An Unsupervised Intelligent System to Detect Fabrication in Photocopy Document using Geometric Moments and Gray Level Co-Occurrence Matrix

Suman V Patgar
P.E.T Research Foundation,
P.E.S College of Engineering,
Mandya, Karnataka, India- 571401

Vasudev T
Maharaja Research Foundation,
MIT, Belawadi, S.R Patna,
Mandya, Karnataka, 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 two level classification system to detect fabricated photocopied document using Geometric moments and Gray Level Co-Occurrence Matrix features. The work in this paper mainly focuses on detecting fabrication of photocopied document in which some contents are manipulated by smearing whitener over the original content and writing new contents above it. 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 94.59%.

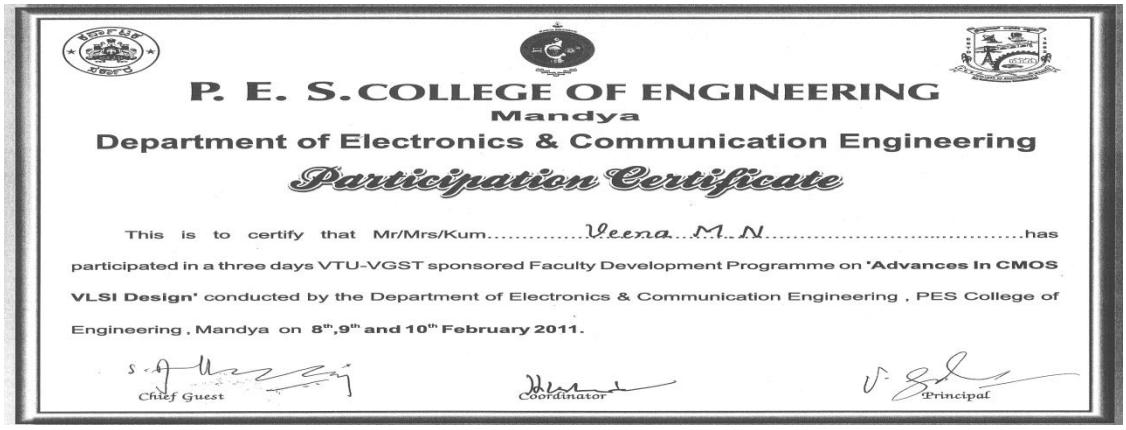
Keywords
Fabricated photocopy document, Geometric Moments, Gray Level Co-occurrence Matrix, GLCM features, text contour.

1. INTRODUCTION
Many authorities 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 the authorities, and indulge in forging/ tampering/ fabricating photocopy document. A fabricated photocopy is the recursive photocopy generated from an intellectually modified photocopy of an authenticated document. These things would be deliberately made at the time of obtaining the photocopy of the document without damaging the original document. It is learned that in majority of the cases fabrications are made by replacing a different photograph in place of photograph of authenticated person, replacing contents in variable regions [1], 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.

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 [2] 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.

Many research attempts are carried out on original documents instead on photocopied documents, like signature verification, detection of forged signature [3], handwriting forgery [4], printed data forgery [5], and finding authenticity of printed security documents [6]. Literature survey in this direction reveals that the above research attempts have been made in the following issues: Discriminating duplicate cheques from genuine ones [6] using Non-linear kernel function; Detecting counterfeit or manipulation of printed document [5] and this work is extended to classify laser and inkjet printouts; Recognition and verification of bank notes of different country [7] using society of neural networks along with a small work addressing on forged bank currencies; Identification of forged handwriting [4] using wrinkles as a feature is attempted along with comparison of genuine handwriting.

Further, in literature to the best of our knowledge no significant effort is noticed towards detecting forgery made while taking photocopy. This domain of research is in its early stage and there is no standard data set is 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 [8] before processing. Fig 1 shows a non fabricated photocopied document and Fig 2 shows a fabricated photocopied document in which the encircled area indicates the fabrication region.

