DETECTION OF FABRICATION IN PHOTOCOPY DOCUMENT USING TEXTURE FEATURES THROUGH K-MEANS CLUSTERING

Suman V Patgar[1], Sharath Kumar Y.H[2] and Vasudev T[2]

[1]P.E.T Research Foundation, P.E.S College of Engineering, Mandya, India, 571401
[2] Maharaja Research Foundation, Maharaja Institute of Technology Mysore,
Belawadi, S.R Patna, Mandya, India, 571438

ABSTRACT

Photocopy documents are very common in our normal life. People are permitted to carry and produce photocopied documents frequently, to avoid damages or losing the original documents. But this provision is misused for temporary benefits by fabricating fake photocopied documents. When a photocopied document is produced, it may be required to check for its originality. An attempt is made in this direction to detect such fabricated photocopied documents. This paper proposes an unsupervised system to detect fabrication in photocopied document using texture features. The work in this paper mainly focuses on detection of fabrication in photocopied documents in which some contents are manipulated by new contents above it through different ways. A detailed experimental study has been performed using a collected sample set of considerable size and a decision model is developed for classification. Testing is performed with a different set of collected testing samples resulted in an average detection rate of 89%.

KEYWORDS

Fabricated photocopy document, Texture Feature, K-Means Clustering, Region of interest

1. INTRODUCTION

Many authorities in India trust and consider the photocopied documents submitted by citizens as proof and accept the same as genuine. Few such applications like to open bank account, applying for gas connection, requesting for mobile sim card, the concerned authorities insist photocopy documents like voter id, driving license, ration card, pan card and passport as proof of address, age, photo id etc to be submitted along with the application form. Certain class of people could exploit the trust of such authorities, and indulge in forging/ tampering/ fabricating photocopy document for short / long term benefits unlawfully. Fabricated photocopy is generated normally by making required changes intelligently in the photocopy obtained from an original document and then taking the recursive photocopy [1] from the modified photocopy. It is learned that in majority cases, fabrications are made by replacing a different photograph in place of photograph of an authenticated person; replacing contents in variable regions [2] through cut-and-paste technique from one or more documents; overlaying new content above actual content; adding

new content into existing content; removing some content from existing; changing content by overwriting; intellectually changing character in contents. It is quite evident from the applications listed above, the fabrication could be mainly made in the variable regions [2] of documents.

The fabricated photocopy documents are generated to gain some short term or long term benefits unlawfully. This poses a serious threat to the system and the economics of a nation. In general, such frauds are noticed in the application areas where photocopy documents are just enough. These types of systems trusting photocopied document raise an alarm to have an expert system [3] that efficiently supports in detecting a forged photocopy document. The need of such requirement to the society has motivated us to take up research through investigating different approaches to detect fabrication in photocopy document.

Further, in literature no significant effort is noticed towards research on fabricated photocopy documents except for the work of detection of fabrication in photocopy document using GLCM features [4]. In this work, attempt is made to detect fabrication of photocopy documents in which text in variable region is fabricated by putting new contents in many ways. Many research attempts are carried out on original documents instead on photocopied documents, like signature verification, detection of forged signature [5], handwriting forgery [6], printed data forgery [7], and finding authenticity of printed security documents [8]. Literature survey in this direction reveals that the above research attempts have been made in the following issues: Discriminating duplicate cheques from genuine ones [8] using Non-linear kernel function; Detecting counterfeit or manipulation of printed document [7] and this work is extended to classify laser and inkjet printouts; Recognition and verification of currency notes of different countries [9] using society of neural networks along with a small work addressing on forged currencies; Identification of forged handwriting [6] using wrinkles as a feature is attempted along with comparison of genuine handwriting.

The domain of research is in its early stage and there is no standard data set available for experimentation. Hence for the purpose of experimentation a considerable size of data samples for training and testing are collected. The samples include conference certificates, attendance certificates, birth certificates, death certificates, degree certificates, transfer certificates, DDs, Cheques and reservation letters etc. The copies were scanned using an hp flat-bed scanner to produce bitmap images at 300dpi. The noise introduced during scanning or photocopying process is cleared using median filter [10] before considering them as inputs. The fabricated samples are obtained through writing new contents over an actual contents, smeared whitener, cut and paste and adding new contents above it. Fig 1a shows a non fabricated photocopied document and Fig 1b shows a fabricated photocopied document in which the encircled area indicates the fabrication region.

