Estimation of Recursive Order Number of a Photocopied Document Through Entropy From Gray Level Co-occurrence Matrix

Suman V Patgar
PET Research Foundation
PES College of Engineering
Mandya, Karnataka, India 581401
sumanpatgar@gmail.com

Vasudev T
Maharaja Research Foundation
Maharaja Institute of Technology Mysore
SR Patna, Mandya, Karnataka, India
vasu@mitmysore.in

Abstract—**Photocopy documents are very common in our normal life. In country like India, people are permitted to carry and present photocopied documents to avoid damages to the original documents. But this provision is misused for temporary benefits by fabricating fake photocopied documents. Fabrication of fake photocopied document is possible only in 2nd and higher recursive order of photocopies. Whenever a photocopied document is submitted in place of original document, it may be required to check its originality. When the document is 1st order photocopy, chances of fabrication can be ignored. On the other hand when the photocopy recursive order is 2 or above, probability of fabrication may be suspected. Hence when a photocopied document is presented instead of original document, the recursive order number of photocopy is to be estimated to ascertain the originality. This requirement demands to investigate a method to estimate order number of photocopy. It is noticed that the degradation in photocopy obtained increases as the recursive order of photocopy increases. Considering the degradation as a feature in this work, a method based on entropy from gray level co-occurrence matrix is proposed to estimate the recursive order number of the photocopied document through computing degradation rate. A detailed experimentation is performed on a generated data set and the method exhibits efficiency close to 95%.**

Keywords—fabricated photocopy documents, recursive order number, gray level co-occurrence matrix, entropy

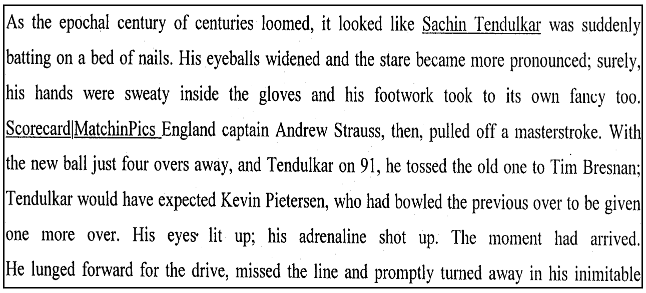
I. Introduction

Many authorities in India trust and accept the self attested photocopied documents submitted by citizens as proof and consider the same as genuine. Few such applications like to open bank account, applying for gas connection, requesting for mobile sim card, concerned authorities insist self attested 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. These things would be deliberately made at the time of obtaining the photocopy of document without damaging the original document. It is learned that in majority of the cases fabrications are made intelligently by changing/ replacing/ overwriting/ removing/ adding contents in place of authenticated content.

The fabricated photocopy documents are generated to gain some short term and long term benefits unlawfully. This poses a serious threat to the system and the economics of a nation. The types of systems trusting photocopied document as proof raise an alarm to have an expert system [1] that efficiently supports in detecting a fabricated 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. It is quite evident from the above discussions that the probability of fabricating a photocopy is zero in the 1st photocopy obtained from the original document, where as the fabrication may be suspected in the higher recursive order photocopy. Hence it is a prerequisite to find the recursive order of a photocopy submitted in place of original. Whenever the estimated order number is 2 or above, further investigation can be carried out to detect the possibility of fabrication in photocopy, which is not within the scope of this paper. Considering the above aspects an attempt is made to estimate the recursive order number of the photocopy as an initial stage leading to further investigation on fabrication in photocopy.

