Chapter 6

Automatic Numeric Data Extraction from Pre-Printed Input Forms: Case Studies

Large numbers of applications are available in real world that can be subjected for automatic data extraction from the input forms. Most of the applications require both numeric and alphabetic data to be extracted from the input forms. In this research work the automatic data extraction work is limited only for extracting numeric data from the input forms. Accordingly two case studies are considered for automatic numeric data extraction. The first case study is automatic reading of printed PINCODE from postal mails. In this case study, two approaches are developed. The first method automatically reads the printed English/Kannada PINCODE through identification of script. The second method reads the printed English/Kannada PINCODE without identification of the script. Both the methods use extended 7-segment projection for bilingual PINCODE recognition. The second case study is automatic reading of denomination entries of cash in bank remittance challan. The method is developed using modified 7-segment projection for handwritten numeral recognition.

In this chapter, Section 6.1 describes the methods developed for automatic reading of printed English/Kannada PINCODE from postal mails. Section 6.2 describes the method developed for automatic reading of denomination entries in bank challan. Section 6.3 gives the summary of the works explained in this chapter.

6.1 Automatic Reading of printed English/Kannada PINCODE

The postal mail service continues to thrive as there is a need to deliver documents, journals, catalogs etc in their original form to the addressee, which is not possible by using other modern communication tools such as email, fax etc. Automation is being introduced to make postal services efficient. In this connection the postal service providers have introduced postal codes to help in efficient and automatic sorting of postal mail. "India Post" the department of posts in India has introduced the Postal Index Number (PIN) system[Ind100]. The terms PIN and PINCODE are interchangeably used to refer to the postal index number. PINCODE system works on an intricate pattern of connections based on geography and transportation modes. India has been divided into eight PIN regions and a six-digit code has been allotted to each delivery post office, except the branch post office[url9]. But the PIN code can be written in any of the various vernacular languages prevalent in India and mentioned in schedule 8 of Indian Constitution [Ind100]. Hence there is a need to develop efficient address reading systems more particularly for PINCODE recognition irrespective of the script.

The all-numeric nature of the PINCODE makes the task of PINCODE interpretation comparatively easy, involving recognition of the ten possible numerals. In a multi-lingual country like India the PINCODE can be written in any of the languages listed in schedule 8 of the Indian constitution [Ind100]. This introduces an additional complexity in PINCODE interpretation as the script of the PINCODE is to be ascertained before recognizing the numerals of PINCODE. Hence an automatic PINCODE interpretation system requires a two-stage process. In the first stage the script of the PINCODE is to be identified which is followed by script dependant numeral recognition stage. Though there are many languages in the country mail sorters in a post office are generally required to handle a limited set of languages. The languages handled by a post office vary from 2 to 5 depending on the location and use of languages in the region. More commonly the languages used for writing the PINCODE are the regional language of the area (for example Kannada in Karnataka

state) and English the accepted common language for communication. The destination postal address and PINCODE can be handwritten, hand printed or machine printed. The statistics available at post offices suggest that 10-12% of the addresses are machine printed and others either hand printed or hand written. Also the use of machine printed addresses is on the rise as the penetration of computers increases. In this context a PINCODE interpretation system that can recognize and interpret PINCODE printed in at least two scripts can be gainfully employed in a practical scenario.

Numeral recognition is at the heart of any PINCODE recognition systems. Numeral recognition is an offshoot of character recognition research. Literature survey on character recognition reveals that sufficient research has been carried out on character recognition [Gov90; Gor98]. Many researchers have come out with robust methods for character recognition based on neural network techniques [Cun89; Jai94; Tan98; Nag96], fuzzy methods [Soo101; May102], zoning/region decomposition[Kim91; Sue95], transformation techniques[Sue92; Cao94; Flu94] graphical and geometric approaches[Ram89; Cha103] etc. Researchers have proposed various techniques for numeral interpretation; a multi-layer neural network using back propagation is presented in [Cun89] for recognition of digits drawn from handwritten zip codes with a recognition accuracy of 92%. A real time system for recognition of 5 digit zip codes is presented in [Gre97]. The zipcode recognition is cross validated with the recognition of city and state names and an accuracy of 73% is reported. An HMM based zip code candidate generator and a city-state zip code verifier is proposed in [Jia98] which obtains a recognition accuracy of 83.5%. All these methods use complex models and are computationally expensive and are specifically designed for handwritten characters and numerals written in a specific language. But, printed PINCODE recognition involves a small set of classes and does not require complex classifiers in the recognition process. This motivated to come out with a proposal to develop simple but robust bilingual numeral classifier to classify printed numerals in PINCODES. In this research work two approaches are proposed for automatic PINCODE interpretation system for recognizing PINCODES printed in English and

Kannada. Subsection 6.1.1 describes a two stage approach for automatic PINCODE interpretation system in which first stage identifies the script using modified invariant moments and in the second stage the PINCODE is interpreted using 7-segment projection. Subsection 6.1.2 describes a single stage approach for automatic PINCODE interpretation system without identifying the script. In section 6.1.3 a brief summary and conclusion on PINCODE interpretation systems is discussed.