The organization of the paper is as follows. In Section 2, the proposed method is explained with the support of a block diagram along with a brief introduction to Gabor texture analysis, Local Binary Patterns and Edge orientation histogram. The experimental results under varying database sizes and a qualitative comparative analysis on the state of the art techniques are discussed in Section 3. The work is concluded in Section 4.


P. E. S. COLLEGE OF ENGINEERING
Mandya
Department of Electronics & Communication Engineering
Participation Certificate
This is to certify that Mr/Mrs/Kum. Veena. M. N .....has participated in a three days VTU-VGST sponsored Faculty Development Programme on 'Advances In CMOS VLSI Design' conducted by the Department of Electronics & Communication Engineering , PES College of Engineering , Mandya on 8th,9th and 10th February 2011.
Chief Guest
Coordinator
Principal

Fig 1a: Photocopy of non fabricated Document


P. E. S. COLLEGE OF ENGINEERING
Mandya
Department of Electronics & Communication Engineering
Participation Certificate
This is to certify that Mr/Mrs/Kum. Santhosh Raikar .....has participated in a three days VTU-VGST sponsored Faculty Development Programme on 'Advances In CMOS VLSI Design' conducted by the Department of Electronics & Communication Engineering , PES College of Engineering , Mandya on 8th,9th and 10th February 2011.
Chief Guest
Coordinator
Principal

Fig1b: Photocopy of Fabricated Document

2. PROPOSED METHOD

The proposed work takes a set of segmented variable regions from a document image as input and texture features (Gabor/LBP/Edge histogram/Combination) are extracted from them. These features are classified into two classes as fabricated region and non fabricated region respectively. When a document is considered for detection of fabrication, the region of interest are identified and the contents are subjected for extraction of texture feature and a process is constructed to reduce the number of related variables. The reduced extracted features are queried to the K-means clustering to classify the same as fabricated or not fabricated. The block diagram of the proposed method is given in Fig. 2.

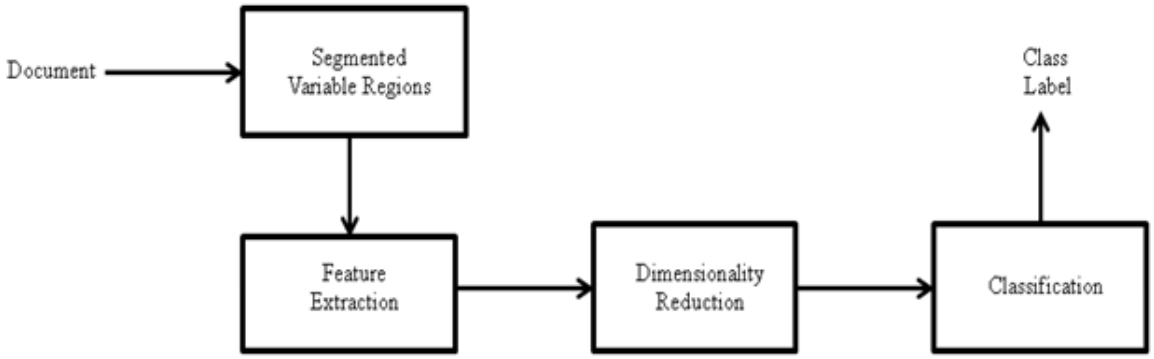

Fig 2: Block diagram of stages involved in the proposed method

2.1 Segmentation of Region of Interest

The first step in detection of fabrication is to segment the Region of Interest (ROI). Most of the variable regions contain handwritten text in the documents of the above mentioned applications. Since the fabrication could be suspected in variable regions, there is a strong need to segment variable regions as ROI from the document under consideration. Once the variable regions in a document are segmented, further investigations could be performed to check for the possibility of fabrication. The segmentation of ROI is performed using central moments [16].