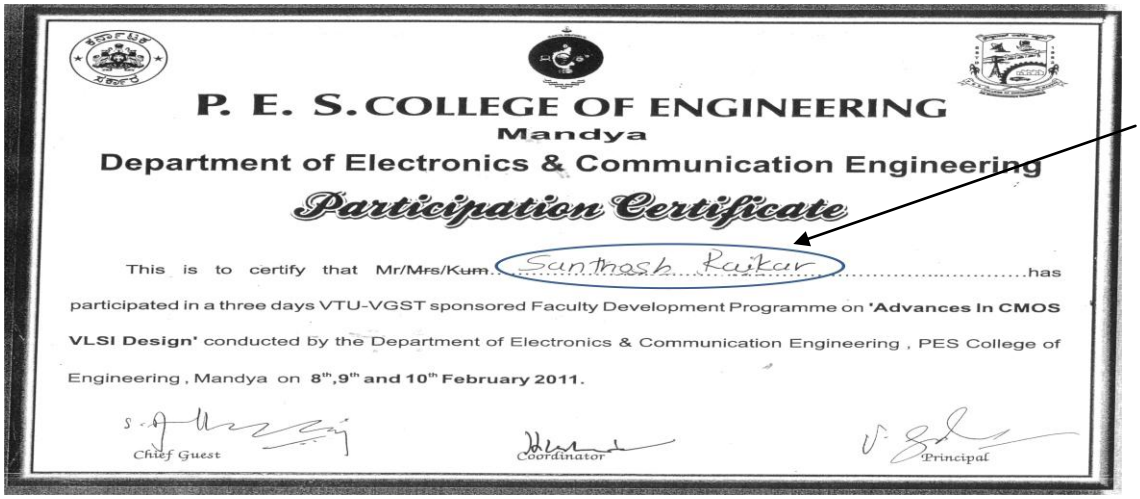


P. E. S. COLLEGE OF ENGINEERING
Mandya
Department of Electronics & Communication Engineering
Participation Certificate

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 1: 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

Fig 2: Photocopy of Fabricated Document

Fig 3 shows few segmented region of interest (ROI) from different photocopied document which are suspected to be fabricated through smearing whitener over original content and writing different content above it.

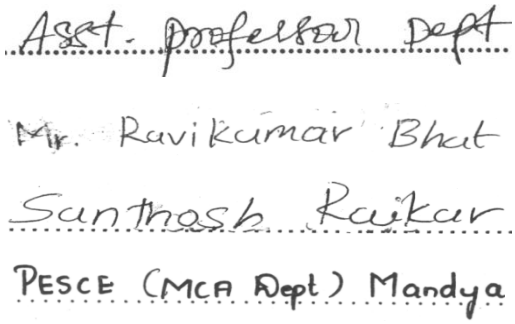


Fig 3: Segmented ROI which are suspected to be fabricated

Whenever the photocopies are submitted as proof, the original document will not be available for verification or comparison. In such situations, an unsupervised model is required for detection of fabrication in photocopy documents. The visual analysis performed on non fabricated and fabricated samples collected, exhibit insignificant texture variations in fabricated photocopy document and no such variations are noticed in non fabricated photocopy document. This led us also to explore the effect on contours of text in photocopy document. Accordingly, a consistent intensity level with smooth and strong edge contour is obtained for non fabricated text. On the other hand inconsistent intensities with rough and relatively weak edge contour is resulted for fabricated photocopy text. In addition, maximum distortion is noticed in contour of fabricated photocopy text. Fig 4 shows contour of non fabricated photocopy text and Fig 5 shows contour of fabricated photocopy text.

Fig 4: Contour of non Fabricated Text

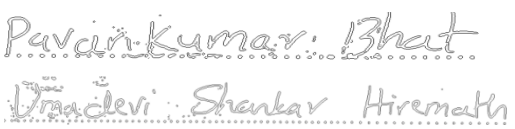


Fig 5: Contour of Fabricated Text

It is quite evident from the principles of digital image processing [9] that the Geometric Moments (GM) and Gray Level Co-occurrence Matrix (GLCM) features are the best analytical approaches for texture analysis in images and this research is carried out to investigate a classification model using GM and GLCM features on ROI.

