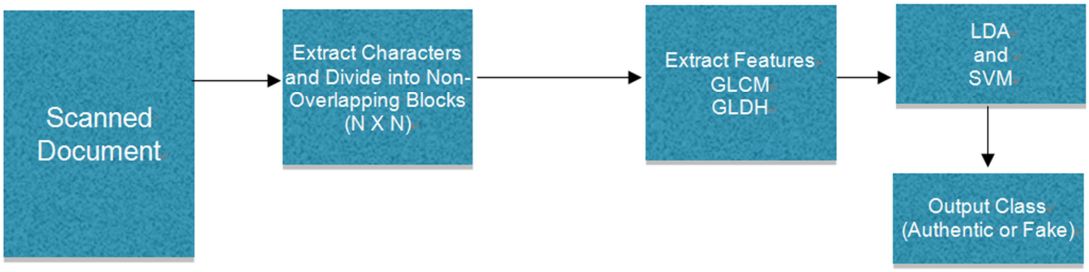

Figure 1. Block Diagram of the Proposed Scanned Document Authenticity Detection.

3. Graylevel Co-occurrence Matrix (GLCM)

Scanned documents are differ than scanned images in the way in which colors are distributed across their pixels, scanned documents generally lack presence of continuous tones and are dominated by "saturated" pixels. That is, most of the pixel values are either close to zero or are close to 255 [16]. This makes it very difficult to accurately use the type of signatures described in earlier source camera forensics techniques [13]. Our proposed forgery detection system is developed based on the identification of the scanner used to scan the tested scanned document, as this technique depends on identifying the signature of the scanner. It is observable that the quality of edges of characters in scanned documents will vary according to the scanner used in the scanning process. In more details, high resolution scanners produce more solid black lines characters with sharper edges, while low resolution scanners produce characters represented by black lines consisting of variations from black to higher graylevels, and the edges of these characters will be more gradual. These differences will result changes in texture features. Focusing on the scanned "e"s, differences in the extracted texture features can be quantified. For documents scanned at low resolution such as 200dpi, each character is very small, about 15 × 20 pixels and is non-convex [16], this causes some difficulties in filtering the scanned image in either the pixel or transform domain. In addition to the type of the scanner sensor, the direction of scanning process can produce specific textures as it can cause vacillation of graylevel in the scanned characters. These textures extracted from the fluctuation in graylevel can be represented by GrayLevel Co-occurence Matrix (GLCM), and they are very robust for identifying scanned documents, and therefore, they are very robust for detecting tampered regions on the scanned documents. In our experiments, to bypass the problem of small size scanned characters, we use a group of 100 "e" characters to be able to produce the GLCM.

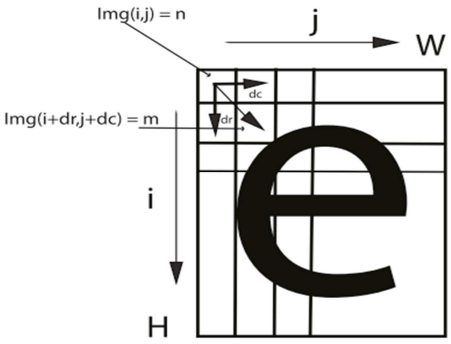

Figure 2. Scanned image for the character "e".

In this proposed work, we suppose that the scanned document texture is related to the direction of the scanning process. Figure 2 shows the scanned image of the character "e", features are extract form *Img(i, j)*. The region of interest (ROI) is the set of all pixels within the rectangular bounding box around the character, the open source OCR "ocrad" [17] is used to determine this ROI.

Equation (1) shows how to calculate GLCM for the ROI mentioned above depending on four parameters, the number of occurrences of pixels with graylevels n and m with a separation of (dr, dc) pixels (Figure 2). In our work, we choose $dc = 0$ and $dr = 1$ to generate the scanned character features.

For each one of the scanned and extracted "e" letter, a GLCM matrix is calculated, then an average

GLCM is calculated for each group of "e" s.

$$glcm(n, m, dr, dc) = \sum_{(i,j),(i+dr,j+dc)\in ROI} 1_{\{Img(i,j)=n, Img(i+dr,j+dc)=m\}} \quad (1)$$

$$glcm_{isotropic}(n, m) = \sum_{dr=-1}^{1} \sum_{\substack{dc=-1 \\ (dr,dc)\neq(0,0)}}^{1} glcm\,(n, m, dr, dc) \quad (2)$$

$$gldh_{isotropic}(k) = \sum_{\substack{0\le n\le 255 \\ 0\le m\le 255 \\ |n-m|=k}} glcm_{isotropic}(n, m), k \in [0, 255] \quad (3)$$

Equations (2) and (3) show how to calculate an isotropic graylevel difference histogram (GLDH) or each block of $N \times N$ pixels. which is used to extract another 246 features, these features are used with the 22-texture features extracted from GLCM to identify the signature of the scanner used. In equation (3), we choose k to be in the range (10.255), as lower values will correspond to completely white or completely black regions which will not be useful in identifying the signature of the scanner used. Moreover, the isotropic GLDH in (3) is normalised to one, then it can be used to identify the scanner.

