ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 5, May 2014

Comparison of GLCM and IQM for Printer Identification Using Printed Documents

Ruby Yadav[1], Kusha Goyal[1], Rachana Panwar[1], Nitin Khanna[2]

PG Student, Department of Electronics and Communication, Graphic Era University, Dehradun, India[1]
Professor, Department of Electronics and Communication, Graphic Era University, Dehradun, India[2]

Abstract: Identification of source of a printed document can be a very important step for the forensic science. Currently there are some techniques like special inks, watermarking, holograms etc. that can be used to secure documents. This paper presents a comparison of gray level occurrence matrix (GLCM) and image quality matrix (IQM) based features for source printer identification using printed pages. The proposed research utilizes texture-based features of the printed document. These features will help us gain knowledge of the printer used for printing that document. We have taken printout of documents from 10 different printers of different brands and model number and obtained average classification accuracies of around 75% and 80%, using GLCM and IQM features respectively.

Keywords: Sensor forensics, Printer identification, Feature Extraction, GLCM, IQM.

I. INTRODUCTION

Identification/verification of a printed document can be a very important step for the forensic science. Currently there are some techniques like special inks, watermarking, holograms etc. that can be used to secure documents. A printed data contains information that conveys a message. To check the authenticity of a document, document examiners are often called to check if there are any forgeries and alterations done to the document. A questioned document can be wills, letters, contracts, bills etc. And in printed document case there is a defect in the copying and printing method that results in the occurrence of unusual characters. Examinations are carried out by forensic document examiners in the following number of areas :

1. Identification of handwriting.
2. To check the authenticity of a signature as a genuine or forged.
3. To check the origin of a document.
4. Markings and dating on a document.
5. If there are any additions, deletions, and alterations made to the printed document.

In figure 1 we can see that a questioned document can be a printed document, a photocopy of a printed document and a hand written document. A printed document can be taken from a laser printer, ink jet printer or a dot matrix printer. To print a character on a sheet of paper the laser printer don't use ink, the characters are made with the help of toner made from finely chopped carbon and binders and a laser. And there are very less chances that characters will become deformed over a period of time. In case if a cartridge containing the toner runs out, then only the print quality shows irregularities. If an unknown document i.e. is the questioned document and exemplars depict these irregularities at the same location on the copy, it can provide information about the source of the questioned document. Most important point is that a number of exemplars must be taken to verify the degree of consistency of the extraneous markings or irregularities.

In ink jet printer the characters are made up by spraying ink onto the paper. The problem that occurs in the case on ink jet printers is that the typescript may become uneven and can also fade as the cartridge runs out of ink. And in case of colour cartridges this problem is much more prominent, even the loss of one colour can distort the other colours. In this case until there is no more ink provided or until the cartridge is not been changed this problem continuous to occur.

Two types of marks generally appear on photocopies that are trash marks and drum or mechanism marks. Trash marks which are transient are produced by dust particles on the glass of the machine and they appear like dots on the copy. It can be noted that if a photocopy bears a trash mark into it then all its daughter copies will also have the same trash marks. And the marks made by drums persist for a longer period; they don't appear in the same position on every page but occurs at consistent intervals so they can be investigated by determining these trashes.

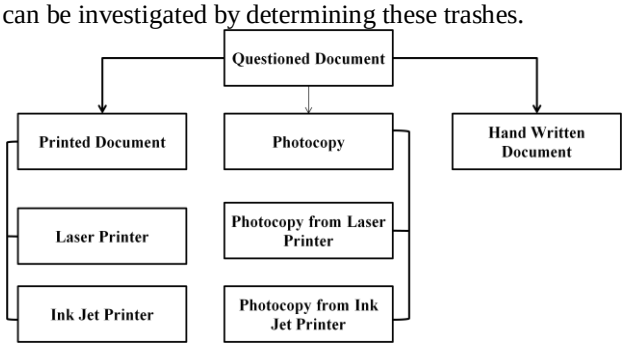

Figure1: Categories of questioned document

So, by identifying a particular printer that printed a particular document will be a major step in the field of printer forensics. The idea of secure printing involves that the printed data is a mean to identify various feature of that printer. If a forgery is present in the document then we

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 5, May 2014

can tell what type of printer was used to create the document. Because of the advancing technologies in world, various image processing tools are available to forge the documentation easily and efficiently. Due to these reasons authentication of printed data is a big challenge. There are two strategies for printer identification passive and active. In the passive strategy we characterize a printer and find its intrinsic features in the printed output that are the characteristic of that particular printer, model, or manufacturer products. It is also called as intrinsic signature; it needs a modelling and understanding of the printing method. Image analysis tool for intrinsic signature detection from data with unknown content can be developed by design based on standard test pattern. In the active method an extrinsic signature is embedded in every page of printed document. The embedded signature in the document can be thought as a watermark.