Initially it was attempted to classify fabricated photocopies and non fabricated photocopies using GM up to 3rd order and GLCM features individually on text in ROI of the document. The approaches did not yield required threshold for classification since the volume of text in the fabricated area is small in ROI. Further, experimental analysis on modified 2nd order GM on text in ROI exhibited 3 ranges of values for classification; 1st range for non fabricated photocopied text, 2nd range for fabricated text and 3rd range for conflicting cases. A second level classification is investigated to resolve conflicting cases using GLCM features. This approach provides a good range of values for classification with relatively little number of misclassifications. In this work it is assumed that, ROI is segmented out from the document image and considered as input for classification.

The remainder of the paper is organized as follows: Section 2 gives introduction to Geometric Moments, Gray Level Co-occurrence Matrix and their applications. Section 3 describes the methodology adopted for developing a two level decision model to find fabrication in photocopy document. The results of experiment are discussed in section 4 and section 5 concludes the work.

2. GEOMETRIC MOMENTS AND GRAY LEVEL CO-OCCURRENCE MATRIX

Geometric moments are geometrical features of pixel distribution in an image [9,10] and extensively applied in many applications of image processing particularly in classification problems. Gray level co-occurrence matrix is one of the most known texture analysis methods [11]. It helps in estimating image properties related to second-order statistics.

2.1 Geometric Moments

The moments are geometrical features obtained from the document image based on pixel distribution. Seven orders of moments can be computed for an image, which are derived from geometrical features that are invariant to translation, scaling and rotation [9, 10]. The main characteristic of these moments is, higher the order of moments greater is the classifying range of values. Because of this characteristic, moments are extensively used in classification problems. First and second order moments are mainly used in computing texture measures in the images [1]. A modified second order moments [12] are used in this application for better discrimination and these are also invariant to scaling and translation. The modified invariant moments are derived from the first and second order geometric moments [9, 10] of the image. The first order geometric moments m_{00}, m_{01}, m_{10} are evaluated using the computation indicated in expressions (2.1.1). The terms $f_{xy,}$ H and W indicate the intensity value of the image *F* at (x, y) coordinate, height and width of the image respectively.

$$m_{00} = \sum_{x=1}^{H} \sum_{y=1}^{W} f_{xy}$$

$$m_{10} = \sum_{x=1}^{H} \sum_{y=1}^{W} x * f_{xy}$$

$$m_{01} = \sum_{x=1}^{H} \sum_{y=1}^{W} y * f_{xy} \qquad (2.1.1)$$

The mean of the image in x and y directions are calculated using expressions (2.1.2).

$$\bar{x} = \frac{m_{10}}{m_{00}} \qquad \bar{y} = \frac{m_{01}}{m_{00}} \qquad (2.1.2)$$

The 2nd order geometric moments μ_{20} and μ_{02} in x and y directions respectively are computed using expressions (2.1.3).

$$\mu_{20} = \sum_{x=1}^{H} \sum_{y=1}^{W} (x - \bar{x}) f_{xy}$$

$$\mu_{02} = \sum_{x=1}^{H} \sum_{y=1}^{W} (y - \bar{y}) f_{xy} \qquad (2.1.3)$$

The 2nd order geometric moments obtained from (2.1.3) for a small area of document is unable to provide distinct values for classification. Hence a modified 2nd order invariant GM are defined [12] to get better range of values for classification. The modified 2nd order moments are obtained through the modification factors x_s and y_s, which are computed as defined in expressions (2.1.4).

$$x_s = \sqrt{\frac{\mu_{20}}{m_{00}}} \qquad y_s = \sqrt{\frac{\mu_{02}}{m_{00}}} \qquad (2.1.4)$$