6.1.1 Recognition of PINCODE Printed in Kannada/ English through Script Identification[10]

A few works on script identification are found in literature to distinguish between Latin and Oriental scripts and also the different languages used in these scripts. The image is first classified into Han and Latin script based on spatial relationship of the features related to upward concavities in [Spi97]. Further, Language identification within Han script class is performed by analysis of optical density in text images, Language identification within Latin script class is performed by a technique based on character shape code and word shape tokens [Spi97]. Cluster based templates are employed to identify thirteen different scripts in [Hoc97]. The extraction of rotation invariant features using a pair of Gabor filters and their use in script identification is proposed in [Tan98]. A neural network approach using spatial characteristics as features is proposed to classify various fonts of English in [Jun99]. Algorithms for skew detection, page segmentation and then script identification using bounding box features and horizontal projection is presented in [Wak98]. Document images are classified into script and language classes using various features such as height distribution of bounding boxes, character density and horizontal projection, in [Wak98]. A few works to distinguish between various Indian language scripts are found in literature. A script identification methodology for two Indian Languages along with English is proposed in [Bas102]. The methodology uses around 50 features

[10]This part of work is published in the International Journal of Intelligent Systems and Research, "Recognition of PINCODE Printed in Kannada/English: Script Identification through Texture Analysis and Recognition based on 7 - Segment Projection", Vol. 1, No. 1, pp. 69-82, June 2007

drawn from original and morphologically operated images. A system for Indian Postal Automation is presented in [Roy105], the paper presents algorithms for Destination Address Block (DAB) location, script identification, numeral recognition for PIN code reading and word recognition for city name reading. The features employed for script identification are derived from water reservoir concept and the existence of shirorekha which is suitable for distinction between Bangla and English script which are the languages considered in the said work. The focus of script identification research is on alphabet-based identification and numeral script identification is generally overlooked.

There are certain areas where numeral script identification is very vital such as automatic interpretation of Identity Numbers, Vehicle registration numbers, PINCODES etc. There are very few works pertaining to numeral script identification. Various methodologies to classify the images into different script classes (including numerals) for line images and word images are proposed in [Pad104]. A texture based method based on Hu's absolute moments is also proposed for script identification in [Pad104]. Hu's absolute moments have been shown to be variant for symmetrical images in [Pal100], hence texture features derived from modified invariant moments have been employed in this work to identify the PINCODE numeral script.

The research work presents an integrated system for recognition of PINCODES printed in Kannada and English after script identification. The proposed methodology uses the PINCODE image extracted from the postal mail as input. The input PINCODE image is processed to remove speckle noise and thinned to get outline image of the PINCODE. The modified invariant moments [Pal100] of the thinned image are extracted and used to derive the newly defined texture feature. This texture feature is used to identify the script of the PINCODE. The thinned PINCODE image is further subjected to recognition using 7-segment projection based on the identified script. The system is robust enough for practical use and an identification accuracy of 96.66% is obtained when tested on a large sample of PINCODE images printed in various font styles and sizes.

This section is organized into four subsections. Subsection 6.1.1.1a describes the architecture of the proposed system. Subsection 6.1.1.2 gives the texture analysis methodology for script identification. Subsection 6.1.1.3 presents the experimental analysis based on 7-segment projection. Subsection 6.1.1.4 presents the conclusion.

6.1.1.1 The PINCODE Interpretation System

The PINCODE interpretation system is devised to handle PINCODES printed in two languages namely, English and Kannada. The PINCODE image is extracted from the mail image by a simple position based algorithm [Nag105]. The image is thinned using the morphological thinning operator and is processed to extract the modified invariant moments as described in Chapter 3, section 3.1. These features are employed to compute the newly devised texture feature γ which is used in script identification. The thinned PINCODE image is segmented into numerals and each numeral is identified using decision trees and 7-segment projection features. The numeral identification is summarized to recognize the PINCODE. The complete process of PINCODE interpretation is brought out by the flowchart in Figure 6.1.1.1a. The texture analysis based script identification for PINCODE is described in next subsection.

According to the system flow-chart depicted in figure 6.1.1.1a, the PINCODE image extracted from DAB is taken as input. The PINCODE image is thinned and thinned image undergoes two parallel tasks. In one task, the invariant moments are computed to derive texture feature γ for identification of script. The second task performs projection process of the digits in PINCODE through which the digits are identified later with the help of script based decision tree. Finally, the recognized digits are summarized to interpret the PINCODE.

