International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V70I2P229

Volume 70 Issue 2, 247-252, February, 2022

Original Article

Combination of DWT And LPQ Features for Document Age Identification

Pushpalata Gonasagi[1], Shivanand S Rumma[2], Mallikarjun Hangarge[3]

[1,2]*Department of P.G. Studies and Research in Computer Science, Gulbarga University Kalaburagi, Karnataka, India.*
[3]*Karnatak Arts, Science and Commerce College, Bidar, Karnataka, India*

[1]gonasagi99@gmail.com, [2]shivanand_sr@yahoo.co.in, [3]mhangarge@yahoo.co.in

Abstract *- This paper presents an algorithm based on DWT(Discrete Wavelet Transform) of RT(Radon Transforms) and LPQ(Local Phase Quantization) for document age identification. Features computed by applying fusion of DWT of RT and LPQ techniques on a dataset of 640 handwritten document images written by 640 writers(320 are original and 320 forged). The classification of forged and original documents was performed using an SVM(Support Vector Machine) classifier. The average classification accuracy of original and forgery documents is 93.9% and 92.5%, respectively.*

Keywords — *RT, DWT, LPQ, Forgery Documents, SVM.*

I. INTRODUCTION

We are living in a digital world. Moreover, digital technologies help to improve career prospects, awareness, digital skills, etc. The documents are forged within a period for malicious purposes because software and hardware tools are easily affordable at a low cost. The new printing technologies are used to produce fake documents such as bank cheques and official documents. Identification of such forged documents is hard to the human eyes [1,2]. The forgery documents detected based on a particular page from the set of documents produced by the sources (printers, fax machines, scanners, mobiles, etc.) can be granted as forgery documents [3]. The forgery document is detected by analyzing the document at the character level. The shape, alignment, or skew of characters are clues for identifying the forgery of documents [4]. The forgery identification of printed documents is highly focused compared to the handwritten forged documents identification. Hence, this paper studied handwritten forged documents identification.

The rest of the paper is as follows: Section-II provides the review of age and source identification of documents, Section-III presents the proposed method, Section-IV discusses the experimental results, and Section-V concludes the paper.

II. RELATED WORK

The literature survey related to the identification of forgery documents is as follows: Halder et al. [5] presented a method based on the colour of the ink used in determining the age of the printed documents. The pixel profile, average intensities and kurtosis features are used for the printed documents forgery identification. Google Life magazine cover pages are considered printed documents for identifying the age of the documents and achieved an accuracy of 74.5% by applying the Neural Network. Raghunandan et al. [6] described a system for the detection of forged documents. Fourier Coefficients were extracted from document images for identification of the forgery documents. The authors have carried out experiments on both handwritten and printed document images. The reported accuracy of handwritten forgery and original documents identification is 78.5% and 77.5%, respectively. Forgery identification accuracy is 78.5% in printed documents, and original documents identification accuracy is 82.0%. Barboza et al. [7] discussed a method for determining the document images through age analysis based on the colour components of the document images. They have extracted the features of normalized RGB components of the document images using Gaussian distribution and the age identifications of documents like birth, wedding, death documents performed and reported outperforming results. Elkasrawi et al. [8] explored a method to identify the source automatically using printers' distinct noise. Each printer produces noise depending on its imperfection. They investigated the forgery documents depending on the scanning resolution. Twenty different printers and four hundred documents are used for the experiment. They have achieved a classification accuracy of 76.75%. Khan et al. [9] demonstrated the handwritten documents' examination for some part of the text was altered or forged. They have been used the hyperspectral image for detecting handwritten notes based on ink mismatch. They have been used sparse and selection techniques for ink match detection. Luo et al. [10] discussed an approach to distinguish similar visual inks automatically. The objective is the detection of forgery documents using unsupervised clustering techniques. It is handled to distinguish inks for unbalanced ink proportion. Alkawaz et al. [11] presented a system for detecting authenticate images fineness using a Discrete Cosine Transform (DCT). They have been used to detect the tampered regions accurately. Block-based copy-move images are detected and duplicated block located by

the Euclidean distance method, and the forged detection performance in precision is evaluated. Vieira et al. [12] proposed an information system for identifying forgery documents. They have used the critical point descriptor algorithm of SURF, SIFT, and ORB to extract document images' features. They achieved an accuracy of 92.46% to the false documents. Berenguel et al. [13] presented a method to detect counterfeit identity card documents by scanning and printing operations. They have extracted texture features using dense SIFT of security background printed in documents as identification for banknotes.