II. LITERATURE REVIEW

There are some methods used previously for printer identification which are described as follows. Aravind K. Mikkilineniet.al proposed a method using GLCM of the alphabet "e"s in the document. Texture based features were extracted and then classified using 5-Nearest-Neighbour (5NN) classifier. For printer identification 10 printers were used and 500 features were extracted by using basic 22 features.

Method described in uses distance transform for printer identification, 45 laser printers were used for identification and 200~300 features were extracted then minimum distance classifier was applied.

In paper geometric distortion method is applied to extract intrinsic features and a method based on distortion is presented. Least square method is used for estimation and Support Vector Machine is used for classification.1496 features were extracted using 10 different printers of 5 different models.

In colour printer identification was done by statistical examination of the HH sub-band; on discrete wavelet transform and 39 noise features were extracted. For identification, 9 models of 4 brands were used and extracted features were applied for training and classification of the support vector machine.

Approach described in proposes a technique for electro-photographic printer identification method based examining the intrinsic half toning effect. Hough transform was used for extracting feature vectors from textured area. Experiments were performed on 9000 images were taken out from 9 different printers.

Choi et.al proposes a method for colour laser printer identification in which invisible noises were estimated using wiener-filter and then gray level co-occurrence matrix is calculated to extract 60 statistical features. In this experiment 2,597 images from 7 laser printers were used.

Avcibas et.al described image quality measure for classification. Total 10 features are used to classify the source camera.

III. PROPOSED METHOD

For identifying the correct printer which is used for printing a particular document, we have used some technique for getting the features of particular printer. By using these features we can easily know the correct printer which is used for printing that document. Image texture analysis is used to identify the printer and a set of features provides the forensic information about a document. We are proposing a comparison between the GLCM method and IQM technique.

A. Test Document

We have collected our data from 10 different printers of different brands with different model and serial number. After data collection the images are scanned at 200 dpi with 8 bits/pixel (gray scale). And then this features like contrast, correlation, homogeneity, and energy has been extracted.

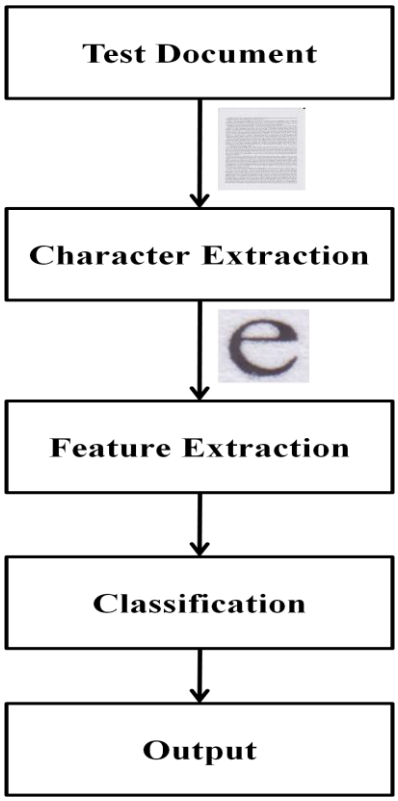

Figure 2: Block Diagram of the Proposed System

B. Character Extraction

Extract all the 'e' in the document because this is the most used alphabet in the English language. It can be noticed from figure 3 that different printers print this particular alphabet differently.

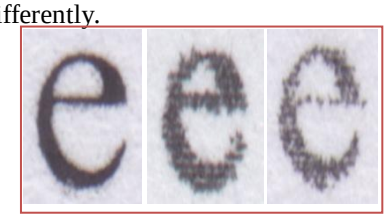

Figure 3: Character 'e' printed from different printers

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 5, May 2014

C. Feature Extraction

I. Gray level co-occurrence matrix

The four basic geometric features are given below:

a. **Contrast**: Intensity contrast between a pixel and its neighbour over the whole image is returned by this particular feature. The value of contrast is zero for a constant image.

$$\sum_{i,j}|i-j|^2 p(i,j) \quad (1)$$

b. **Correlation**: It is a measure that tells how correlated a pixel is to its neighbour over the whole image. The value of correlation is 1 for positive image and -1 for a negative image.

$$\sum_{i,j}\frac{(i-\mu i)(j-\mu j)p(i,j)}{\sigma_i \sigma_j} \quad (2)$$

c. **Energy**: This property returns the sum of squared elements in the gray-level co-occurrence matrix. And the value of energy is 1 for a constant image.