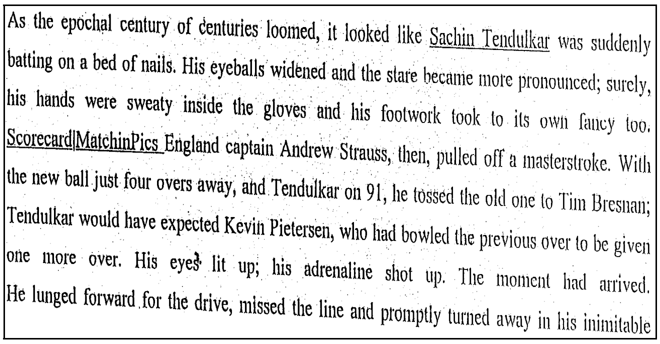
Further, in literature to the best of our knowledge no significant effort is noticed towards detecting fabrication made while taking photocopy. Many research attempts are carried out on original documents like signature verification, detection of forged signature [2], handwriting forgery [3], printed data forgery [4], and finding authenticity of printed security documents [5]. Literature survey in this direction reveals that the above research attempts are made in the following issues: Discriminating duplicate cheques from genuine ones [5] using non-linear kernel function; Detecting counterfeit or manipulation of printed document [4] and this work is extended to classify laser and inkjet printouts; Recognition and verification of currency notes of different countries [6] using society of neural networks along with a small work addressing on fake currencies; Identification of forged handwriting [3] using wrinkles as a feature is attempted along with comparison of genuine handwriting.

As the domain under consideration is new for research, research literature and standard data set are not available to compare our work with results of others. Hence for the purpose of experimentation sufficient numbers of sample photocopies are obtained from different copier machines to generate different recursive order copies. The photocopies were scanned using a scanner to produce bitmap images at 300dpi. Fig. 1 and Fig. 2 show the samples of 1st order and 5th order recursive photocopies of a document respectively.



As the epochal century of centuries loomed, it looked like Sachin Tendulkar was suddenly batting on a bed of nails. His eyeballs widened and the stare became more pronounced; surely, his hands were sweaty inside the gloves and his footwork took to its own fancy too. Scorecard|MatchinPics England captain Andrew Strauss, then, pulled off a masterstroke. With the new ball just four overs away, and Tendulkar on 91, he tossed the old one to Tim Bresnan; Tendulkar would have expected Kevin Pietersen, who had bowled the previous over to be given one more over. His eyes lit up; his adrenaline shot up. The moment had arrived. He lunged forward for the drive, missed the line and promptly turned away in his inimitable

Fig. 1 1st order photocopy



As the epochal century of centuries loomed, it looked like Sachin Tendulkar was suddenly batting on a bed of nails. His eyeballs widened and the stare became more pronounced; surely, his hands were sweaty inside the gloves and his footwork took to its own fancy too. Scorecard|MatchinPics England captain Andrew Strauss, then, pulled off a masterstroke. With the new ball just four overs away, and Tendulkar on 91, he tossed the old one to Tim Bresnan; Tendulkar would have expected Kevin Pietersen, who had bowled the previous over to be given one more over. His eyes lit up; his adrenaline shot up. The moment had arrived. He lunged forward for the drive, missed the line and promptly turned away in his inimitable

Fig. 2 5th order photocopy

Visual analysis performed on the recursive photocopied documents exhibits a relative degradation in the texture of the document. The degradation keeps relatively increasing on each recursive photocopy i.e., more the order of recursion, higher is the degradation in texture which is quite clear from Fig. 1 and Fig. 2. This directed us to explore a texture analysis method to study the relative degradation in the recursive photocopies of documents. Earlier, a method was proposed by us to study the texture degradation using Geometric Moments [7] and the approach showed an efficiency of 65%. As the previous work showed comparatively low performance, a different approach is proposed through this paper. In this new approach, texture degradation analysis is made using the texture feature entropy from Gray Level Co-occurrence Matrix (GLCM) of an image and the approach performed better than earlier method.

The remainder of the paper is organized as follows: Section II gives introduction to GLCM, theory of texture feature entropy and application of the study of texture degradation. Section III describes methodology adopted for estimation of recursive order number of photocopy using texture feature entropy. The experiments conducted along with analysis of results are discussed in section IV. Conclusion on the work is presented in section V.

