International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 06 | June-2018 www.irjet.net p-ISSN: 2395-0072

CLASSIFICATION OF DOCUMENTS AS OLD OR NEW USING FREQUENCY DOMAIN

Annapurna A[1], Veccham Ashritha[2], Vindhya C[3], Vishnupriya B. K[4] , Harini S[5]

[1,2,3,4] *VTU & Information Science and Engineering, BNMIT, Karnataka, India*
[5]*Assistant Professor, Dept. of Information Science and Engineering, BNMIT, Karnataka, India*

Abstract - *Document classification is the task of grouping documents into categories based upon their contrast and content. The proposed approach studies the quality of handwritten and printed documents by taking into account both foreground and background information to classify it as old or new. Fourier coefficients are used to identify a given handwritten document is old or new. The printed document classification deals with effective separation of foreground and background. The proposed technique shows an excellent adaptability to tackle with problems such as uneven illumination, tampering, distortions, non-uniformity in background and foreground colors. To evaluate the proposed approach, experiments are conducted on the dataset which is collected from various sources such as college records, blue books and rough sheets. The proposed approach shows better results in terms of classification rate compared to other existing approaches.*

Key Words: **Fourier Coefficient, foreground, background, edge detection, Peak-signal-to-noise-ratio (PSNR).**

1. INTRODUCTION

A large amount of data is still represented in the form of paper documents. The official documents even though are secured with techniques such as printed patterns, paper documents suffer from a lack of security [1]. Classification of the document is an important subject for forensic purposes. The paper aging leads to changes in color due to various reasons such as its original colour, storage conditions, environment temperature, humidity etc. technological advances have made the printed document easy to modify for malicious purposes.

Document forgery could involve changing and adding some information or replacing an entire page with a counterfeited page in a document set. Location distortion will occur when a page is replaced in a genuine document. By computing the matching quality of both the images, the forged page or tampered region will be detected [2].

Common lifelong security documents are legal deeds, certificates, university/college mark sheets, bank notes etc. Forgery of such documents has increased substantially due to the advancement in technologies such as printing, scanning, and photocopying techniques. These types of fraud cannot be detected easily through untrained human eyes. Authenticity of these documents is ascertained from forensic experts by checking the relative or absolute age of a questioned document [3].

2. RELATED WORK

Lot of research is carried out in the field of classification of documents as original or fraud on both handwritten and printed documents. Some of the major works are as follows:

- Romain Bertrand *et al.* [1] found a novel approach which focuses on the detection of binarized low resolution documents which are grossly frauded. The approach uses a set of intrinsic features that are computed at character level. The main idea of the approach is to identify characters that are very similar which is the case of a copy and paste forgery or an imitation forgery.

- Amr Ahmed *et al.* [4] proposed an automatic approach for identifying static parts of a document and has introduced the DocAlign algorithm and presented different approaches for improving accuracy of DocAlign using automatic identification of static parts of documents. A matrix of pair wise matching results are produced by matching the documents in the training set using the RAST algorithm. Afterwards, the test document is matched to all documents in the training set in a pair wise fashion using the RAST algorithm. A summed score is calculated and a variant of Grubbs test was run to detect the outliers.

- Paul Luo *et al.* [5] proposed an approach which aims to detect document forgery in handwritten notes based on ink spectral information. Multiple criterion methods, based on local information of the target window, to determine ink numbers were conducted and the comparison of the results generated from different anomaly methods were presented.

- Sarah Elkasrawi *et al.* [6] proposed a semi-automatic approach which assists the news-consumers the credibility of accessing the information by means of meta-data and feature analysis of images in hyper text news articles instantaneously. For checking the authenticity of an image, a hybrid approach is proposed which includes image and text clustering techniques and also for checking the alteration in an image a hierarchical feature analysis technique is used , where different sets of features such as edges and speeded up robust features (SURF) are used.

- Biswajit Halder *et al.* [7] proposed an approach for extracting a suitable set of color features and use them for analysis to properly associate them with the ink age and to determine the ages of unknown samples a neural net is designed.