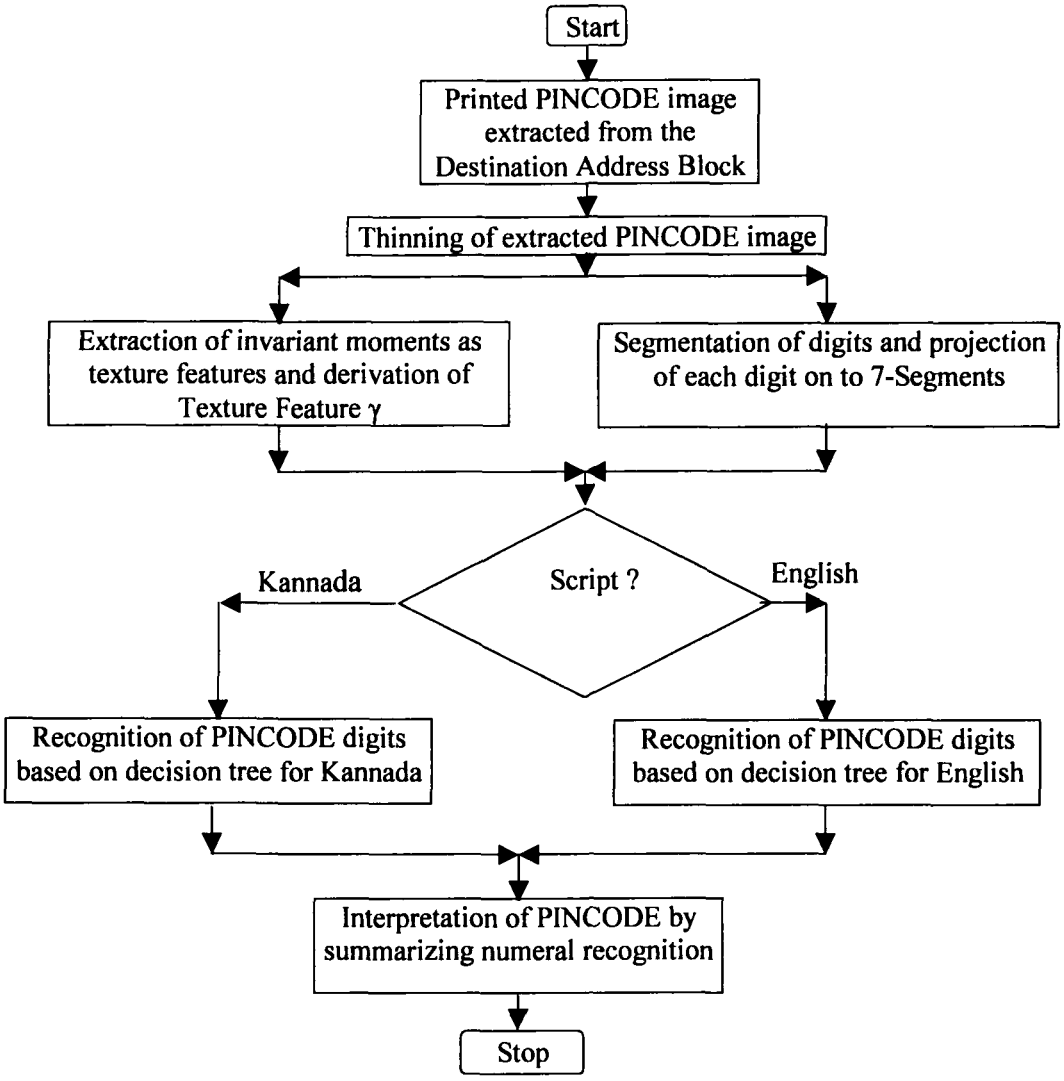


Figure 6.1.1.1a Flowchart for the PINCODE Interpretation System

6.1.1.2 The Texture based Script Identification

The script of the PINCODE numerals taken as one image can be found by the overall texture of the image as different scripts have different texture/ appearance. Also the PINCODE of a given script has the same texture irrespective of the numerals employed. Hence texture features of the PIN Segment can be employed gainfully for identification of the PINCODE script. Various features can represent the texture of an image; the prominent among them are moments of the image. Hu's absolute moments have been employed in several character recognition works, but it is shown in

[Pal100] that Hu's moments are not invariant particularly for symmetric image objects. The numerals of the various scripts considered have several such characters, hence Hu's absolute moments are not suitable for this work. The current work uses the scaling and translation invariant modified invariant moments suggested in [Pal100], after normalization to represent the texture of the PIN segment. The modified invariant moments are found after thinning the PIN segment to overcome the effect of varying thickness of the machine printed PINCODES. The PIN Segment is thinned using the thinning tool supported by MATLAB and is further processed to find the normalized and modified invariant moments. The modified invariant moments are derived from the first and second order geometric moments of the PIN Segment. The intelligent script identification strategy presented in this paper utilizes a texture feature based on modified invariant moments. The modified invariant moments φ_{20}, φ_{02} and φ_{11} are computed as explained in chapter-3 through equations 3.3.1 to 3.3.5. These modified invariant moments are adjusted to get significant values and are normalized to get the texture features as given by equation (6.1.1.2a).

$$\phi_{20} = \frac{|\log(\varphi_{20})|}{H * W}$$
$$\phi_{02} = \frac{|\log(\varphi_{02})|}{H * W}$$
$$\phi_{11} = \frac{|\log(\varphi_{11})|}{H * W} \qquad (6.1.1.2a)$$

Experimentation with large samples of PINCODE images printed in English, and Kannada has yielded statistical average values for each of the three features with very little variance. The features ϕ_{20}, ϕ_{02} and ϕ_{11} differentiate between the two language classes, but the class separation becomes more distinct by using the texture feature γ derived from ϕ_{20}. The feature γ is evaluated as given in equation (6.1.1.2b). The statistical average values and the standard deviation of the feature γ is listed in table-6.1.1.2a.

$$\gamma = \phi_{20} * H * W \qquad (6.1.1.2b)$$

Table-6.1.1.2a: Average and Standard Deviation of Texture Feature γ

Script	γ	
	Average	**Std Dev**
English	0.354592	0.191276
Kannada	1.288075	0.164425

A decision methodology to identify the script of the PIN segment is devised by using the feature γ, taking into account the average and standard deviation values given in table-6.1.1.2a. The script of the input image is identified as Kannada if γ value is greater than 1 and it is identified as English if the γ value is less than 1. The identified script along with PINCODE image is passed to the PINCODE interpretation stage.

In the PINCODE interpretation stage PINCODE image is first segmented into individual numerals by simple X cut segmentation technique [Gor98]. Each numeral is then subjected for projection followed by classification and finally recognized numerals are combined to interpret the PINCODE. The process of projection and classification of numerals is as described in chapter 5 section 5.1.

6.1.1.3 Experimental Analysis

The proposed PINCODE interpretation system was tested thoroughly using PINCODE images printed in different fonts of English and Kannada. The script identification accuracy is found to be about 95% for English script and 100% for Kannada script PINCODES. It is found that the system is robust enough to identify the script of the PINCODE image. The font-wise script identification accuracy is listed in table-6.1.1.3a.

Amongst the fonts tested Century Gothic and Tahoma in English script have resulted in certain amount of script misclassification. This is due to rounded nature of numerals in these font styles and the misclassification is more pronounced when the PINCODE consists of more number of 0's. The PINCODE is subjected to interpretation after script identification.

The result of PINCODE interpretation using 7-segment projection show high efficiencies for both English and Kannada and is dependent on the efficiency of script identification. Testing is performed on various script identified samples of different font style and font size. The summary of recognition for English PINCODES is shown table-6.1.1.3b and for Kannada PINCODES in table-6.1.1.3c.

