A VOTING BASED APPROACH TO DETECT RECURSIVE ORDER NUMBER OF PHOTOCOPY DOCUMENTS USING PROBABILITY DISTRIBUTIONS

Suman V Patgar[1], Rani K[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 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 order recursive order of photocopies. Whenever a photocopied document is submitted, it may be required to check its originality. When the document is 1st order photocopy, chances of fabrication may be ignored. On the other hand when the photocopy order is 2nd or above, probability of fabrication may be suspected. Hence when a photocopy document is presented, the recursive order number of photocopy is to be estimated to ascertain the originality. This requirement demands to investigate methods to estimate order number of photocopy. In this work, a voting based approach is used to detect the recursive order number of the photocopy document using probability distributions exponential, extreme values and lognormal distributions is proposed. A detailed experimentation is performed on a generated data set and the method exhibits efficiency close to 89%.

KEYWORDS

fabricated photocopy documents, recursive order number, probability distributions, voting.

1. INTRODUCTION

Many authorities trust and accept the 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 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 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 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 fabrication is zero in the 1st photocopy obtained from the original document, where as the fabrication may be suspected in the higher order photocopy. In this direction, an attempt is made to estimate the recursive order number of the photocopy submitted. Further, based on the estimation of order number, investigations can be explored to detect the possibility of fabrication in photocopy document.

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. One of the interesting features is a measure of the variability of the handwriting on a small scale. Although one can copy the shape of another's handwriting, it is difficult to mimic the dynamic aspects, such as speed and acceleration. Because forged handwriting tends to be drawn slowly, when scanned, it might be more wiggly than the authentic handwriting. Forgery handwriting shows more wrinkliness than natural handwriting does. This wrinkliness feature can be measured using the *fractal* dimension measure.

While preparing any security document, the designers embed certain features that are considered as security features. It is generally assumed that these features are difficult to replicate or copy. Duplicity of a document is identified by checking these security features. Security features in documents like cheques, legal deeds, certificates, etc. are embedded by three attributes namely (i) color features, (ii) background artwork and logo, and (iii) paper quality. After extracting features from a cheque document, its authentication is to be done. This is modeled as a 2-class pattern recognition problem, i.e. whether the document belongs to the genuine document class or not. Support Vector Machines (SVMs) [5] are used to verify authenticity of these cheques.

Further, in literature to the best of our knowledge no significant effort is noticed towards detecting forgery made while taking photocopy. As the domain under consideration is new for research, no standard data set is available. Hence for the purpose of experimentation sufficient numbers of sample photocopies are obtained from the different brand copier machine to generate different recursive order copies. The photocopies were scanned using a scanner to produce bitmap images at 300dpi. Fig 1a and 1b show the samples of 1st order and 5th order recursive photocopies of a document respectively.

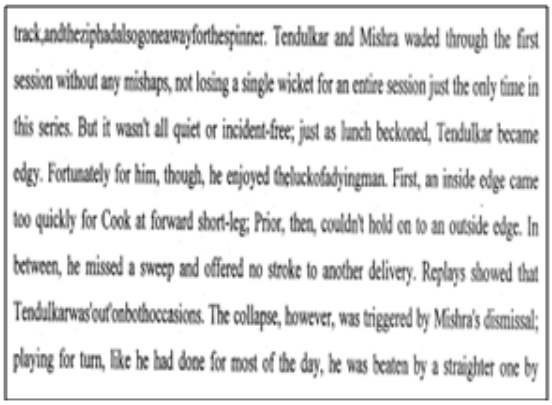
track,andtheziphadalsogoneawayforthespinner. Tendulkar and Mishra waded through the first session without any mishaps, not losing a single wicket for an entire session just the only time in this series. But it wasn't all quiet or incident-free; just as lunch beckoned, Tendulkar became edgy. Fortunately for him, though, he enjoyed theluckofadyingman. First, an inside edge came too quickly for Cook at forward short-leg; Prior, then, couldn't hold on to an outside edge. In between, he missed a sweep and offered no stroke to another delivery. Replays showed that Tendulkarwas'out'onbothoccasions. The collapse, however, was triggered by Mishra's dismissal; playing for turn, like he had done for most of the day, he was beaten by a straighter one by