The modified 2nd order moments invariant to scale and translation are computed as given in expressions (2.1.5a, 2.1.5b, 2.1.5c).

$$\varphi_{20} = \frac{\sum_{x=1}^{H} \sum_{y=1}^{W} (x - \bar{x} + x_s)^2 f_{xy}}{m_{00}^2} \qquad (2.1.5a)$$

$$\varphi_{02} = \frac{\sum_{x=1}^{H}\sum_{y=1}^{W}(y-\bar{y}+y_s)^2 f_{xy}}{m_{00}^2} \quad (2.1.5b)$$

$$\varphi_{11} = \frac{\sum\sum(x-\bar{x}+x_s)*(y-\bar{y}+y_s)* f_{xy}}{m_{00}^2} \quad (2.1.5c)$$

φ_{20}, φ_{02}, and φ_{11} are the modified 2nd order moments in x, y and x-y directions respectively. The modified 2nd order moment φ_{20} in x direction from 2.1.5a is found suitable feature as it provides good classification range of values compared to other two features.

2.2 Gray Level Co-occurrence Matrix and GLCM features

Texture is an important characteristic used in identifying objects or regions of interest in an image. A statistical method of examining texture that considers the spatial relationship of pixels is the gray level co-occurrence matrix and is also known as the gray level spatial dependence matrix. The approach has been used in many applications, like carpet wear assessment [13], texture feature extraction [14] and image texture segmentation [15]. A GLCM of an image is a square matrix in which the number of rows and columns is equal to the number of gray levels in that image [11]. GLCM of an image [9] is constructed for an image *F* with K possible intensity levels and Q being an operator that defines the position of two pixels relative to each other. In other words *G* is the matrix obtained in which each element g_{ij} is the number of times that pixel pairs with intensities z_i and z_j occur in *F* at the relative position specified by Q, where 1≤ i, j≤ k. GLCM is the basis for extracting texture features of an image. The texture features of an image are characterized by a set of descriptors known as Haralick features [11]. Each texture feature is obtained from the normalized probability density matrix *P* of GLCM *G* is given by

$$p_{ij} = g_{ij} /n \quad (2.2.1)$$

where n is the total number of pixel pairs that satisfy Q in *G* which is same as the sum of all the elements of *G*. The probability values p_{ij} are in the range [0, 1] and their sum is defined as

$$\sum_{i=1}^{K}\sum_{j=1}^{K} p_{ij} = 1 \quad (2.2.2)$$

Equations 2.2.3 to 2.2.7 define the computations of different Haralick texture features of an image *F* obtained through *P*

$$\text{Contrast } (c_1) = \sum_{i=1}^{K}\sum_{j=1}^{K}(i-j)^2 p_{ij} \quad (2.2.3)$$

$$\text{Correlation } (c_2) = \sum_{i=1}^{K}\sum_{j=1}^{K}\frac{(i-m_r)(j-m_c)p_{ij}}{\sigma_r\sigma_c} \quad (2.2.4)$$

where $m_r = \sum_{i=1}^{K} i\sum_{j=1}^{K} p_{ij}$ $\quad m_c = \sum_{i=1}^{K} j\sum_{j=1}^{K} p_{ij}$

and $\sigma_r^2 = \sum_{I=1}^{K}(i-m_r)^2\sum_{j=1}^{K} p_{ij}$

$\sigma_c^2 = \sum_{I=1}^{K}(j-m_c)^2\sum_{i=1}^{K} p_{ij}$

$$\text{Energy } (e_1) = \sum_{i=1}^{K}\sum_{j=1}^{K} p_{ij}^2 \quad (2.2.5)$$

$$\text{Homogeneity } (h) = \sum_{i=1}^{K}\sum_{j=1}^{K}\frac{p_{ij}}{1+|i-j|} \quad (2.2.6)$$

$$\text{Entropy } (e2) = -\sum_{i=1}^{K}\sum_{j=1}^{K} p_{ij}\log_2 p_{ij} \quad (2.2.7)$$

The modified 2nd order GM and GLCM features are applied for the text in ROI and contour of text in ROI in order to analyze the texture distortions for the purpose of classification. The methodology followed to develop classification model is described in the next section.