Table-6.1.1.3a: The Font-wise Script Identification

Sl No	Script	Font Style (All font sizes are tested)	Number of PIN code images tested	Number of images with correct script identification	Recognition Accuracy in percentage
1	English	Arial	20	20	100
2	English	Times New Roman	20	20	100
3	English	Tahoma	20	16	80
4	English	Verdana	20	020	100
5	English	Century Gothic	20	18	90
6	English	Courier	20	20	100
7	English	All fonts	120	114	95
8	Kannada	Sirigannada	25	25	100
9	Kannada	Kasturi	25	25	100
10	Kannada	Vijaya	25	25	100
11	Kannada	Kannada Extended	25	25	100
12	Kannada	All fonts	100	100	100

Experiments conducted indicate that English PINCODES are recognized efficiently with an average recognition rate of 100% for most of the commonly used font styles. The efficiency is very much sensitive to artistic font styles like Calligraph, Monotype Cursiva, Broadway, Gothic etc. Efficiency of recognition for Kannada PINCODES is also good except for Kasturi font style since it is artistic in appearance. The average recognition rate for Kannada PINCODES is 98.33% with Kasturi font style and 100% without Kasturi font style.

Table-6.1.1.3b: Result of English PINCODE recognition

Font Style	Font Size	No. of Samples Tested	No. of Samples Recognized	% of Recognition
Arial	8	56	56	100.00
	10	54	54	100.00
	12	59	59	100.00
	14	52	52	100.00
Times New Roman	8	56	56	100.00
	10	54	54	100.00
	12	59	59	100.00
	14	52	52	100.00
Tahoma	8	56	56	100.00
	10	54	54	100.00
	12	59	59	100.00
	14	52	52	100.00
Century Gothic	8	56	56	100.00
	10	54	54	100.00
	12	59	59	100.00
	14	52	52	100.00
Courier	8	56	56	100.00
	10	54	54	100.00
	12	59	59	100.00
	14	52	52	100.00
Verdana	8	56	56	100.00
	10	54	54	100.00
	12	59	59	100.00
	14	52	52	100.00
Average Recognition				**100.00**

The combined script identifier and PINCODE recognizer was tested and the results are enlisted in table-6.1.1.3d. The PINCODE interpretation for correct script identification is brought out. The system has achieved a PINCODE recognition accuracy of 99.14% when only correctly identified scripts are input to the PINCODE recognition stage. When the PINCODE recognition stage was tested irrespective of the correctness of script identification, an accuracy of 96.66% was obtained.

Table-6.1.1.3c: Result of Kannada PINCODES recognition

Font Style	Font Size	No. of Samples Tested	No. of Samples Recognized	% of Recognition
Siri	8	58	58	100.00
	10	52	52	100.00
	12	55	55	100.00
	14	52	52	100.00
Vijaya	8	58	58	100.00
	10	52	52	100.00
	12	55	55	100.00
	14	52	52	100.00
Kasturi	8	58	50	86.21
	10	52	49	94.23
	12	55	52	94.54
	14	52	49	94.23
Kannada Extended	8	58	58	100.00
	10	52	58	100.00
	12	55	55	100.00
	14	52	52	100.00
Average Recognition				**98.38**

Table-6.1.1.3d: Combined script identification and PINCODE Interpretation Results

Sl No	Script	Images Tested	Script Correctly Identified	% Script Identification	PINCODE correctly identified after script identification	Recognition accuracy after script identification
1	English	120	114	95.00	114	100.00
2	Kannada	120	120	100.00	118	98.33
3	Both	240	234	97.50	232	99.14

In case the PINCODE script is incorrectly recognized the numeral recognition strategy rejects such script identification as the projection pattern obtained for the numerals are not defined. But if the PINCODE with misclassified script has only those numerals, which have valid projection in both the scripts, (for example English 0, 2 and Kannada 0, 3, 7) then the numeral recognition methodology wrongly recognizes the PINCODE.

6.1.1.4 Conclusion

The PINCODE interpretation system developed is based on the decision tree after projecting the test image on to seven segments for numeral recognition. Before the numeral recognition stage, the script of the printed PINCODE is identified using texture parameter defined here. The recognition accuracy is good enough for a practical mail sorting application when printed mail images are considered. The proposed methodology is very sensitive to variations and is applicable to machine printed PINCODE only, however the method can be extended for hand printed PINCODEs with slight modifications like building knowledge base for hand printed numerals.

6.1.2 Recognition of PINCODE Printed in English/ Kannada without Script Identification[11]

If the text documents are made up of more than one language/script, one would expect that the recognition of the text contents would be preceded by a language recognition stage. However in particular instances such as in the recognition of numerals in PINCODE of Indian postal address, this two-stage requirement can be dispensed with by integrating language and numeral recognition. This is because, for the postal mail originating from a specific geographical area in India, PINCODE could be predominantly written either in a regional language (such as Kannada in Karnataka) or in English language (which is the accepted common language at national level). Hence a system, which can recognize PINCODE in two or more languages, has practical relevance and the two stage process can be eliminated by combining script and numeral recognition.

[11]This part of work is communicated to VIVEK: A Journal of Artificial Intelligence, 'A New 7 Segment Projection Model for the Recognition of PINCODE Printed in English and Kannada', Dec' 2006.

Integrated PINCODE recognition for two scripts requires identification of numerals only and the total classes are relatively less. In addition, when the numerals are machine printed or hand printed a complex recognition model is not required. Hence a new 7-segment projection approach is devised for the purpose. Two simple and direct methods for recognizing printed numerals are available and they are (i) Template matching [Mor92] and (ii) Histogram technique [Mor92]. These methods can also be extended to PINCODE recognition printed in two scripts/ languages. Though these methods show high accuracy, template matching method suffers from high time and space requirements whereas histogram technique requires a large amount of memory space. Keeping these points in mind, a less complex and less expensive model suitable for recognition of PINCODE printed in English and Kannada is proposed in this research work.