2.2 Feature extraction

As work focus on to study the texture features only for document classification, Gabor response matrix, Local binary patterns and Edge orientation histogram are extracted from ROI for texture analysis. The following subsections give a brief introduction to Gabor texture features, Local binary patterns and Edge orientation histogram.

2.2.1 Gabor filter responses

Texture analysis using filters based on Gabor functions falls into the category of frequency-based approaches [11]. These approaches are based on the premise that the texture is an image pattern containing a repetitive structure that can be effectively characterized in a frequency domain, such as the Fourier domain. An attractive mathematical property of Gabor functions is that they minimize the joint uncertainty in space and frequency. They achieve the optimal trade-off between localizing the analysis in the spatial and frequency domains. Also, a Gabor filter is a linear filter whose impulse response is defined by a harmonic function multiplied by a Gaussian function. Because of the multiplication–convolution property (Convolution theorem), the Fourier transform of a Gabor filter's impulse response is the convolution of the Fourier transform of the harmonic function and the Fourier transform of the Gaussian function, and it is given by

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\left(-\frac{x^2+\gamma^2y^2}{2\sigma^2}\right)\cos\left(2\prod\frac{x^1}{\lambda}+\psi\right)$$

Here $x' = x\cos\theta + y\sin\theta$ and $y' = x\sin\theta + y\cos\theta$, and λ represents the wavelength of the cosine factor, θ represents the orientation of the normal to the parallel stripes of a Gabor function,

ψ is the phase offset, σ is the Gaussian envelope, and γ is the spatial aspect ratio specifying the ellipticity of the support of the Gabor function. A filter bank of Gabor filters with various scales and rotations is created. In this work, we have considered scales of 0, 2, 4, 6, 8, and 10 and orientations of 0°, 45°, 90°, and 135°. For each response image obtained we extract the first three moments as features.

2.2.2 Edge Orientation Histogram (EOH)

The general idea of EOH [14], is to represent an image by a histogram obtained from the predominant gradient orientations of its edge pixels. More specifically, EOH applies a Canny operator to obtain contour pixels from the image, and then it constructs a histogram of directions, counting for each bin the number of contour pixels whose gradient falls within its specific orientation interval. EOH is based on the image edge of the statistical features and invariant to scaling. It is able to accurately reflect the image edge and texture information, and extract features fast. EOH describes spatial distribution of four edges and one non-directional edge in the image.

2.2.3 Local Binary Patterns (LBP)

LBP has been widely used in texture classification because of its simplicity and efficiency [12]. LBP is a simple but efficient operator to describe local image patterns. The LBP based methods have achieved good classification results on representative texture databases. LBP is a gray-scale texture operator that characterizes the local spatial structure of the image texture. Given a central pixel in the image, a pattern code is computed by comparing its value with those of its neighborhoods:

$$\text{LBP}= \sum_{p=0}^{N-1} s(g_p - g_c)2^p, s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$$

where g_c is the gray value of the central pixel, g_p is the value of its neighbors, P is the total number of involved neighbors and R is the radius of the neighborhood. Suppose the coordinate of g_c is (0, 0), then the coordinates of g_p are given by $(R\cos 2\pi p/P), (R\sin 2\pi p/P)$. The gray values of neighbors that are not in the center of grids can be estimated by interpolation.

2.3 Dimensionality Reduction

Principal component Analysis (PCA) is used for dimensionality reduction [17]. PCA involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The goal of PCA is to reduce the dimensionality of the data while retaining as much as possible of the variation present in the original dataset. It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. In the proposed methodology feature set is reduced approximately to 50%.

2.4 Classification

K-Means algorithm is an unsupervised classification technique, where the user initiates the algorithm by specifying the number of clusters to be created from feature sets of an image [15].

This algorithm splits the given image into different clusters of features in the feature space, each of them defined by its center. Initially each feature in the image is allocated to the nearest cluster. Then the new centers are computed with the new clusters. These steps are repeated until convergence. Basically we need to determine the number of clusters K first. Then the centroid will be assumed for these clusters. We could assume random objects as the initial centroids or the first K objects in sequence could also serve as the initial centroids. In the proposed method we consider two clusters because only two classes are required. One is fabrication and another one is non fabricated.