III. PROPOSED METHOD

Forgery is the process of creating, changing or possessing a false document with the purpose to commit duplicity. Forgery can be the creation of a false document or alteration of an actual one. A person can use the same pen, paper, or ink to create forged documents; however, the writing time is different, which leads to affect the texture of the document. It gives a clue to identify the forgery documents based on the variations in textures between the original and forgery documents irrespective of text, logo, equations or any graphic symbols. In other words, it can be estimated the age of documents through the classification of forgery documents and original documents. Further, consider the documents as original if they are more than six months older than other age groups. The limitation to estimating the document's age is that if the created documents are more than six months, we have considered them as original documents. If document images were created in less than six months, then consider them as forgery documents. To distinguish between the authentic and falsified documents in terms of texture, employed a combination of DWT of RT and LPQ techniques to extract the document's features. Based on the discriminating feature values, document images are classified as original and forgery document images using an SVM classifier. Figure 1 shows a schematic diagram of the proposed method.

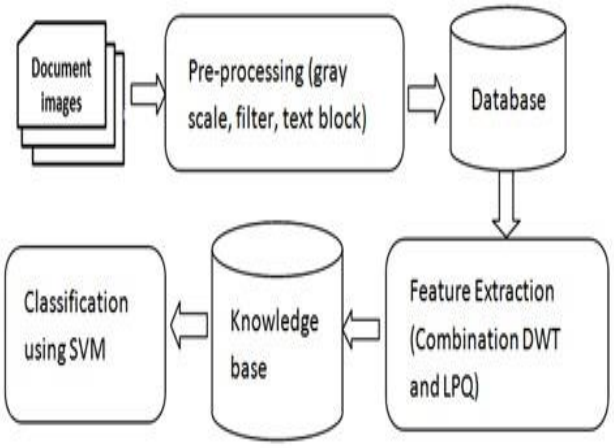

Fig. 1 Schematic diagram of the proposed system

A. Dataset

To assess the efficacy of the proposed approach, consider handwritten document images for the experiment. The dataset includes the original and forgery document images. The original documents are collected from the students' notes. The original document images contain four classes with a gap of six months (2 years, 1 ½ year, one year and six months) by referring to April 2016. The entire original documents are 320 pages. The forgery documents are created by writing the same document from the respective four classes: 2 years, 1 ½ year, one year and six months. These documents are written by using different pens, paper and persons. Figure 2 shows sample document images of a handwritten document. A total of 640 documents images was considered for the experiment. The details of the dataset are shown in Table 1 (this dataset is borrowed from [6]). The original documents are 320 pages, and the forgery documents are 320 pages. It is a two-class problem for the classification of original and forgery documents from their respective classes, as discussed in algorithm 1 and 2 as follows:

B. Feature Extraction

The extraction of features is an important step, and here extracted the features using DWT based on energy coefficients of Radon transform and LPQ techniques.

a) Radon Transform*:* It is the process of computing the image matrix's projection in a specific direction [14]. The projection of the two-dimensional image of f(x, y) is the line integral of a specific direction. The image's resulting projection is the sum of the pixels' intensities or directional energy in each direction and represents a new image g(ρ, θ). Formally it can be expressed as follows:

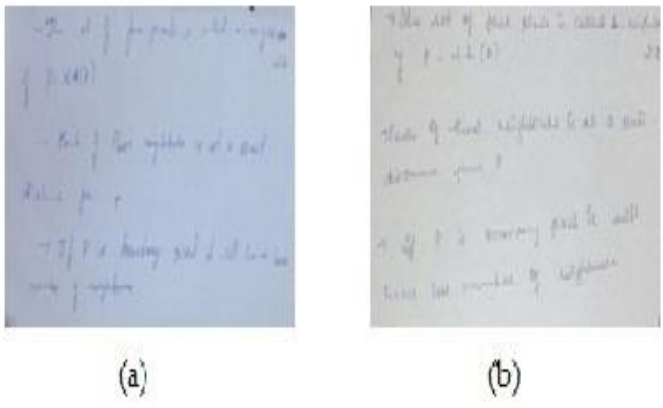

Fig. 2. (a) sample Original document (b)Sample Forgery document of (a)

Table 1. Number of handwritten documents and their text blocks

Documents classes ⇩		Original document images (pages)	Text block images (512X512 size)	Text block images with the complete covered text
6 months	Original	80	653	279
	Forgery	80	658	283
1 year	Original	80	469	159
	Forgery	80	453	153
1 ½ year	Original	80	433	139
	Forgery	80	365	100
2 year	Original	80	488	171
	Forgery	80	418	134

$$\rho = x\cos\theta + y\sin\theta \quad (1)$$

Where x and y are x-axis and y-axis in the two-dimensional plane, respectively.

The radon transform function can thus be written like this:

$$g(\rho, \theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,\delta(\rho - x\cos\theta - y\sin\theta)\,dxdy \quad (2)$$

In our approach, a feature vector is generated using DWT based on the Radon Transform function. The Radon Transform from equation (2) generates the handwritten document image projection's energy co-efficient with fixed angular direction denoted by θ in the range [0, 179] degrees along each row distance ρ. Then energy coefficients of projection of document images are input images to DWT.

b) Discrete Wavelet Transform*:* It is a very effective and adaptable method for signal subband decomposition. It is an inherent multi-resolution nature, scalability and tolerable degradation. Images are decomposed into a set of basic functions. Wavelets are the name for these basic functions. Dilations and shifting are used to create wavelets from a single prototype wavelet termed mother wavelet [15]. Discrete Wavelet Transformation, which serves as an orthonormal basis for the DWT expansion, uses wavelets with single scaling functions. The two-dimensional scaling function ϕ(a, b) that is further decomposed into three sub-bands: ΨH(a, b), ΨV(a, b), ΨD(a, b) which are two-dimensional wavelets. These are the product of one-dimensional scaling and wavelet transformation, which are represented by:

$$\Psi H(a,b) = \phi(a)\,\Psi(b) \quad (3)$$
$$\Psi V(a,b) = \Psi(a)\,\phi(b) \quad (4)$$
$$\Psi D(a,b) = \phi(a)\,\phi(b) \quad (5)$$

These wavelets measure the functional changes in the intensity of images encoded as input information in different directions such as horizontal, vertical and diagonal from Equations 3, 4 and 5. Digital filters are used to implement two-dimensional DWT. For multi-resolution analysis, DWT Daubechies9 is employed.
To extract the features, used four decomposition levels, resulting in four DWT sub-bands at each level. As a result, have a total of 4 x 4 = 16 sub-bands. Then, used entropy and standard deviations were to measure each sub-band property. Thus, obtained 32 features (16 sub-bands*2 statistical measures) for documenting age discrimination.

c) Local Phase Quantization*:* As a texture descriptor, it is analogous to the LBP approach [16, 17]. Image blurring does not affect the LPQ descriptor. By distributing different codewords in the image region, the LPQ descriptor uses the local phase spectrum to differentiate the underlying texture. This work explores the LPQ descriptor's capability to handle blurred images and sharp images. Around each pixel's four frequency points, the local Fourier coefficients G(x) are determined. After that, binary scalar quantization is used to quantify the signs of each coefficient's natural and imaginary parts to acquire phase information for each pixel in the superpixel region. Finally, each coefficient's quantization result is encoded as an 8-bit binary string and forms 256 features. Mathematically express [18] the LPQ as:

$$q_i(x) = \begin{cases} 1, & g_i(x) \le 0 \\ 0, & \text{Otherwise} \end{cases} \quad (6)$$

