Machine Authentication of Security Documents

Utpal Garain
Computer Vision & Pattern Recognition Unit
India Statistical Institute
203, B.T. Road, Kolkata 700108, India
Email: utpal@isical.ac.in

Biswajit Halder
Dept. of IT, Greater Kolkata College of
Engineering and Management
Baruipur, 24 Pgs (South), WB, India.
Email: biswajithalder88@gmail.com

Abstract

This paper presents a pioneering effort towards machine authentication of security documents like bank cheques, legal deeds, certificates, etc. that fall under the same class as far as security is concerned. The proposed method first computationally extracts the security features from the document images and then the notion of 'genuine' vs. 'duplicate' is defined in the feature space. Bank cheques are taken as a reference for conducting the present experiment. Support Vector Machines (SVMs) and Neural Networks (NN) are involved to verify authenticity of these cheques. Results on a test dataset of 200 samples show that the proposed approach achieves about 98% accuracy for discriminating duplicate cheques from genuine ones. This strongly attests the viability of involving machine in authenticating security documents.

1. Introduction

With the advent of cheap and sophisticated scanning and printing technologies, the security document frauds are on rise. This poses a serious threat to the society and the economics of a nation. At present, normally the human experts in forensic departments check the authenticity any security document, in question. They in general involve different devices (e.g., UV lamp, magnifying glass, IR detector, etc.) [1] to check certain properties of the document in question to come up with a decision. However, execution of such a process is quite difficult while dealing with a huge number of such documents. Involvement of forensic people often requires several steps like filing a legal case on finding a document in question, sending the document in original to the police people, waiting for expert's view, etc. making the entire process a lengthy one.

Therefore, design of an automatic means for authentication of security documents would be of enormous help for the communities that everyday deal with a large number of security documents. The present is motivated by this research need. Bank cheques are considered as reference security documents. The approach is based on image processing and pattern recognition principles by which relevant image-level features are initially extracted from cheques and then using these features an algorithm for discrimination between genuine and duplicate cheques is outlined.

1.1. Security features in bank cheques

Depending upon the importance of a document, nature of security features varies from one class of document to others. For instance, currency notes are having very high level of security aspects and security features in documents like lottery tickets, postal stamps, etc. are somewhat less complicated. Certain features in bank cheques, certificates, etc. are fall in the middle of this complicacy scale. Security features in bank cheques [2, 3] are mostly incorporated in three distinct areas.

Security design or background artwork: Generally light fine-line printing or others security patterns are appear on background of cheques. This type of printing is difficult to reproduce on scanning equipment or replicate by other printing methods. The use of an intricate fine-line/patterns background design (for example, Guilloche module [4]) is an essential part of any security document. Apart from using fine-line artwork, repetition of micro-sized typeset characters through out the background can also be included in the design. Sometimes special security icons or descriptive markers are used so that verification of cheques becomes an easy task for the bank people. Watermarks, silver bullets, MICR fonts, hidden symbols, company logo, etc. generally fall under this category.

Use of color inks: Color ink pigments contribute substantially to the security of cheques. For example, fluorescence of color pigment is often used to counter

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.234

IEEE computer society

color photocopying of negotiable documents. Similarly, key areas of the security design may use fugitive inks that are highly sensitive to a variety of solvents. A relatively new breed of inks known as thermo chromic inks is also used for security measures.

*Paper and printing process***:** The quality of the paper also plays as security feature. Sensitized paper stock, fluorescent fiber (invisible or visible) based paper, etc. are often used and different security aspects are addressed. Sometimes paper manufacturers use their own watermarks that may provide additional visual protection. Printing process also provides security to documents like bank cheques. For instance, intaglio printing as used for printing bank cheques is special kind of offset printing that gives a document a very high quality look that is difficult to reproduce by using scanners, color copiers or computers with color laser printers. Use of MICR (magnetic ink character recognition) characters at bottom of the bank cheques is also considered as a security means.