3. EXPERIMENTAL RESULTS

In this section, we intend to study the classification accuracy under varying features of PCA. We pick samples randomly from the database and experimentation is conducted on database of more than 300 samples. The Figure 3 to 5 shows accuracy using only individual features like LBP, EOH, Gabor and their combination under varying reduction PCA features from 10% to till results is saturated. The experimentation is conducted more than five times and best one is picked from two classes. From Fig3 we can understand that the LBP features are normalized at 79.55 by 60% reductions of features, EOH features are normalized at 70.15 by 50% reductions of features, Gabor features are normalized at 82.19 by 50% reductions of features. From Fig4 we can understand that the LBP+EOH features are normalized at 80.12 by 60% reductions of features, EOH+Gabor features are normalized at 89.1 by 60% reductions of features, LBP+Gabor features are normalized at 70.15 by 50% reductions of features, Gabor features are normalized at 86.34 by 60% reductions of features. From Fig5 we can understand that the LBP+EOH+Gabor features are normalized at 88.54 by 60% reductions of features.

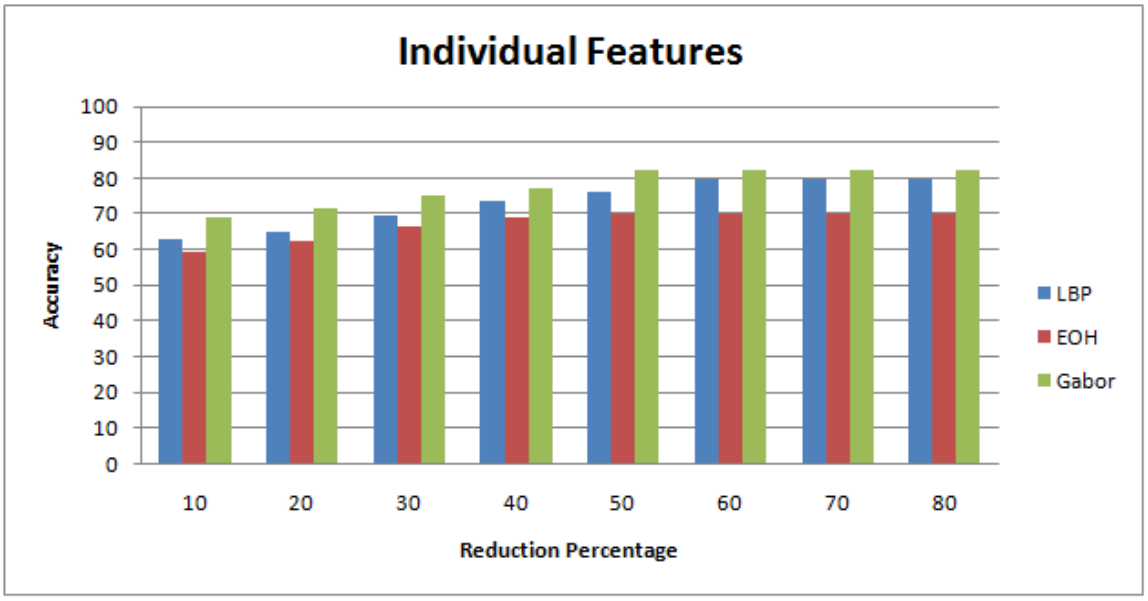

Fig 3: Graphical Representation of Individual Feature (LBP, EOH, Gabor)