3. METHODOLOGY

The method considers the segmented ROI from the photocopy document as input. The modified 2nd order GMs φ_{20} and φ^1_{20} in x direction are obtained using 2.1.5a for the text in ROI and contour of text in ROI respectively. The rate of degradation in texture is computed as

$$\boldsymbol{\Phi} = \frac{\varphi_{20}}{\varphi^1_{20}} \quad (3.1)$$

In continuation to the discussion made in section-1, the contour of non fabricated photocopy text is smooth and consistent. The GM φ^1_{20} value of such text is normally high and makes degradation rate $\boldsymbol{\Phi}$ to remain lesser than a threshold. Further the contour of fabricated photocopy text is rough and inconsistent. The GM φ^1_{20} of such text is normally low and keeps the degradation rate $\boldsymbol{\Phi}$ much higher than a threshold value. Based on the training set, two thresholds T_{g1}=5.0 and T_{g2}=7.0 are identified for degradation rate $\boldsymbol{\Phi}$. A first level classification is made using these thresholds T_{g1} and T_{g2}. The values *of* $\boldsymbol{\Phi}$ less than or equal to T_{g1} are classified as non fabricated photocopy text and the values of $\boldsymbol{\Phi}$ greater than or equal to T_{g2} are classified as fabricated photocopy text. Any values of $\boldsymbol{\Phi}$ between T_{g1} and T_{g2} are considered as the cases of overlap/conflict. Table 1 shows the decision table for 1st level classification.

Table 1: Decision table for 1st level classification

$\boldsymbol{\Phi} = \frac{\varphi_{20}}{\varphi^1_{20}}$	**Fabricated**	**Non-Fabricated**	**Conflicts**
$\leq T_{g1}$	N	Y	N
$\geq T_{g2}$	Y	N	N
$>T_{g1}$ and $<T_{g2}$	N	N	Y

The conflict/overlap cases during first level classification are further subjected to a second level classification using GLCM features to resolve conflicts. The GLCM features Correlation (c_1), Contrast (c_2), Energy (e_1), Homogeneity (h) and Entropy (e_2) obtained from 2.2.3 to 2.2.7 do not exhibit considerable variations for fabricated and non fabricated photocopy text. Whereas Correlation (c_1^1), Contrast (c_2^1), Energy (e_1^1), Homogeneity (h^1) and Entropy (e_2^1) for the contour of text shows considerable variations between fabricated and non fabricated photocopy text. For a fabricated photocopy text values c_1^1, c_2^1, e_2^1 decreases and the values e_1^1, h^1 increases. Based on the above texture features five degradation parameters are defined as follows:

$$\phi_{c1} = \frac{c_1}{c_1^1} \qquad \phi_{c2} = \frac{c_2}{c_2^1} \qquad \phi_{e1} = \frac{e_1}{e_1^1}$$

$$\phi_h = \frac{h}{h^1} \qquad \phi_{e2} = \frac{e_2}{e_2^1} \qquad (3.2)$$

Again, using the same training sample set, thresholds are identified for the above five degradation parameters as T_{c1}=13.6, T_{c2}=5.8, T_{e1}=0.80, T_h=0.95 and T_{e2}=7.6 respectively. A voting procedure is adopted to decide the photocopied text is non fabricated or fabricated. The decision tree for second level classification is shown in Fig 6. The notations F3, F4, F5, F6 and F7 used in decision tree represent the following conditions:

F3 $\Rightarrow \phi_{c1} \geq T_{c1}$, F4$\Rightarrow \phi_{c2} \geq T_{c2}$, F5$\Rightarrow \phi_{e1} \leq T_{e1,}$

F6$\Rightarrow \phi_h \leq T_{h,}$ F7$\Rightarrow \phi_{e2} \geq T_{e2}$

In the decision tree shown in fig 6, the square boxes indicate conditions and circles indicate the decisions. The terms F and NF denote the decision as fabricated and non fabricated photocopy text respectively. A true in conditions branch to the left and a false in condition takes right branch.