The rest of the paper describes the development of the proposed system. Section-2 elaborates the features extracted from bank cheques and at the same time, explains why these features are considered important in authenticating the bank cheques. Section-3 describes the proposed authentication scheme. The support vector machines (SVM) and neural Networks (NN)-based classifiers are considered for discriminating genuine vs. duplicate cheques. Section-4 presents the experimental results to confirm that the security features captured computationally are sufficiently robust for the said purpose. Section 5 concludes the paper and outlines the issues to be considered in future.

2. Feature Extraction

As discussed earlier, there are three aspects of the security features in bank cheques: (i) color features, (ii) background artwork and logo, and (iii) paper quality. In this study, we capture features related to the first two categories. Four different attributes are investigated for color related security and two more features are computed from the background artwork or texture patterns. Computation of these six features and rationale behind choosing them to discriminate genuine vs. duplicate documents are discussed below.

Computation of this feature requires registration of two images (reference and target images). In the present study, there exists a rectangular box at the right of each cheque. This box is given for writing the amount in numerals. We have considered this rectangular box for registration. Detection of this box in the images is rarely missed and this process needs less amount of computational effort. Once this box is detected, its four (outer) corner points are chosen as *control* or *tie* points for registration.

Average Image Hue (f_h): In printing color theory, hue defines the quality of a color. Perceptually two colors may look alike but if they are printed with different ink pigmnets they will occupy different position in the color cube. Actually, dominant wavelength is a physical analog to this perceptual attribute, i.e. hue. Because of this reason, comparison of hue values in two images may give a significant clue to decide whether color quality in those images are same. Since the proposed system takes an RGB image (of documents, e.g. bank cheques) as input, hue for an individual pixel (p), h_p is computed from its RGB values r_p, g_p, and b_p as follows:

$$h_p = \begin{cases} \theta & if \quad b_p \le g_p \\ 360-\theta & if \quad b_p > g_p \end{cases} \tag{1}$$

where θ is the angle measured with respect to the red axis of the HSI color space. For each pixel, h_p is measured and then an average hue is computed for the entire image. Let f_h denote this average hue.

Gray level variation (f_{gv}): The standard deviation of the gray level distribution (of the image pixels) is considered as a feature and denoted by f_{gv}:

$$f_{gv} = \sqrt{\frac{\sum\left(g_p - \bar{g}\right)^2}{N-1}} \tag{2}$$

where N is total number of pixels, g_p be the gray value of the pixel p and $\bar{g}$ be the mean gray value.

Binary correlation (f_{bc}): Computation of this feature assumes the knowledge of authenticity (or genuineness). A given (or target) document (d_T) is compared with the genuine one (d_R). Here a correlation between two binary images (gray images are converted into binary images using Otsu's thresholding method [5]) is measured. The correlation coefficient (i.e. a similarity measure), r between the reference (i.e. genuine) and the target image is measured as,

$$r(d_R, d_T) = \frac{1}{2} - \frac{s_{10}s_{01} - s_{00}s_{11}}{2\sqrt{(s_{11}+s_{10})(s_{01}+s_{00})(s_{11}+s_{01})(s_{10}+s_{00})}} \tag{3}$$

where s_{00}, s_{11}, s_{01}, and s_{10} denote the number of zero matches, one matches, zero mismatches, and one mismatches, respectively. This coefficient lies in [0, 1] and gives the value for the feature, f_{bc}.

Kurtosis of image colors (f_{kr}, f_{kg}, and f_{kb}): Apart from measuring standard deviation of gray values we also measure kurtosis to analysis whether variations is due to infrequent extreme deviations. R, G, and B channels are separately considered to measure respected kurtosis. For instance, let f_{kr} denote the kurtosis of red channel and it is measured as

$$f_{kr} = \left\{ \frac{N(N+1)}{(N-1)(N-2)(N-3)} \sum \left(\frac{r_p - \bar{r}}{\sigma_r} \right)^4 \right\} - \frac{3(N-1)^2}{(N-2)(N-3)} \quad (4)$$

where r_p be red value of pixel *p*, $\bar{r}$ and σ_r be the mean and standard deviation of red values of the image. Similarly, f_{kg} and f_{kb} are computed to denote kurtosis of green and blue channels, respectively.