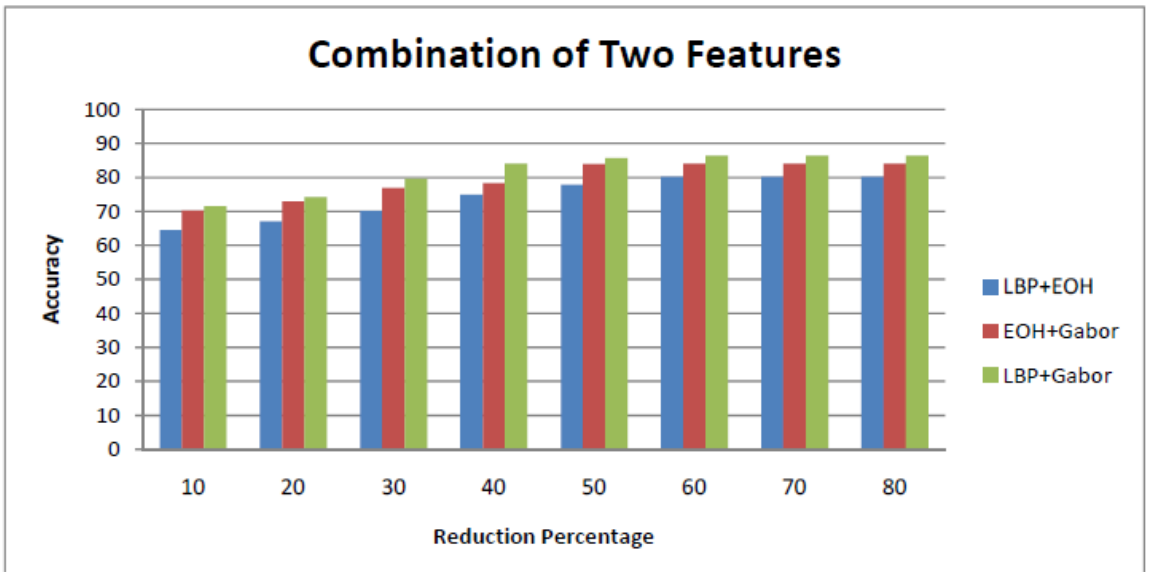

Fig4: Graphical Representation of Combinational Feature (LBP+EOH, EOH+Gabor, LBP+Gabor)

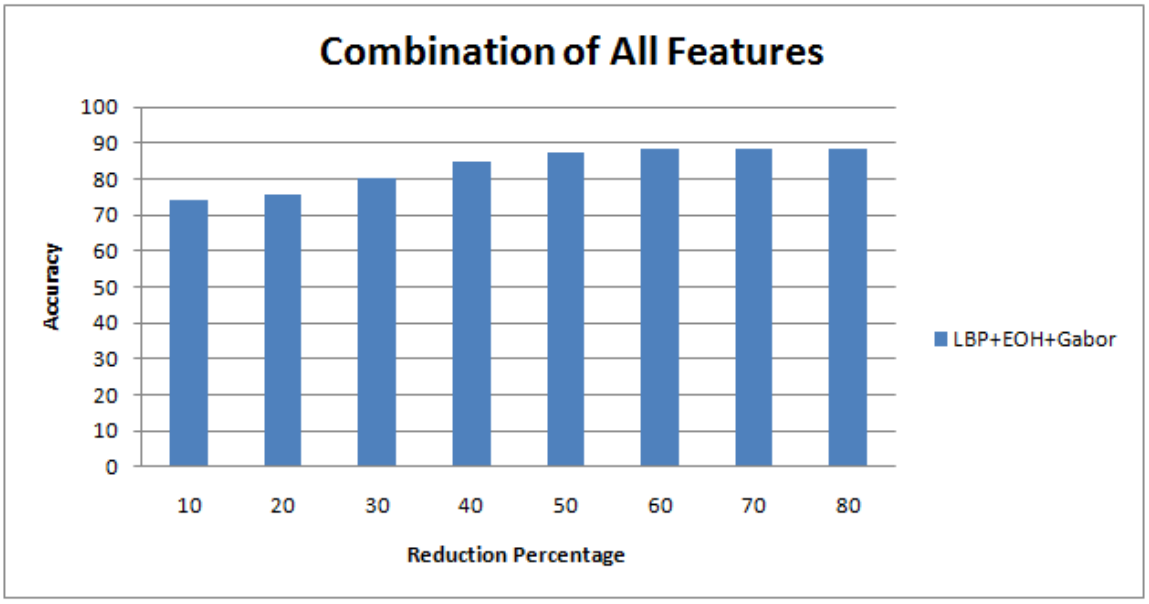

Fig5: Graphical Representation of Combination of all Feature(LBP+EOH+Gabor)