Where $g_i(x)$ is the i^{th} element of G(x) and the phase information of 8-bit of the image is as in equation (7).

$$f_{LPQ}=\sum_{n=1}^{8} q_n 2^{n-1} \quad (7)$$

d) DWT and LPQ Descriptors*:* The features extracted from DWT based on radon transform and LPQ are 32 and 256, respectively. Finally, 288 features were generated by combining the DWT and LPQ.

Algorithm 1: Classification of original and forgery documents using text blocks of handwritten document images**:**

Step 1: Input the colour scanned document images.

Step 2: Pre-processing

a) Convert colour document images to grayscale document images.
b) Apply image enhancement techniques, namely the Wiener filter.
c) Segment the page into a text block of size 512×512 and store them individually

Step 3: Repeat the Steps from 1–2 for all classes of handwritten document images, namely two years, 1 ½ years, one year and six months.

Step 4: Extract the features from DWT and LPQ descriptors from Step 3. Then combine these features and store them as a knowledge base.

Step5: Apply an SVM classifier to identify forgery or original documents from their respective classes.

Algorithm 2: Classification of original and forgery pages of handwritten document images:

Step 1: Input the colour scanned document images.

Step 2: Pre-processing

a) Convert colour document images to grayscale document images.
b) Apply image enhancement techniques, namely the Wiener filter.

Step 3: Repeat Step 2 for handwritten document images of four classes 2 years, 1 ½ year, one year and six months.

Step 4: Extract the features using DWT and LPQ techniques from Step 3. Then combine these features and store them as a knowledge base.

Step 5: Apply an SVM classifier to identify forgery or original documents from their respective classes.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The above-proposed algorithms are implemented on Intel Core i3, @2.10 GHz system using MATLAB tool. A total of 640 handwritten document images were used for the experiment. By referring to April 2016, documents images with a gap of two years, one and half years, one year, and six months were collected. These (original) are categorized into four classes based on their ages, as mentioned above. Using these documents, created forged document images by writing the same content from their classes, as discussed in Section III. This is a two-class problem where the forgery document classes are compared with the original document classes individually. For instance, a six-month original document image is compared to 6-month forgery document images to identify the given document as an original or forgery document. Similarly, other classes of original document images compared to their forgery document images and noted the results.

Two experiments have been conducted on the above dataset. Experiment 1 is on the text block images, and experiment 2 is on the whole document pages.

A. Experiment-1

A detailed description of the dataset is available in [6]. This dataset consists of 640 pages. In this experiment, instead of using the complete pages, we applied an automatic segmentation algorithm on 640 pages to segment 3937 text blocks of 512x512 pixels. Aimed to investigate the impact of the size of the image on classification accuracy and computing time. Out of 3937 text blocks, 1418 are considered for an experiment that is entirely text-included blocks(see Table 1). To compute these text blocks' features of all four classes, applied fusing of DWT based on RT and LPQ techniques, as discussed in section III. It is a two-class problem for comparing original text blocks to forgery text blocks with a six-month gap into their respective four classes. It means that forgery text blocks are compared to the original text block individually. Based on the traditional SVM classifier [19], original text blocks and forgery text blocks are classified. The average classification accuracy of four classes with a six-month gap is given in Table 2. Our experimental results exhibit a unique fact that as the documents' age increases, the handwritten document images' texture quality decreases [5]. As the age gap of handwritten document images increases from 6 months to one year, 1 ½ year and two years, the handwritten document images' classification accuracy decreases into their classes. The older documents consist of less consistent information than new documents because of the ageing effect on documents. Table 1 presents the age identification accuracies of class-1 (6-Months), class-2 (1-Year), class-3(1 ½ - Year) and class-4(2-years). It is exciting to note that the accuracy of age identification of class-1 to 4 is decreasing. It indicates that texture degrades when the document becomes older, resulting in lower results**.** The classification accuracy of all original text blocks is 93.9%, and all forgery text blocks are

92.5%, shown in Table 3. Our algorithm's performance is compared with [6] and Table 4 indicates the excellent performance of the proposed method over [6].