$$\sum_{i,j} p\,(i,j)^2 \quad (3)$$

d. **Homogeneity:** It is a measure that tells the closeness of the distribution of elements in the diagonal of GLCM. For a diagonal GLCM the value of homogeneity is 1.

$$\sum_{i,j}\frac{p(i,j)}{1+|i-j|} \quad (4)$$

II. Image Quality Measure

Image Quality Measure provides the measure of deteriorated image, when compared with a perfect image. The features extracted falls under three classes namely, the pixel difference, spectral distance and the correlation based measure. Here, *i, j,* and in band *x* as $A_x(i,j)$, where X=1…..3 for three different colours namely red, green and blue. $A(i,j)$ and $\hat{A}(i,j)$ are multispectral pixel vectors at position (i,j). $\Gamma_x(u,v)$ and $\acute{\Gamma}_x(u,v)$ is the Discrete Fourier Transform (DFT) of x[th] band of the original and embedded image respectively.

a. The pixel difference measures dissimilarity between two images

$$I_{P.D} = \frac{1}{X}\sum_{x=1}^{X}\left\{\frac{1}{N^2}\sum_{i,j=1}^{N}\left|A_x(i,j)-\hat{A}_x(i.j)\right|^2\right\}^{1/2} \quad (5)$$

b. Correlation based measures the statistics of the angles between two images pixel vectors

$$I_{C.B} = \frac{1}{X}\sum_{x=1}^{X}\frac{\sum_{i,j=0}^{N-1}A_x(i,j)\hat{A}_x(i,j)}{\sum_{i,j=0}^{N-1}A_x(i,j)^2} \quad (6)$$

c. Spectral Measure is used to measure the spectral magnitude distance between two pixels

$$I_S = \frac{1}{XN^2}\sum_{x=1}^{3}\sum_{u,v=0}^{N-1}\left|\left|\Gamma_x(u,v)\right|-\left|\acute{\Gamma}_x(u,v)\right|\right|^2 \quad (7)$$

D. Classification

The texture features obtained are used as inputs in WEKA for classification. Different classifiers present in WEKA are used to classify the printer. The classifier which gives best accuracy is used for classification.

Weka(Waikato Environment for Knowledge Analysis) is used for data mining tasks that is a collection of various machine learning algorithms. These algorithms can be applied directly on the dataset or called from your own java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

After applying various classifiers best results were obtained using Rotation Forest, IB1, and Attribute Selected classifiers.

I. Rotation Forest

This classifier is based on ensemble learning method i.e. if many learning schemes are available instead of choosing the best scheme it is more advantageous to use them all and then combine the result. Rotation forest combines bagging and random subspace approaches and for the construction of an ensemble of decision trees, principle component feature generation is performed on the dataset.

II. IB1

IB1 finds the training instance nearest to the Euclidean distance for the given test instance. Therefore, is called an instance based classifier. First instance found among several closest instances is chosen.

IV. EXPERIMENTAL RESULTS

The comparison between the results of classification using GLCM and IQM technique has been done using our database. 10 different printers of different brands with different model and serial number are used.

TABLE 1
PRINTER DATASET

S.NO.	Brand	Model No.	Serial No.
P_1	HP	1020	PLH2612
P_2	EPSON	L100	NFUKO23425
P_3	EPSON	L210	S2AK017568
P_4	XEROX	50212	898E86230RevA
P_5	HP	1020	Q2612A
P_6	HP	P2035	CNCO605771
P_7	HP	P2035	CNCO420745
P_8	HP	P2035	CNC0420776
P_9	HP	400	VNC86C04132
P_{10}	HP	P2035	CNC0602861

A. *Results using GLCM*

After collecting the dataset, features were extracted using GLCM method and different classifiers were applied to classify our data. Among various classifiers best results were from Rotation Forest and IB1 classifier. From Rotation forest the accuracy was 75.97% and from IB1 it was 74.16%.In the table 2 classification of GLCM and IQM is shown.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 5, May 2014

TABLE 2: CLASSIFICATION WITH GLCM AND IQM

S.No.	Name of Classifier	Accuracy (%)	
		GLCM	IQM
1	Rotation Forest	75.97	81.66
2	IB1	74.16	75
3	J48	65	70
4	Attribute Selected	57.22	80

In Figure 4 a demonstration has been done to show classification result of different classifiers used for GLCM. From here we can see that the best average results were from Rotation Forest classifier.