II. GLCM AND TEXTURE FEATURE ENTROPY OF IMAGES

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, also known as the gray-level spatial dependence matrix. GLCM are extensively used in many applications like carpet wear assessment [8], texture feature extraction [9] and image texture segmentation [10]. A GLCM of an image is a square matrix in which the number of rows and columns are equal to the number of gray levels in the image [11]. GLCM ***G*** is constructed [12] for an image ***F*** with K possible intensities by considering an operator Q that defines the position of two pixels relative to each other. In other words, ***G*** is the matrix obtained for image ***F*** 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 $1 \leq$ i, j $\leq$ K. GLCM of an image is the basis for extracting texture features of the image. The texture features of an image are mainly characterized by a set of descriptors known as Haralick features [11]. Six texture features are extracted from GLCM viz correlation, contrast, energy, homogeneity and entropy. Each texture feature is obtained from the normalized probability density matrix ***P*** from ***G*** is given by

$$p_{ij} = g_{ij} / \mathrm{n} \tag{2.1}$$

where n being the total number of pixel pairs that satisfy Q in ***G*** which is 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

$$\sum_{i=1}^{K} \sum_{j=1}^{K} p_{ij} = 1 \tag{2.2}$$

It is quite evident from the statistical theory described for Haralick texture feature that significant variations are noticeable in all the six texture features which are defined through equations 2.3 to 2.7.

$$\text{Contrast} = \sum_{i=1}^{K} \sum_{j=1}^{K} (i-j)^2 p_{ij} \tag{2.3}$$

$$\text{Correlation} = \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{(i-m_r)(j-m_c)p_{ij}}{\sigma_r \sigma_c} \tag{2.4}$$

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

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

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

$$\text{Energy} = \sum_{i=1}^{K}\sum_{j=1}^{K}p_{ij}^2 \qquad (2.5)$$

$$\text{Homogeneity} = \sum_{i=1}^{K}\sum_{j=1}^{K}\frac{p_{ij}}{1+|i-j|} \qquad (2.6)$$

$$\text{Entropy(E)} = -\sum_{i=1}^{K}\sum_{j=1}^{K}p_{ij}\log_2 p_{ij} \qquad (2.7)$$

Though all the texture features of GLCM show significant variations with respect to degradation in texture, the feature Entropy (E) in particular shows better classification range compared to other features when they are applied on recursive photocopies. This made us to focus our research on texture feature entropy of GLCM to estimate the order number of recursive photocopies.

From the theory defined in (2.7), the entropy (E) of an image keeps increasing as the degradation of the texture increases. Next section describes the model designed to find the rate of texture degradation in recursive photocopies based on texture feature entropy.

III. Methodology

Since the original copy is not available for comparison and verification, an unsupervised intelligent system is required to find the rate of degradation in texture of a photocopy document. It is quite difficult to design a system that finds the rate of degradation based only on entropy of the text in the photocopy. Hence for a given photocopy, entropy of text (E_t) and entropy of text contour (E_c) is obtained first using (2.7) as,