Figure 1a: 1st order photocopy

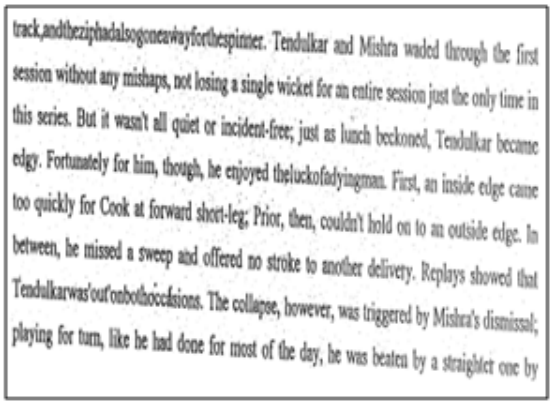
track,andtheziphadalsogoneawayforthespinner. Tendulkar and Mishra waded through the first session without any mishaps, not losing a single wicket for an entire session just the only time in this series. But it wasn't all quiet or incident-free; just as lunch beckoned, Tendulkar became edgy. Fortunately for him, though, he enjoyed theluckofadyingman. First, an inside edge came too quickly for Cook at forward short-leg; Prior, then, couldn't hold on to an outside edge. In between, he missed a sweep and offered no stroke to another delivery. Replays showed that Tendulkarwas'out'onbothoccasions. The collapse, however, was triggered by Mishra's dismissal; playing for turn, like he had done for most of the day, he was beaten by a straighter one by

Figure 1b: 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 1a and 1b. This directed us to explore a texture analysis method to study the relative degradation in the recursive photocopies of documents. Earlier, two methods were proposed to find texture degradation. The first method uses Geometric Moments [7] to find texture degradation and the second method was proposed using entropy from Gray Level Co-occurrence Matrix [8]. In order to achieve higher efficiency exploration of another method was attempted on this problem to analyze texture degradation using probability distributions.

The remainder of the paper is organized as follows: Section 2 gives introduction to Probability distributions used in the methods. Section 3 describes methodology adopted for estimation of order number of photocopy using texture feature. The experiments conducted along with analysis of results are discussed in section 4. Conclusion on the work is presented in section 5.

2. PROBABILITY DISTRIBUTIONS

In the proposed system, the recursive order of the photocopy documents is estimated based on measure of texture degradation using different distribution methodologies like Exponential distribution, Extreme value distribution and Lognormal distribution [9]. These distribution methods are applied to maximum peak values which are extracted from the distribution graphs and brief introduction to the same is given subsequently.

2.1 Exponential Distribution

The exponential distribution(ED) is the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. It is the continuous analogue of the geometric distribution, and it has the key property of being memory less. In addition to being used for the analysis of Poisson processes, it is found in various other contexts [10]. The probability density function (pdf) of an exponential distribution is,

$$f(x:\lambda) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x < 0 \end{cases} \quad (1)$$

Alternatively, this can be defined using the Heaviside step function [13], H(x).

$$f(x;\lambda) = \lambda e^{-\lambda x} H(x) \quad (2)$$

Here $\lambda > 0$ is the parameter of the distribution, often called the rate parameter. The distribution is supported on the interval $[0, \infty]$. If a random variable X has this distribution, we write X ~ Exp (λ). The exponential distribution exhibits infinite divisibility [10].

2.2 Extreme value Distribution

The extreme value distribution(EVD) has two forms. One is based on the smallest extreme and the other is based on the largest extreme. These are the minimum and maximum cases, respectively. The extreme value distribution is also referred o as the Gumble distribution [11].

The general formula for the probability density function of the Gumbel (minimum) distribution is,

$$f(x) = \frac{1}{\beta} e^{\frac{x-\mu}{\beta}} e^{-e^{\frac{x-\mu}{\beta}}} \quad (3)$$

Where μ is the location parameter and β is the scale parameter. The case where $\mu = 0$ and $\beta = 1$ is called the standard Gumbel distribution [11]. The equation for the standard Gumbel distribution (minimum) reduces to

$$f(x) = e^{-x} e^{-e^{x}} \quad (4)$$

2.3 Lognormal Distribution

Consider a N (μ, σ^2) random variable Z, then the random variable X = exp (Z) is said to have a lognormal distribution. In other words X is lognormal if its logarithm log X has a normal distribution [12]. It is easy to see that the pdf of a lognormal distribution(LD) associated to an N (μ, σ^2) distribution is,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi\sigma^2}} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}} \quad (5)$$

The median of the above distribution is exp (μ), while its mean is exp ($\mu+\sigma^2/2$). The mean is larger than the median which indicates that the lognormal distribution is right skewed. In fact the larger the variance σ^2 of the associated normal distribution, the more skewed the lognormal distribution is. Lognormal distributions are particularly important in mathematical finance, as it appears in the modeling of returns, where geometric Brownian motion appears [13].

Mean and standard deviation are the essentials to apply probability distribution to the images. The mean is the average of the numbers and mean is used to calculate the central value of a set of numbers [14]. The mean is denoted by the symbol '$\overline{X}$'. The formula for mean is,

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (6)$$

Standard deviation is the measure of how the numbers are spread out. The standard deviation is denoted by the symbol 'σ'. The formula for standard deviation is,

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \quad (7)$$

Probability distribution function are applied to several set of samples and get a range of values through which we can predict the recursive order of the given photocopy documents.