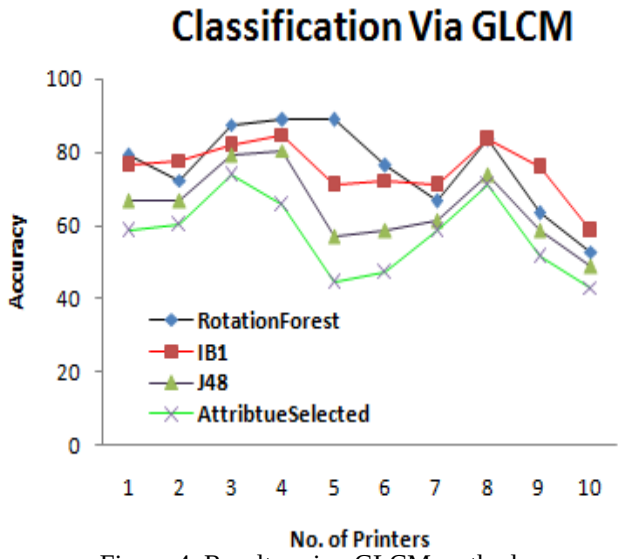

Figure 4: Results using GLCM method

Table 3 depicts a confusion matrix of GLCM that is used to represent the correctly and incorrectly instances. Diagonal element represented in red shows the correct classification and rest shows the incorrect classification.

TABLE 3
CONFUSION MATRIX OF GLCM

P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	
79.1	4.1	4.1	9.7	0	0	0	2.7	0	0	P_1
12.5	72.2	15.2	0	0	0	0	0	0	0	P_2
5.5	4.1	87.5	2.77	0	0	0	0	0	0	P_3
5.5	0	4.1	88.8	0	0	0	1.3	0	0	P_4
0	0	0	0	88.8	4.1	1.3	0	2.7	2.7	P_5
0	0	0	0	2.7	76.3	9.7	0	2.7	8.3	P_6
0	0	0	2.7	1.3	12.5	66.6	0	6.9	9.7	P_7
6.9	1.3	0	8.3	0	0	0	83.3	0	0	P_8
0	0	0	0	6.9	6.9	2.7	0	63.8	19.4	P_9
0	0	0	0	11.1	9.7	2.7	0	23.6	52.7	P_{10}

B. *Results using IQM*

When we applied IQM technique and classified our data using different classifiers the best results were from Rotation Forest and Attribute Selector classifier. And the accuracy from Rotation Forest was 81.6% and from Attribute Selector was 80%.

Table 4 depicts a confusion matrix of IQM that is used to represent the correctly and incorrectly instances. Diagonal element represented in red shows the correct classification and rest shows the incorrect classification.

TABLE 4: CONFUSION MATRIX OF IQM

P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	
83.3	0	0	16.6	0	0	0	0	0	0	P_1
0	100	0	0	0	0	0	0	0	0	P_2
0	0	66.6	33.3	0	0	0	0	0	0	P_3
0	0	0	100	0	0	0	0	0	0	P_4
0	0	0	0	100	0	0	0	0	0	P_5
0	0	0	0	0	66.6	33.3	0	0	0	P_6
0	0	0	0	0	16.6	83.3	0	0	0	P_7
0	0	0	0	0	0	0	100	0	0	P_8
0	0	0	0	16.6	0	0	0	33.3	50	P_9
0	0	0	0	0	0	0	0	16.6	83.3	P_{10}

In figure 5 a demonstration has been done to show classification result of different classifiers used for IQM method. From here we can see that the best average results were from Rotation Forest classifier.

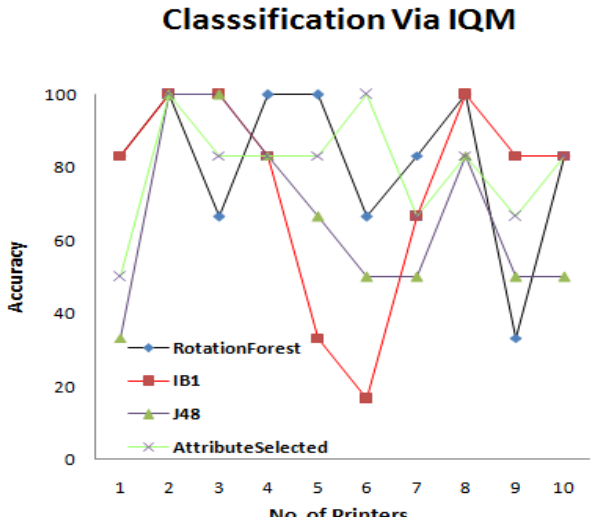

Figure 5: Results using IQM method

C. *Comparison between accuracy of GLCM and IQM*

Figure 6 shows the comparison between the common classifier used for both the methods. From the graph we can see that the accuracies from IQM method is much better than the GLCM method.