$$E_t = -\sum_{i=1}^{K}\sum_{j=1}^{K}p_{ij}\log_2 p_{ij} \qquad (3.1)$$

$$E_c = -\sum_{i=1}^{K}\sum_{j=1}^{K}p_{ij}^1\log_2 p_{ij}^1 \qquad (3.2)$$

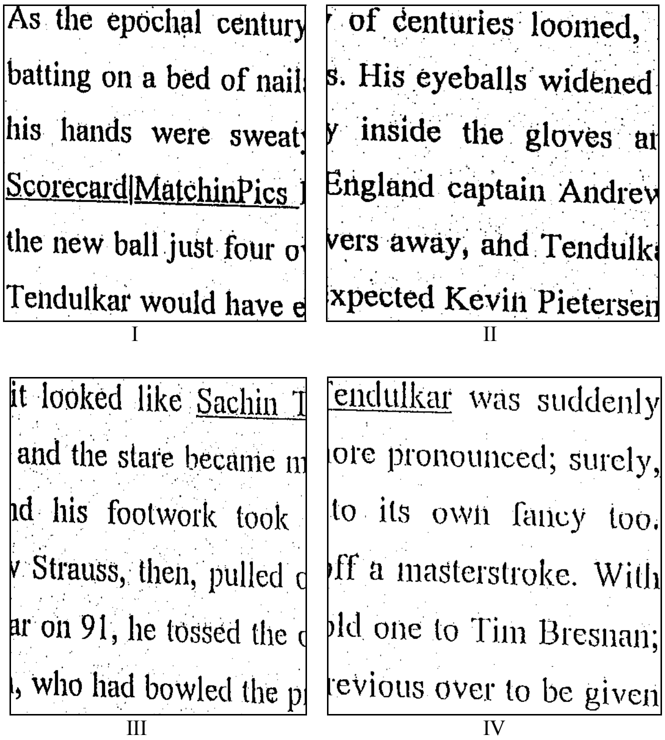
where ***P*** and ***P^1*** are normalized GLCM for text and text contour of a photocopied document. Next, texture degradation rate *Φ* is computed as

$$\Phi = E_c/E_t \qquad (3.3)$$

The distortions in degraded texture make the contour of text normally rough and inconsistent. Such rough edges with inconsistencies in contour of text have higher entropy value compared to entropy value of the text in a document. This makes the value of *Φ* to increase as degradation in texture increases.

Further, it is noticed from Fig. 2, the degradation is more towards the right side than left side of the 5th order photocopy. Also the thicknesses in the characters keep reducing from left to right and the orientation of lines change in the photocopy. To correlate the above issues in calculation of rate of degradation, an absolute rate of degradation ***δ*** is computed by considering left part and right part of the photocopy through dividing the document shown in Fig. 2 into 4 equal vertical parts as shown in Fig. 3. The first part to left with minimum degradation and 4th part towards the right with maximum degradation are considered to compute Φ_1 and Φ_4 using (3.3). The absolute degradation rate for a photocopy is finally given by

$$\boldsymbol{\delta} = \Phi_4 - \Phi_1 \qquad (3.4)$$



As the epochal century
batting on a bed of nail
his hands were sweat
Scorecard|MatchinPics
the new ball just four o
Tendulkar would have e

of centuries loomed,
s. His eyeballs widened
y inside the gloves a
England captain Andrev
vers away, and Tendulk
xpected Kevin Pietersen

it looked like Sachin T
and the stare became m
ld his footwork took
v Strauss, then, pulled c
ar on 91, he tossed the c
, who had bowled the p

endulkar was suddenly
ore pronounced; surely,
to its own fancy too.
ff a masterstroke. With
old one to Tim Bresnan;
revious over to be given

Fig. 3 Division of photocopy into 4 equal vertical parts

The tabulation of ranges of absolute degradation rates for recursive order photocopy upto 5th order is shown in Table I. As there is no scope for fabrication of photocopy documents beyond 4th order due to high texture degradation, the ranges for higher order photocopies are not presented in this work. The value of ***δ*** less than 3.70E-02 is considered as copy with no texture degradation and the document is rejected for estimation of order number. The range of absolute rate of degradation (*δ*) for the order of photocopies is shown in the form of graph in Fig. 4. The experimental study with test samples is discussed in next section.

TABLE I Range values of ***δ*** for classification

Photocopy order	*Range of absolute degradation rate (δ)*
1st	3.70E-02 – 5.09E-02
2nd	5.10E-02 - 6.99E-02
3rd	7.00E-02 - 8.39E-02
4th	8.40E-02 – 9.99E-02
5th	1.00E-01 - 1.25E-01
6th and above	>1.25E-01