3. METHODOLOGY

Recursive order number is the order number of photocopy which is obtained recursively. Texture degradation is one of the noticeable features of the recursive photocopy document. Degradation of text increases as the recursive order number of the photocopy document increases. From the Fig 1b, it is evident that degradation in the recursive photocopy is maximum on right side of the document. The work focus on measuring texture degradation only on the right edge part of the recursive photocopy. In order to estimate the recursive order of the photocopy document, different distribution functions are applied to the given photocopy document and based on the values obtained by these functions the order of the photocopy document is estimated. In the proposed work three different distribution functions Exponential distribution, Extreme value distribution and Lognormal distributions are computed. These distribution functions are applied to the training samples, range of values for every distribution function are generated and tabulated as shown in Table 1.

Table 1. Range of distribution functions

Distribution Functions	**Order Number**					
	1st	**Overlapping between 1st& 2nd**	**2nd**	**3rd**	**4th**	**5th**
Exponential	0.0211 - 0.0250	0.0251 - 0.0274	0.0275 - 0.0308	0.0208 - 0.0210	0.0189 - 0.0207	0.0165 – 0.0188
Extreme value	0.3401 - 0.3622	0.3623 - 0.3994	0.3995 - 0.4235	0.2735 - 0.3400	0.2009 - 0.2734	0.0845 - 0.2008
Lognormal	0.2013 - 0.2207	0.2208 - 0.2544	0.2545 - 0.2890	0.1083 - 0.2012	0.0162 - 0.1082	0.0001 - 0.0161

Initially for the given photocopy document the three distribution functions are applied and their values are computed. In order to estimate the recursive order number of the given photocopy document with these distribution values a voting procedure is used.Voting procedure focus on the aggregation of individuals, preferences to produce collective decisions.

From the table 1 it is noticeable that there are considerable overlapping values for order numbers one and two. This is because the texture degradation is quite narrow in these two recursive orders where as other order number classification has distinct range values. In voting procedure, if any

two distribution values among three values falls in overlapping range and other one gives the distinct value, then it is quite difficult to decide the order number of the given photocopy document. Hence, a decision system is developed to make the classification in case of overlap to resolve the conflicts. Figure 2 shows the flow while making the decision.