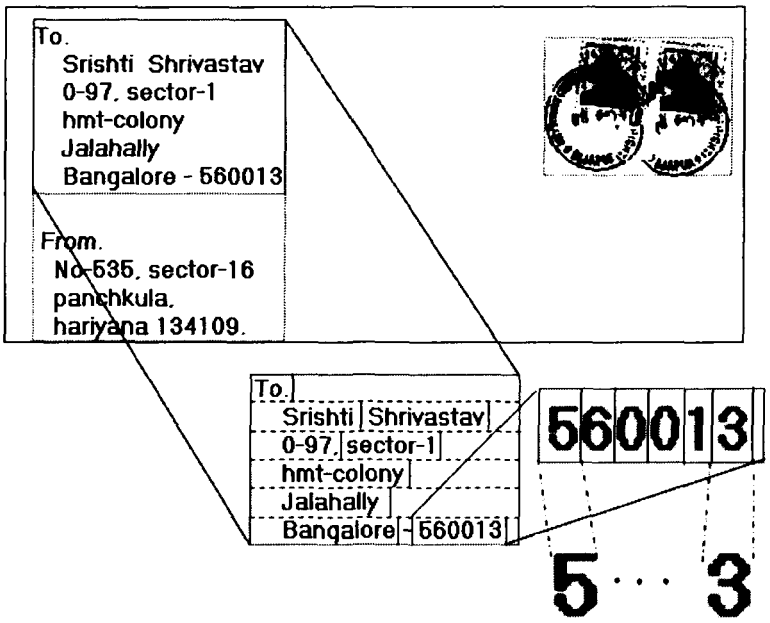
The proposed work employs a simple X-Y cut methodology for extracting the PINCODE segment from the destination address block (DAB) of the mail image. The extracted PINCODE segment is subjected to X-cut to separate the numeral images. Each numeral image is projected onto 7 segments from 7 non-overlapping regions (referred to as first level projection in this work) to identify the numeral along with the language. If the first level projection cannot uniquely identify the numeral, the numeral image is projected onto 7 segments from 7 overlapping regions (referred to as second level projection in this work) to uniquely decide on the numeral along with the script of the PINCODE. The identifications of all the numerals are combined to interpret the PINCODE. The methodology is tested on samples of PINCODE image segments extracted from mail envelopes and other PINCODE images prepared for the purpose of testing. The methodology was tested on 2032 PINCODE images and a combined script and PINCODE recognition efficiency of 99.02% is obtained.

The work is organized into five subsections. Subsection 6.1.2.1 describes the algorithms employed for PINCODE segment and numeral extraction. The experimental results are discussed in subsection 6.1.2.2. Subsection 6.1.2.3 gives the conclusion.

6.1.2.1 PINCODE Segment and Numeral Extraction

India Post has issued a guideline to include PINCODE in all destination addresses to help in speedier sorting of postal mail. India Post also insists that the PINCODE be written on the last line of the address all by itself or as the last part in the last line [Ind100], whatever be the script used. The proposed PINCODE extraction strategy assumes that the PINCODE of the destination address is written according to this guideline and is free from skew. The PINCODE is extracted from the mail image by first extracting the destination address block (DAB) and then the PINCODE segment is derived by using simple X-Y cut strategy. The DAB extraction methodology is applicable only for mail images of square/rectangular shape. But it is independent of the script/ language of the address information and also whether the address is handwritten or printed. It is resilient to various positional variations of different information blocks on the mail image. The destination address block identification is carried out based on simple position based approach using bounding rectangles. The positional information about the mails is gathered by observing a large sample of plain envelopes.

The image of destination address block is further processed to obtain the PINCODE segment. The PINCODE is extracted by separating the destination address into separate lines and the lines into tokens by using horizontal space between lines (Y cut) and vertical space (X cut) between tokens. The PINCODE extraction strategy assumes that the PINCODE lies in the last line of the destination address if the last line contains only one word and the PINCODE is the last word if the last line contains more than one word. The PINCODE segment is further divided into numerals by employing X-cut to separate the numerals. The methodology is depicted in figure(6.1.2.1a). The separated numerals are subjected to 7-segment projection to identify the script and the numeral. The complete 7-segment projection methodology is described in the subsequent section.



Figure(6.1.2.1a): Preprocessing Tasks to Extract PINCODE Image from a Mail Envelope

Each extracted digit of the PINCODE is subjected to 7-segment projection process and the recognition of all the digits is combined to interpret the PINCODE along with the script. The process of printed English/Kannada numerals without script identification is performed based on the method explained in chapter 5 section 5.2.

Results of all the 6 digits of a PINCODE subjected for recognition as described are obtained and analyzed to interpret the PINCODE. English '0' and Kannada '0' are not considered for identification of script as they do not contribute anything to identify the script. If the recognized digits of the PINCODE other than '0's are all English numerals then the PINCODE is of English script and if the recognized digits of the PINCODE other than '0's are all Kannada numerals then it is Kannada script PINCODE. Whenever, there is a combination of English and Kannada numerals in recognition of digits other than '0's in PINCODE then there is an error in recognition and decision is rejected. Failure occurs when the PINCODE interpretation is made incorrectly within a script and this arises due to misclassification in recognition process. Analysis of experiments conducted on the methodology is discussed in the next section.

6.1.2.2 Experimental Analysis

The methodology has been implemented using MATLAB 6.0 on Windows platform. The DAB and PINCODE segment extraction strategy has been tested on 80 envelope mail images with printed addresses and various positions for DAB. The DAB has been correctly identified in all the 80 cases and PINCODE segment has been correctly extracted from 78 mail images (ie 97.5% accuracy). The two mail images from which PINCODE segment were not correctly extracted was because of space between two parts of the PINCODE.

The PINCODE recognition strategy is tested on PINCODE segments extracted from mail images and other PINCODE images prepared for the purpose of testing the strategy. In all about 1000 English and 1000 Kannada printed PINCODE samples are used to test the method. The samples considered are from different font styles and font sizes. The different font styles considered in English are shown figure(5a) and font styles in Kannada are shown in figure(5b). The font sizes considered vary from 14 points to 28 points. Table-6.1.2.3a shows the results of experiments conducted on the samples.