- Ricardo da Silva Barboza *et al.* [8] proposed a model based on the color components to determine the age of documents by taking into account the background of its scanned image. The images having the age gap of one to fifty-two years were taken and scanned by a HP scanner model G2410 at 300 dpi and were saved in lossless format TIFF. The proposed system considered normalized RGB-components of the samples studied which presented Gaussian like distributions.

- K. S. Raghunandan *et al.* [3] proposed an approach which studies the quality of images by considering both textual and non-textual information irrespective of content by extracting the contrast of images to determine the age of handwritten document. The main aim of the proposed system is to analyze Fourier coefficients and use contrast features for identifying the handwritten document as old or new.

- Garima Chutani *et al.* [9] proposed a new binarization approach in which the binarization algorithm is neither applied on whole images nor on sub images but instead uses the concept of bounding box and edge detection method where region localization is done and background and foreground of images are separated.

3. PROPOSED WORK:

3.1 System design for handwritten documents

The System design of proposed system, as shown in figure 1, explains the overall process of handwritten document.

First, the input color images are collected from various sources like records, the input color images which will be in RGB form are converted to grayscale images.

The fast fourier transform is applied to convert grayscale images to obtain fourier images.

The obtained fourier images are divided into positive and negative co-efficient images.

Equation 1 gives the study of distribution of positive and negative coefficients.

If(FFT(x,y)>0) Positive coefficient
If(FFT(x,y)>0) Negative coefficient……Eq.(1)

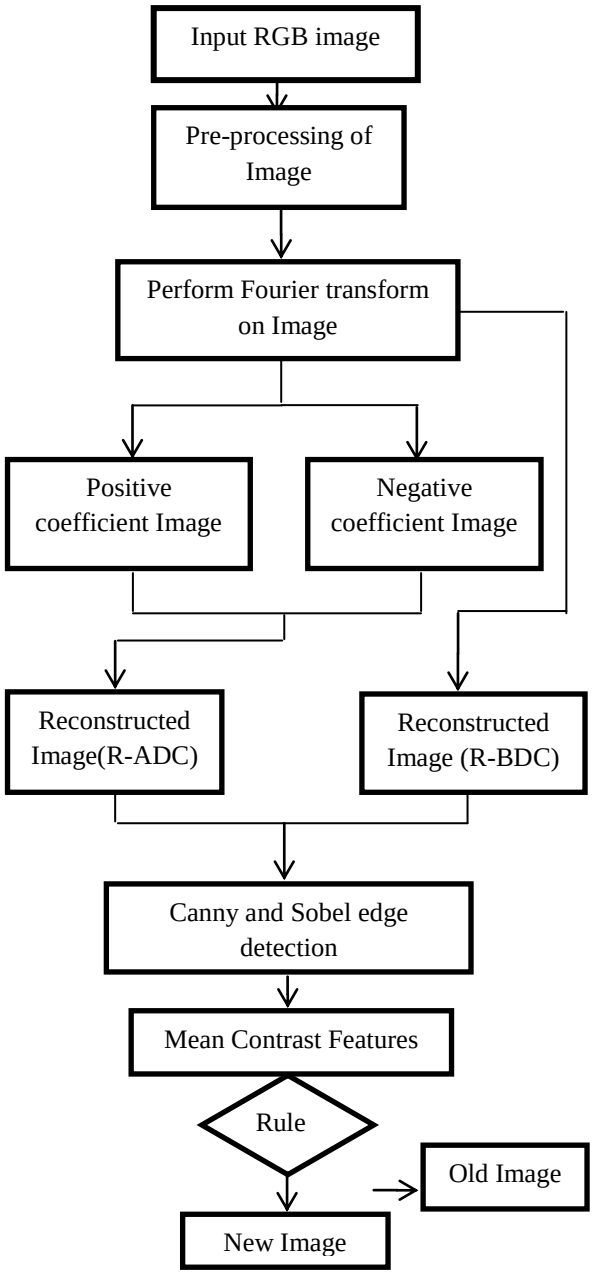

Figure 1: Work flow diagram for handwritten document

The text pixels which will represent positive frequency coefficient goes to the positive coefficients image, while the non-text pixels which represents the negative frequency coefficient goes to the negative coefficient image.