4. CONCLUSION

The implemented method serves as an unsupervised intelligent system for detection of fabricated photocopy in which fabrication/forgery is made through different ways on a photocopy document. The method is essentially based on texture analysis using Gabor features, LBP and EOH. It shows an average classification efficiency of 89%. Suitable texture features such as LBP,

EOH, and Gabor responses are explored for the purpose of document classification. It is observed that using the proposed textural features one can achieve relatively a good classification accuracy when compared to any other available features. The method can be used on photocopied document having two or more ROI. The method cannot be extended effectively for the photocopied document having only one ROI. The experimental results have shown that using combined features outperforms any individual feature. It can be used without a complex hardware setup to detect fabrication in photocopy document in applications where only photocopy documents are sufficient. The misclassification is due to dirt and background art in photocopy document. A small amount of fabrication like changing a character or a part of character in the ROI also accounts for misclassification

.

REFERENCES

[1] Suman Patgar, Vasudev T, 2013, Estimation of Recursive Order Number of a Photocopied Document Through Entropy From Gray Level Co-occurrence Matrix, ICSEC2013, pp 313-317

[2] Vasudev T, 2007, Automatic Data Extraction from Pre-Printed Input Data Forms: Some New Approaches, PhD thesis supervised by Dr. G. Hemanthakumar, University of Mysore, India.

[3] Rich Kevin Knight, Artificial Intelligence, 2nd Edition, McGraw-Hill Higher Education.

[4] Suman Patgar, Vasudev T, 2013 An Unsupervised Intelligent System to Detect Fabrication in Photocopy Document Using Geometric Moments and Gray Level Co-Occurrence Matrix, IJCA(0975-8887) volume- 74/N0. 12, July 2013.

[5] Madasu Hanmandlu, Mohd. Hafizuddin Mohd. Yusof, Vamsi Krishna Madasu, off-line signature verification and forgery detection using fuzzy modeling Pattern Recognition Vol. 38, pp 341-356, 2005

[6] Cha, S.-H., & Tapert, C. C., Automatic Detection of Handwriting forgery, Proc. 8thInt.Workshop Frontiers Handwriting Recognition(IWFHR-8), Niagara, Canada, pp 264-267, 2002

[7] Christoph H Lampert, Lin Mei, Thomas M Breuel, Printing Technique Classification for Document Counterfeit Detection Computational Intelligence and Security, International Conference, Vol. 1, pp 639-644, 2006

[8] Utpal Garian, Biswajith Halder, On Automatic Authenticity Verification of Printed Security Documents, IEEE Computer Society Sixth Indian Conference on Computer vision, Graphics & Image Processing, pp 706-713, 2008

[9] Angelo Frosini, Marco Gori, Paolo Priami, A Neural Network-Based Model For paper Currency Recognition and Verification IEEE Transactions on Neural Networks, Vol. 7, No. 6, Nov 1996

[10] Kuo Chin Fan, 2001, Marginal Noise Removal of Document Image, ICDAR 01, pp 317-321.

[11] S.D. Newsam, C. Kamath, in: SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology, pp. 21–32.

[12] T. Ojala, M. Pietikäinen, and T. T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with Local Binary Pattern, IEEE Trans. on PAMI 24(7), pp. 971-987, 2002.

[13] K. Somasundaram, N. Kalaichelvi,, Feature Identification in Satellite Images using K-Means Segmentation , National Conference on Signal and Image Processing (NCSIP-2012)

[14] A. Pinheiro, Image descriptors based on the edge orientation, in: Semantic Media Adaptation and Personalization, 2009. SMAP '09. 4th International Workshop on, 2009, pp. 73–78.

[15] K. Somasundaram1 and N. Kalaichelvi, Feature Identification in Satellite Images using K-Means Segmentation, NCSIP-2012, pp.264-269

[16] Suman Patgar, Vasudev T, Segmentation of Handwritten Text from Underlined Variable Regions in Documents, IJIP-2014,accepted

[17] Chris Ding and Xiaofeng He, "K-Means Clustering via Principal Component Analysis", In proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004