Hence, given a tested scanned document of size WxH pixels, we get 22-features from each group of letter of "e", 22-GLCM features from each of blocks of size NxN, and 246-isotropic GLDH features from each block of size NxN pixels. Combining these features together will lead to the final decision about whether the tested document is an authentic or a forged one.

4. Experiments

In our experiments, we need to generate a test and train dataset by scanning 20 different test documents using two different scanners at 200dpi and 300dpi resolution with 8 bits/pixel (grayscale), afterward we made several different modifications on these scanned documents using Adobe Photoshop, these modifications include: adding some text using text tool in Photoshop, erasing some text using eraser tool in Photoshop, and copying scanned text from another scanned document to the targeted one.

Three separate classifiers (LDA + SVM) are trained for each class of features, namely GLCM features from groups of "e" s, GLCM features from each of the blocks of size $N \times N$ pixels, and isotropic GLDH features from each of the blocks of size $N \times N$ pixels. Results show that the accuracy for classifying tested documents is over 90%. This indicates that it can be very reliably to use the features used to indicate source scanner in scanned document tampering detection process.

5. Conclusion

In this paper we proposed a method for detecting scanned text documents forgery, this detection method is based on source scanner identification by using texture features. As shown by the experiments, the proposed method is robust to JPEG compression and gives over 90% detection accuracy. The proposed features are also robust to the document tampering techniques, as detection process gives good results regardless the applied modification technique.

Acknowledgment

This work was supported by Palestinian Scientific Research Council Scholarship - 2013

References

[1] J. A. Redi, W. Taktak, and J.-L. Dugelay, "Digital image forensics: A booklet for beginners," Multimedia Tool Appl., Vol. 51, no. 1, pp. 133-62, Jan. 2011.

[2] J. Wang, G. Liu, Z. Zhang, Y. Dai, and Z. Wang, "Fast and robust forensics for image region-duplication forgery," Acta Automatica Sinica, Vol. 35, no. 12, pp. 1488-95, Dec. 2009.

[3] V. Tyagi, "Detection of forgery in images stored in digital form," Project report submitted to DRDO, New Delhi, 2010.

[4] Ansari, Mohd Dilshad, S. P. Ghrera, and Vipin Tyagi. "Pixel-Based Image Forgery Detection: A Review." IETE Journal of Education 55.1, 2014.

[5] Mauro Barni, Christine I. Podilchuk, Franco Bartolini and Edward J. Delp, Watermark embedding: hiding a signal within a cover image, IEEE Communications Magazine, 39, 102, 2001.

[6] R. W. Wolfgang, C. I. Podilchuk and E. J. Delp, Perceptual watermarks for digital images and video, in Proceedings of the IEEE, pp. 1108–1126 1999.

[7] C. I. Podilchuk and E. J. Delp, Digital watermarking: Algorithms and applications, IEEE Signal Processing Magazine, 18, 33 2001.

[8] Edward J. Delp, Is your document safe: An overview of document and print security, in Proceedings of the IS&T's NIP18: International Conference on Digital Printing Technologies, 2002.

[9] Mikkilineni, Aravind K., et al. "Printer identification based on texture features." NIP & Digital Fabrication Conference. Vol. 2004. No. 1. Society for Imaging Science and Technology, 2004.

[10] P.-J. Chiang, N. Khanna, A. K. Mikkilineni, M. V. O. Segovia, S. Suh, J. P. Allebach, G. T.-C. Chiu, and E. J. Delp, "Printer and scanner forensics," IEEE Signal Processing Magazine, vol. 26, no. 2, pp. 72–83, March 2009.

[11] J. Fridrich, "Digital image forensics," IEEE Signal Processing Magazine, vol. 26, no. 2, pp. 26–37, March 2009.

[12] N. Khanna, G. T. Chiu, J. P. Allebach, and E. J. Delp, "Forensic tech- niques for classifying scanner, computer generated and digital camera images," Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, NV, March 2008, pp. 1653–1656.

[13] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Determining image ori- gin and integrity using sensor noise," IEEE Transactions on Information Forensics and Security, vol. 3, no. 1, pp. 74–90, March 2008.

[14] N. Khanna, A. K. Mikkilineni, and E. J. Delp, "Scanner identification using feature-based processing and analysis," IEEE Transactions on Information Forensics and Security, vol. 4, no. 1, pp. 123–139, March 2009.

[15] Khanna, Nitin, and Edward J. Delp. "Source scanner identification for scanned documents." Information Forensics and Security, 2009. WIFS 2009. First IEEE International Workshop on. IEEE, 2009.

[16] N. Khanna and E. J. Delp, "Source scanner identification for scanned documents," First IEEE International Workshop on Information Foren- sics and Security, London, United Kingdom, December 2009, to appear.

[17] A. K. Mikkilineni, P.-J. Chiang, G. N. Ali, G. T.-C. Chiu, J. P. Allebach, and E. J. Delp, "Printer identification based on textural features," Proceedings of the IS&T's NIP20: International Conference on Digital Printing Technologies, vol. 20, Salt Lake City, UT, October/November 2004, pp. 306–311.

[18] Khanna, Nitin, and Edward J. Delp. "Intrinsic signatures for scanned documents forensics: effect of font shape and size." Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on. IEEE, 2010.