Table 2. Age identification accuracy of text blocks based on SVM classifier (DWT of RT and LPQ)

Age of the Document	6-Months	1-Year	1 ½ -Year	2-Year
Accuracy	98.1%	97.2%	95.0%	84.3%

Table 3. Classification accuracy of all original text blocks and all forgery text blocks using SVM classifier

Text blocks	Identification rate (%)	Error rate (%)
All original text blocks	93.9%	6.1%
All forgery text blocks	92.5%	7.5%

Table 4. Comparative analysis

Methods	All Original documents	All Forgery documents
Raghunandan K. S[6] approach	77.5%	78.5%
Proposed approach	93.9%	92.5%

B. Experiment-2

In the second experiment, all the steps mentioned in algorithm 2 from section III are followed. The difference between experiment 2 from experiment 1 is that experiment 1 works on text blocks and experiment 2 on whole pages. Original and forgery sample document pages can be seen in Figure 3. The average classification accuracy of four classes with a six-month gap is given in Table 5. It is noticed from Table 5 that the document pages' age increases, the texture quality of the pages decreases. The average classification accuracy of all original pages have got 91.3%, and all forgery pages have got 94.7%, as shown in Table 6. Compared to the existing method [6], the proposed method has outstanding performance, as shown in Table 7.

Table 2 and Table 5 of experiments 1 and 2 respectively show that the age identification of text blocks of 2-years older is 84.3%, and the whole document age identification of 2-years older is 92.5%. An exciting observation evolved from these two tables is that when the area of the texture understudy decreases, the results also decreases because of the non-availability of sufficient properties of the texture. Hence the error rate is 8.2%

Table 5. Age identification accuracy of the whole document based on SVM classifies (DWT of RT and LPQ)

Age of the Document	6-Months	1-Year	1 ½ -Year	2-Year
Accuracy	98.1%	96.9%	94.4%	92.5%

Table 6. Classification accuracy of all whole original documents and all forgery original whole documents using SVM classifiers

Document images	Identification rate (%)	Error rate (%)
All Original	91.3%	8.7%
All Forgery	94.7%	5.3%

Table 7. Comparative analysis

Methods	All Original documents	All forgery documents
Raghunandan K. S [6] approach	77.5%	78.5%
Proposed approach	91.3%	94.7%

Fig. 3. The first row: contains sample forgery pages of 2-Years, 1 ½-Year, 1-Year and 6-Months. The second row: contains original sample pages of 2-Years, 1 ½-Year, -Year and 6-Months

V. CONCLUSION

This paper concludes by demonstrating the efficacy of blending DWT of RT with LPQ to capture the texture's global and local properties. Experimentally proved that this combination of features works well on text block and whole pages of handwritten documents for age identification. It is pretty interesting to know that though the image's size is

small (text blocks), age identification accuracy remains good. In both the experiments, the proposed method is excellent in identifying the documents' age compared to [6]. In future, we proposed to employ deep learning architecture to enhance and generalize handwritten documents' age identification problems.

ACKNOWLEDGEMENT

We thank Prof. K.S Raghunandan, Department of Studies in Computer Science, University of Mysore, Karnataka, India, for providing the dataset used in this work [6].

REFERENCES

[1] Chim, J. L. C., Li, C. K., Poon, N. L., & Leung, S. C., Examination of counterfeit banknotes printed by all-in-one colour inkjet printers. Journal of the American Society of Questioned Document Examiners, 7(2) (2004) 69-75.