The block diagram in Fig 7 shows computational process involved in the proposed methodology.

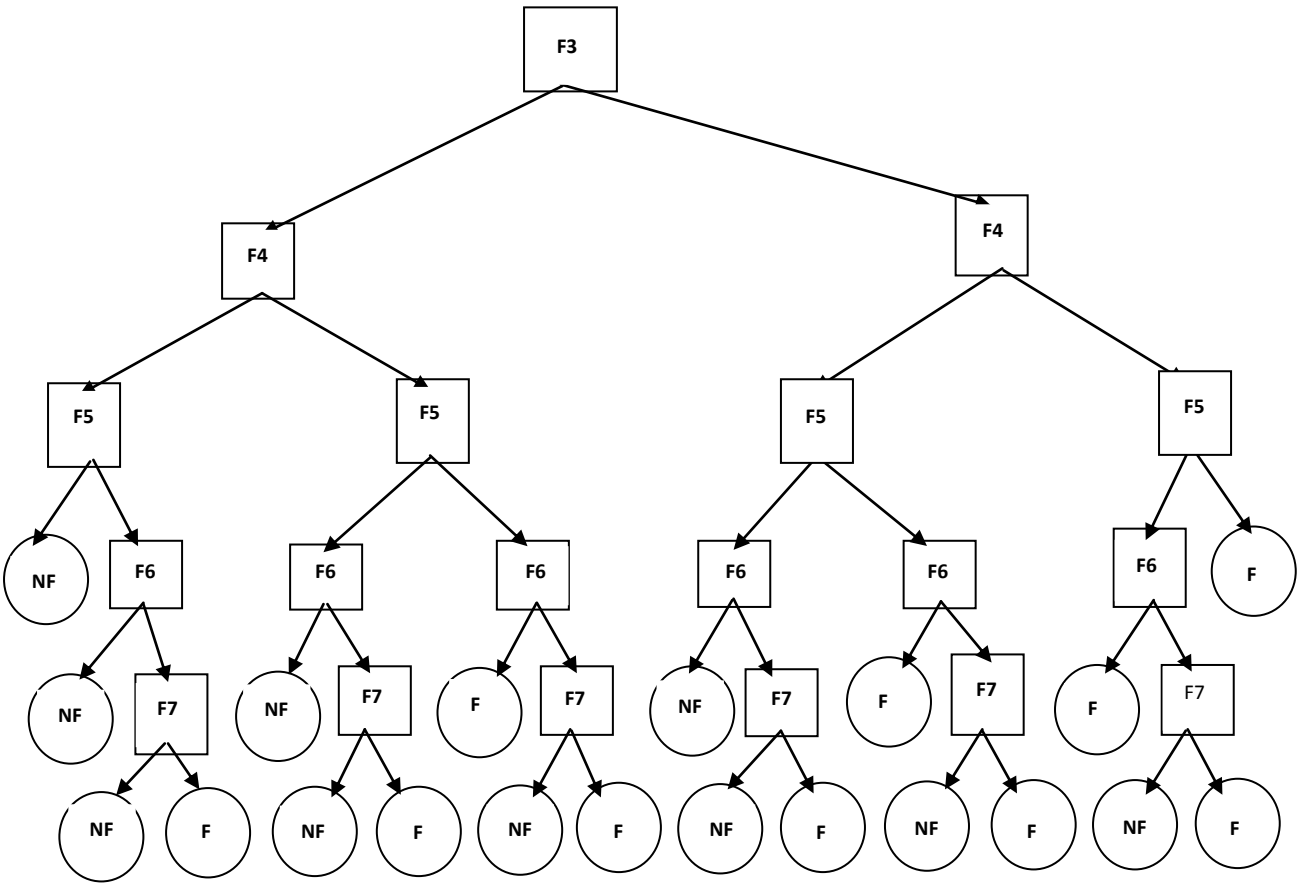


Fig 6: Decision tree for second level classification

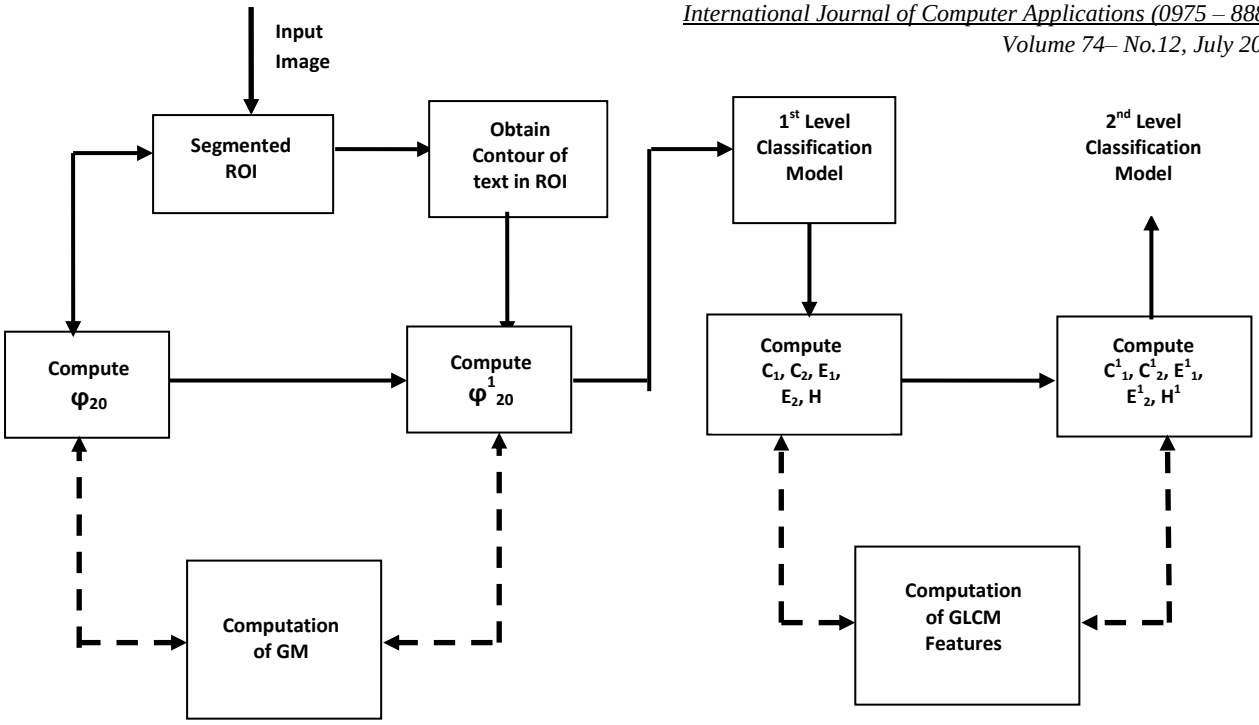


Fig 7: Block diagram of computational process

4. EXPERIMENTAL RESULTS

Testing of the proposed method is carried out using collected samples of photocopy documents. Experimentation is carried out with 185 number of test samples in which 105 samples are fabricated photocopy documents obtained by smearing whitener on some part of text and writing different text above it. The remaining samples are non fabricated photocopy samples. The result of the first level classification is shown in table 2 and result in the second level classification for conflicting cases of level-1 is shown in table 3. The overall result of the developed method is given in table 4.