The following two features are considered to investigate the changes in the background artwork. This artwork is basically a texture pattern and if the forgery were done very carefully it would be difficult to identify the changes with open eyes. Forensic people use magnifying glass or microscope to detect changes, if any. Here, features are also carefully extracted so that they would be able to report the unexpected changes, if any exists, in a given document image.

Measure of Line Quality (f_l): Since the background artwork is basically a line drawing, it has been experienced by the forensic community that scanning and subsequent printing of the scanned image results in broken lines in the artwork part (though perceptually the artwork may look like one in the genuine document). To capture this attribute, we measure the line quality of a given image as follows.

In the binary image, straight lines originating from a pixel in three directions namely horizontal (left to right), vertical (top to bottom) and diagonal (in the South-East direction) are identified. Length of each such line is recorded. Isolated pixels are not counted. Detection of such lines started from the leftmost top pixel and the algorithm scans pixels in row major order. Once a line is identified, its containing pixels are (virtually) deleted so that the same straight line or its part is not counted repeatedly. Let *L* be the number of lines identified in the above process and l_i be the pixel length of the *i*-th line. The average length of these lines gives the value of the feature, f_l and is computed as,

$$f_l = \frac{1}{L} \sum_{i=1 to L} l_i \quad (5)$$

Fourier Power Spectrum (f_{ps}): Since the background artwork is a result of repetitive pattern (or texture), Fourier analysis of the image provides significant clues to identify unexpected changes in the artwork pattern. Gray version of the image is considered and its centered Fourier Spectrum is investigated. In this spectrum as the dominant frequencies are distributed over a very small region around the center, it's difficult to differentiate a nicely forged duplicate document from its corresponding genuine copy. However, if we consider the log transformation of this spectrum, differences become more evident and significant. This attribute is captured in the feature f_{ps} and computed as

$$f_{ps} = \log(1 + FS) \quad (6)$$

where *FS* is the centered Fourier spectrum, $FS = \sqrt{R^2 + I^2}$. *R* and *I* are the real and imaginary parts of the Fourier Transform, *F*.

3. Automatic Authentication

Authentication of a cheque document is modeled as a 2-class pattern recognition problem, i.e. whether the document belongs to the genuine class or not. Let *m* be the number of samples known as genuine and *n* be the number of samples known as duplicate. In the feature space, it is expected that these *m* samples would form a cluster (C_G) and *n* duplicate samples would form another cluster, C_D. To check whether these two clusters are linearly separable, we implement a k-means algorithm and cluster the *m*+*n* labeled samples into two classes. Selecting two samples randomly initializes the centers in k-means algorithm. Since k-means results get affected by this initialization phase, k-means is executed more than once (three times) and each time the clustering results are investigated. This investigation reveals that the clusters always overlap i.e. not linearly separable and therefore, it is difficult to find a linear decision boundary.

Next, support vector machines (SVM) are used with an aim of determining the location of decision boundaries that produce the optimal separation of classes. Two types of common non-linear kernel functions namely, polynomial and radial basis function (RBF) are considered. The whole sample set consisting of genuine as well as duplicate samples is divided into four subsets. A four-fold test is conducted so that each subset appears at least once as in training, validation and testing. The proportion in which samples appear in training, validation and test data is 2:1:1 (training: 50%, validation: 25% and testing: 25%).

The classification accuracy is also checked with a Neural Network (NN)-based classifier. An MLP (Multi-Layer Perceptrons) consisting of 8 input nodes correspond to eight dimensions of a feature vector is used. The output consists of only one node to gives binary output (genuine or duplicate). Hidden layer, in the present experiment, contains 2 nodes. A logistic function as explained in the next section is used as the activation function of the network. Like SVM-based classifier a four-fold test is conducted for NN-based classification. Samples appear in training, validation and test data following the ratio 2:1:1.