[2] Makris, J.D., Krezias, S.A. and Athanasopoulou, V.T., Examination of newspapers. Journal of the American Society of Questioned Document Examiners (ASQDE), 9(2) (2006)71-75.

[3] Gebhardt, J., Goldstein, M., Shafait, F. and Dengel, A., , August. Document authentication using printing technique features and unsupervised anomaly detection. In 2013 12th International conference on document analysis and recognition (2013). 479-483. IEEE.

[4] Bertrand, R., Gomez-Krämer, P., Terrades, O.R., Franco, P. and Ogier, J.M., A system based on intrinsic features for fraudulent document detection. In 2013 12th International conference on document analysis and recognition ., (2013) 106-110. IEEE.

[5] Halder, B. and Garain, U., Colour feature-based approach for determining ink age in printed documents. In 2010 20th International conference on pattern recognition., (2010) 3212-3215. IEEE.

[6] Raghunandan, K. S., Shivakumara, P., Navya, B. J., Pooja, G., Prakash, N., Kumar, G. H. & Lu, T., Fourier coefficients for fraud handwritten document classification through age analysis. In 2016 15th International conference on frontiers in handwriting recognition (ICFHR) ., IEEE, (2016) 25-30.

[7] da Silva Barboza, R., Lins, R. D., & de Jesus, D. M., A colour-based model to determine the age of documents for forensic purposes. In 2013 12th International conference on document analysis and recognition (2013) 1350-1354. IEEE.

[8] Elkasrawi, S., & Shafait, F., Printer identification using supervised learning for document forgery detection. In 2014 11th IAPR International Workshop on Document Analysis Systems,. (2014). 146-150. IEEE.

[9] Khan, Z., Shafait, F., & Mian, A., Automatic ink mismatch detection for forensic document analysis. Pattern Recognition, 48(11) (2015) 3615-3626.

[10] Luo, Z., Shafait, F., & Mian, A., Localized forgery detection in hyperspectral document images. In 2015 13th International Conference on Document Analysis and Recognition (ICDAR) (pp. 496-500). IEEE.

[11] Alkawaz, M. H., Sulong, G., Saba, T., & Rehman, A, 2018. Detection of copy-move image forgery based on discrete cosine transform. Neural Computing and Applications, 30(1) (2018)183-192.

[12] Vieira, R., Silva, C., Antunes, M., & Assis, A., Information system for automation of counterfeited documents images correlation. Procedia Computer Science, 100 (2016) 421-428.

[13] Berenguel, A., Terrades, O. R., Lladós, J., & Cañero, C,. e-Counterfeit: a mobile-server platform for counterfeit document detection. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)., 9 (2017) 15-20. IEEE.

[14] Nawade S.A., Rumma S., Pardeshi R., Hangarge M., Old Handwritten Music Symbol Recognition Using Radon and Discrete Wavelet Transform. In: Chiplunkar N., Fukao T. (eds) Advances in Artificial Intelligence and Data Engineering. Advances in Intelligent Systems and Computing, Springer, Singapore. https://doi.org/10.1007/978-981-15-3514-7_86., 1133 (2021).

[15] Mallat, S. G., A theory for multiresolution signal decomposition: the wavelet representation. In Fundamental Papers in Wavelet Theory., (2009) 494-513. Princeton University Press.

[16] Jiao, J., & Deng, Z., Deep combining of local phase quantization and histogram of oriented gradients for indoor positioning based on a smartphone camera. International Journal of Distributed Sensor Networks, 13(1) (2017) 1550147716686978.

[17] Gonasagi, P., Pardeshi, R., & Hangarge, M., Classification of Documents based on Local Binary Pattern Features through Age Analysis. In Ambient Communications and Computer Systems, 265-271 (2020) Springer, Singapore.

[18] Veershetty, C., & Hangarge, M. Logo retrieval and document classification based on LBP features. In Data Analytics and Learning , Springer, Singapore, (2019) 131-141.

[19] Cortes, C., & Vapnik, V. Support-vector networks. Machine learning, 20(3) (1995) 273-297.