Table-6.1.2.3a: Results of experiments on samples

Script	No. of samples	Recognized	Failure	Rejection
English	1020	1009 (98.92%)	4 (00.40%)	7 (00.68%)
Kannada	1012	1003 (99.11%)	0 (00.00%)	9 (00.89%)
Combined	2032	2012 (99.02%)	4 (00.20%)	16 (00.78%)

It is evident from the experiments that the proposed methodology gives an average recognition efficiency of about 99%. Failures and rejections are mainly due to degradation in digits i.e., erosion, incompleteness and disconnections. Failure and rejection occur only due to misclassification of a maximum of one digit with in a PINCODE. The method exhibits 100% recognition efficiency for synthetically generated PINCODES as they do not suffer from any degradation. However the developed methodology is very much sensitive to stylish/artistic font style. Efficiency

reduces considerably with the inclusion of fonts like monotype corsiva, Georgia in English and Kasturi in Kannada.

6.1.2.3 Conclusion

The methodology described in this paper extracts and interprets the PINCODE of postal mail printed in English and Kannada scripts. The method exhibits an average recognition efficiency of about 99% for PINCODES printed in both the scripts. The method is sensitive to artistically printed font styles. The approach does not require the stage of thinning of numerals required by most of the other recognition approaches. The time complexity of the projection processes is $O(n^2)$. The proposed approach shows performance efficiency, which is similar to other standard approaches that do not make use of any mathematical or statistical models, but uses lesser memory and CPU time. In addition, the proposed method is not sensitive to thickness of the numerals. Thus the methodology is a better alternative for recognition of printed English and Kannada numerals.

The failure rates can be further reduced, by building up a database of all valid PINCODES. The PINCODE interpreted can be looked up in the database as a validation process by finding the nearest match and subjecting incorrect digit to re-recognition using some other features. This reduces the rate of failures to certain extent. It is also noticed that rejection cases are due to misclassification of script in a maximum of one digit. Such misclassified digit can be resubmitted for correct recognition using some other features of the digits. The adaptation of this technique can reduce the failure and rejection rates and the process is under investigation.

In banks every remittance transaction requires a remittance-challan to be filled up, which is most of the times compelled to be prepared manually. A typical remittance challan consists of three most important columns, containing the preprinted currency denomination as an apriori designated first column. In the second column against each denomination, number of currency notes being remitted is entered followed by the total amount per denomination. The counter assistant has to interactively process the

counting operations managing between reading of the challan by himself and automatic counting of the currency by the machine. This research is to integrate a challan reading machine, which can automatically decipher denominations and other entries to perform the cross validation with the currency counting machine, enabling a smooth, quick and an error-free remittance transaction, relieving the cash-counter assistant from unnecessarily being burdened. In such a transaction, the accurate numeral reading being of paramount importance, a modified 7-segment projection approach for numeral recognition is proposed in this research. The efficacy of the newly proposed approach is established by rigorously testing the algorithm with MNIST[url10] test data set.

The second case study for automatic handwritten numeric data extraction from input forms is from a Banking application and is explained in the next section.

6.2 Automatic Reading of Denomination Entries in Bank Challans as a Support to Currency Counting Machine[12]

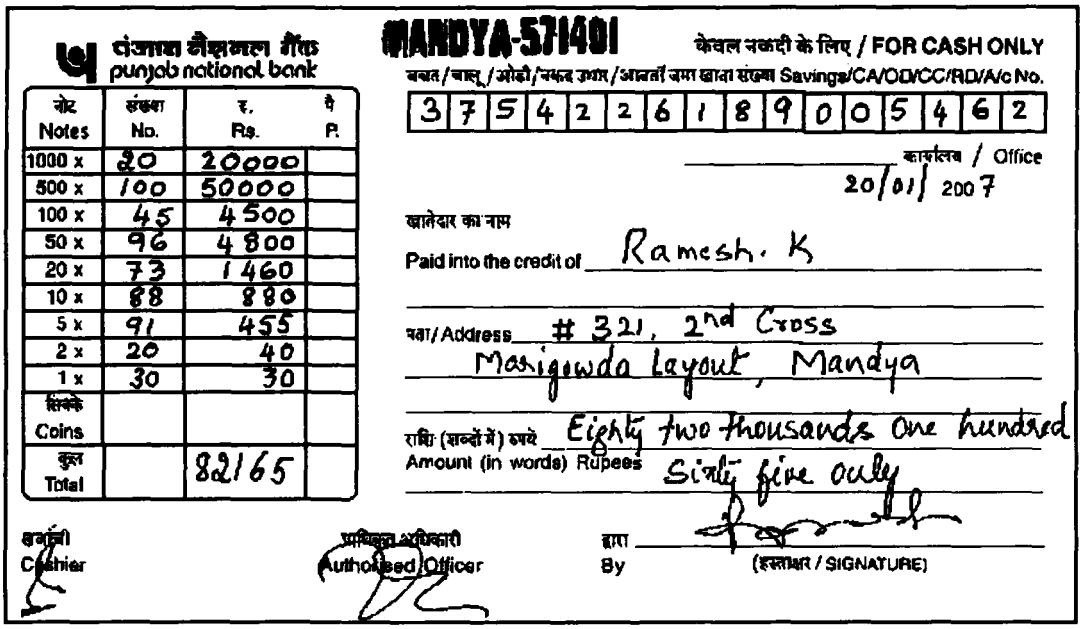
Banks in India at present are equipped with currency counting machines. The cash is counted using currency counting machine while cash is remitted as well as the cash is withdrawn. But, process of remitting cash follows a sequence of operations. The cash along with a challan filled with necessary details and the denomination entries of the currencies used for remittance are submitted at cash counter. The cashier collects the cash, puts the cash into counting machine denomination wise and verifies the same with the denomination entries in the challan. Finally, a manual counter check is made with the final total in the denomination entry. This nature of distributed work between man and machine has motivated us to come out with a proposal through which the reading of denomination and verification is also automated along with counting.