4. Experimental Results

Professionals from printing press designed sample cheques for our study. All the security measures [2] that are normally taken care of during designing and printing of real bank cheques were also followed in designing sample bank cheque. Prints of cheques are taken from a printer specially designed for security printing. Fake or duplicate cheques have been generated using the means that are commonly used in forgery of such documents. Forensic experts were consulted for this purpose to know the process of forgery in details. Hi-end scanners, printers (mostly offset), image editing software packages [6] and experts, etc. were involved in generating duplicate versions of bank cheques. The kind of paper used for taking print out is very similar in quality to ones used in printing of cheques in press. After printing, duplicate cheques are manually checked for their perceptual similarity with the authentic ones. In total, 200 cheques (100 *genuine*, 100 *duplicate* samples) are considered in the present study. A flatbed scanner is used to scan these cheques at 300 dpi, true color images, which are stored as uncompressed .tiff format. Each cheque takes about 16MB storage spaces.

Table 1. K-means results for clustering of samples in two clusters.

	Distribution of samples in clusters				Clustering Accuracy (%)
	#Samples in Genuine (G)		#Samples in Duplicate (D)		
	G	D	D	G	
Iteration 1	90	12	88	10	178 (89%)
Iteration 2	88	14	86	12	174 (87%)
Iteration 3	90	10	90	10	180 (90%)
Average	89.3	12	88	10.7	177.3 (88.7%)

Results of the k-means: The k-means clustering (k=2) is done to analyze the distribution of samples in the feature space. The algorithm finds two clusters one corresponding to *genuine* samples and another for *duplicate* samples. Initialization is done by choosing two samples randomly to initialize two cluster centers.

The *k*-means results are evaluated by computing the number of similar samples grouped together vs. the number of dissimilar samples contained in that group. Since all samples are tagged with their classes (*genuine* or *duplicate*) evaluating clustering results in this way is straightforward. Table-1 presents the evaluation of k-means results. Since cluster centers are initialized randomly, k-means were executed three times to get an average result. From Table 1, overlapping of samples in the feature space can easily be visualized.

SVM-based Classification: SVM-based classification makes use of two different types of non-linear kernel functions namely polynomial and radial basis function (RBF). These two kernel functions are defined as:

$$\text{Polynomial: } K(x, x') = (x.x' + 1)^d \quad (7)$$

$$\text{RBF: } K(x, x') = \exp\left(-\gamma\|x - x'\|^2\right), \text{ for } \gamma > 0 \quad (8)$$

Where x denotes training vectors and d, γ are the kernel parameters.

The set of 200 samples are divided into 4 sets to realize a four-fold experiment. In each run, two sets are considered as training sets, the remaining two sets serve as validation and test sets. Each set appears at least once as a test set and a validation set. Four different runs were executed such that each set appears twice as training set. Table-2 reports the result of this four-fold experiment. It is to be noted that the following values were estimated for the kernel parameters for polynomial: d = 3 and for RBF: γ =1 and they do not change with changing of training sets.

Table-2 shows some important observations. For the present problem, a polynomial kernel function performs better than an RBF based kernel. Moreover, the number of support vectors used by the polynomial kernel is far less than the number used by the RBF based kernel. However, values for mean squared error (MSE) show that RBF kernel gives very low MSE when compared to a polynomial kernel.

Classification using Neural Network: As mentioned before an MLP is used to design a Neural Network-based classifier. Well-known back propagation algorithm is used to train the network. The network does use of the following logistic function as transfer or activation function.

$$f(x) = e^x / (1 + e^x) \quad (9)$$

A gradient descent method is used to find the optimized set of connection weights that are updated as per the following equation. Update of weights does consider both the t-th and (t-1)-st weights to compute the (t+1)-st weight as follows.

$$W_{(t+1)} = W_t + \alpha * (\frac{\partial E}{\partial W})|_{W(t)} + \beta * (W_{(t)} - W_{(t-1)}) \quad (10)$$

Where α is the learning parameter, β is known as the momentum and E is the error term. In the present experiment, α is set to 0.9 and β is assigned 0.1. The same dataset as used for SVM-based classifier is also used here to train and test the MLP. Like SVM a four-fold experiment is conducted and results are reported in Table-3. The training, validation and test sets used in different runs of experiments are exactly the same as they were in SVM-based classification. Table-3 shows NN-based classification gives about 97.5% accuracy in classifying test documents as genuine or duplicate. The accuracy is slightly less than that of SVM-based classifier but both the results are definitely comparable. Integration of these two classifiers will be considered in next step of this study.