The average operation is performed on the reconstructed coefficient images as defined in equation (2), which is represented as the reconstructed image after divide and conquer (R-ADC).

RI = IFFT of Positive + IFFT of Negative.....Eq.(2)
2

The proposed method similarly reconstructs another image using Fourier image by applying Inverse Fourier transform, which is denoted as the reconstructed image before divide and conquer (R-BDC).

The difference between old and new documents can be obtained with the help of reconstructed images by applying canny and sobel edge detection on them .

The contrast features (CF) for each pixel in the Canny and Sobel edge images of R-BDC and R-ADC of the new and old documents as defined in equation(3).

$CFw(i)=\frac{1}{MN}\sum_{x=1}^{M}\sum_{Y=1}^{N}|g(x,y)$ –WindowMedian|.....Eq.(3)

Where g(x , y) represents gray image of pixel, and median of the window is given by Window Median, w is 3*3 size.

The average contrast (Avg-CF) values are calculated for each pixel in the respective images as given in equation (4).

$$\text{Avg-CF(I)}=\frac{\Sigma CF}{Number\ of\ pixels\ in\ edge\ image}\text{.....Eq.(4)}$$

The equation (5) and (6) gives the rule which classify the images into old or new document image

$$NEW=\begin{cases}CCD<SCD & 1\\ Else & 0\end{cases}\text{.....Eq.(5)}$$

$$OLD=\begin{cases}CCD>SCD & 1\\ Else & 0\end{cases}\text{.....Eq.(6)}$$

The contrast difference between Canny of R-ADC and R-BDC is given by CCD and SCD represents the contrast difference between Sobel of R-ADC and R-BDC.

3.2 System design for printed documents

As shown in figure 2, the steps for printed documents are as follows.

The input images are collected from various sources like printed notes etc.

The background is extracted to get the RGB components of the selected document and the values are subtracted with the foreground values of the image as shown in figure 2.

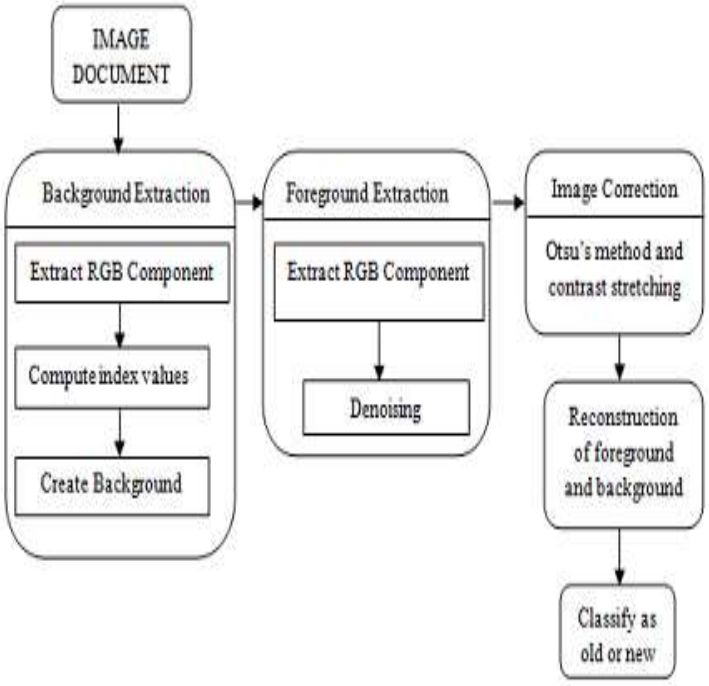

Figure 2: System design for printed documents

The new background for the image is obtained from the previously calculated values. The foreground is extracted to get the RGB components of the selected document and the values are subtracted with the background values of the image.

The Kovesi filter is applied on image to remove the unwanted noise and to preserve the image detail. To achieve consistency in dynamic range for set of image, normalization is applied on denoised image. Binarization of the original image using otsu's method is performed in order to calculate the threshold values based on which each pixel is set to either 0 or 1 that is background or foreground.