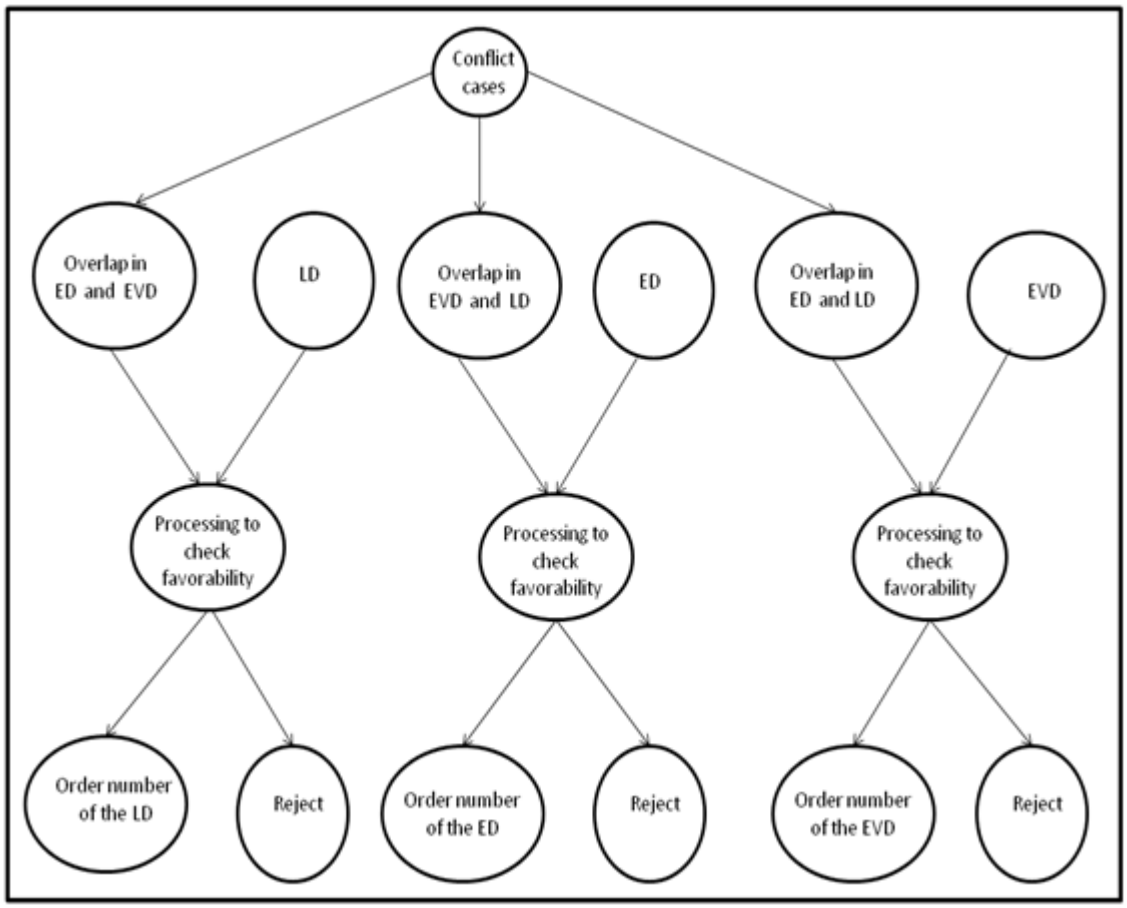

Figure 2. Flow Diagram to resolve the overlapping conflict

The diagram in figure 2 indicates the flow decision making under conflict situation. The conflicts are resolved through examining the favorability of the values in the distribution values. If any conflicting values is favorable to the distinct decision, then weightage is given towards the distinct decision. In case the conflict cases donot favor the distinct decision then value is considered as rejection and suggested for manual verification. In case if all the three distributions indicate conflict then it is considered as rejection and suggested for manual verification.

4. 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 5^{th} order since higher order copies are not suitable for fabrication. The test samples include different sizes, different contents, figures, tables etc. The results of the testing are tabulated in Table 2. The experiments conducted using test samples

show an average classification efficiency of 88.6% with the misclassification of 7.85% and rejection is 4.64%. 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.

Table 2. Results of testing

Order Number	No.of samples	Classification			Efficiency
		Correct	Incorrect	Rejection	
1st	50	44	03	03	88.8%
2nd	60	50	07	03	83.3%
3rd	65	55	06	04	84.6%
4th	50	46	04	02	92%
5th	55	52	02	01	94.5%
Total	**280**	**247**	**22(7.85%)**	**13(4.64%)**	**88.6%**

5. CONCLUSION

The implemented method provides a supervised system for estimating the recursive order of photocopy submitted. The method is essentially based on the probability distribution methods. The method shows average classification efficiency close to 89%. 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 is a base work to continue research for better efficiency through investigating methods to find rate of degradation using variations in character thickness and line orientations. The rejection case indicates the proposed model cannot resolve the conflict and suggested for physical verification. The proposed work is prerequisite to continue the research to detect the photocopy under consideration is a fabricated or not.

REFERENCES

[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 Vol. 38, pp 341-356, 2005

[3] 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

[4] 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

[5] 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

[6] 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
[7] Suman Patgar, Vasudev T, 2012, Estimation of order number from successively photocopied document using Geometric moments, SACAIM 2012.
[8] Suman. V. Patgar, Vasudev. T, An unsupervised intelligent system to detect fabrication in photocopy document using Geometric Moments and Gray Level Co-Occurrence Matrix, 2013.
[9] Alvina Goh and Ren´e Vidal ,2007, Riemannian Analysis of Probability Density Functions with Applications in Vision. IEEE Conference on Computer Vision and Pattern Recognition
[10] http://mathworld.wolfram.com/ExponentialDistribution.html.
[11] www.mathworks.in/help/stats/extreme-value-distribution.html.
[12] J. Martín & C.J. Pérez , Application of a generalized lognormal distribution to engineering data fitting
[13] L. E. Wagner, D. Ding, representing aggregate size distributions modified Lognormal distributions
[14] Erwin Kreyszing, "Advanced Engineering Mathematics" 8th edition.

AUTHORS

Vasudev T, is Professor in the Department of Computer Applications, Maharaja Institute of Technology, Mysore. He obtained his Bachelor of Science and Post Graduate diploma in Computer Programming with two Masters Degrees one in Computer Applications and other one is Computer Science and Technology. He was awarded Ph.D. in Computer Science from University of Mysore. He is having 30 years of experience in academics and his area of research is Digital Image Processing specifically document image processing.

Suman V Patgar, is Research Scholar, P.E.T Research Center Mandya. She obtained her Bachelor of Engineering from Kuvempu University in 1998. She obtained her Masters degree in Computer Science and Engineering from VTU Belgaum in 2004. She is pursuing doctoral degree with the supervision of Vasudev T under University of Mysore.

Rani K, obtained her Bachelor degree in Computer Science and Engineering from P.E.S college of Engineering, Mandya in 2012. She is pursuing Master degree in Computer Science and Engineering under VTU Belgaum.