Table 2. Classification of bank cheques using SVM.

	#Support Vectors		Classification accuracy (%) on Test Dataset		MSE	
	Polynomial	RBF	Polynomial	RBF	Polynomial	RBF
Run 1	6	65	100	100	4.806	0.06
Run 2	7	56	96	96	0.982	0.169
Run 3	5	61	98	96	2.05	0.154
Run 4	5	60	98	96	1.139	0.169
Avg.	5.75	60.5	98	97	2.24	0.138

Table 3. Classification of bank cheques using Neural Network.

	Classification accuracy (%)			
	Training Dataset		Test Dataset	
	#Samples*	Correct Classification	#Samples	Correct Classification
Run 1	100 (G: 54, D: 46)	100% (G: 54, D: 46)	50 (G: 23, D: 27)	98% (G: 23, D: 26)
Run 2	100 (G: 46, D: 54)	98% (G: 45, D: 53)	50 (G: 30, D: 20)	96% (G: 29, D: 19)
Run 3	100 (G: 60, D: 40)	98% (G: 60, D: 38)	50 (G: 22, D: 28)	98% (G: 22, D: 27)
Run 4	100 (G: 52, D: 48)	97% (G: 51, D: 46)	50 (G: 20, D: 30)	98% (G: 19, D: 30)
Avg.		98.25%		97.5%

* G: Genuine and D: Duplicate

5. Conclusions and Future Scope of Study

This paper provides an automatic means for verification of authenticity of security documents in particular, bank cheques. To the best of our knowledge, this is the first attempt for authentication of the bank cheque. The present could serve an as additional module of a bank cheque reading system [7]. Moreover, the proposed framework considers design of an efficient but low-cost solution so that mass scale deployment of such systems could be feasible.

The future extension of the study would conduct experiment on a larger dataset to test the generality and scalability of the proposed method. Analysis of error cases has not been reported here and such a study will be considered as the next part of the present study. Finally, experiments with other kinds of security documents [8] like legal deeds, lottery tickets, tickets for watching different outdoor games, certificates, mark sheets, postal stamps, etc. are to be conducted to establish the well acceptance of the method for authenticating security documents.

6. References

1. "Procedure Manuals" prepared by Directorate of Forensic Science, Ministry of Home Affairs, Govt. of India, http://www.dfs.gov.in.
2. U. Garain and B. Halder, "On automatic authenticity verification of printed security documents," in Proc. of the 6th *Indian Conference on Computer Vision, Graphics, and Image Processing* (ICVGIP), Bhubaneswar, India, Dec. 2008.
3. Mechanised Cheque Processing Using MICR Technology-Procedural Guidelines (Abridged), by Reserve Bank of India, Dept. of IT, central Office, Mumbai, India. Web: http://www.rbidocs.rbi.org.in/ rdocs/ Publications/DOCs/33140.doc.
4. Fortuna Users' Guide, © Barco Graphics, Belgium.
5. N. Otsu, "A threshold selection method from gray-level histograms," IEEE SMC, vol. 9: 62–66, 1979.
6. SecuriDesign tool of CorelDRAW 11, © Corel Corp.
7. C.Y. Suen, Q. Xu and L. Lam, "Automatic recognition of handwritten data on cheques – Fact or Fiction?" Pattern Recognition Letters. 20: 1287-1295, 1999.
8. US Patent 5682103, "Infrared detection of authenticity of security documents comprising electromagnetic particles".