Table 2: First level classification results

Type of Samples	No. of samples	No. of correct classification	Conflicts	No. of Misclassification	Efficiency
Fabricated	105	78 (74.29%)	25 (23.81%)	02(1.90%)	74.29%
Non Fabricated	80	67 (83.75%)	12 (15.00%)	01(1.25%)	83.75%
Total	185	145(78.38%)	37 (20.00%)	03(1.62%)	78.38%

Table 3: Second level classification results

Type of Samples	No. of samples	No. of correct classification in	No. of Misclassification	Efficiency
Fabricated	25	20 (80.00%)	5(20.00%)	80.00%
Non Fabricated	12	10 (83.33%)	2(16.67%)	83.33%
Total	37	30 (81.08%)	7(18.92%)	81.08%

Table 4: Overall efficiency of method

Type of Samples	No. of samples	No. of correct classification	No. of Misclassification	Efficiency
Fabricated	105	98(93.33%)	7(6.67%)	93.33%
Non Fabricated	80	77(96.25%)	3(33.75%)	96.25%
Overall efficiency	185	175(94.59%)	10(5.41%)	94.59%

In first level classification, for fabricated photocopied text, efficiency is 74.29%, conflicts are 23.81% and misclassifications are 1.90%. In case of non fabricated photocopied text, the first level classification efficiency is 83.75%, conflicts are 15% and misclassifications are 1.25%. First level classification results a total efficiency of 78.38% with 1.62% of misclassifications and 20% of conflicts. In second level classification, for the conflicting cases of first level, the efficiency for fabricated photocopy text is 80.00% and misclassification is 20.00%. Similarly, for non fabricated photocopy text classification efficiency is 83.33% with 16.67% of misclassification. Second level classification results a total efficiency of 81.08% with 18.92% of misclassification.

The efficiency of the combined levels is 93.33% for fabricated photocopied text and 96.25% for fabricated text. The overall efficiency of the proposed method is 94.59%. The main reasons for misclassification in non fabricated photocopy document is due to the presence of noise, quality of writing, background art, more usage and folding in the input document. Misclassification in fabricated document is noticed due to very small or little alteration/changes during fabrication.

5. CONCLUSION

The implemented method serves as an unsupervised intelligent system for detection of fabricated photocopy in which fabrication/forgery is made through smearing whitener on some text and overwriting different text on a photocopy document. The method is essentially based on statistical moments of intensity values. It shows an average classification efficiency of 94.59%. 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. There is much scope to enhance the performance efficiency through exploring higher order geometric moments and also preparing the document free from noise, dirt and background art. A work is under investigation to have a classification based on single approach rather than a hybrid approach to have a better computational efficiency.

6. ACKNOWLEDGMENTS

We thank PET Research Foundation and Maharaja Research Foundation for providing the facilities required to carry out research in the area of Document Image Analysis (DIA). We also thank Dr. S Murali for his constant support, suggestions and guidance in this research.

7. REFERENCES

[1] 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.

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

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

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

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

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

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

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

[9] Rafael C Gonzales & Richard E Woods, 2002, Digital Image Processing, 2nd Edition, Pearson Education Publication.

[10] Jain A K, 1998, Fundamentals of Digital Image Processing, Prentice Hall, Englewood Cliffs, NJ.

[11] R Haralick K Shanmugam and I Dinstein, 1973 "Textural Features for Image classification", IEEE Trans on system Man and Cybernetics SMC-3(6): 610-621,.

[12] Angadi S A, 2007,Postal Automation, PhD thesis submitted under the supervision of Dr.Nagabhushan P, University of Mysore, India

[13] L.H. Siew, R.H. Hodgson, and E.J. Wood, 1988, Texture Measures for Carpet Wear Asessment, IEEE Trans. on Pattern Analysis and Machine Intell., Vol. PAMI-10, pp. 92-105.

[14] D.C He, L Wang and J Juibert, Texture Feature extraction, Pattern Recognition Letter, Vol 6 pp 269-273.

[15] Rishi Jobanputra, David A Clausi, 2006, Preserving boundaries for image texture segmentation using greylevel co-occurring probabilities Pattern Recognition 39 234-245.