[12]This part of work is communicated to the Journal of Engineering and Technology, "Automatic Bank Challan Reading as a support to Currecny Counting Machine", April, 2007.

Literature survey in this direction reveals that considerable quantum of work is reported in extracting data from document images especially on extracting data from printed forms. All the works reported in literature are application or domain specific and no generic models for extracting data from document images are proposed. Data processing applications obtain data manually from the corresponding forms like data from response sheets, data from marks sheets, data from train/bus reservation request forms, data from bank forms etc. Majority of such applications have the data provided through manual entry on pre-printed forms. In 1997 Jainchan, Raymond and Mohiuddin[Jai97] have developed a system for automatic reading of IATA flight coupans. Stephan et al.[Ste97] have attempted to extract messages from printed documents. Yaun and Liu[Yau97] have proposed a system for information acquisition and storage of forms in document processing. In 1999 Mahadevan and Shrihari[Mah99] developed a model for parsing and recognition of City, State and Zip codes of handwritten address. Suen, Xu and Lam[Sue99] suggested a model for automatic recognition of handwritten data on cheques. In 2001, Chen and Lee [Che101] proposed data extraction for form document processing using a gravitational-based algorithm. In 2001, Vasudev, Hemantha kumar and Nagabhushan[Vas102] have proposed automatic reading of response sheets through a heuristic approach. The same team[Vas102] in 2002, proposed data extraction from pre-printed documents and automatic updation of database. In 2006, Veena et al. [Vee106] suggested a model for automatic evaluation of answer sheets using domain knowledge. To best of our efforts, we could not trace the work on automatic reading of denominations from a bank challan. A work on automatic reading of denomination is attempted by us on a bank challan of Punjab National Bank, Mandya Branch, Karnataka, India, and a sample challan is shown in figure(6.2a).

The main contribution in this work is the approach developed for recognition of numerals in the denomination of the challan. The recognition process is a modified version of the earlier work[Vas102] based on 7-segment projection. In the earlier approach the horizontal and vertical projection from overlapping regions method was employed for projection and the display segments were not projected in some of the

cases since the density of points due to projection fall below the threshold fixed. Hence, a radial projection is designed from non-overlapping regions which produce better population of projected points during projection and the method showed considerable improvement in the recognition rate.



Figure(6.2a): Sample Bank Challan

The input image is obtained by scanning the document through hp7890 scanner at 300 dpi. The image is assumed to be preprocessed for skew correction and noise removal. In addition, it is also assumed that the digits in the denomination entries are not connected with each other and the digits are not crossing the boundary lines of the boxes provided in the table. Segmentation of denomination entries are made by isolating the parts of the image based on the apriori knowledge of the structure of the challan. Further, the images of individual digits with in each entry are obtained using Y-cut technique and are subjected for recognition. Finally the recognized digits are summarized and interpreted to represent the value in the entry.

This chapter is organized into the following sections. Recognition process of numerals in denomination entries is performed using modified projection which is explained in chapter 5 section-5.3 with experimental results and analysis of results. Automatic bank

challan reading as a support to currency counting machine is explained in section 6.2.1 with experimental results. In section 6.2.2, a brief conclusion on the work is presented. Recognition of numerals segmented out from denomination table with the knowledge of challan structure is explained in the next section.

6.2.1 Automatic reading as a support to currency counting machine

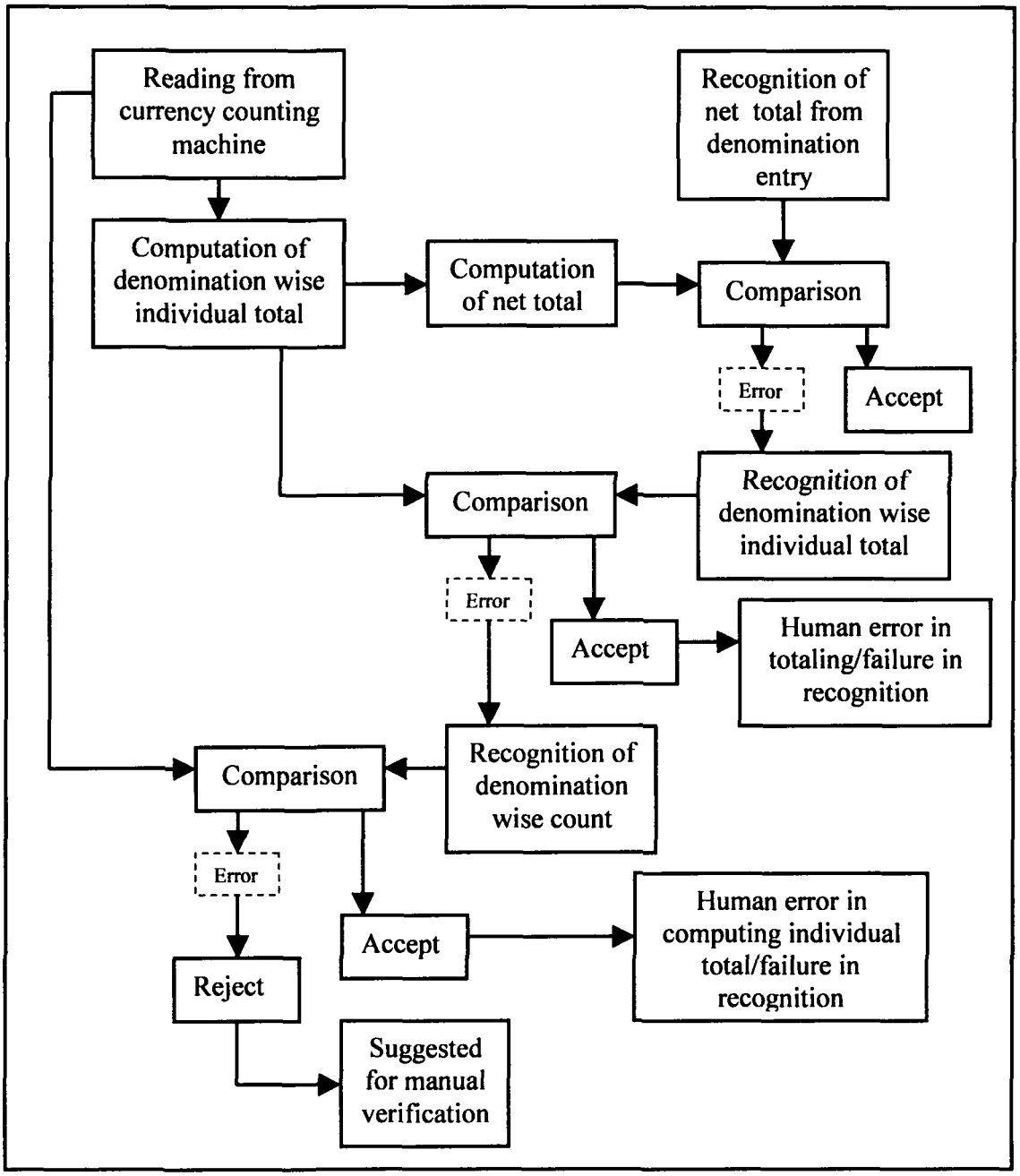
This section consists of two subsections, the first subsection explains the proposed model for automatic reading of denomination entries from bank challan and in the second subsection the results of experimental analysis are discussed.