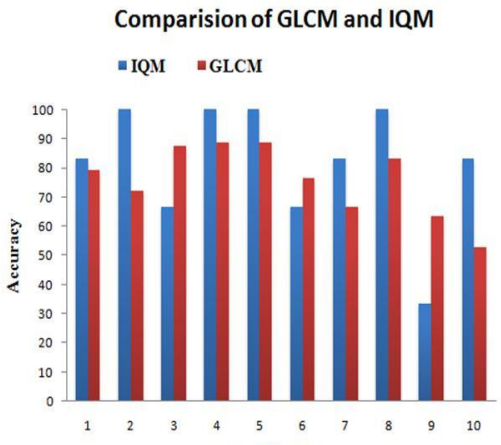

Figure 6: Comparison of Accuracy of GLCM and IQM

V. CONCLUSION

In this paper, techniques for printer identification are described and a comparison between GLCM and IQM method is discussed. In this research paper printer identification using image texture analysis have been investigated. From this system we can automatically classify a particular printed document from the 10 kinds of printer used in this research.

Texture features are extracted with the help of GLCM for every printer they are classified using Weka in which the dataset is classified using different classifiers, and a similar procedure is applied for IQM also. The Rotation Forest classifier gives most promising results amongst all and is used for classification. Printer Identification using GLCM shows that the method is workable with an accuracy of 75.9% on 10 kinds of printer. And when the IQM technique was implied on our dataset the accuracy has been increased to 81.6% using the Rotation Forest classifier. Our future approach includes improving the accuracy by extracting more features and increasing the dataset.

REFERENCES

[1] Ian H Written, Eibe Frank, and Mark A Hall, Data Mining:Practical Machine Learning Tools and Technique, 3rd ed., Eibe Frank, Mark A Hall Ian H Written, Ed. Burlington,Massachusetts, USA: Morgan Kaufman.

[2] Peter White, Crime Scene to Court-The Essentials of Forensic Science, 2nd ed., Peter White, Ed. Norfolk, UK: The Royal Society of Chemistry, 2005.

[3] Jan Seaman Kelly and Brian S Linblom, Scientific Examination of Questioned Documents, 2nd ed., Jan S Kelly Brian S Linblom, Ed. Boca Raton, Florida: CRC Press , 2006.

[4] Max M Houck and Jay A Siegel, Fundamentals of Forensic Science , 2nd ed.: Elsevier , 2010.

[5] Rafael C Gonzalez and Richard E Woods, Digital Image Processing, 3rd ed.: Pearson Prentice Hall, 2008.

[6] Jung Ho Choi, Heung Kyu Lee, Hae Yeoun Lee, and Young Ho Suh, "Color Laser Printer Forensics with Noise Texture Analysis," in Proceedings of the 12th ACM workshop on Multimedia and security, 2010, pp. 19-24.

[7] Wei Deng, Qinghu Chen, Feng Yuan, and Yuchen Yan, "Printer Identification Based on Distance Transform," in Intelligent Networks and Intelligent Systems, Wuhan, 2008, pp. 565-568.

[8] Aravind K Mikkilineni et al., "Printer Identification Based on Texture Features," in Society for Imaging Science and Technology, 2004, pp. 306-311.

[9] Seung Jin Ryu, Hae Yeoun Lee, Dong -Hyuck Im, Jung Ho Choi, and Heung Kyu Lee, "Electrophotographic Printer Identification By Halftone Texture Analysis," in Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference, 2010, pp. 1846-49.

[10] Yubao Wu, Xiangwei Kong, Xin'gang You, and Yiping Guo, "Printer Forensics Based On Page Document's Geometric Distortion," in Image Processing (ICIP), 2009 16th IEEE International Conference , Cairo, 2009, pp. 2902-2912.

[11] Jung Ho Choi, Dong Hyuck Im, Hae Yeoun Lee, and Taek Jun Oh, "Color Laser Printer Identification by Analyzing Statistical Features on Discrete Wavelet Tansform," in Image Processing (ICIP), 2009 16th IEEE International Conference , Cairo, 2009, pp. 1505-1508.

[12] John G Cleary and Leonard E Trigg, "K^*: An Instance-based Learner Using an Entropic Distance Measure," in ICML, 1995, pp. 108-114.

[13] Anil k Jain, Fundamentals of Digital Image Processing.: Pearson Prentice-Hall, 2004.