Gamma correction is performed on the original image in order to control the overall brightness of the image. To improve the contrast of the image, contrast stretching is used which helps in stretching the range of intensity values it contains to a desired range of values. Reconstructed image is obtained by taking the complement of the otsu's normalized denoised image and subtracting it with the background of the image.

The Peak-signal-to-noise-ratio (PSNR) is used for quality measurement and is applied on both the original image and the reconstructed image. The Structural similarity index (SSIM) is also applied on both the original and the reconstructed image in order to quantify the level of image degradation which could have been caused due to compression or by losses in data transmission. The threshold value is set based on SSIM for the images which will help in classifying the printed document images as old or new.

4. COMPARISON AND ACCURACY

The image dataset consisting of 100 old and 100 new documents were tested with other edge detection types along with Canny and Sobel such as Canny and Robert, Roberts and sobel, Sobel and prewitt, Canny and prewitt. The comparison and accuracy obtained among the different edge detection types are as shown in Table 1.

Table 1: Comparison and accuracy for handwritten documents

TYPES OF EDGE DETECTION	OLD	NEW
CANNY AND SOBEL	84	58
CANNY AND ROBERTS	82	52
SOBEL AND PREWITT	57	43
CANNY AND PREWITT	82	54
SOBEL AND ROBERT	40	50

Table 2: Accuracy for printed documents

PRINTED DOCUMENTS	OLD	NEW
PRINTED DOCUMENT DATASET	90	88

5. CONCLUSIONS AND FUTURE ENHANCEMENTS

The proposed system introduces an approach to classify the given document as new or old, which is further applied for identification of fraud document in forensic crime. To achieve the objective behind handwritten documents, fourier coefficients is explored to obtain reconstructed images by divide and conquer process along with edge detection, which helps in finding the difference between new and old documents.

The proposed approach based on mean contrast values defines a new rule for classifying a given image as old or new.

To achieve the objectives of printed documents, peak signal to noise ratio (PSNR) is used which measures the quality of reconstructed printed documents and also Otsu's method to automatically evaluate image thresholding which helps in finding new or old printed documents.

Restoration can be done in future as it is necessary to maintain and preserve the information contained in old documents.

ACKNOWLEDGEMENT

We consider it a privilege to express a few words of gratitude towards our college for supporting and encouraging us throughout our work.

REFERENCES

[1] R. Bertrand, P. G. Kramer, O. Terrades, P. Franco and J. M. Ogier, "A System based on Intrinsic Features for Fraudulent Document Detection", In Proc. ICDAR, 2013, pp 106-110.

[2] Shang et al. "Detecting documents forged by printing and copying", Journal on advanced signal processing", 2014, pp 10-1186.

[3] K.S. Raghunandan, B.J. Navya and G. Pooja, "Fourier Coefficients for Fraud Handwritten Document Classification Through Age Analysis", In Proc. ICFHR, 2016, pp 2167-6445.

[4] A. Ahmed and F. Shafait, "Forgery Detection based on Intrinsic Document Contents", In Proc. DAS, 2014, pp 252-256

[5] Paul Luo , "Localized Forgery Detection in Hyperspectral Document Images", In Proc. University of Science and Technology of China, August 28, 2014.

[6] S. Elkasrawi and F. Shafait, "Printer Identification using Supervised Learning for Document Forgery Detection", In Proc. DAS, 2014, pp 146-150.

[7] B. Halder and U. Garain, "Color Features based Approach for Determining Ink Age in Printed Documents", In Proc. ICPR, 2010, pp 3212-3215.

[8] R. D. S. Barboza, R. D. Lins and D. M. D. Jesus, "A Color-based Model to Determine the Age of Documents for Forensic Purposes", In Proc. ICDAR, 2013, pp 1350-1354.

[9] Garima Chutani et al, "An improved approach for automatic denoising and binarization of degraded document images based on region location", In Proc. IEEE, 2015, pp 978-1-4799.