6.2.1.1 Proposed model for automatic reading of denomination entries

The model is developed assuming that the currency counting machine is interfaced into the system in which recognition process is performed and also assumed that the counting machine counts the currencies denomination wise twice and count enters the system as input. A system diagram for the proposed model is shown in figure(6.2.1.1a).

The proposed model for automatic reading of denomination from bank challan performs the operations in the following sequence. To start with, the denomination wise currency count is obtained as input from counting machine. Based on the counting, denomination wise individual currency total and net total are computed. Next, the computed net total is compared with recognized net total. A match between the two is an acceptance case and the reading is accepted for further processing. A mismatch in this stage obtains the denomination wise computed individual total and compared with the recognized denomination wise individual total entries. A match in comparison is again a case of acceptance with indication of a human error in totaling or misclassification in recognition process due to poor/bad quality of writing. A mismatch at this stage looks for the denomination wise count of currencies and is compared with the recognized denomination counts from the entries. A match in comparison at this stage is an acceptance case with an indication of error in manual

individual total computation or misclassification of recognition process due to poor quality of writing. A mismatch at this stage signifies failures at all three levels and the case is rejected and directed to proceed with manual verification.



Figure(6.2.1.1a): System diagram of model for automatic reading of denomination entries

The error indication modules in the system are slightly strengthened with a process of supervised learning. The misclassified digit is looked in the set of overlapping /conflicting digits. The digit is reclassified to nearest matching class and validation is performed through computed values. If the digit in overlapping set does not get validated with the result then human error in computation/writing is concluded.

6.2.1.2 Experimental Results

The proposed model is tested by considering 46 numbers of challans. The samples include with neatly written, badly written denominations and denomination entries with error in totals. Cases of connected digits, denominations overlapping and crossing the boxes are included in small numbers to study the behavior of the system. The result of testing is tabulated in the table(6.2.1.2a).

Table-6.2.1.2a: Stage wise results of testing the proposed model

Stage of acceptance	Stage–1 (Net total verification)	Stage – 2 (Individual total verification)	Stage – 3 (currency count verification)
No. of samples accepted	18 (39.13%)	05 (10.87%)	08 (17.39%)
No. of samples sent To next stage	28 (60.87%)	23 (50.00%)	15 (32.61%)
Rejection	---	---	15 (32.61%)
Total efficiency	---	---	31 (67.39%)

Table(6.2.1.2a) shows the results of testing conducted on the samples. About 39.13% of acceptance is noticed at stage-1 .i.e., verification with net total stage. 10.87% of acceptance is observed in stage-2 i.e., verification of denomination wise individual totals. At stage-3 i.e., verification of currency counting stage 17.39% of acceptance is noticed. With these results an overall efficiency shown by the model is 67.39% with a rejection rate of 32.61%. The reasons for rejections are due to misclassification in recognition of badly written numerals, connection between the digits, incomplete digits, crossing and overlapping of numerals on the boundaries of boxes, extra strokes in the digits, overwriting the numerals.

6.2.2 Conclusion

In this chapter a model for automatic reading of denominations from bank challan is proposed. The model shows an acceptance rate of 67.39% with a rejection rate of 32.61%. Automatic reading of bank challan uses three stages for verification and the reading may be accepted at any of the three different stages otherwise reading is rejected directing the cash counter assistant for manual verification. A small model for reclassification of numerals is also included in some cases of misclassification of numerals based on computed values.

The automatic reading model has certain limitations like the denominations are to be written legibly, not with connected numerals, writing not to cross over the boundary boxes. In addition the input is assumed as noise free. There is scope to improve the efficiency of acceptance by strengthening the reclassification model through a supervised learning for accurate classification of numeral.

6.3 Chapter Summary

In this chapter two case studies are discussed. In the first case study, two methods were proposed. The first method reads automatically printed English/Kannada PINCODE through the identification of script. The second method reads automatically printed English/Kannada PINCODE without the identification of script. Experiment on both the methods show recognition rate above 90%. The second case study is for automatic reading of handwritten denomination entries in bank challans. In the case study, the automatic reading is proposed as a support to currency counting machine.