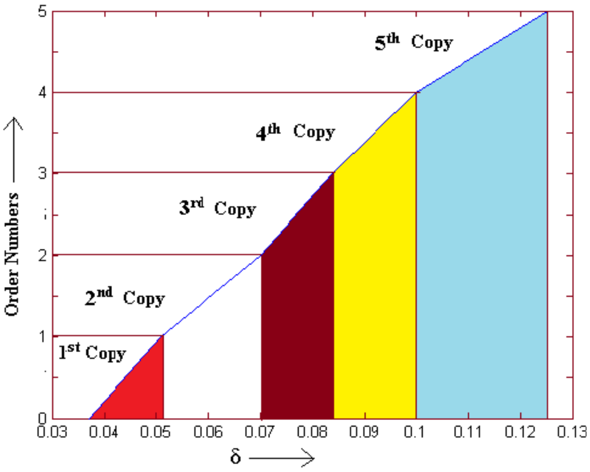


Fig. 4 Graph of order number v/s range of δ

IV. Experimental Results

Experimentation is performed through testing the proposed method using synthetically generated samples of photocopy documents from different photocopying machines. Testing is carried out with sufficient number of test samples up to 5th order. The test samples include different sizes, different contents, figures, tables etc. The results of the testing are tabulated in Table II. The experiments conducted using test samples show an average classification efficiency of 94.92% with 4.18% misclassification and 0.9% of rejection. The reason for rejection is mainly due to a very low value of ***δ*** indicating minor/little degradation in photocopy and unable to calculate distinct texture degradation rate. It is noticed from the Table II, rejection is found only in 1st order photocopy document and not in other orders. The misclassifications are mainly due to the presence of noise and dirt in the document, toner quality and machine's quality used in production of photocopies. The efficiency of the system with respect to the recursive order numbers of the photocopies is shown as pie chart in Fig. 5.

The experiments are not carried out beyond 5th order, as there is no scope for fabrication in such high order photocopies since the rate of texture degradation is too high for fabrication.

TABLE II Results of testing

Photocopy order	**No. of samples**	**classification**			**Efficiency**
		Correct	***Incorrect***	***Rejection***	
1st	125	114	05	06	91.2%
2nd	140	130	10	00	92.85%
3rd	115	111	04	00	96.52%
4th	150	145	05	00	96.66%
5th	140	136	04	00	97.14%
Total	**670**	**636**	**28**	**06**	**94.92%**

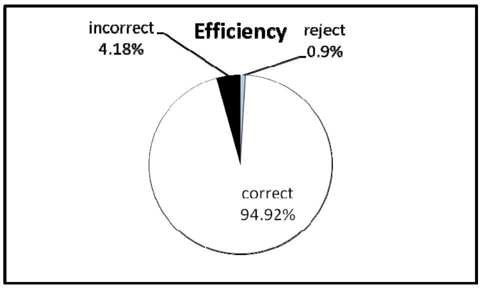


Fig. 5 Pie chart showing efficiency of the method

V. Conclusion

In India, the self attested photocopy documents can be submitted as proof in place of original document. Certain classes of people misuse this provision to gain advantages illegally through fabricating fake photocopy documents. This demands an expert system to detect fabricated photocopy documents. Since fabrication of photocopy documents are possible only in photocopies of recursive order 2 or above, there is a need to find whether the photocopies of order one or high. Such requirements of obtaining the recursive order number of photocopy lead our research to estimate the same. This work proposes a method to provide an unsupervised intelligent system for estimating the recursive order number of photocopy submitted based on entropy from GLCM texture features. The methodology is designed to compute the rate of degradation in recursive photocopy document using texture feature entropy. The method shows average classification efficiency close to 95%. The misclassification is due to photocopies obtained from different machines and their quality. The proposed work is design of an efficient method to estimate the recursive order of photocopy and this becomes an initial stage to detect fabrication in photocopy. Avenues are open to explore methods to find rate of degradation using variations in character thickness and line orientations which is under investigation.

Acknowledgment

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.

REFERENCE

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

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

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

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

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

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

[7] Suman Patgar, Vasudev T, Estimation of order number from successively photocopied document using Geometric moments, 2012, India, SACAIM 2012.

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

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

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

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

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